On Some Connections between Nonlinear Filtering, Information Theory and Statistical Mechanics

Sanjoy K. Mitter

Laboratory for Information and Decision Systems
Massachusetts Institute of Technology
Cambridge, MA, USA

Abstract

1. Some basic concepts in Probability Theory and Information Theory
2. Filtering, Control and Thermodynamics
   Interaction of Communication and Control
3. Stochastic Linear Dynamical Systems. Kalman Filtering from the Innovations viewpoint.
   Separation Theorem
4. Dissipative Systems (J.C. Willems). Kalman Filter as an Informally Dissipative System. Information Optimality of the Kalman Filter.
   Nonlinear Generalization
5. Introduction to Statistical Mechanics. The Ising Model. Variational Description of Gibbs Measures. Bayesian Inference as Free Energy Minimization. The Duality between Estimation and Control

Introduction

- Problems of Control where Sensors, Actuators and Controllers are linked via Noisy, Communication Channels

- Information Theory

- Stochastic Control with Partial Observations (Dynamics Programming)

- Fundamental Limitations: Noisy Channel Coding Theorem

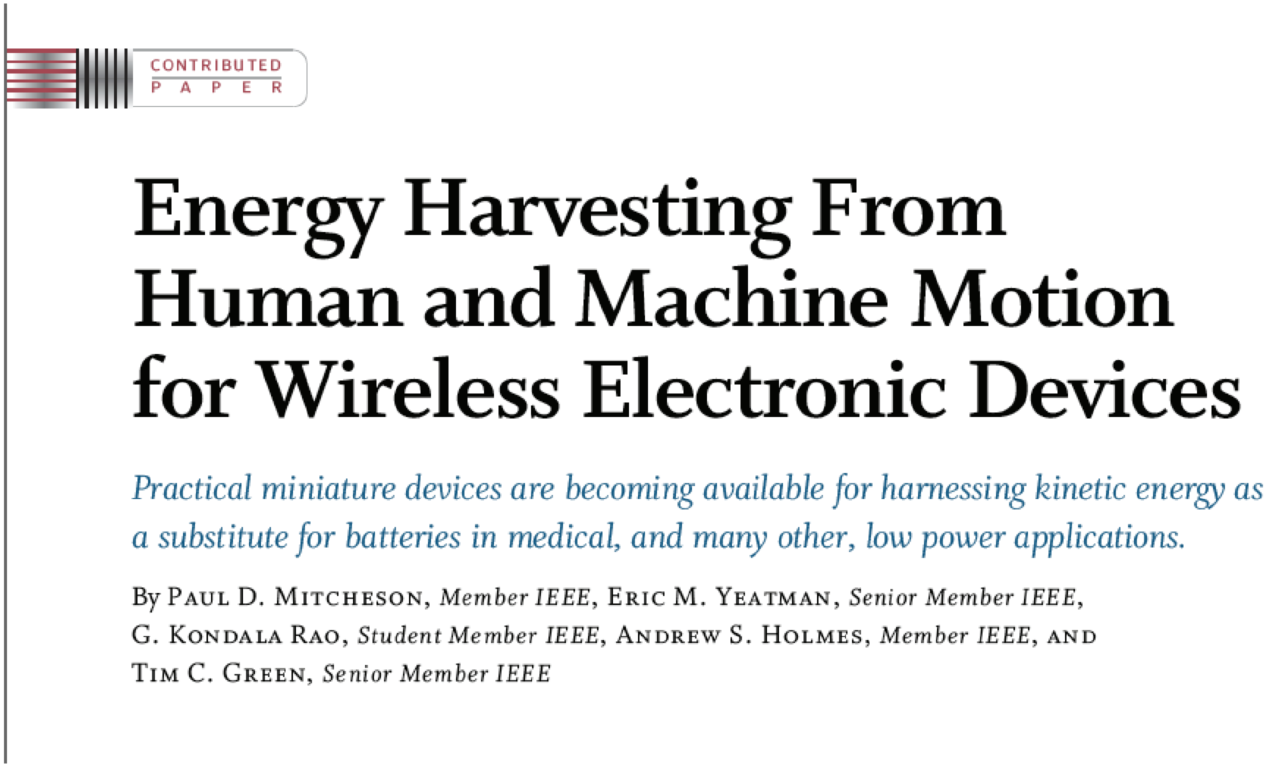
- LQG: Irreducible error

- Stablilization

- Energy Harvesting

- Recover Energy from Vibrating Motion

- Monitoring and Sensor Networks



CONTRIBUTED PAPER

Energy Harvesting From Human and Machine Motion for Wireless Electronic Devices

Practical miniature devices are becoming available for harnessing kinetic energy as a substitute for batteries in medical, and many other, low power applications.

By Paul D. Mitcheson, *Member IEEE*, Eric M. Yeatman, *Senior Member IEEE*, G. Kondala Rao, *Student Member IEEE*, Andrew S. Holmes, *Member IEEE*, and Tim C. Green, *Senior Member IEEE*

- Integration of Physics, Information, Control and Computation

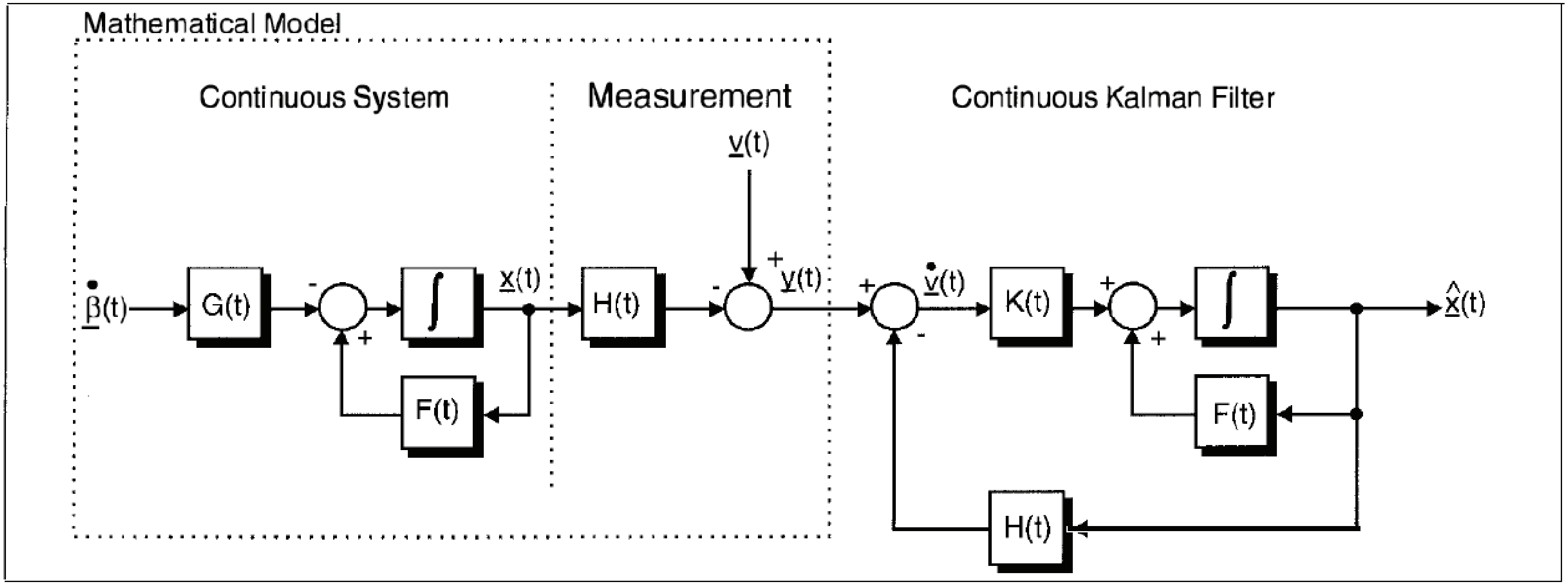


Fig. 1. System model and continuous Kalman filter.

Dissipative Dynamical System (Willems)

Σ: Dynamical Systems

(Soln. of state space system with read out map)

$w(t) = w(u(t), y(t))$, integrable Supply rate

Definition:

A dynamical system Σ with supply rate w is said to be *dissipative* if $\forall$ nonnegative function $S : X \to \mathbb{R}^{\dagger}$, called the "storage function" such that $\forall$ $(t_1, t_0) \in \mathbb{R}^{\dagger}_2$, $x_0 \in X$ and $u \in U$

$$S(x_0) + \int_{t_0}^{t_1} w(t)dt \geq S(x_1)$$

where $x_1 = \varphi(t_1, t_0; x_0, u)$ and $w(t) = w(u(t), y(t))$ and $y = y(t_0, x_0, u)$.

Maximum Work Extraction and Implementation Costs for Non-equilibrium Maxwell's Demons

by H. Sandberg, J.-C. Delvenne,
N.J. Newton and S.K. Mitter

in *Physical Review E* **90** 042119 (2014)

Maximum work extraction and implementation costs for non-equilibrium Maxwell's demons

Ever since Maxwell [1] put forward the idea of an abstract being (a demon) apparently able to break the second law of thermodynamics, it has served as a great source of inspiration and helped to establish important connections between statistical physics and information theory [2–6]. In the original version, the demon operates a trapdoor between two heat baths, such that a seemingly counterintuitive heat flow is established.

Today, more generally, devices that are able to extract work from a single heat bath by rectifying thermal fluctuations are also called "Maxwell's demons" [7]. Several schemes detailing how the demon could apparently break the second law have been proposed, for example Szilard's heat engine [2]. More recent schemes are presented in Refs. [7–12], where measurement errors are also accounted for.

A classical expression of the second law states the following: Maximum (average) work extractable from a system in contact with a single thermal bath cannot exceed the free energy decrease between the system's initial and final equilibrium states.

However, as illustrated by Szilard's heat engine, it is possible to break this bound under the assumption of additional information available to the work-extracting agent.

Free Energy = Average Energy − Entropy

To account for this possibility, the second law can be generalized to include transformations using *feedback control* [8,13–19]. In particular, in Ref. [18] it is shown that under feedback control, the extracted work W must satisfy

$$W \leq kTI_c, \tag{1}$$

where k is Boltzmann's constant, T is the temperature of the bath, and I_c is the so called *transfer entropy* from the system to the measurement.

Note that in (1) we have assumed there is no free energy decrease from the initial to the final state. Related generalizations of the second law are stated in Refs. [20–22]. It is possible to construct feedback protocols that saturate (1) using reversible and quasistatic transformations [18,23,24]. Reversible feedback protocols may be optimal in terms of making (1) tight, but they are also infinitely slow, and in Refs. [17,25–29] some related finite-time problems are addressed.

Our new contribution is to state an explicit *finite-time* counterpart to (1), characterizing the maximum work extractable using feedback control, in terms of the transfer entropy.

To explain our result, consider a system modeled by an overdamped Langevin equation. We show that the maximum amount of extractable work over a duration t, $W_{\mathrm{max}}(t)$, can be expressed by the integral

$$W_{\mathrm{max}}(t) = k\int_0^t T_{\mathrm{min}} \dot{I}_c \, dt' \le kTI_c(t). \qquad (2)$$

Here $T_{\mathsf{min}}(t)$ has an interpretation as the lowest achievable system temperature after t time units of continuous feedback control, assuming an equilibrium initially $T_{\mathsf{min}}(0) = T$. Since $T_{\mathsf{min}}(t) \leq T$, for all t, the upper bound in (2) follows trivially, implying (1).

The transfer entropy $I_c(t)$ measures the useful amount of information transmitted to the controller from the partial observations in the time interval $[0, t]$.

Therefore, every bit of transfer entropy, if optimally exploited, allows us to retrieve between $kT_{\text{min}} \ln 2$ and $kT \ln 2$ units of work. We furthermore provide a novel expression for the transfer entropy $I_c(t)$, applicable to a large class of systems in both continuous and in discrete time. In particular, the new expression yields closed-form solutions of the transfer entropy and shows its independence of the applied feedback law.

Our other contribution is to use control theory to characterize and interpret the feedback protocol the demon should apply to reach the upper limit $W_{\text{max}}(t)$.

The protocol is a linear feedback law based on the optimal estimate of the system state, which can be recursively computed using the so-called Kalman–Bucy filter. The found feedback law also offers a simple electrical implementation.

A proper physical implementation is also shown to require an external work supply to maintain the noise on the wires at an acceptable level. The cost of this noise suppressing mechanism can be evaluated by standard thermodynamic arguments, or through a non-equilibrium extension of Landauer's memory erasure principle.

System model

The system we first consider is an electric capacitor C, a resistor R with thermal noise (the heat bath), and a feedback controller (the demon) with access to noisy voltage measurements, see Fig. 1. The resistor is subjected to Johnson–Nyquist noise [39,40].

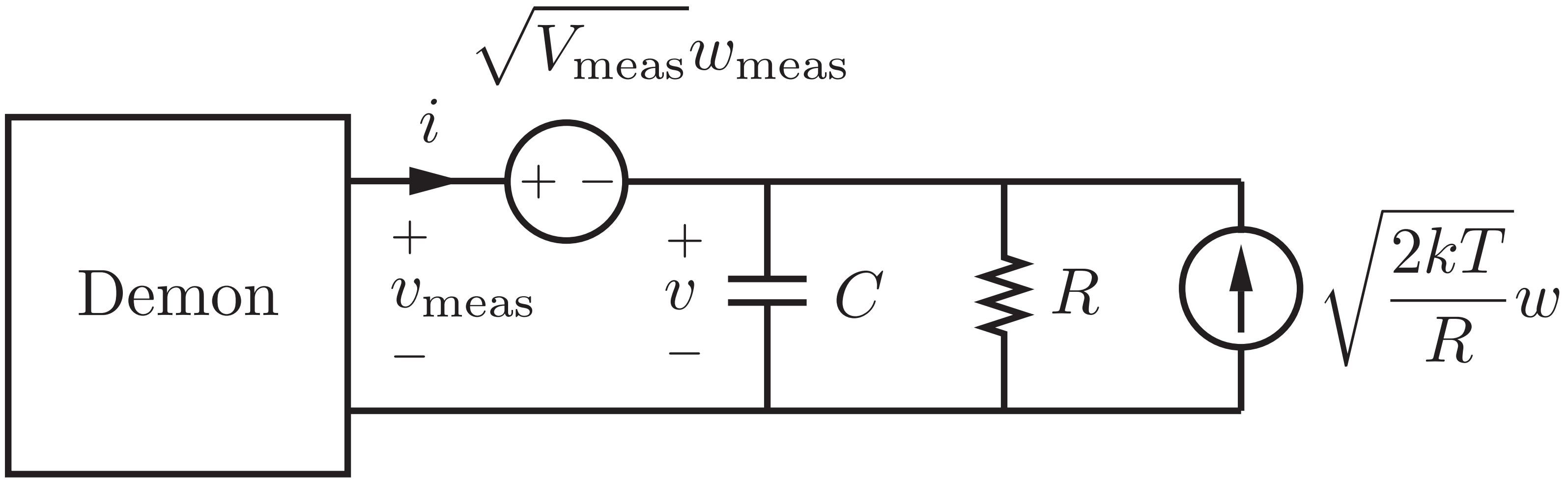


FIG.1: The demon (the feedback controller) connected to a capacitor, a heat bath of temperature T, and a measurement noise source of intensity V_{meas}. The demon may choose the current i freely, and has access to the noisy voltage measurement v_{meas}.

The circuit is modeled by an overdamped Langevin equation

$$\begin{aligned} \tau\dot{v} &= -v + Ri + \sqrt{2kTR}w, \ \langle v(0)\rangle = 0, \\ v_{\text{meas}} &= v + \sqrt{V_{\text{meas}}}w_{\text{meas}}, \ \langle v(0)^2\rangle = \tfrac{kT}{C}, \end{aligned} \qquad (3)$$

with $v(0)$ Gaussian, w and w_{meas} uncorrelated Gaussian white noise ($\langle w(t)w(t')\rangle = \langle w_{\text{meas}}(t)w_{\text{meas}}(t')\rangle = \delta(t-t')$), V_{meas} the intensity of the measurement noise, and $\tau = RC$ being the time constant of the open circuit.

The measurement noise $\sqrt{V_{\text{meas}}}w_{\text{meas}}$ can be thought of as the Johnson–Nyquist noise of the wire between the capacitor and the demon, whose resistance for simplicity is incorporated in the demon. The heat flow to the capacitor is $\dot{Q}$ and the work-extraction rate of the demon is $\dot{W}$, and satisfy the first law of thermodynamics,

$$\dot{U} = \dot{Q} - \dot{W} \tag{4}$$

where

$$\begin{aligned} U &= \tfrac{1}{2}C\langle v^2\rangle \equiv \tfrac{1}{2}kT_C, \\ \dot{Q} &= \tfrac{k}{\tau}(T - T_C), \quad \dot{W} = -\langle vi\rangle. \end{aligned} \tag{5}$$

We denote the *effective* instantaneous temperature (“kinetic temperature”) of the capacitor by T_C, and its internal energy by U. For detailed derivations of (4)–(5), see Ref. [37]. Furthermore, we assume the capacitor initially is in thermal equilibrium with the heat bath, i.e., $T_C(0) = T$.

Just as in Ref. [17], we can justify calling $T_C(t)$ a temperature since it appears in a Fourier-like heat conduction law (see $\dot{Q}$). Also, since our applied controls will maintain a Gaussian distribution of v, $T_C(t)$ will be the true temperature of the capacitor if it were to be disconnected from all the other elements at time t. The voltage v_{meas} is the measurement that supplies the demon with information, and can be seen as a noisy measurement of the fluctuating capacitor voltage v.

We will show how a demon can optimally control the work extraction by carefully exploiting the measurements v_{meas} and properly choosing the injected current i. Intuitively, the demon can create a positive work rate $\dot{W}$ if it chooses $i < 0$ when it correctly estimates $v > 0$, and vice versa. But how the demon should estimate v, and how to optimally choose i may be less obvious.

If we know the trajectory of the effective temperature T_C, it is from (4) and (5) possible to solve for the amount of extracted work,

$$W(t) = \int_0^t \frac{k}{\tau}(T - T_C)\, dt' + \frac{1}{2}k(T - T_C(t)). \qquad (6)$$

In particular, if we can characterize a lower bound on the effective temperature under all allowed controls, $T_{\text{min}}(t') \leq T_C(t')$ for $0 \leq t' \leq t$, we get an upper bound on the work that a demon can extract,

$$W_{\text{max}}(t) := \int_0^t \frac{k}{\tau}(T - T_{\text{min}})\, dt' + \frac{1}{2}k(T - T_{\text{min}}(t)), \tag{7}$$

so that $W(t) \leq W_{\text{max}}(t)$. In the following, we characterize T_{min}, and thereby W_{max}, using optimal control theory.

Demon model and optimal continuous-time feedback

Optimal control theory [41] teaches how to compute T_{min}, and to characterize the corresponding feedback law. In particular, for linear systems the *separation principle* [42] says we can achieve the goal in two steps: First, we should continuously and optimally estimate the voltage $v(t)$ of the capacitance, given the available measurements $(v_{\text{meas}})_0^t \equiv \{v_{\text{meas}}(t'), \quad 0 \leq t' \leq t\}$.

Second, we should continuously use the found optimal estimate to update the current $i(t)$ using a suitable linear feedback law.

The best possible estimate $\widehat{v}(t)$ of $v(t)$, given the measurement trajectory $(v_{\text{meas}})_0^t$, can be recursively constructed by the celebrated *Kalman–Bucy filter* [43], which leads to a minimum variance estimation error [41] and exploits as much of the information contained in v_{meas} as is possible [44].

We give a brief background to the Kalman–Bucy filter and its properties next. The filter state is denoted $\widehat{v}$ and satisfies the differential equation

$$\tau \frac{d}{dt}\widehat{v} = -\widehat{v} + Ri + K(v_{\mathsf{meas}} - \widehat{v}), \quad \widehat{v}(0) = 0, \quad (8)$$

where K is a time-varying function to be specified.

Guided by optimal control theory and the separation principle, we let the demon use the simple linear causal feedback

$$i(t') = -G\widehat{v}(t'), \quad 0 \leq t' \leq t, \tag{9}$$

where $0 \leq G < \infty$ is a fixed scalar feedback gain. We may think of the feedback gain G as the "conductance" of the demon: If the demon believes the voltage of the capacitor to be $\widehat{v}$, it will admit the current $G\widehat{v}$. If $v \approx \widehat{v}$, the demon will indeed look like an electric load of conductance close to G.

While $G = 0$ (open circuit) creates a demon that only (optimally) observes, $G \to \infty$ also removes energy from the capacitance at the highest possible rate, achieving the minimum effective temperature T_{min}. This can be seen as follows: Inserting (9) in (8) we can compute the evolution of the variance $\widehat{V} \equiv \langle \widehat{v}^2 \rangle$ of the filter estimate as

$$\tau \frac{d}{dt} \widehat{V} = -2\,(1 + GR)\,\widehat{V} + \frac{\sigma k T_{\text{min}}^2}{2CT}, \qquad \widehat{V}(0) = 0. \tag{10}$$

We note that since T_{min} is bounded, $\hat{V}$ can be made arbitrarily close to zero by increasing the feedback gain G. From (paper14) and (paper15), it follows that

$$\frac{kT_C}{C} = \langle v^2 \rangle = \langle \widehat{v}^2 \rangle + \langle \Delta v^2 \rangle = \hat{V} + \frac{kT_{\mathrm{min}}}{C}. \qquad (11)$$

Since T_{min} is independent of G, and $\hat{V}$ can be made arbitrarily close to zero, we realize that the demon through its policy is cooling the capacitor and for all t,

$$T_C(t) \searrow T_{\mathrm{min}}(t) \quad \text{as} \quad G \to \infty. \qquad (12)$$

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

For a general feedback gain $G \geq 0$ in (9), the effective temperature of the capacitor will drop exponentially from $T_C(0) = T$ to

$$T_C^{\text{NESS}} = \frac{1}{1+GR}T + \frac{GR}{1+GR}T_{\text{min}}^{\text{NESS}}. \qquad (13)$$

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

$$\dot{W}^{\mathsf{NESS}} = \frac{k}{\tau}(T - T_{\mathsf{min}}^{\mathsf{NESS}})\frac{GR}{1+GR}. \qquad (14)$$

Thus the continuous feedback protocol in (9) can realize any NESS work rate between 0 and the maximum $\dot{W}_{\mathsf{max}}^{\mathsf{NESS}} = \frac{k}{\tau}(T - T_{\mathsf{min}}^{\mathsf{NESS}})$ by proper choice of gain G.

The above optimal controller can be generalized to any system with linear dynamics.

Information flow and maximum work theorem

To establish the maximum work theorem in (2), we need to quantify the information flow from the uncertain part of the voltage v to the measurement v_{meas}, under continuous feedback. This is the *transfer entropy*, as is explained in Ref. [18], for example.

We show that the appropriate continuous-time limit of the transfer entropy is

$$I_c(t) = I((v(0), (w)_0^t); (v_{\mathrm{meas}})_0^t). \tag{15}$$

This is the *mutual information* between the uncertain initial voltage $v(0)$ and noise trajectory w from the bath, and the measurement trajectory v_{meas}.

Mutual information [47] between two stochastic variables ξ and θ is as usual defined as

$$I(\theta;\xi) \equiv \int \ln \left(\frac{d\mathbb{P}_{\theta\xi}}{d(\mathbb{P}_\theta \otimes \mathbb{P}_\xi)} \right) d\mathbb{P}_{\theta\xi} \geq 0, \tag{16}$$

and is equal to the amount the (differential) Shannon entropy of ξ decreases with knowledge of θ, and vice versa.

Here $\mathbb{P}_{\theta\xi}$, $\mathbb{P}_\theta$, and $\mathbb{P}_\xi$ are joint and marginal probability measures of the stochastic variables θ and ξ. We prove that the transfer entropy in fact has the following explicit form:

$$I_c(t) = \frac{\sigma}{4\tau} \int_0^t \frac{T_{\text{min}}}{T} \, dt'. \tag{17}$$

Note that I_c *does not* otherwise depend on the details of the demon, for example the feedback gain G.

It now follows from Eqs. (7), (paper11), and (17) that the maximum extracted work must satisfy

$$\begin{aligned} W_{\mathrm{max}}(t) &= \int_0^t \frac{k}{\tau}(T - T_{\mathrm{min}})\, dt' + \frac{1}{2}k(T - T_{\mathrm{min}}(t)) \\ &= \int_0^t \frac{\sigma k T_{\mathrm{min}}^2}{4\tau T}\, dt' = k \int_0^t T_{\mathrm{min}} \dot{I}_c\, dt', \end{aligned} \qquad (18)$$

which proves the equality in (2). The inequality trivially follows since $T_{\mathrm{min}} \leq T$.

The expressions for I_c and W_{max} provide interesting insights concerning information and work flow in the feedback loop. Since T_{min} decreases monotonically, the transfer entropy rate $\dot{I}_c$ is largest just when the measurement and feedback control start, and then decreases until it stabilizes at

$$\dot{I}_c^{\text{NESS}} = \frac{\sigma T_{\text{min}}^{\text{NESS}}}{4\tau T} = \frac{\sqrt{1+\sigma}-1}{2\tau}. \tag{19}$$

In NESS, the fresh measurements are no longer able to improve the quality of the estimate, i.e., to decrease the error variance $\langle[v(t) - \widehat{v}(t)]^2\rangle$ any further. Since $\dot{W}_{\text{max}} = kT_{\text{min}}\dot{I}_c$, the work-extraction rate also decreases until it stabilizes at

$$\dot{W}^{\text{NESS}} = kT_{\text{min}}\dot{I}_c^{\text{NESS}}\frac{GR}{1+GR}, \qquad (20)$$

see (14).

As in related studies [18,23,26], we can now define and study the *information efficiency* η of the demon,

$$\eta \equiv \frac{W}{kTI_c} \in [0,1]. \tag{21}$$

It measures the amount of extracted work per unit of received useful information. An $\eta \approx 1$ means that the demon is close to saturating (1), and is operating at the limit of the generalized second law of thermodynamics.

For our demons in NESS, we obtain the efficiency

$$\eta^{\text{NESS}} = \frac{T_{\text{min}}}{T}\frac{GR}{1+GR}, \tag{22}$$

using (20). Hence, only a maximum work demon ($G \to \infty$) with $T_{\text{min}} \approx T$ will operate at an information efficiency close to 1. This corresponds to the poor measurement limit ($\sigma \approx 0$), and a very small maximum work rate. A demon with access to almost perfect measurements ($\sigma \to \infty$) has $T_{\text{min}} \approx 0$, and a very low information efficiency, $\eta \approx 0$.

Note also that a less aggressive demon (small G) has a lower efficiency, but that this is by choice: The transfer entropy rate is independent on G, and a smaller G decreases $\dot{W}$, leading to a lower efficiency.

REFERENCES

[1] Maxwell, J. C., *Theory of Heat*, (Longmans, London, 1871).

[2] Szilard, L. *Z. Phys.* **53**, 840 (1929).

[3] Landauer, R., *IBM Journal of Research and Development* **5**, 183 (1961).

[4] Bennett, C. H., *International Journal of Theoretical Physics* **21**, 905 (1982).

[5] Penrose, O., *Foundations of Statistical Mechanics: A Deductive Treatment*, Dover Books on Physics Series (Dover, London, 2005).

[6] Leff, H. and Rex, A., *Maxwell's Demon 2 Entropy, Classical and Quantum Information, Computing*, Maxwell's Demon (CRC Press, Boca Raton, 2010).

[7] D. Mandal, H. T. Quan, and C. Jarzynski, Phys. Rev. Lett. 111, 030602 (2013).

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

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

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

[11] D. Mandal and C. Jarzynski, PNAS 109, 11641 (2012).

[12] A. C. Barato andU. Seifert, Phys. Rev. Lett. 112, 090601 (2014).

[13] H. Touchette and S. Lloyd, Phys. Rev. Lett. 84, 1156 (2000).

[14] H. Touchette and S. Lloyd, Physica A 331, 140 (2004).

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

[16] Y. Fujitani and H. Suzuki, J. Phys. Soc. Jpn. 79, 104003 (2010).

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

[18] T. Sagawa and M. Ueda, Phys. Rev. E 85, 021104 (2012).

[19] S. Ito and T. Sagawa, Phys. Rev. Lett. 111, 180603 (2013).

[20] H.-H. Hasegawa, J. Ishikawa, K. Takara, and D. Driebe, Phys. Lett. A 374, 1001 (2010).

[21] M. Esposito and C. V. den Broeck, Europhys. Lett. 95, 40004 (2011).

[22] S. Deffner and C. Jarzynski, Phys. Rev. X 3, 041003 (2013).

[23] J. M. Horowitz and J.M. R. Parrondo, Europhys. Lett. 95, 10005 (2011).

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

[25] T. Schmiedl and U. Seifert, Phys. Rev. Lett. 98, 108301 (2007).

[26] M. Bauer, D. Abreu, and U. Seifert, J. Phys. A: Math. Theor. 45, 162001 (2012).

[27] G. Diana, G. B. Bagci, andM. Esposito, Phys. Rev. E 87, 012111 (2013).

[28] P. R. Zulkowski and M. R. DeWeese, Phys. Rev. E 89, 052140 (2014).

[29] M. Bauer, A. C. Barato, and U. Seifert, J. Stat. Mech. (2014) P09010.

[37] J.-C. Delvenne and H. Sandberg, Physica D 267, 123 (2014).

[39] J. B. Johnson, Phys. Rev. 32, 97 (1928).

[40] H. Nyquist, Phys. Rev. 32, 110 (1928).

[41] K. J. A strom, Introduction to Stochastic Control Theory, Dover Books on Electrical Engineering Series (Dover, London, 2006).

[42] W. Wonham, SIAM J. Control 6, 312 (1968).

[43] R. S. Bucy and P. D. Joseph, Filtering for Stochastic Processes with Applications to Guidance (Interscience, New York, 1968).

[44] S. K. Mitter and N. J. Newton, J. Stat. Phys. 118, 145 (2005).

Towards a Unified View of Communication and Control

Sanjoy K. Mitter

Laboratory for Information and Decision Systems

Massachusetts Institute of Technology

INTRODUCTION

1. Partially observable stochastic control problem

2. Feedback communication problem

3. Interaction of information and control

4. A general view of interconnection

5. Are communication problems really different from control problems?

1. Partially Observable Stochastic Control

1. Partially observable stochastic control

1.1 Formulation

$$\left.\begin{array}{lrcll} (1.1) & X_{n+1} & = & F_n(X_n, u_n, \xi_n)\ , & n = 0, 1, 2, \dots \\ (1.2) & Y_{n+1} & = & G_n(X_n, \eta_n)\ , & n = 0, 1, 2, \dots \end{array}\right\} \begin{array}{l}\text{Hidden}\\ \text{Markov}\\ \text{Process}\end{array}$$

$$\begin{aligned} E &= \text{State space} \\ U &= \text{Control space} \\ E_1 &= \text{Observation space} \end{aligned}$$

$$\xi_i \in S_1 \ ; \qquad \eta_i \in S_2$$

Assume: $X_0, (\xi_i), (\eta_i)$ are independent

Admissible strategy $\pi := u_0(y_0), u_1(y_0, y_1), \dots$ sequence of measurable mappings

Cost function:

$$(1.3) \quad J_N(\pi, X_0) = \mathbb{E}(\sum_{n=0}^{N-1} q_n(X_n, u_n) + r_N(X_n))$$

J_N to be *minimized* by choice of π.

Bellman Equation

Fully Observable Case:

$$\begin{aligned} V_N(x) &= \int_E r_N(x)\mu(dx) \\ V_n(x) &= \inf_{u\in U}[\widehat{q}_n(x,u) + P_n^u \widehat{V}_n(x)] \end{aligned}$$

In Partially Observation Case: we want

$$\begin{aligned} \widehat{V}_N(c) &= \widehat{r}_N(c) \\ \widehat{V}_n(c) &= \inf_{u\in U}[\widehat{q}_n(c,u) + \widehat{P}_n^u \widehat{V}_n(c)] \end{aligned}$$

where c is a sufficient statistic

Reduce the Partially Observable Case to the Fully Observable Case

$(\Omega, \mathcal{F}, P)$ fixed probability space

Idea: Define *sufficient statistics* (*appropriate conditional distributions*), describe their evolution and rewrite the cost function in terms of these sufficient statistics.

$$(1.4)\ Y_n = \sigma(Y_0, \ldots, Y_n)$$

$= \ $ Information contained in observations up to time n

1.2 Conditional distributions and their evolution

Let C_n, $n = 0, 1, \ldots$ be the conditional distribution of X_n given Y_n. This means

$$(1.5)\ \ \mathbb{E}(\varphi(X_n, Y_n)(\omega) = \int_E \varphi(x, Y_n(\omega)) C_n(\omega, dx)$$

$\varphi : E \times E_1 \to \mathbb{R}_+ \ P -$ a.s. ω

Subtlety: The observations $(Y_n)_{n=0,1,\ldots}$ depend on the control sequence $(u_n(y_0,\ldots,y_n))_{n=0,1,\ldots}$ and, in turn, the control sequence depends on the observations.

It can be shown that the sequence $(C_n)_{n=0,1,\ldots}$ of conditional distributions is a Controlled Markov Chain with a time-dependent and control-dependent transition function:

(1.6) $\widehat{P}_n^u\psi(c) =$

$$\mathbb{E}\left[\int_E \psi[\widehat{F}_n(c,u,G_{n+1}(F_n(x,u,\xi_n),\eta_{n+1}))]c(dx)\right]$$

where $\widehat{E} \subset \mathcal{P}(E)$ measurable

$$\begin{aligned} \widehat{F}_n &: \widehat{E} \to \mathbb{R}_+ \\ c &\in \widehat{E} \\ \psi &\in B_b(\widehat{E}) \end{aligned}$$

Furthermore, in many situations, one can show that there exists a sequence of *independent random* variables $(\widehat{\xi}_L)_{i=0,1,...}$ taking values in $\widehat{S}_1$ (*Innovations sequence*) and measurable mappings:

$$\widehat{F}_n : \widehat{E} \times U \times \widehat{S}_1 \to \widehat{E} \text{ s.t.}$$

$$(1.7) \qquad C_{n+1} = \widehat{F}_n(C_n, u_n, \widehat{\xi}_n) \ , \quad n = 0, 1, \ldots$$

Note: Loop between control and observation has been eliminated

Define:

$$\begin{aligned} (1.8)\ \widehat{q}_n(c,u) &= \int_E q_n(x,u)c(dx) \ , \quad n = 0, 1, \ldots \\ (1.9)\quad \widehat{r}_N(c) &= \int_E r_N(x)c(dx) \ , \quad \begin{array}{l} c \in \widehat{E} \\ u \in U \end{array} \end{aligned}$$

Bellman Equations

$$(1.10)\quad \widehat{V}_N(c) \;=\; \widehat{r}_N(c)$$

$$(1.11)\quad \widehat{V}_n(c) \;=\; \inf_{u\in U}[\widehat{q}_n(c,u)+\widehat{P}_n^u\widehat{V}_n(c)]$$

$$=\; \widehat{q}_n(c,\widehat{u}_n(c))+\widehat{P}_n^{\widehat{u}_n}\widehat{V}_n(c)$$

$$n=0,1,2,\ldots$$

Theorem: Optimal strategy π is determined by the sequence:

$$\widehat{u}_0(\widehat{c}_0),\widehat{u}_1(\widehat{c}_1,\widehat{c}_0),\ldots$$

where $\widehat{c}_0,\widehat{c}_1,\ldots$ are functions of $y_0,(y_0,y_1),\ldots$ given recursively by

$$(1.12)\quad \widehat{C}_{n+1} \;=\; \widehat{F}_n(\widehat{c}_n,\widehat{u}_n(\widehat{c}_n),y_{n+1})$$

$$\widehat{c}_0 \;=\; \mathbb{E}(X_0|Y_0)$$

$$(1.13)\qquad \text{Min. Cost} = \mathbb{E}(\widehat{V}_0(c_0))$$

Separation Theorem

Estimation and Control Separate

Average Cost Problem

$$\min \lim_{N\to\infty} \frac{1}{N}\mathbb{E}\left(\sum_{n=0}^{N-1} q_n(x_n, u_n) + r_N(x_N)\right)$$

Minimization is over all strategiess

$$\pi := u_0(y_0), u_1(y_0, y_1), \ldots$$

Feedback Communication

Ref: S. Tatikonda and S.K. Mitter:

The Capacity of Channels with Feedback, *IEEE Trans. on Information Theory*, Vol. 55, Jan. 2009, pp. 323–349.

2. Feedback Communication Problem

2. Feedback communication problem

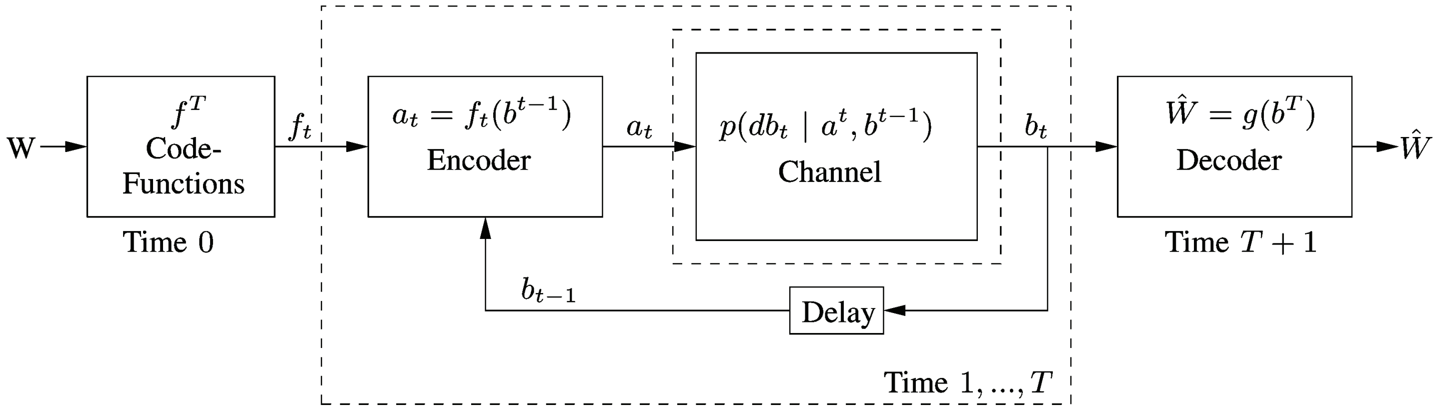


Figure 1. Interconnection

Choose encoder and decoder to transmit message over the channel to minimize the probability of error

Channel at time t: $P(db_t | a^t, b^{t-1})$ stochastic kernel

$$a^t = (a_0, \ldots, a_t)$$

(2.1) *Channel* $=$ Sequence of $P(db_t | a^t, b^{t-1})\Big|_{t=1}^{t}$

Time ordering: Message $= W, A_1, B_1, \quad , A_T, B_T, \hat{W} =$ Decoded message

$$W = (1, 2, \ldots, M)$$

Code function:

$$\begin{aligned} (2.2) \qquad \mathcal{F}_t &= \{f_t : B^{t-1} \to A : \text{measurable}\} \\ \mathcal{F}_T &= \prod_{t=1}^{T} \mathcal{F}_t \end{aligned}$$

Channel code function: $f^T = (f_1, \ldots, f_t)$

Distribution on code functions: $P(df_t | f^{t-1})\Big|_{t=1}^{T}$

Channel code = list of M channel code functions

Code functions are introduced to reduce the feedback communication problem to a no feedback communication problem.

Average Measure of Dependence

Mutual Information

$$(2.3) \qquad \begin{aligned} I(A^T;B^T) &= \mathbb{E}_{P_{A^T,B^T}} \log\left(\frac{P_{A^T,B^T}}{P_{A^T}P_{B^T}}\right) \\ &= \mathbb{E}_{P_{A^T,B^T}} \log\left(\frac{P_{B^T|A^T}}{P_{B^T}}\right) \end{aligned}$$

$$I(A^T;B^T) = \sum_{t=1}^{T} I(A^T;B_t|B^{t-1})$$

Information transmitted to the receiver depends on future $(A_{t+1}, \ldots, A_T)$.

Directed Mutual Information (Causal)

$$(2.4) \qquad I(A^T \to B^T) = \sum_{t=1}^{T} I(A^t;B_t|B^{t-1})$$

To compute Mutual Information (Directed Mutual Information), need joint distribution

$$\mathbb{P}_{A^T,B^T}(da^T, db^T)$$

This can be done if we are given the channel

$$P(db^t|a^t, b^{t-1})\Big|_{t=1}^{T}$$

and channel input distributions

$$\mathcal{D}_t := \mathbb{P}(da_t|a^{t-1}, b^{t-1}\Big|_{t=1}^{T} \tag{2.5}$$

Interconnection of channel input to channel

Channel Capacity

$$C_T = \sup_{\mathcal{D}_T} \frac{1}{T} I(A^T \to B^T) \tag{2.6}$$

(Note: Optimization over original input codes, not on space of code functions.)

Markov Channel

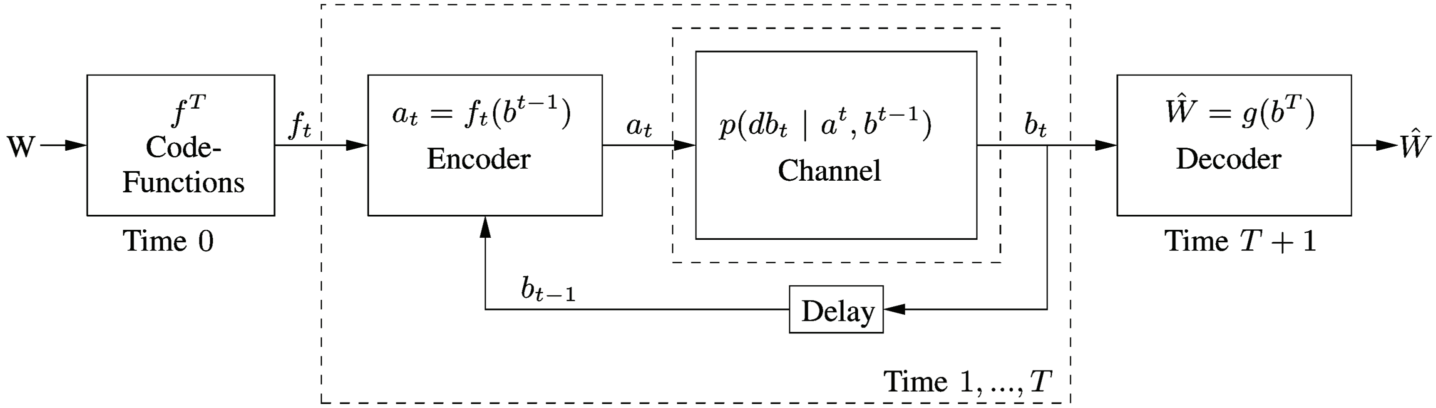


Figure 2. Markov channel

$$P(ds_{t+1}|s_t, a_t, b_t)\Big|_{t=1}^{T} \quad : \quad \text{state transition}$$

$$P(db_t|s_t, a_t)|_{t=1}^{T} \quad : \quad \text{channel output}$$

Capacity of Markov Channels

$$\sup_{\mathcal{D}_\infty} \lim_{T\to\infty} \frac{1}{T} I(A^T \to B^T) \qquad (2.7)$$

It turns out that by appropriately defining sufficient statistics (π_t) (conditional distributions of the state given information from encoder to decoder) and controls $u_t(da_t|\pi_t)$, and state $X_t = (\pi_{t-1}, A_{t-1}, B_{t-1})$ and instantaneous cost $c(x_t, u_t, u_{t+1})$, (2.7) can be formulated as a Partially Observed Stochastic Control Problem.

In turn, as shown in Part 1, this can be reformulated as a fully-observable stochastic control problem.

This problem is more like a *dual* control problem since the choice of the channel input can help the decoder identify the channel.

This is also an example where the *information pattern is nested*: The encoder has more information than the decoder.

The Interaction of Control and Communication

Sanjoy K. Mitter
MIT

Big questions about communication.

- Are all communication problems asymptotically alike?

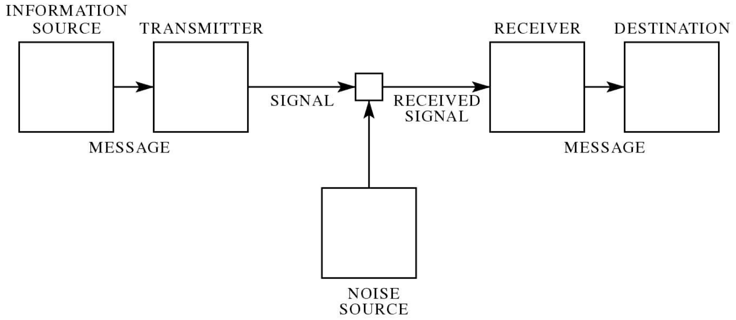
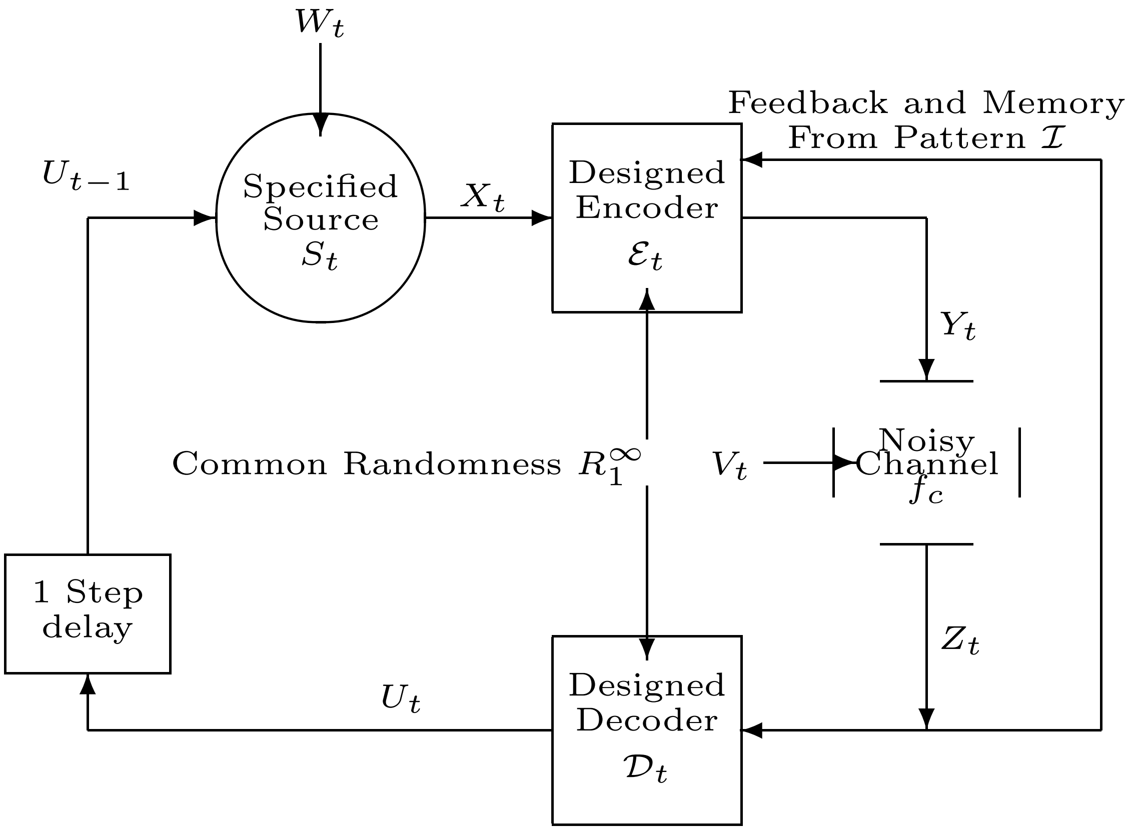


Fig. 1—Schematic diagram of a general communication system.

- How does delay interact with capacity issues?
- Can we find examples that let us explore these questions in an asymptotic setting?

“. . . can be pursued further and is related to a duality between past and future and the notions of control and knowledge. Thus we may have knowledge of the past and cannot control it; we may control the future but have no knowledge of it.” — Claude Shannon 1959

An abstract model of single channel problems


W_t
Specified Source S_t
X_t
Designed Encoder $\mathcal{E}_t$
Feedback and Memory From Pattern $\mathcal{I}$
U_{t-1}
Y_t
Common Randomness R_1^∞
V_t
Noisy Channel f_c
1 Step delay
Z_t
U_t
Designed Decoder $\mathcal{D}_t$

- Problem: Source S, Information pattern $\mathcal{I}$, and Objective $\mathcal{V}$.
- Constrained resource: Noisy channel f_c
- Designed solution: “Encoder” $\mathcal{E}$, “Decoder” $\mathcal{D}$

Focus: what channels are "good enough"

- f_c *solves* the problem if $\exists \mathcal{E}, \mathcal{D}$ so system satisfies $\mathcal{V}$
- Problem A is *harder* than problem B if any f_c that solves A solves B.
- Information theory is an *asymptotic* theory
  - Pick $\mathcal{V}$ family with an appropriate "slack" parameter
  - Consider the set of channels that solve the problem.
  - Take union over slack parameter choices.

The Shannon problems $A_{R,\epsilon,d}$

- Source: *noninteractive* X_i (R bits): fair coin tosses
- Information pattern: $\mathcal{D}_i$ has access to Z_1^i
  - A^f With feedback: $\mathcal{E}_i$ gets X_1^i and Z_1^{i-1}
  - A^{nf} Without feedback: $\mathcal{E}_i$ gets only X_1^i
- Performance objective: $\mathcal{V}(\epsilon, d)$ is satisfied if $\mathcal{P}(X_i \neq U_{i+d}) \leq \epsilon$ for every $i \geq 0$.
  - Slack parameter: permitted delay d
  - Natural orderings: larger ϵ, d is easier but larger R is harder.
- Classical capacity

$$\mathcal{C}_R^f = \bigcap_{\epsilon>0} \bigcap_{R'<R} \bigcup_{d>0} \{f_c | f_c \text{ solves } A^f_{R',\epsilon,d}\}$$

$$C_{\text{Shannon}}(f_c) = \sup\{R > 0 | f_c \in \mathcal{C}_R\}$$

Classical relationships

- Feedback doesn't change capacity for memoryless channels $\mathcal{C}^m$

$$\mathcal{C}_R^{nf} \cap \mathcal{C}^m = \mathcal{C}_R^f \cap \mathcal{C}^m$$

- Zero-error capacity

$$\mathcal{C}_{0,R}^f = \bigcap_{R'<R} \bigcup_{d>0} \{f_c | f_c \text{ solves } A_{R',0,d}^f\}$$

$$C_0(f_c) = \sup\{R > 0 | f_c \in \mathcal{C}_{0,R}\}$$

  - Can change with feedback even for memoryless channels

$$(\mathcal{C}_{0,R}^{nf} \cap \mathcal{C}^m) \subset (\mathcal{C}_{0,R}^f \cap \mathcal{C}^m)$$

  - Zero-error problem is fundamentally harder

$$(\mathcal{C}_{0,R}^{nf} \cap \mathcal{C}^m) \subset (\mathcal{C}_{0,R}^f \cap \mathcal{C}^m) \subset (\mathcal{C}_R \cap \mathcal{C}^m)$$

Estimation with distortion: $A_{(F_X,\rho,D,d)}$

- Source: *noninteractive* X_i drawn iid from F_X
- Same information patterns: with/without feedback.
- Performance objective: $\mathcal{V}(\rho, D, d)$ is satisfied if $\lim_{n\to\infty} \frac{1}{n} E[\sum_{i=1}^{n} \rho(X_i, U_{i+d})] \leq D$.
  - Slack parameter: permitted delay d
  - Natural orderings: larger D, d is easier
- Channels that are good enough

$$\mathcal{C}^{f}_{e,(F_X,\rho,D)} = \bigcap_{D'>D} \bigcup_{d>0} \{f_c | f_c \text{ solves } A^{f}_{(F_X,\rho,D',d)}\}$$

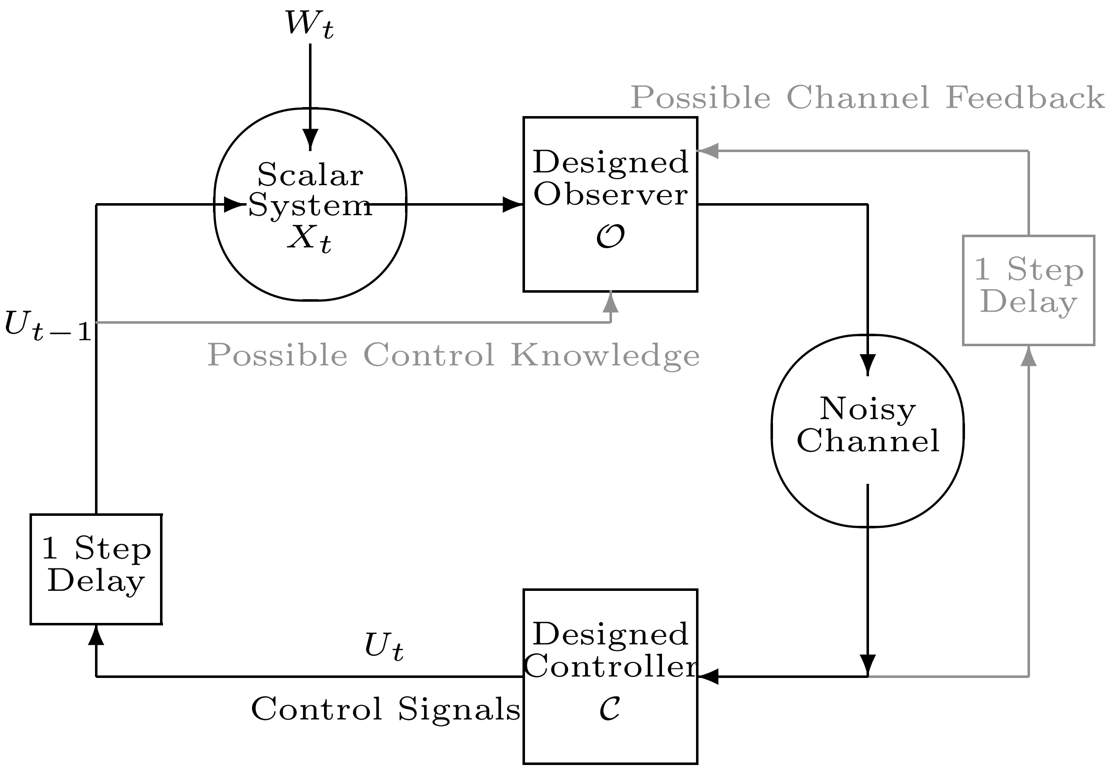
- "Separation Theorem" if ρ is finite.

$$(\mathcal{C}_{R(D)} \cap \mathcal{C}^{m}) = (\mathcal{C}^{nf}_{e,(F_X,\rho,D)} \cap \mathcal{C}^{m}) = (\mathcal{C}^{f}_{e,(F_X,\rho,D)} \cap \mathcal{C}^{m})$$

Stabilization and anytime communication

- Simple control problem
- Why classical capacity.is not enough.
- Why anytime (delay-universality) is needed
- Some simple implications (power laws, etc.)

A simple scalar distributed control problem


W_t
Scalar System X_t
Designed Observer $\mathcal{O}$
Possible Channel Feedback
1 Step Delay
U_{t-1}
Possible Control Knowledge
Noisy Channel
1 Step Delay
U_t
Control Signals
Designed Controller $\mathcal{C}$

$$X_{t+1} = \lambda X_t + U_t + W_t$$

- Unstable $\lambda > 1$, bounded initial condition and disturbance W.
- Goal: Stability = $\sup_{t>0} E[|X_t|^\eta] \leq K$ for some $K < \infty$.

Is Shannon capacity all we need?

- Consider a system with
  - $\lambda = 2$ for the dynamics
  - noisy channel that sometimes drops packets but is otherwise noiseless (Real erasure channel)

$$Z_t = \begin{cases} Y_t & \text{with Probability } \frac{1}{2} \\ 0 & \text{with Probability } \frac{1}{2} \end{cases}$$

- No other constraints, so design is obvious: $Y_t = X_t$ and $U_t = -\lambda Z_t$
- Resulting closed loop dynamics:

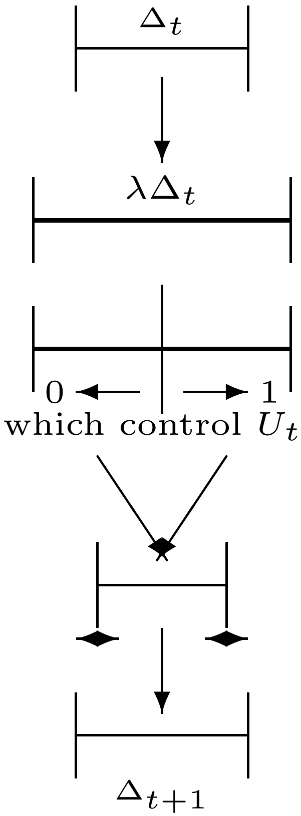
$$X_{t+1} = \begin{cases} W_t & \text{with Probability } \frac{1}{2} \\ 2X_t + W_t & \text{with Probability } \frac{1}{2} \end{cases}$$

Is the closed-loop system stable?

$$X_{t+1} = \begin{cases} W_t & \text{with Probability } \frac{1}{2} \\ 2X_t + W_t & \text{with Probability } \frac{1}{2} \end{cases}$$

- i.i.d. erasures mean arbitrarily long stretches of erasures are possible, though unlikely.
  - System is not guaranteed to stay inside any box.
  - Under stochastic disturbances, the variance of the state is asymptotically infinite.
- For worst case disturbances $W_t = 1$, the tail probability is dying off as $P(|X| > x) \approx \frac{K}{x}$.
- Meanwhile, $C = \infty$!

Run same plant X over noiseless channel



Window known to contain X_t

will grow by factor of $\lambda > 1$

Sending R bits, cut window by a factor of 2^{-R}

Encode which control U_t to apply

grows by $\frac{\Omega}{2}$ on each side

giving a new window for X_{t+1}

As long as $R > \log_2 \lambda$, we can have Δ stay bounded forever.

What is needed: key intuition

- Break state X into sum of $\check{X}$ (response to disturbance) and $\tilde{X}$ (response to control)

- Suppose $\lambda = 2$ and so $\check{X}_t = \sum_{i=0}^{t} 2^i W_{t-1}$

- Assume W_j either 0 or 1

- In binary notation: $\check{X}_t = W_0 W_1 W_2 \ldots W_{t-1}.00000\ldots$

- If $-\tilde{X}_t$ is close to $\check{X}_t$, their binary representations likely agree in all the high-order bits.
  - High-order bits represent earlier disturbances.
  - Typically, to get a difference at the W_{t-d} level, we have to be off by about 2^d.

Stabilization implies communicating bits reliably in a special fashion.

Anytime communication problems: $A_{R,\alpha,K}$

- Same as Shannon problem in source and information pattern.
- Performance objective different:
  - Reinterpret $U_t = 0.\hat{X}_0(t), \hat{X}_1(t), \hat{X}_2(t), \ldots$ in binary
  - $\mathcal{V}_{(K,\alpha)}$ is satisfied if $\mathcal{P}(X_i \neq \hat{X}_i(i+d)) \leq K2^{-\alpha d}$ for every $i \geq 0, d \geq 0$.
  - Slack parameter: constant factor K
  - Natural orderings: larger K is easier, but larger R, α are harder.
- Capacity

$$\mathcal{C}^f_{a,(R,\alpha)} = \bigcap_{R'<R} \bigcap_{\alpha'<\alpha} \bigcup_{K>0} \{f_c | f_c \text{ solves } A^f_{(R',\alpha',K)}\}$$

$$C_{\text{any}}(f_c, \alpha) = \sup\{R > 0 | f_c \in \mathcal{C}^f_{a,(R,\alpha)}\}$$

Separation theorem for control

Necessity: If a scalar system with parameter $\lambda > 1$ can be stabilized with finite η-moment across a noisy channel, then the **channel with noiseless feedback** must have

$$C_{\text{any}}(\eta \log_2 \lambda) \geq \log_2 \lambda$$

In general: If $P(|X| > m) < f(m)$, then $\exists K : P_{\text{error}}(d) < f(K\lambda^d)$

Sufficiency: If there is an $\alpha > \eta \log_2 \lambda$ for which the **channel with noiseless feedback** has

$$C_{\text{any}}(\alpha) > \log_2 \lambda$$

then the scalar system with parameter $\lambda \geq 1$ with a bounded disturbance can be stabilized across the noisy channel with finite η-moment.

What does all this imply?

- If we want $P(|X_t| > m) \leq f(m) = 0$ for some finite m, we require zero-error reliability across the channel.
- For generic DMCs, anytime reliability with feedback is upper-bounded:

$$\begin{aligned} f(K\lambda^d) &\geq \zeta^d \\ f(m) &\geq \zeta^{\frac{\log_2(\frac{m}{K})}{\log_2 \lambda}} \\ f(m) &\geq K' m^{-\frac{\log_2 \frac{1}{\zeta}}{\log_2 \lambda}} \end{aligned}$$

A controlled state can have at best a power-law tail.

- If we just want $\lim_{m \to \infty} f(m) = 0$, then just Shannon capacity $> \log_2 \lambda$ is required for DMCs.
- Almost-sure stabilization for $W_t = 0$ follows by time-varying transformation.

Stabilization and Anytime Equivalence

- With nested information: $A^f_{\lambda,\eta,K}$. Without: A^{nf}
  - Slack parameter: K (Performance)
  - Natural ordering: larger η, λ are harder, but larger K is easier.

$$\mathcal{C}^f_{s,(\lambda,\eta)} = \bigcap_{\lambda'<\lambda} \bigcap_{\eta'<\eta} \bigcup_{K>0} \{f_c | f_c \text{ solves } A^f_{(\lambda',\eta',K)}\}$$

- Equivalences

$$\mathcal{C}^{nf}_{s,(\lambda,\eta)} \subseteq \mathcal{C}^f_{s,(\lambda,\eta)} = \mathcal{C}^f_{a,(\log_2 \lambda, \eta \log_2 \lambda)}$$

$$(\mathcal{C}^{nf}_{s,(\lambda,\eta)} \cap \mathcal{C}^{\text{finite}}) = (\mathcal{C}^f_{s,(\lambda,\eta)} \cap \mathcal{C}^{\text{finite}}) = (\mathcal{C}^f_{a,(\log_2 \lambda, \eta \log_2 \lambda)} \cap \mathcal{C}^{\text{finite}})$$

Asymptotic communication problem hierarchy

- The easiest: Shannon communication
  - Asymptotically: a single figure of merit C
  - Equivalent to most estimation problems of stationary ergodic processes with bounded distortion measures.
  - Feedback does not matter.
- Intermediate families: Anytime communication
  - Multiple figures of merit: $(\vec{R}, \vec{\alpha})$
  - Feedback case equivalent to stabilization problems
  - Related nonstationary estimation problems fall here also
  - Does feedback matter?
- Hardest level: Zero-error communication
  - Single figure of merit C_0
  - Feedback matters.

Language of Probability Theory

Probability Space

$(\Omega, \mathcal{F}, P)$

$\mathcal{F} = \sigma$-field

$=$ class of subsets of Ω, closed under complementation, countable intersections, countable unions.

A nonnegative set function $P(\cdot)$ defined on $\mathcal{F}$ is a Probability Measure if

(i) $P(\Omega) = 1$

(ii) $\forall$ finite, countable collection $\{B_k\}$ of subsets of $\mathcal{F}$ s.t.

$$B_k \cap B_j = \phi \ , \ k \neq j$$

$$P(\bigcap_k B_k) = \sum_k P(B_k)$$

Given $(\Omega, \mathcal{F}, P)$ a Probability space

$f\ :\ \Omega \to \mathbb{R}$ measurable

$L^2(\Omega, \mathcal{F}, P) = \{f | \int |f|^2 dP < \infty\}$

Finite Energy signal

$\mathcal{G} \subset \mathcal{F}$, sub σ-field

$L^2(\Omega, \mathcal{G}, P)$

$E(\mathcal{F}|\mathcal{G}) =$ Projection of $L^2(\Omega, \mathcal{F}, P)$ onto $L^2(\Omega, \mathcal{G}, P)$

Conditional Expectation

“Nonlinear” Projection

$(X_t)_{t\geq 0}$: Stochastic Process

$E(X_t|X_s, s < t) = X_s$ a.s.

Martingale

Independence and Conditional Independence

σ-field generated by $(X_t)_{t \in T} =$ smallest σ-field such that $\forall$ X_t are measurable

Markov Process

Past and Future Conditionally Independent given the present

$$\sigma(X_s|s<t) \perp \sigma(X_s|s>t) \mid \sigma(X_t)$$

Past Future Present

Why we need this language:

See J.C. Willems, "Open Stochastic Systems," *IEEE Trans. on Automatic Control*, 2013.

Wiener and Kalman Filtering

"Filtering and Stochastic Control:
A Historical Perspective"
by S.K. Mitter

in *IEEE Control Systems*
pp.67-76, June 1996

Wiener and Kalman Filtering

In order to introduce the main ideas of nonlinear filtering, we first consider linear filtering theory. A rather comprehensive survey of linear filtering theory was undertaken by Kailath in [1] and therefore we shall only expose those ideas which generalize to the nonlinear situation.

Suppose we have a signal process (z_t) and an orthogonal increment process (w_t), the noise process and we have the observation equation

$$y_t = \int_0^t z_s ds + w_t \ . \tag{23}$$

Note that if (w_t) is Brownian motion then this represents the observation

$$\widehat{y} = z_t + \eta_t \ . \tag{24}$$

where η_t is the formal (distributional) derivative of Brownian motion and hence it is white noise.

We make the following assumptions.

(A1) (w_t) has stationary orthogonal increments

(A2) (z_t) is a second-order q.m. continuous process

(A3) For $\forall s$ and $t > s$

$$(w_t - w_s) \perp H_s^{w,z}$$

where $H_s^{w,z}$ is the Hilbert space spanned by $(w_\tau, z_\tau | \tau \leq s)$.

The last assumption is a causality requirement but includes situations where the signal z_s may be influenced by past observations as would typically arise in feedback control problems. A slightly stronger assumption is

(A3$'$) $H^w \perp H^z$

which states that the signal and noise are uncorrelated, a situation which often arises in communication problems.

The situation which Wiener considered corresponds to (24), where he assumed that (z_t) is a stationary, second-order, q.m. continuous process.

The *filtering* problem is to obtain the best linear estimate $\widehat{z}_t$ of z_t based on the past observations $(y_s|s \leq t)$. There are two other problems of interest, namely, *prediction*, when we are interested in the best linear estimate $\widehat{z}'_r, r > t$ based on observations $(y_s|s \leq t)$ and *smoothing*, where we require obtaining the best linear estimate $\widehat{z}'_r, r < t$ based on observations $(y_s|s \leq t)$.

Abstractly, the solution to the problem of filtering corresponds to explicitly computing

$$\widehat{z}_t = P_t^y(z_t) \tag{25}$$

where P_t^y is the projection operator onto the Hilbert space H_t^y. We proceed to outline the solution using a method originally proposed by Bode and Shannon [2] and later presented in modern form by Kailath [3]. For a textbook account see Davis [4] and Wong [5], which we largely follow.

Let us operate under the assumption (A3)$'$, although all the results are true under the weaker assumption (A3). The key to obtaining a solution is the introduction of the innovations process

$$\nu_t = y_t - \int_0^t \widehat{z}_s ds \ . \tag{26}$$

The following facts about the innovations process can be proved:

(F1) ν_t is an orthogonal increment process.

(F2) $\forall s,\ \forall t > s$

$$\nu_t - \nu_s \perp H_s^y \quad \text{and} \quad \text{cov}(\nu_t) = \text{cov}(w_t)$$

(F3) $H_t^y = H_t^\nu$.

The name “innovations” originates in the fact that the optimum filter extracts the maximal probabilistic information from the observations in the sense that what remains is essentially equivalent to the noise present in the observation. Furthermore, (F3) states that the innovations process contains the same information as the observations. This can be proved by showing that the linear transformation relating the observations and innovations is causal and causally invertible.

As we shall see later, these results are true in a much more general context. To proceed further, we need a concrete representation of vectors residing in the Hilbert space H_t^y. The important result is that every vector $Y \in H_t^y$ can be represented as

$$Y = \int_0^t \beta(s) dy_s \tag{27}$$

where β is a deterministic square integrable function and the above integral is a stochastic integral.

For an account of stochastic integrals see the book of Wong [5]. Now using the Projection Theorem, (27), and (F1)–(F3) we can obtain a representation theorem for the estimate $\widehat{z}_t$ as:

$$\widehat{z}_t = \int_0^t \left(\frac{\partial}{\partial s} E(z_t \nu_s') d\nu_s \right) \ . \tag{28}$$

What we have done so far is quite general. As we have mentioned, Wiener assumed that (z_s) was a stationary q.m. second-order process, and he obtained a linear integral representation for the estimate where the kernel of the integral operator was obtained as a solution to an integral equation, the Wiener–Hopf equation.

As Wiener himself remarked, an effective solution to the Wiener–Hopf equation using the method of spectral factorization (see, for example, Youla [6]) could only be obtained when (z_S) had a rational spectral density.

In his fundamental work, Kalman [7,8,9] made this explicit by introducing a Gauss–Markov diffusion model for the signal

$$\begin{cases} dx_t = Fx_t dt + G d\beta_s \\ z_t = Hx_t \end{cases} \tag{29}$$

where x_t is an n-vector-valued Gaussian random process, w_t is m-dimensional Brownian motion, z_t is a p-vector-valued Gaussian random process, and F, G, and H are matrices of appropriate order.

We note that (29) is actually an integral equation

$$x_t = x_0 + \int_0^t F x_s ds + \int_0^t G d\beta_s \tag{30}$$

where the last integral is a stochastic integral. The Gauss–Markov assumption is no loss of generality since in Wiener's work the best linear estimate was sought for signals modeled as second-order random processes. The filtering problem now is to compute the best estimate (which is provably linear)

$$\widehat{x}_t = P_t(x_t) \ . \tag{31}$$

Moreover, in this new setup no assumption of stationarity is needed. Indeed the matrices F, G, and H may depend on time. The derivation of the Kalman filter can now proceed as follows. First note that

$$\widehat{x}_t = \int_0^t \frac{\partial}{\partial s} E(x_t \nu_s') d\nu_s \ , \tag{32}$$

(See (28).)

Now we can show that

$$\widehat{x}_t - \widehat{x}_0 - \int_0^t F\widehat{x}_s ds = \int_0^t K(s) d\nu_s \ . \qquad (33)$$

where $K(s)$ is a square integrable matrix-valued function. This is analogous to the representation theorem given by (27). Eq. (33) can be written in differential form as

$$d\widehat{x}_t = F\widehat{x}_t dt + K(t) d\nu_t \qquad (34)$$

and let us assume that $\widehat{x}_0 = 0$.

The structure of (34) shows that the Kalman Filter incorporates a model of the signal and a correction term, which is an optimally weighted error $= K(t)(dy_t - \widehat{z}_t dt)$ (see Fig. 1).

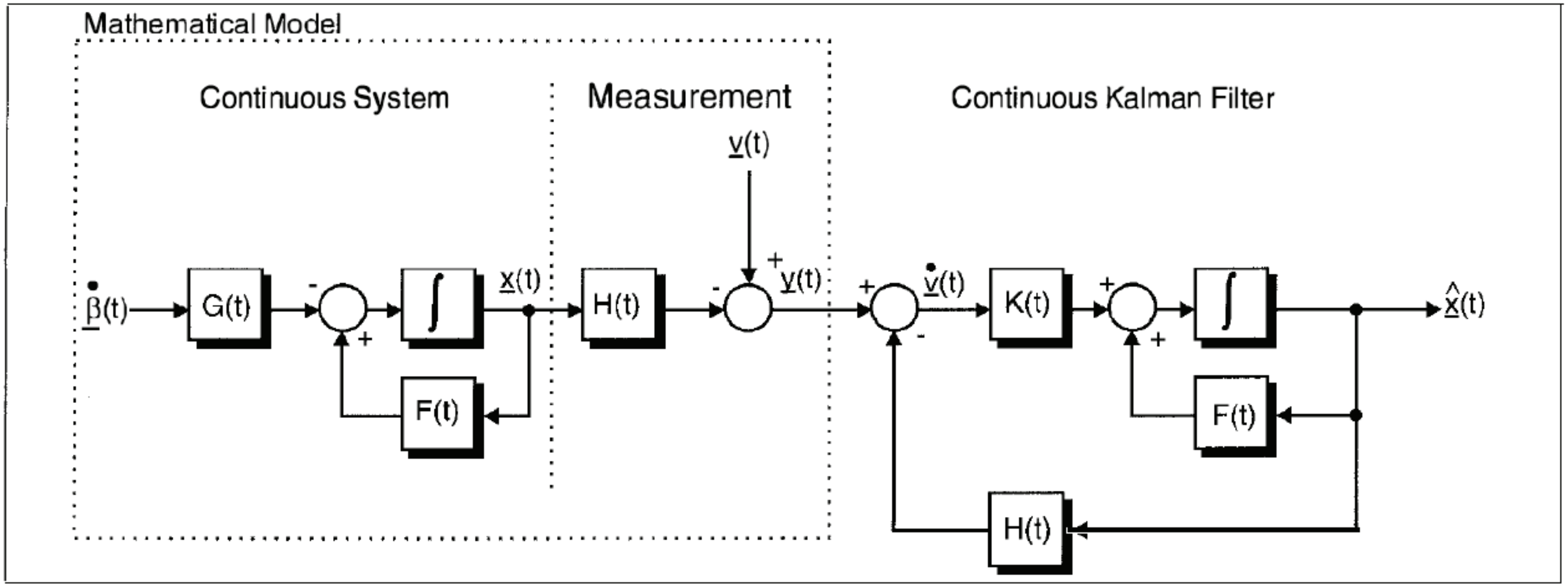


Fig. 1. System model and continuous Kalman filter.

It remains to find an explicit expression for $K(t)$. Here we see an interplay between filtering theory and linear systems theory. The solution of (34) can be written as

$$\widehat{x}_t = \int_0^t \Phi(t,s)K(s)d\nu_s \tag{35}$$

where $\Phi(t,s)$ is the transition matrix corresponding to F.

From (32) and (35)

$$\Phi(t,s)K(s) = \frac{\partial}{\partial s}E(x_t \nu_s')$$

and hence

$$K(t) = \frac{\partial}{\partial s}E(x_t \nu_s')\,|_{s=t} \quad .$$

Some further calculations, using the fact that $x_t \perp H_s^w$, show that

$$K(t) = P(t)H' \ ,$$

where $P(t) = E(\tilde{x}_t \tilde{x}_t')$, $\tilde{x}_t = x_t - \widehat{x}_t$.

Finally, using the representation of solutions of the linear stochastic differential equations (29) and using (34), we can write a linear stochastic differential equation for $\tilde{x}_t$ and write down a representation for $P(t) = E(\tilde{x}_t \tilde{x}_t')$ as

$$\begin{aligned} P(t) &= \psi(t,0)P(0)\psi'(t,0) + \int_0^t \psi(t,s)GG'\psi'(t,s)ds \\ &\quad + \int_0^t \psi(t,s)P(s)H'HP(s)\psi'(t,s)ds \qquad (36) \end{aligned}$$

where $\psi(t,s)$ is the transition matrix corresponding to $(F - PH'H)$.

There is again a role of linear systems theory evident here. Differentiating w.r. to t, we get a matrix differential equation for $P(t)$, the matrix Ricatti equation

$$\begin{aligned} \frac{dP}{dt} &= GG' - P(t)H'HP(t) + FP(t) + P(t)F' \\ P(0) &= \mathrm{cov}(x_0) = P_0 \ . \end{aligned} \tag{37}$$

Note that $K(t) = P(t)H'$ is deterministic and does not depend on the observation process y_t, and hence can be pre-computed.

The approach to the solution of the Wiener Filtering Problem consists in studying the equilibrium behavior of $P(t)$ as $t \to \infty$. There is again a beautiful interplay between the infinite time behavior of the filter and the structural properties of (29). One can prove that if the pair (F, G) is stabilizable and (H, F) is detectable then $P(t) \to \overline{P}$ as $t \to \infty$ where $\overline{P}$ is the unique non-negative solution to the algebraic Ricatti equation corresponding to (37) and that $F - \overline{P}H'H$ is a stability matrix.

Thus the filter is stable, in the sense that the error covariance converges to the optimal error covariance for the stationary problem even if F is not a stability matrix. For the linear systems concepts introduced here and the proof of the above results the reader may consult Wonham [10].

In a Control context, the controls enter the filter as a separate input and one needs to study the controlled filtering problem. This is important for proving the Separation Principle.

On Some Connections between Nonlinear Filtering, Information Theory, and Statistical Mechanics

Sanjoy K. Mitter

Laboratory for Information and Decision Systems
Massachusetts Institute of Technology

Joint work with Nigel Newton, University of Essex

Some Connections between Information Theory, Filtering and Statistical Mechanics

Variational Approach to Bayesian Estimation

Stochastic Control Interpretation of Nonlinear Filtering

Preliminaries

X, Y discrete random variables with joint distribution P_{XY} and marginals P_X and P_Y

$$I(X;Y) = E_{P_{XY}}\left(\log \frac{P_{XY}}{P_X \otimes P_Y}\right) \quad : \quad \text{Mutual Information}$$

Average measure of dependence of two random variables

Mutual Information is an example of the general notion of relative entropy between two measures μ and ν on some probability space $(\Omega, \mathcal{F}, P)$ (discrete for the moment)

$$h(\mu|\nu) = E_\mu \log\left(\frac{\mu}{\nu}\right)$$

Properties:

(i) $h(\mu|\nu) \geq 0$

(ii) $h(\mu|\nu) = 0 \Leftrightarrow \mu = \nu$

(iii) $h(\mu|\nu)$ jointly convex in μ, ν

(But, not symmetric). Defines a pseudo-distance between two measures μ and ν.

We will have to deal with random variables in a more general setting.

Nonlinear Dynamical Systems forced by (scaled) white noise

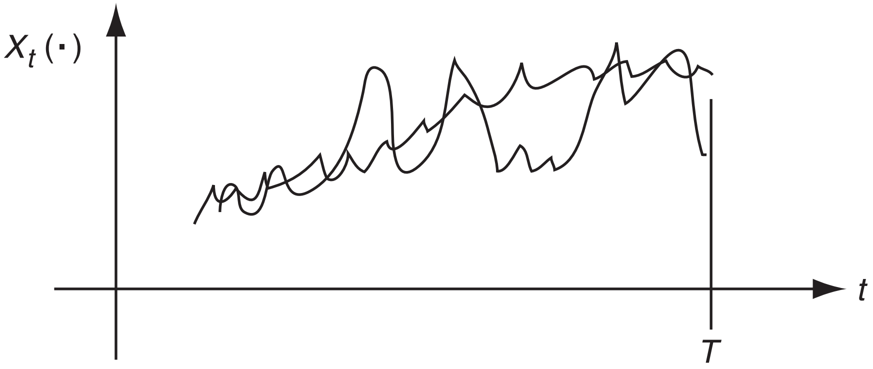
$$\frac{dx_t}{dt} = b(x_t) + \sigma(x_t)\dot{v}_t \ ,$$

where v_t: Brownian motion and $\dot{v}_t =$ white noise, formal derivative of Brownian motion

Rewrite as Integral equation

$$\begin{aligned} x_t &= x_0 + \int_0^t b(x_s)ds + \int_0^t \sigma(x_t)\dot{v}_t dt \\ &= x_0 + \int_0^t b(x_s)ds + \int_0^t \sigma(x_t)dv_t \ \leftarrow \text{Ito integral} \end{aligned}$$

We want to think of $x_{(\cdot)} := X$ as a map (random variable) from $(\Omega, \mathcal{F}, P)$ to $(\mathcal{X}, \mathcal{B}(\mathcal{X})$ where $\mathcal{X} = \mathcal{C}(0, \mathcal{T}; \mathbb{R})$ and $\mathcal{B}(\mathcal{X})$ is the Borel field associated with $\mathcal{X}$. We call the probability measure of $X \in \mathcal{P}(\mathcal{X})$ the path space measure



X is a random trajectory

Sometimes, we would want to look at these random trajectories "through" a different measure $\widehat{P}$ (instead of P) in order for it to "appear" differently, for example, trajectories of Brownian Motion.

Gibbs Measures:

Variational Characterization for Finite Systems

(H.O. Georgii: *Gibbs Measures and Phase Transitions*, Chapter 15)

Let $S =$ finite set, and $E =$ state space, finite set and let $\Omega = E^S$, finite.

Let Φ be any potential, and $H = \sum_{A \subset S} \Phi_A(w)$ be the associated Hamiltonian

The unique Gibbs measure for Φ is given by

$$\nu(\omega) \ = \ Z^{-1} \text{exp}[-H(\omega)] \ , \ \omega \in \Omega$$

where

$$Z \ = \ \sum_{\omega \in \Omega} \text{exp}[-H(\omega)] \ : \ \text{Partition function}$$

For each probability measure μ on Ω,

$$\mu(H) = \sum_{\omega\in\Omega} \mu(\omega)H(\omega) \text{ and } h(\mu) = -\sum_{\omega\in\Omega} \mu(\omega)\log\mu(\omega)$$

be the Energy and Entropy associated with μ

Then

$$\mu(H) - h(\mu) + \log Z = h(\mu|\nu) \geq 0$$

$$h(\mu|\nu) = 0 \Leftrightarrow \mu = \nu \qquad \square$$

$$F(\mu) = \mu(H) - h(\mu) \;\;:\;\; \text{Free Energy}$$

$$F(\nu) = -\log Z$$

Generalization of these ideas to infinite systems leads to characterization of translation-invariant Gibbs measures as minimization of Specific Free Energy. A modification of these ideas (using Exchangeability) leads to a proof of the Noisy Channel Coding Theorem (BSC).

Variational Bayes and a Problem of Reliable Communication, Part II, N. Newton, S.K. Mitter, *J. Stat. Mech.: Theory and Experiment*, Iss. 11, pp. 111008, 2012

Information Theory, Filtering and Statistical Mechanics

$(X_t)_{t\geq 0}$ Markov Process, time homogeneous

$$P(X_t \in B | X_r, r \in [0,s]) = \pi(t-s, X_s, B) \quad 0 \leq s \leq t \leq T$$

P_t is the distribution of X_t with density p_t

$$P_t(B) = P(X_t \in B) = \int_B p_t(x)\lambda_x(dx) \quad \lambda_x \text{ : ref. measure}$$

Diffusion

$$\begin{aligned}
(Ap)(x) &= \frac{1}{2}\sum_{i,j} \frac{\partial^2 (a_{i,j}p)}{\partial x_i \partial x_j}(x) - \sum_i \frac{\partial}{\partial x_i}(b_i p)(x) \text{ on } \mathbb{R}^d \\
X_t &= X_0 + \int_0^t b(X_s)dt + \int_0^t \sigma(X_s)dv_s \\
a &= \sigma\sigma'
\end{aligned}$$

Relative Entropy

$$\begin{aligned} h(\mu|\lambda) &= \int_X q(x)\log q(x)\lambda(dx) \quad \mu \text{ has density } q \text{ w.r.t. } \lambda \\ &= +\infty \quad \text{otherwise} \\ \langle f, \lambda\rangle &= \int_X f(x)\lambda(dx) \end{aligned}$$

Σ_x: statistical mechanical system, associated with $(X_t)_{t\geq 0}$

P_t: state of Σ_x at time t

P_{SS}: unique invariant measure with density p_{SS}

Internal Energy $\mathcal{E}_X(P_t) = \langle H_x, P_t\rangle$

Entropy $S_x(P_t) = -h(P_t|\lambda_x)$

Free Energy $\mathcal{F}_X(P_t) = \mathcal{E}_x(P_t) - S_x(P_t)$

Energy Function $H_x(x) = -\log p_{SS}(x)$

Choice assures Energy Function is a Gibbs measure for Σ_x

Proposition:

(i) Unique minimizer of Free Energy of Σ_x is P_{SS}

(ii) $\mathcal{F}_x(P_{SS}) = 0$

(iii) Free Energy of Σ_x is non-increasing

Proof.

$$\mathcal{F}(x)(P_t) = h(P_t|P_{SS}) \Rightarrow \text{(i) and (ii)}$$

To prove (iii), $P^{(2)}_{s,t}$ = two point joint distribution

$P^{(2)}_{s,t}(B, C) = P(X_s \in B, X_t \in C) = \int_B \pi(t - s, X, C) P_s(dx)$

$P^{(2)}_{s,t,SS}$ = joint distribution when $P_s = P_{SS}$

Chain rule for Relative Entropy □

$$\begin{aligned} & h(P^{(2)}_{s,t}|P^{(2)}_{s,t,SS}) \\ & \quad = h(P_t|P_{SS}) + \int h(\tilde{\Pi}(t,s,x,\cdot\,)|\tilde{\Pi}_{SS}(t-s,x,\cdot\,))P_t(dx) \\ & \quad \geq h(P_t|P_{SS}) \end{aligned}$$

(Chain Rule)

where $\tilde{\Pi}(t,s,x,\cdot\,)$ = regular $(X_t = x)$-conditional distribution for X_s under the joint distribution $P^{(2)}_{s,t}$ and $\tilde{\Pi}_{SS}(t-s,x,\cdot\,)$ is the equivalent under the joint distribution $P^{(2)}_{s,t,SS}$.

Σ_x: one component of a two-component energy conserving system that includes a unit temperature heat bath with which Σ_x interacts

If Entropy of system $=$ Entropy of the sum of two components then any change in this entropy resulting from the evolution of P_t $=$ neg. of corresponding change in $\mathcal{F}_x(P_t)$

P_{SS}: unique invariant measure with density p_{SS}

Proposition: Entropy of closed system is maximized by P_{SS} and non-decreasing

Assertion (iii) in Proposition can be thought of as a Second Law of Thermodynamics for Σ_x

Observations (Interaction with Measurements)

$$Y_t = \int_0^t g(X_s)ds + W_t$$

$$E\left[\int_0^t |g(X_t)|^2 dt < \infty\right.$$

$(Z_t | t \in [0, T])$: regular conditional probability of X_t given $(Y_s | 0 \leq s \leq t)$

ξ_t: density

$$\xi_t(x) = \xi_0(x) + \int_0^t (\mathcal{A}\xi_s)(x)ds + \int_0^t \xi_s(x)(g(x) - \langle g, Z_s\rangle)' d\nu_s \tag{1}$$

$$\nu_t = Y_t - \int_0^t \langle g, Z_s\rangle) ds \quad \text{Innovations}$$

We want to study the Information flow from the initial state and running observations $(Y_s | 0 \leq s \leq t)$ into the regular conditional distribtution

$$P_{X_t|(Y_s, 0 \leq s \leq t)}\ (\,\cdot\,, y)$$

(the filter).

Is this flow, conservative, dissipative?

Information Theoretic Quantities

$$
\begin{aligned}
S(t) &= I((X_s, s \in [0,T]); Y_s, s \in [0,t]) = \text{supply} \\
C(t) &= I((X_s, s \in [t,T]); Y_s, s \in [0,t]) = \text{storage} \\
D(t) &= S(t) - C(t) = \text{dissipation}
\end{aligned}
$$

Proposition

$$
\begin{aligned}
S(t) &= C(0) + \frac{1}{2} E \int_0^t |g(X_s) - \langle g, Z_s \rangle|^2 ds \\
C(t) &= I(X_t; Z_t) = Eh(Z_t|P_t) \\
D(t) &= EI((X_s, s \in [0,t]); Y_s, s \in [0,t]|X_t)
\end{aligned}
$$

$$\dot{S}(t) = \frac{1}{2}E|g(X_t) - \langle g, Z_t \rangle|^2 \tag{2}$$

$$\dot{D}(t) = E\left(\frac{Ap_t}{p_t}\log p_t - \frac{A\xi_t}{\xi}\log \xi_t\right)(X_t) \tag{3}$$

Sensitivity of Mutual Information $C(t)$ to the randomization in the dynamics of the signal

For Diffusions

$$\dot{D}(t) = \frac{1}{2}E\nabla \log\left(\frac{\xi_t}{p_t}\right)' a\nabla \log\left(\frac{\xi_t}{p_t}\right)(X_t)$$

Rate of change of storage can be found by application of Ito's rule to

$$\xi_t \log\left(\frac{\xi_t}{p_t}\right)(X_t)$$

Equations (2) and (3) show that the supply of information is associated with the second integral in (1)

$$\int_0^t \xi_s(x)(g(x) - \langle g, Z_s\rangle)' d\nu_s$$

and the dissipation associated with the first integral in (1)

$$\int_0^t (\mathcal{A}\xi_s)(x) ds$$

$\dot{S}(t) =$ signal to noise power ratio of the observations and $\dot{D}(t) =$ measure of the rate at which X forgets its past

Notes on Proof:

$$\begin{aligned} C(t) &= I(X_t; Y_s; s \in [0,t]) = I(X_t; Z_t) \\ S(t) &= E \log M_t \ , \end{aligned}$$

where

$$\begin{aligned} M_t &= \frac{dZ_0}{dP_0}(x_0) \exp\left(\int_0^t g(x_s) - \langle g, Z_s \rangle \right)' dw_s \\ &\quad + \frac{1}{2} \int_0^t |g(x_s) - \langle g, Z_s \rangle|^2 ds) \end{aligned}$$

Interactive Statistical Mechanics

The conditional distribution Z_t takes into account the partial observations available up to time t. Define an energy function for $\Sigma_{X|Z}$ in such a way that Z_t is the minimum free-energy state at time t.

Let $(\tilde{Z}_t)$ be a stochastic process that satisfies the filter equation $(\tilde{Z}_t \neq Z_0)$ with density $(\tilde{\xi}_t)$.

$E\tilde{\xi}_t$ corresponds to a state of Σ_X and satisfies the Fokker–Planck equation.

Define energy function

$$\begin{aligned}
H_{X|Z}(x,t) &= -\log \xi_t(x) \\
E_{X|Z}(\tilde{Z}_t, t) &= \langle H_{X|Z}(\,\cdot\,, t), \tilde{Z}_t \rangle \\
S_{X|Z}(\tilde{Z}_t) &= S_X(\tilde{Z}_t) = -h(\tilde{Z}_t|\lambda_X) \\
\mathcal{F}_{X|Z}(\tilde{Z}_t, t) &= \mathcal{E}_{X|Z}(\tilde{Z}_t, t) - S_{X|Z}(\tilde{Z}_t)
\end{aligned}$$

Proposition

(i) Unique minimizer of the free energy of the conditional system $\Sigma_{X|Z}$ at time t in the state Z_t

(ii) $\mathcal{F}_{X|Z}(Z_t, t) = 0 \ \forall \ t$

(iii) If $E\mathcal{F}_{X|Z}(\tilde{Z}_t, t) < \infty$ and $h(\tilde{\Phi}_0|\Phi_0) < \infty$, where $\tilde{\Phi}_0$ and Φ_0 are the distributions of Z_0 and $\tilde{Z}_0$, then the Free Energy of $\Sigma_{X|Z}$ as state $\tilde{Z}_t$ evolves in a positive $(Y_s, s \in [0, t])$ supermartingale.

Item (iii) is like a Conditional Second Law.

We can study the statistical mechanics of the joint system (X, Z). Connection to Bayesian Inference as Free-Energy Minimization

Data Assimilation $\equiv$ Path Estimation or Filtering
or Prediction

Nonlinear Filtering: The Innovations Viewpoint

Stochastic Partial Differential Equation for the Evolution
of the Conditional Density

The Variational Viewpoint:
Information-theoretic Interpretation

Connections to Stochastic Control

Non-equilibrium Statistical Mechanics

2. A Variational Formulation of Bayesian Estimation

Let $(\Omega, \mathcal{F}, P)$ be a probability space, $(\mathbf{X}, \mathcal{X})$ and $(\mathbf{Y}, \mathcal{Y})$ Borel spaces, and $X : \Omega \to \mathbf{X}$ and $Y : \Omega \to \mathbf{Y}$ measurable mappings with distributions P_X, P_Y and P_{XY} on $\mathcal{X}$, $\mathcal{Y}$ and $\mathcal{X} \times \mathcal{Y}$, respectively. Suppose that:

(H1) there exists a σ-finite (reference) measure, λ_Y, on $\mathcal{Y}$ such that $P_{XY} \ll P_X \otimes \lambda_Y$. (This could be P_Y itself.)

Let $Q : \mathbf{X} \times \mathbf{Y} \to [0, \infty)$ be a version of the associated Radon-Nikodym derivative, and

$$\bar{\mathbf{Y}} = \left\{ y \in \mathbf{Y} : 0 < \int_{\mathbf{X}} Q(x, y) P_X(dx) < \infty \right\}; \qquad (4)$$

then $\bar{\mathbf{Y}} \in \mathcal{Y}$ and $P_Y(\bar{\mathbf{Y}}) = 1$. Let $H : \mathbf{X} \times \mathbf{Y} \to (-\infty, +\infty]$ be defined by

$$\begin{array}{rcll} H(x,y) & = & -\log(Q(x,y)) & \text{if } y \in \bar{\mathbf{Y}} \\ & & 0 & \text{otherwise :} \end{array} \qquad (5)$$

then $P_{X|Y} : \mathcal{X} \times \mathbf{Y} \to [0,1]$, defined by

$$P_{X|Y}(A,y) = \frac{\int_A \exp(-H(x,y)) P_X(dx)}{\int_{\mathbf{X}} \exp(-H(x,y)) P_X(dx)}, \qquad (6)$$

is a *regular conditional probability distribution* for X given Y; i.e.

$P_{X|Y}(\,\cdot\,, y)$ is a probability measure on $\mathcal{X}$ for each y,

$P_{X|Y}(A, \cdot\,)$ is $\mathcal{Y}$-measurable for each A, and

$P_{X|Y}(A, Y) = P(X \in A\,|\,Y)$ a.s.

Eqs. (4)–(6) constitute an ‘outcome-by-outcome’ abstract Bayes formula, yielding a posterior probability distribution for X for each outcome of Y.

Let $\mathcal{P}(\mathcal{X})$ be the set of probability measures on $(\mathbf{X}, \mathcal{X})$, and $\mathcal{H}(\mathbf{X})$ the set of $(-\infty, +\infty]$-valued, measurable functions on the same space. For $\tilde{P}_X, \hat{P}_X \in \mathcal{P}(\mathcal{X})$ and $\tilde{H} \in \mathcal{H}(\mathbf{X})$, we define

$$h(\tilde{P}_X \,|\, \hat{P}_X) = \int_{\mathbf{X}} \log\left(\frac{d\tilde{P}_X}{d\hat{P}_X}\right) d\tilde{P}_X \quad \text{if } \tilde{P}_X \ll \hat{P}_X \text{ and the integral exists} \tag{7}$$

$$+\infty \qquad \text{otherwise,}$$

$$i(\tilde{H}) = -\log\left(\int_{\mathbf{X}} \exp(-\tilde{H}) dP_X\right) \quad \text{if } 0 < \int_{\mathbf{X}} \exp(-\tilde{H}) dP_X < \infty \tag{8}$$

$$-\infty \qquad \text{otherwise,}$$

$$\langle \tilde{H}, \tilde{P}_X \rangle = \int_{\mathbf{X}} \tilde{H} d\tilde{P}_X \qquad \text{if the integral exists} \tag{9}$$

$$+\infty \qquad \text{otherwise.}$$

It is well known that the relative entropy $h(\tilde{P}_X \,|\, \hat{P}_X)$ can be interpreted as the *information gain* of the probability measure $\tilde{P}_X$ over $\hat{P}_X$. In fact, any version of $-\log(d\tilde{P}_X/d\hat{P}_X)$ is a generalisation of the Shannon information for X. For almost all x, it is a measure of the 'relative degree of surprise' in the outcome $X = x$ for the two distributions $\tilde{P}_X$ and $\hat{P}_X$. Thus, $h(\tilde{P}_X \,|\, \hat{P}_X)$ is the average *reduction* in the degree of surprise in this outcome arising from the acceptance of $\tilde{P}_X$ as the distribution for X, rather than $\hat{P}_X$.

If we interpret exp$(-\tilde{H})$ as a likelihood function for X, associated with some (unspecified) observation, then $\tilde{H}(x)$ is the 'residual degree of surprise' in that observation if we already know that $X = x$, and $i(\tilde{H})$ is the 'total degree of surprise' in that observation, i.e. the information in the unspecified observation if all we know about X is its prior P_X. In what follows we shall call $\tilde{H}(X)$ the X*-conditional information* in the unspecified observation, and $i(\tilde{H})$ the information in that observation. (Of course, $H(X, y)$ and, respectively, $i(H(\,\cdot\,, y))$ are the X-conditional information and, respectively, information in the observation that $Y = y$.)

Theorem 1

(i) $i\,((H(\,\cdot\,,y)) =$ min$_{\tilde{P}_X}[h(\tilde{P}_X|P_X) + \langle H(\,\cdot\,,y), \tilde{P}_X\rangle]$

(ii) $h(P_{X|Y}(\,\cdot\,,y)|P_X) =$ max$_{\tilde{H}}\left\{i(\tilde{H}) - \langle \tilde{H}, P_{X|Y}(\,\cdot\,,y)\right\}$

(iii) $P_{X|Y}(\,\cdot\,,y)$ *is the unique minimizer in* (i)

(iv) *If* H^* *is a maximizer in* (ii), *then* $\exists K \in \mathbb{R}$ *s.t.* $H^*(X) = H(\mathbf{X}, y) + K$

Conceptualization

Information Processing over and above that in prior P_X

In (i): Source of additional information is $Y = y$

Bayes Formula: Extracts info. pertinent $h(P_{X|Y}(\cdot\,, y)|P_X)$ and leaves *residual* $\langle H, P_{X|Y}\rangle$.

Input information is held in likelihood exp$(-H(\cdot\,, y))$ and extracted information in $P_{X|Y}(\cdot\,, y)$

Arbitrary Information procedure that postulates $\tilde{P}_X$ as post-obs. distribution has access to additional information. Hence: the notion Apparent Information.

In (ii): Source of additional information in Posterior Distribution $P_{X|Y}(\,\cdot\,, y)$. The aim now is to postulate an observation, i.e. a likelihood function $\exp(-\tilde{H})$ which gives rise to this observation.

Input Information

$$h\left(P_{X|Y}(\,\cdot\,,y)|P_X\right)$$

is *merged* with the residual information of the postulated observation

$$\begin{aligned} \langle \tilde{H}, P_{X|Y}(\,\cdot\,,y)\rangle \quad &: \\ \text{Result} \;&\geq\; i(\tilde{H}) \\ \text{With equality} \;&\Leftrightarrow\; \text{Obs. is compatible with } P_{X|Y} \end{aligned}$$

$$\begin{aligned} i(\tilde{H}) - \langle \tilde{H}, P_{X|Y}(\,\cdot\,,y)\rangle \\ &= \text{Inf. in Postulated Obs.} \\ &\quad\; \text{compatible with } P_{X|Y}(\,\cdot\,,y) \end{aligned}$$

Compatible Inf. of exp($-\tilde{H}$)

3. Path estimation and the Stochastic Control View

3.1. Path estimators

The techniques of Section 2 are specialized here for the case in which the estimand, X, and observation, Y , are, respectively, continuous $\mathbb{R}^n$- and $\mathbb{R}^d$-valued processes governed by the following Itô integral equations:

$$\begin{aligned} X_t &= X_0 + \int_0^t b(X_s, s)\, ds + \int_0^t \sigma(X_s, s)\, dV_s, \quad \text{for } 0 \le t \le T, \\ X_0 &\sim \mu, \end{aligned} \tag{10}$$

$$Y_t = \int_0^t g(X_s)\, ds + W_t \quad \text{for } 0 \le t \le T, \tag{11}$$

where $X_t, V_t \in \mathbb{R}^n$, μ is a law on $(\mathbb{R}^n, \mathcal{B}^n)$, $Y_t, W_t \in \mathbb{R}^d$, and b, σ and g are measurable mappings.

Under suitable regularity conditions, these equations will be unique in law and have a weak solution

$$[\Omega, \mathcal{F}, (\mathcal{F}_t), P, (V, W), (X, Y)] \ ,$$

i.e., a filtered probability space supporting an $(n + d)$-dimensional Brownian motion (V, W) and an $(n + d)$-dimensional semimartingale (X, Y) such that (10) and (11) are satisfied for all t. The abstract spaces $(\mathbf{X}, \mathcal{X})$ and $(\mathbf{Y}, \mathcal{Y})$ now become the spaces $(C([0, T]; \mathbb{R}^n), \mathcal{B}_T)$ and $(C([0, T]; \mathbb{R}^d), \mathcal{B}_T)$ of continuous functions, topologized by the uniform norm. We continue to use the notation $(\mathbf{X}, \mathcal{X})$ and $(\mathbf{Y}, \mathcal{Y})$, though, for the sake of brevity.

Let λ_Y be Wiener measure on $(\mathbf{Y}, \mathcal{Y})$. Under suitable conditions on μ, b, σ and g, we might expect the technical hypothesis for Theorem 1 to be satisfied and the mutual information, $\mathsf{E}\,\log[dP_{XY}/d(P_X \otimes \lambda_Y)(X,Y)]$, to be finite. This will allow us to proceed as in Section 2 to construct a function H on $X \times Y$, and a corresponding regular conditional probability, $P_{X|Y}$, holds for all y. Furthermore, if we can show that $P_{X|Y}(\,\cdot\,, y) \sim P_X$, then we shall be able to construct a continuous, strictly positive martingale M_y on Ω such that

$$M_{y,t} = \mathsf{E}\left(\frac{dP_{X|Y}(\,\cdot\,,y)}{dP_X}(X) \,|\, \mathcal{F}_t^X\right) \quad \text{for } 0 \le t \le T,$$

where $(\mathcal{F}_t^X)$ is the filtration generated by the process X. It will then follow from the Cameron–Martin–Girsanov theory that

$$\begin{aligned} M_{y,t} \;=\; & M_{y,0} \exp\Big(\int_0^t U'_{y,s}(dX_s - b(X_s,s)\,ds) \\ & -\frac{1}{2}\int_0^t |\sigma(X_s,s)'U_{y,s}|^2\,ds\Big) \end{aligned} \tag{12}$$

for some progressively measurable, $\mathbb{R}^n$-valued process U_y. $P_{X|Y}(\,\cdot\,,y)$ will then be the distribution of a *controlled* process, X_y, satisfying an equation like (10), but with a different initial law, and with a control term, $\sigma\sigma'(X_s,s)U_{y,s}$, entering the drift coefficient.

$$\tilde{X}_t = X_0 + \int_0^t b(\tilde{X}_s, s)ds + \int_0^t \sigma\sigma'(\tilde{X}_s, s)U_{y,s}ds + \int_0^t \sigma(\tilde{X}_s, s)dv_s$$

with $0 \le t \le T$.

The use of the progressively measurable control $\tilde{U}$ instead of U_y will result in a process $\tilde{X}$ having a distribution whose apparent information relative to $[P_X, H(\,\cdot\,, y)]$ is greater than or equal to that of X_y. Thus, at least in part, the variational characterization of Section 2 will become a problem in stochastic optimal control.

It turns out that the Path Estimation Problem can be solved in the following way:

Run a backward likelihood filter starting at the end time to estimate the initial distribution of the state. In the process, some information is dissipated at an optimal rate governed by the Fisher Information[†].

The dissipated information is recovered by running a forward optimal stochastic control problem. The resulting optimal path-space measure is the conditional path estimator.

[†] Mitter, S.K. and Newton, N.J., "Information and Entropy Flow in the Kalman-Bucy Filter," *J. of Stat. Phys* **118** (2005), pp. 145-176.

3.2. Stochastic Control Problem

Consider the following controlled equation

$$\tilde{X}_t = \theta + \int_0^t \Big(b(\tilde{X}_s, s) + a(\tilde{X}_s, s) u(\tilde{X}_s, s) \Big) \, ds + \int_0^t \sigma(\tilde{X}_s, s) \, d\tilde{V}_s, \quad (13)$$

where the initial condition, θ, is non-random. Let $\mathbf{U}$ be the set of measurable functions $u : \mathbb{R}^n \times [0, T] \to \mathbb{R}^n$ with the following properties:

(U1) u is continuous;

(U2) $\mathbf{E}\Gamma^u = 1$, where

$$\Gamma^u = \exp \left(\int_0^T u' \sigma(X_t^{\theta,0}, t) \, dV_t - \frac{1}{2} \int_0^T |\sigma' u(X_t^{\theta,0}, t)|^2 \, dt \right), \quad (14)$$

and $(\Omega, \mathcal{F}, P)$, V and $X^{z,s}$ are the corresponding martingales (Girsanov).

Lemma. *If b and σ satisfy the technical hypothesis and $u \in \mathbf{U}$ then* (13) *has a weak solution and is unique in law.*

Let $(\tilde{\Omega}, \tilde{\mathcal{F}}, (\tilde{\mathcal{F}}_t), \tilde{P}, \tilde{X}, \tilde{V})$ be a weak solution of (13) for some $u \in \mathbf{U}$. We define the cost for controls in $\mathbf{U}$ as the apparent information of the resulting distribution of $\tilde{X}$, $\tilde{P}_X$. This is measured relative to the prior $P_X^{\theta,0}$ (the distribution of $X^{\theta,0}$), and $H_p(0, T, \theta, \cdot, y)$ [the Hamiltonian: see Section 3].

$$
\begin{aligned}
J(u,\theta,y) &= h(\tilde{P}_X \,|\, P_X^{\theta,0}) + \langle H_p(0,T,\theta,\,\cdot\,,y), \tilde{P}_X\rangle \\
&= \frac{1}{2}\tilde{\mathbf{E}}\int_0^T |\sigma' u(\tilde{X}_t,t)|^2\,dt - y_T' g(\theta) + \frac{1}{2}\tilde{\mathbf{E}}\int_0^T |g(\tilde{X}_t)|^2\,dt \\
&\quad - \tilde{\mathbf{E}}\int_0^T (y_T - y_t)'(\mathcal{L}g + \mathcal{D}g)(\tilde{X}_t,t)\,dt \\
&\qquad\qquad \text{if the integrals exist} \\
&\quad +\infty \qquad\quad \text{otherwise,}
\end{aligned} \tag{15}
$$

where $\mathcal{L}$ is the differential operator associated with X,

$$
\mathcal{L} = \sum_i b_i \frac{\partial}{\partial z_i} + \frac{1}{2}\sum_{i,j} a_{i,j}\frac{\partial^2}{\partial z_i \partial z_j},
$$

and $\mathcal{D}$ is the row-vector jacobian operator, $\mathcal{D} = [\partial/\partial z_1\ \partial/\partial z_2 \cdots \partial/\partial z_n]$. The cost functional has a more appealing form in the special case that the observation path, y, is everywhere differentiable:

$$
J(u,\theta,y) = \frac{1}{2}\tilde{\mathbf{E}}\int_0^T \left(|\sigma' u(\tilde{X}_t,t)|^2 + |\dot{y}_t - g(\tilde{X}_t)|^2\right)\,dt - \frac{1}{2}\int_0^T |\dot{y}_t|^2\,dt. \tag{16}
$$

This involves an ‘energy’ term for the control and a ‘least-squares’ term for the observation path fit. These correspond to the two terms in Bayes’ formula representing the degrees of match with the prior distribution and the observation path. The optimal control problem (13), (16) can be thought of as a type of energy-constrained *tracking* problem. The optimal control, under which the distribution of $\tilde{X}$ is the regular conditional probability distribution $P_{X|Y}(\,\cdot\,, y)$, is derived in the following theorem.

Theorem 2 *Suppose that b, σ and g satisfy the usual technical hypotheses, and let the function $u_* : \mathbb{R}^n \times [0, T] \times \mathbf{Y} \to \mathbb{R}^n$ be defined by*

$$u_* = -(\mathcal{D}v)', \tag{17}$$

where v is the value function. Then, for each $y \in \mathbf{Y}$, $u_*(\,\cdot\,, \,\cdot\,, y)$ belongs to $\mathbf{U}$, and for all $\theta \in \mathbb{R}^n$, $y \in \mathbf{Y}$ and $\tilde{P}_X \in \mathcal{P}(\mathcal{X})$ (not necessarily the distribution of a controlled process),

$$J(u_*(\,\cdot\,, \,\cdot\,, y), \theta, y) \leq h(\tilde{P}_X \,|\, P_X^{\theta,0}) + \langle H_p(0, T, \theta, \,\cdot\,, y), \tilde{P}_X \rangle. \tag{18}$$

We now consider the special case in which y is differentiable with Hölder continuous derivative, b and g are bounded, and there exists an $\epsilon > 0$ such that

$$\tilde{z}' a(z) \tilde{z} \geq \epsilon |\tilde{z}|^2 \qquad \text{for all } z, \tilde{z} \in \mathbb{R}^n. \tag{19}$$

In this case ρ is continuously differentiable with respect to s, twice continuously differentiable with respect to z, and by a standard extension of the Feynman–Kac formula satisfies the following p.d.e.

$$\frac{\partial \rho}{\partial s} + \mathcal{L}\rho + \left(\dot{y} - \frac{1}{2} g \right)' g \rho = 0 \quad \text{on } \mathbb{R}^n \times (0, T), \qquad \rho(\,\cdot\,, T, y) = 1. \tag{20}$$

Since $v = -\log(\rho)$, the value function, v, satisfies

$$\frac{\partial v}{\partial s} + \mathcal{L}v - \frac{1}{2}\mathcal{D}va(\mathcal{D}v)' - \left(\dot{y} - \frac{1}{2}g\right)' g = 0$$

$$\text{on } \mathbb{R}^n \times (0,T), \quad v(\,\cdot\,,T,y) = 0. \tag{21}$$

3.3. The Inverse Problem

The variational characterization of the inverse problem [parts (ii) and (iv) of Theorem 1, Section 3] can also be applied to the path estimator. This involves choosing a likelihood function to be compatible with the (given) regular conditional probability distribution, $P_{X|Y}(\,\cdot\,, y)$. Earlier, we minimized apparent information over probability measures corresponding to weak solutions of the controlled equation. Here, we maximize compatible information over (negative) log-likelihood functions, $\tilde{H}$, that give rise to posterior distributions of this type.

Let $(\Omega, \mathcal{F}, P)$, μ, V, and X be as defined previously. For each probability measure on $\mathbb{R}^n$, $\tilde{\mu}$, with $\tilde{\mu} \ll \mu$, and each continuous u satisfying (U2) for all θ, let $\tilde{H}$ be a measurable function such that

$$\begin{aligned} \tilde{H}(X) &= -\log\left(\frac{d\tilde{P}_X}{dP_X}(X)\right) + K \\ &= -\log\left(\frac{d\tilde{\mu}}{d\mu}(X_0)\right) - \int_0^T u'\sigma(X_t, t)\, dV_t \\ &\qquad + \frac{1}{2}\int_0^T |\sigma' u(X_t, t)|^2\, dt + K, \end{aligned} \tag{22}$$

where $K \in \mathbb{R}$ and $\tilde{P}_X$ is as defined previously.

We shall assume that $\mu_Y(\,\cdot\,, y) \ll \tilde{\mu}$. The term K in (22) is the information in the associated (unspecified) observation.

Integral log-likelihood functions of the form (22) can be thought of as being associated with observations that are ‘distributed in time’, in that information from them gradually becomes available as t increases.

The characterization of $P_{X|Y}$ in terms if stochastic control can be used to express the compatible information corresponding to $\tilde{H}$, as follows:

$$\begin{aligned} G(\tilde{H}, y) &= K - \langle \tilde{H}\,, P_{X|Y}(\,\cdot\,, y)\rangle \\ &= K + h(\mu_Y(\,\cdot\,, y)\,|\,\mu) - h(\mu_Y(\,\cdot\,, y)\,|\,\tilde{\mu}) \qquad (23) \\ &\qquad + \int_0^T \int_{\mathbb{R}^n} \left(u_* - \frac{1}{2}u\right)' au(z, t, y) \\ &\qquad \cdot\, P_{X|Y}(\chi_t^{-1}(dz), y)\, dt. \end{aligned}$$

Log-likelihood functions of the form (22) could come from many different types of observation.

The only constraints placed on u here are that it be continuous and satisfy (U2) for all θ. We could further constrain it to take the form

$$u(z,s) = -(\mathcal{D}\tilde{v})'(z,s,\tilde{y}),$$

where

$$\tilde{v}(z,s,\tilde{y}) = -\log \mathbf{E} \exp\left(\int_s^T \left(\dot{\tilde{y}}_t - \frac{1}{2}\tilde{g}(X_t^{z,s}) \right)' \tilde{g}(X_t^{z,s})\, dt \right),$$

for appropriate $\tilde{g}$ and $\tilde{y}$. This would correspond to observations of the ‘signal-plus-white-noise’ variety similar to (11), but with ‘controlled’ observation function and path, $\tilde{g}$ and $\tilde{y}$.

This would show the effects of errors in the observation function or approximations of the observation path. Under appropriate regularity conditions $\tilde{v}$ will satisfy the following partial differential equation:

$$-\frac{\partial \tilde{v}}{\partial t} = \mathcal{L}\tilde{v} - \frac{1}{2}\mathcal{D}\tilde{v}a(\mathcal{D}\tilde{v})' - \left(\dot{\tilde{y}}_t - \frac{1}{2}\tilde{g}\right)' \tilde{g}; \qquad \tilde{v}(\,\cdot\,, T) = 0. \quad (24)$$

Thus one interpretation of the inverse problem involves the infinite-dimensional, deterministic optimal control in reversed time, with control $(\tilde{g}, \tilde{y})$, and payoff

$$\begin{aligned} \Pi(\tilde{g}, \tilde{y}) \;=\; & \int_0^T \int_{\mathbb{R}^n} \mathcal{D}\tilde{v}a \left(u_* - \frac{1}{2}(\mathcal{D}\tilde{v})'\right)(z, t, y) \\ & \cdot\; P_{X|Y}(\chi_t^{-1}(dz), y)\, dt. \end{aligned} \quad (25)$$

The optimal trajectory for this dual problem, $v(\cdot, \cdot, y)$ is a time-reversed likelihood filter for X given Y, and the measure, $\exp(-v(z, s, y))P_X(\chi_s^{-1}(dz))$ is an un-normalized regular conditional probability distribution for X_s given observations $(Y_t - Y_s, s \leq t \leq T)$, which coincides with that provided by the Zakai equation for the time-reversed problem. This provides an information-theoretic explanation of the connection between nonlinear filtering and stochastic optimal control used in †, as well as widening its scope. A detailed account of this, and the information processing aspects of nonlinear filters and interpolators can be found in *. For a somewhat different problem involving optimization over observation functions, see #.

[†] W.H. Fleming and S.K. Mitter,
"Optimal control and nonlinear filtering for nondegenerate diffusion processes,"
Stochastics **8** (1982), pp. 63–77.

[*] Mitter, S.K. and Newton, N.J., "Information and Entropy Flow in the Kalman-Bucy Filter,"
J. of Stat. Phys **118** (2005), pp. 145-176.

[#] B.M. Miller and W.J. Runggaldier, "Optimization of observations: a stochastic control approach,"
SIAM J. Control Optim. **35** (1997), pp. 1030–1052.

Extensions to State Process described by
Partial Differential Equation (Lattice)

See recent work of R. Von Handel and collaborators

Infinite-Lattice, Infinite-time Behavior of Filter

CONCLUSION

- Integration of Stochastic Control and

  Information Theory

- Nonequilibrium Statistical Mechanics

- Application Areas

  Sensor Networks and Monitoring

  Energy Harvesting