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

Lecture at TIFR, Mumbai on January 4, 2020

Some Connections between Information Theory, Filtering and Statistical Mechanics

Variational Approach to Bayesian Estimation

Stochastic Control Interpretation of Nonlinear Filtering

Fokker Planck Equation

Variational Interpretation: Free Energy Minimization

Relation between Shannon Entropy, Fisher Information, and Optimal Transport

Biology is a perfect example of a system which is purposeful. Reliable communication of signals and information is not an end in itself. Communication takes place in order to convey information by which the organism can carry out purposeful action, such as maintaining order and equilibrium with its environment. There are currently no models of communication which are relevant to Biology and Neuroscience. It has been suggested that the organism maintains order by minimizing Free Energy. A theoretical and experimental verification of this Ansatz would be a major accomplishment.

Although I have devoted quite a bit of space to issues arising in Biology, the research themes I have cited have much broader relevance, for example in gaining a fundamental understanding of the functioning of energy and transportation systems and the design of future systems.

At the other extreme, our ability to intervene at the molecular and atomic levels has necessitated that the laws of quantum mechanics have to be invoked in the building of future circuits, devices and systems.

The Nobel Prize winning Biologist, Sydney Brenner, has argued that Turing's idea of a Universal Turing Machine which can simulate any other Turing machine has its embodiment in Biology where every organism contains an internal description of itself. The concept of the gene as symbolic representation of the organism—a code script—is a fundamental feature of the living world and must form the kernel of Biological Theory. The limits of what can be inferred from data alone need to be characterized.

Transfer Entropy and Directed Information

Transfer entropy and directed information are measures of directional dependency between the **past** of one process and the **future** of another process $:\Longrightarrow$ Causality

Directed Information (Massey)

$$D^{M}_{X\to Y}(N) = \sum_{n=1}^{N} I[(X_1,\ldots,X_n);Y_n|Y_1,\ldots,Y_{n-1})]$$

$(X_1, X_2, \ldots)$: Input Sequence to a Channel

$(Y_1, Y_2, \ldots)$: Output Sequence

Transfer Entropy (Schreiber)

$$T_{X\to Y}(k,\ell,n) = I[(X_{n-\ell},\ldots,X_{n-1});Y_n|Y_{n-k},\ldots,Y_{n-1})]$$

Measure of the disambiguation on the future of Y over and above that provided by the past of Y afforded by the past of X.

For $k = \ell = n - 1$,

$$D_{X \to Y}(N) = \sum_{n=1}^{2} I[(X_1, \ldots, X_{n-1}); Y_n | Y_1, \ldots, Y_{n-1}]$$

similar to $D^M_{X \to Y}(N)$

Does not involve mutual information between values of X and Y at the same instant.

For continuous time definitions, see

"Transfer Entropy and Directed Information in Gaussian Diffusion Processes,"

Nigel Newton, University of Essex

arXiv:1604.01969v1

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].

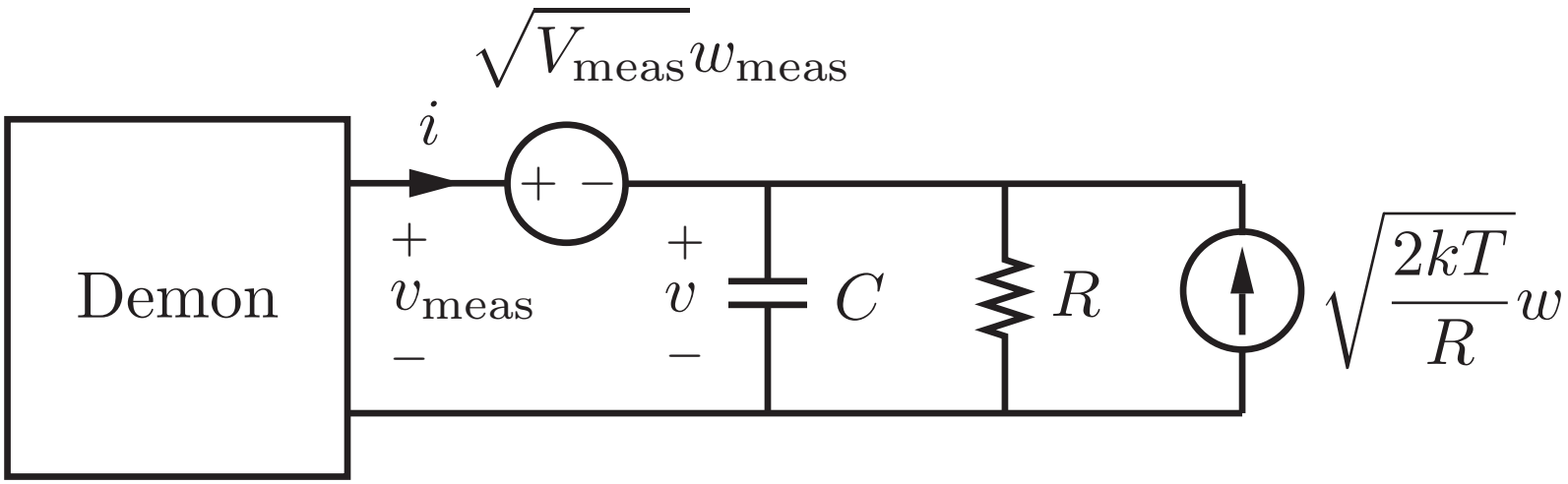


FIG.1: The demon (the feedback controller) connected to a capacitor, a heat bath of temperature T, and a measurement noise source of intensity V_{meas}. The demon may choose the current i freely, and has access to the noisy voltage measurement v_{meas}.

The circuit is modeled by an overdamped Langevin equation

$$\begin{aligned} \tau\dot{v} &= -v + Ri + \sqrt{2kTR}w,\ \langle v(0)\rangle = 0, \\ v_{\text{meas}} &= v + \sqrt{V_{\text{meas}}}w_{\text{meas}},\ \langle v(0)^2\rangle = \frac{kT}{C}, \end{aligned} \tag{1}$$

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.

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

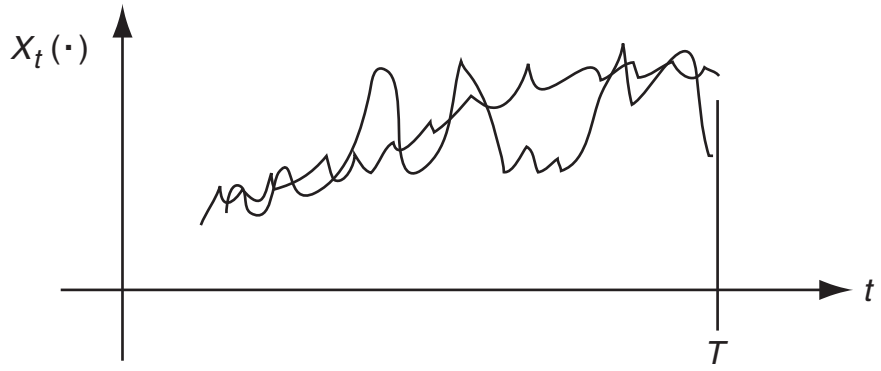
$$\frac{dx_t}{dt} = b(x_t) + \sigma(x_t)\dot{v}_t \ ,$$

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

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

where

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

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, to appear in *J. Stat. Mech.*, 2012

$(X_t)_{t\geq 0}$ 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

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

Σ_x: statistical mechanics 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 $\square$

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

(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

$$\xi_t(x) = \xi_0(x) + \int_0^t (\mathcal{A}\xi_s)(x)ds + \int_0^t \xi_s(x)(g(x) - \langle g, Z_s\rangle)' d\nu_s \tag{2}$$

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

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

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

$$\dot{S}(t) = \frac{1}{2}E|g(X_t) - \langle g, Z_t\rangle|^2 \tag{3}$$

$$\dot{D}(t) = E\left(\frac{Ap_t}{p_t}\log p_t - \frac{A\xi_t}{\xi}\log \xi_t\right)(X_t) \tag{4}$$

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 (3) and (4) show that the supply of information is associated with the second integral in (2)

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

and the dissipation associated with the first integral in (2)

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

$\dot{S}(t) =$ 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

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

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

$$\bar{\mathbf{Y}} = \left\{ y \in \mathbf{Y} : 0 < \int_{\mathbf{X}} Q(x, y) P_X(dx) < \infty \right\}; \qquad (5)$$

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

$$\begin{array}{rcll} H(x,y) & = & -\log(Q(x,y)) & \text{if } y \in \bar{\mathbf{Y}} \\ & & 0 & \text{otherwise :} \end{array} \tag{6}$$

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

$$P_{X|Y}(A,y) = \frac{\int_A \exp(-H(x,y)) P_X(dx)}{\int_{\mathbf{X}} \exp(-H(x,y)) P_X(dx)}, \tag{7}$$

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. (5)–(7) 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

$$h(\tilde{P}_X \,|\, \hat{P}_X) = \begin{cases} \displaystyle\int_{\mathbf{X}} \log\left(\frac{d\tilde{P}_X}{d\hat{P}_X}\right) d\tilde{P}_X & \text{if } \tilde{P}_X \ll \hat{P}_X \text{ and the integral exists} \\ +\infty & \text{otherwise,} \end{cases} \tag{8}$$

$$i(\tilde{H}) = \begin{cases} -\log\left(\displaystyle\int_{\mathbf{X}} \exp(-\tilde{H}) dP_X\right) & \text{if } 0 < \displaystyle\int_{\mathbf{X}} \exp(-\tilde{H}) dP_X < \infty \\ -\infty & \text{otherwise,} \end{cases} \tag{9}$$

$$\langle \tilde{H}, \tilde{P}_X \rangle = \begin{cases} \displaystyle\int_{\mathbf{X}} \tilde{H} d\tilde{P}_X & \text{if the integral exists} \\ +\infty & \text{otherwise.} \end{cases} \tag{10}$$

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

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

(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

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

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

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:

$$\begin{aligned} X_t &= X_0 + \int_0^t b(X_s, s)\, ds + \int_0^t \sigma(X_s, s)\, dV_s, \quad \text{for } 0 \le t \le T, \\ X_0 &\sim \mu, \end{aligned} \tag{11}$$

$$Y_t = \int_0^t g(X_s)\, ds + W_t \quad \text{for } 0 \le t \le T, \tag{12}$$

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 (11) and (12) 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, $\mathbf{E}\log[dP_{XY}/d(P_X \otimes \lambda_Y)(X,Y)]$, 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

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

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

$$M_{y,t} = M_{y,0} \exp\left(\int_0^t U'_{y,s}(dX_s - b(X_s, s)\, ds) - \frac{1}{2} \int_0^t |\sigma(X_s, s)' U_{y,s}|^2 \, ds \right) \tag{13}$$

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 (11), but with a different initial law, and with a control term, $\sigma\sigma'(X_s, s)U_{y,s}$, entering the drift coefficient.

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

$$\tilde{X}_t = \theta + \int_0^t \left(b(\tilde{X}_s, s) + a(\tilde{X}_s, s) u(\tilde{X}_s, s) \right) ds + \int_0^t \sigma(\tilde{X}_s, s)\, d\tilde{V}_s, \quad (14)$$

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

$$\Gamma^u = \exp\left(\int_0^T u' \sigma(X_t^{\theta,0}, t)\, dV_t - \frac{1}{2} \int_0^T |\sigma' u(X_t^{\theta,0}, t)|^2\, dt \right), \quad (15)$$

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* (14) *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 (14) 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].

$$
\begin{aligned}
J(u,\theta,y) &= h(\tilde{P}_X \,|\, P_X^{\theta,0}) + \langle H_p(0,T,\theta,\,\cdot\,,y), \tilde{P}_X\rangle \\
&= \frac{1}{2}\tilde{\mathbf{E}}\int_0^T |\sigma' u(\tilde{X}_t,t)|^2\,dt - y_T' g(\theta) + \frac{1}{2}\tilde{\mathbf{E}}\int_0^T |g(\tilde{X}_t)|^2\,dt \\
&\quad - \tilde{\mathbf{E}}\int_0^T (y_T - y_t)'(\mathsf{cl}g + \mathcal{D}g)(\tilde{X}_t,t)\,dt \\
&\qquad\qquad \text{if the integrals exist} \\
&\quad +\infty \qquad \text{otherwise,}
\end{aligned}
\tag{16}
$$

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:

$$
J(u,\theta,y) = \frac{1}{2}\tilde{\mathbf{E}}\int_0^T \left(|\sigma' u(\tilde{X}_t,t)|^2 + |\dot{y}_t - g(\tilde{X}_t)|^2\right)\,dt - \frac{1}{2}\int_0^T |\dot{y}_t|^2\,dt. \tag{17}
$$

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 (14), (17) 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*

$$u_* = -(\mathcal{D}v)', \tag{18}$$

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),*

$$J(u_*(\,\cdot\,,\,\cdot\,,y),\theta,y) \leq h(\tilde{P}_X \,|\, P_X^{\theta,0}) + \langle H_p(0,T,\theta,\,\cdot\,,y), \tilde{P}_X \rangle. \tag{19}$$

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

$$\tilde{z}' a(z) \tilde{z} \geq \epsilon |\tilde{z}|^2 \qquad \text{for all } z, \tilde{z} \in \mathbb{R}^n. \tag{20}$$

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.

$$\frac{\partial \rho}{\partial s} + \mathcal{L}\rho + \left(\dot{y} - \frac{1}{2} g\right)' g\rho = 0 \quad \text{on } \mathbb{R}^n \times (0, T), \quad \rho(\,\cdot\,, T, y) = 1. \tag{21}$$

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

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

$$\text{on } \mathbb{R}^n \times (0,T), \quad v(\,\cdot\,, T, y) = 0. \tag{22}$$

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

$$\begin{aligned} \tilde{H}(X) &= -\log\left(\frac{d\tilde{P}_X}{dP_X}(X)\right) + K \\ &= -\log\left(\frac{d\tilde{\mu}}{d\mu}(X_0)\right) - \int_0^T u'\sigma(X_t, t)\, dV_t \\ &\quad + \frac{1}{2}\int_0^T |\sigma' u(X_t, t)|^2\, dt + K, \end{aligned} \tag{23}$$

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 (23) is the information in the associated (unspecified) observation.

Integral log-likelihood functions of the form (23) 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:

$$\begin{aligned} G(\tilde{H}, y) &= K - \langle \tilde{H}\,,\, P_{X|Y}(\,\cdot\,, y)\rangle \\ &= K + h(\mu_Y(\,\cdot\,, y)\,|\,\mu) - h(\mu_Y(\,\cdot\,, y)\,|\,\tilde{\mu}) \\ &\qquad + \int_0^T \int_{\mathbb{R}^n} \left(u_* - \frac{1}{2}u\right)' a u(z,t,y) \\ &\qquad\quad \cdot\, P_{X|Y}(\chi_t^{-1}(dz), y)\, dt. \end{aligned} \tag{24}$$

Log-likelihood functions of the form (23) 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 (12), 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:

$$-\frac{\partial \tilde{v}}{\partial t} = \mathcal{L}\tilde{v} - \frac{1}{2}\mathcal{D}\tilde{v}a(\mathcal{D}\tilde{v})' - \left(\dot{\tilde{y}}_t - \frac{1}{2}\tilde{g}\right)' \tilde{g}; \quad \tilde{v}(\,\cdot\,, T) = 0. \tag{25}$$

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

$$\begin{aligned} \Pi(\tilde{g}, \tilde{y}) \;=\; & \int_0^T \int_{\mathbb{R}^n} \mathcal{D}\tilde{v}a \left(u_* - \frac{1}{2}(\mathcal{D}\tilde{v})'\right)(z, t, y) \\ & \cdot\; P_{X|Y}(\chi_t^{-1}(dz), y)\, dt. \end{aligned} \tag{26}$$

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)

Infinite-Lattice, Infinite-time Behavior of Filter

Implication for Optimization involving Neural Networks

Approximation of Conditional Distributions using Data

Statistical Invariants (Vapnik)