Duality Between Estimation and Control

Sanjoy K. Mitter

Laboratory for Information and Decision Systems

Massachusetts Institute of Technology

Joint work with Nigel Newton, University of Essex

September 2012

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

Duality of a Source and a Channel

There is a curious and provocative duality between the properties of a source with a distortion measure and those of a channel. This duality is enhanced if we consider channels in which there is a "cost" associated with the different input letters, and it is desired to find the capacity subject to the constraint that the expected cost not exceed a certain quantity. Thus input letter i might have cost a_i and we wish to find the capacity with the side condition $\sum_i P_i a_i \leq a$, say, where P_i is the probability of using input letter i. This problem amounts, mathematically, to *maximizing* a mutual information under variation of the P_i with a linear inequality as constraint. The solution of this problem leads to a capacity cost function $C(a)$ for the channel. It can be shown readily that this function is *concave* downward. Solving this problem corresponds. in a sense, to finding a source that is just right for the channel and the desired cost.

In a somewhat dual way, evaluating the rate distortion function $R(d)$ for a source amounts, mathematically, to *minimizing* a mutual information under variation of the $q_i(j)$, again with a linear inequality as constraint. The solution leads to a function $R(d)$ which is *convex* downward. Solving this problem corresponds to finding a channel that is just right for the source and allowed distortion level. This duality 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 but cannot control it; we may control the future but have no knowledge of it.

Claude E. Shannon, *Coding Theorems for a Discrete Source with a Fidelity Criterion: Collected Works of Claude E. Shannon*, IEEE Press, 1993, pp. 325–350.

1. Nonlinear Filtering

1.1. Model

Our basic model is the observation equation

$$y_t = \int_0^t z_s ds + w_t \tag{1}$$

with the assumptions

(H1) $y_1$is real-valued process

(H2) w_t is standard Brownian motion

(H3) $E\left[\int_0^T z_s^2 ds\right] < \infty$

(H4) (z_t) is independent of w_t.

Consider the innovations process

$$\left.\begin{array}{rcl} \nu_t & = & y_t - \int_0^t \widehat{z}_s ds \\ & & \\ \widehat{z}^s & = & E(z_s | F_s^y) \ . \end{array}\right\} \qquad (2)$$

where

It can now be shown that:

The process (ν_t, F_t^y) is standard Brownian motion and F_s^y and $\sigma(\nu^u - \nu_t | 0 \leq s \leq t < u \leq T)$ are independent. This result is proved by showing that ν_t is a square integrable martingale with continuous sample paths with quadratic variation t, and the result follows from the Levy characterization of Brownian motion.

Now analogous to the linear case, one can prove that

$$\mathcal{F}_t^y = \mathcal{F}_t^\nu \ ; \tag{3}$$

that is, the innovations contains the same information as the observation, This rather delicate result was proved by Allinger–Mitter[†].

[†] D. F. Allinger and S. K. Mitter, *Stochatics* **4**, pp. 339-348. 1981.

Now combining this with the representation of square integrable martingales as stochastic integrals due to Kunita and Watanabe, we obtain the following:

Every square-integrable martingale (m_t, F_t^y) can be represented as

$$m_t = E(m_0) + \int_0^t \eta_s ds \ . \tag{4}$$

$\int_0^T E(\eta_s^2) ds < \infty$ and η_t is adapted to F_t^ν.

It should be remarked that Fujisaki–Kallianpur–Kunita, in their important paper, proved the same result without (3) holding but with η_t adapted to $\mathcal{F}_t^y$.

To proceed further let us assume that

$$z_t = h(x_t) \tag{5}$$

and x_t satisfies a stochastic differential equation

$$x_t = x_0 + \int_0^t f(x_s)ds + \int_0^t G(x_s)d\beta_s \ . \tag{6}$$

Suppose we want to obtain the estimate

$$\pi_t(\varphi) \stackrel{\triangle}{=} E[\varphi(x_t)|\mathcal{F}_t^y] \ . \tag{7}$$

We want to obtain a recursive equation for $\pi_t(\varphi)$. We need some preliminaries.

Let L be the second-order elliptic operator defined by

$$L(\Psi) = \sum_{i=1}^{n} f^i(x) \frac{\partial^i \Psi}{\partial x^i} + \sum_{i,j=1}^{n} a^{ij}(x) \frac{\partial^2 \Psi}{\partial x^i \partial x^j} \qquad (8)$$

and

$$A(x) = [a^{ij}(x)]_{i,j=1}^{n} = G(x)G'(x) \ .$$

Then we can write Itô's differential rule as:

$$\Psi(x_t) - \Psi(x_0) - \int_0^t L\Psi(x_s)ds = \int_0^t [\nabla \Psi(x_s)]' G(x_s) d\beta_s \qquad (9)$$

where ∇ is the gradient operator, and the last term

$$M_t^{\Psi} = \int_0^t [\nabla \Psi(x_s)] G(x_s) d\beta_s$$

is a $\mathcal{F}_t^{\beta}$-martingale (being a stochastic integral).

To obtain the recursive equation for $\pi_t(\varphi)$, one shows that $M_t^\varphi = \pi_t(\varphi) - \pi_0(\varphi - \int_0^t \pi_s(L\varphi)$ is a square integrable $\mathcal{F}_t^y$, hence $\mathcal{F}_t^\nu$-martingale. Therefore from the representation theorem $M_t^\varphi = \int_0^t \eta_s d\nu_s$ where ν_s is square integrable and adapted to $\mathcal{F}_t^\eta$. Therefore

$$\pi_t(\varphi) = \pi_0(\varphi) + \int_0^t \pi_s(L\varphi)ds + \int_0^t \eta_s d\nu_s \ . \tag{10}$$

It remains to identify ν_s. This can be obtained as follows, by the Itô differential rule (9)

$$\varphi(x_t) = \varphi(x_0) + \int_0^t L\varphi(x_s)ds + M_t^\varphi \ . \tag{11}$$

Also

$$y_t = y_0 + \int_0^t h(x_s)ds + w_t \ . \tag{12}$$

Now, using the Itô-differential rule for semi-martingales,

$$\begin{aligned}\varphi(x_t)y_t &= \varphi(x_0)y_0 + \int_0^t y_s d\varphi(x_s) + \int_0^t \varphi(x_s)dy_s + \langle M^\varphi, w\rangle_t \\ &= \varphi(x_0)y_0 + \int_0^t y_s[L\varphi(x_s)ds + dM_s^\varphi] \\ &\quad + \int_0^t \varphi(x_s)[h(x_s)ds + dw_s) \qquad (13)\end{aligned}$$

since $\langle M^\varphi, w\rangle_t = 0$ from the independence of (x_t) and (w_t).

From the innovations representation

$$y_t = y_0 + \int_0^t \pi_s(h)ds + \nu_t \ . \qquad (14)$$

Therefore

$$\begin{aligned}\pi_t(\varphi)y_t &= \pi_0(\varphi)y_0 + \int_0^t \pi_s(\varphi)[\pi_s)(h)ds + d\nu_s] \\ &\quad + \int_0^t y_s[\pi_s(L\varphi)ds + \eta_s d\nu_s] - \langle N, \nu\rangle_t\end{aligned}$$

where

$$\begin{aligned}N_t = \int_0^t \eta_s d\nu_s &= \pi_0(\varphi)y_0 + \int_0^t \pi_s(\varphi)[\pi_s(h)ds + d\nu_s] \\ &\quad + \int_0^t y_s[\pi_s(L\varphi)ds + \eta_s d\nu_s] + \int_0^t \eta_s d\nu_s \ . \quad (15)\end{aligned}$$

Now noting that

$$E[\varphi(x_t)y_t - \pi_t(\varphi)y_t|\mathcal{F}_s^y] = 0 \ , \quad (16)$$

from (14) and (15) we get

$$\eta_t = \pi_t(h\varphi) - \pi_t(\varphi)\pi_t(h) \quad (17)$$

and hence, from (10), we get

$$\pi_t(\varphi) = \pi_0(\varphi) + \int_0^t \pi_s(L\varphi)ds + \int_0^t [\pi_s(h\varphi) - \pi_s(h)\pi_s(\varphi)]d\nu_s \ . \quad (18)$$

This is one of the fundamental equations of nonlinear filtering. If the conditional distribution π_t has a density given by $\tilde{p}(t,x)$, then $\tilde{p}$ satisfies the stochastic partial differential equation

$$d\tilde{p}(t,x) = L^*\tilde{p}(t,x)dt + \tilde{p}(t,x)[h(x) - \pi_t(h)]d\nu_t \qquad (19)$$

where $\pi_t(h) = \int h(x)\tilde{p}(t,x)dx$. The question of existence of a conditional density can be discussed using the Malliavin calculus[†]. Eq. (19) in this form, where the Itô calculus is involved, was first derived by Kushner[*].

[†] E. Pardoux, Filtrage Non Lineaire at Equations Aux Derivees Partielles Stochastique Associees, *Ecole d'Ete de Probabilites de Saint-Flour XIX*, ed. P.L. Hennequin. Spring Lecture Notes in Mathematics 1464, 1991.

[*] H.J. Kushner, *SIAM J. Control*, **2** pp. 106–119, 1964.

The difficulty in deriving a solution for a conditional statistic $\widehat{x}_t \stackrel{\triangle}{=} \pi_t(x) = \int x\tilde{p}(t,x)dx$ is the so-called closure problem. The equation is

$$d\widehat{x}_t - \pi_t(f)dt + [\pi_t(hx)] - \pi_t(h)\widehat{x}d\nu_t \ . \qquad (20)$$

Note that computation of $\widehat{x}_t$ requires computing

$$\pi_t(f) = \int f(x)\tilde{p}(t,x)dx \ , \quad \pi_t(hx) = \int h(x)x\tilde{p}(t,x)dx \qquad (21)$$

and $\pi_t(h) \stackrel{\triangle}{=} \int h(x)\tilde{p}(t,x)$, and this requires solving stochastic differential equations for each of these above quantities, which in their turn involve higher moments. Hence nonlinear filters are almost always infinite-dimensional.

There are only a few known examples where the filter is known to be finite-dimensional. The first is the linear-Gaussian situation leading to the Kalman filter. The second is the finite-state case, first considered in an important paper by Wonham[†].

[†] W.M. Wonham, *SIAM J. Control* **2**, pp. 347–369, 1965.

Let x_t be a finite-state Markov process taking values $S = (s_1, \ldots, s_N)$. Let $p_t = (p_t^1, \ldots, p_t^N)$ be the probability vector where $p_t^i - \text{Prob}(x_t = s_i)$. Then the evolution of p_t is given by the forward Kolmogoroff equation

$$\frac{dp_t}{dt} = Ap_t \ .$$

If we denote by $\tilde{p}_t = \text{Prob}(x_t = s_i | \mathcal{F}_t^y)$

$$B = \text{diag}[h(s_1), \ldots, h(s_N)] \quad \text{and} \quad b' = [h(s_1), \ldots, h(s_N)],$$

then $\tilde{p}_t$ satisfies

$$d\tilde{p}_t = A\tilde{p}_t dt + [B - (b'\tilde{p}_t)I]\tilde{p}_t(dy_t - (b'\tilde{p}_t)dt] \ . \qquad (22)$$

One of the difficulties with Eq. (19) is that it is a nonlinear stochastic partial differential equation. An important idea due to Zakai[†]], Duncan[*], and Mortensen[#] is to write $\pi_t(\varphi)$ as

$$\pi_t(\varphi) = \frac{\rho_t(\varphi)}{\rho_t(1)} \tag{23}$$

where $\rho_t(\varphi)$ satisfies

$$\rho_t(\varphi) = \rho_0(\varphi) + \int_0^t \rho_s(L\varphi)ds + \int_0^t \rho_s(h\varphi)dy_s \ . \tag{24}$$

ρ_t is an un-normalized version of π_t. Note that this is a linear stochastic partial differential equation.

[†] M. Zakai, *Z. Wahr. Verw. Geb.* **11**, pp. 230–243,1969

[*] T.E. Duncan, Ph.D. dissertation, Stanford University, 1967.

[#] R.E. Mortenson, Ph.D. Dissertation, Univ. of California, Berkeley, 1967.

This is intimately related to the Feynman–Kac formula for integrating linear parabolic equations with a potential term. For a discussion between the analogies between nonlinear filtering and quantum mechanics see Mitter[†]. Recall that the original probability space is $(\Omega, \mathcal{F}, P)$ on which there is an increasing family of σ-fields $(\mathcal{F})_{t\geq 0}$ and the process (x_t) is adapted to it.

[†] S.K. Mitter, *Ricerche di Automatice*, Vol. 10, no. 2, pp. 163–216,1979.

Define a new probability measure P_0 on $(\Omega, \mathcal{F})$ in terms of the Radon–Nikodym derivative

$$\frac{dP_0}{dP} = \exp\left(-\int_0^T h(x_2)dy_s - \frac{1}{2}\int_0^t h^2(x_s)ds\right) \stackrel{\triangle}{=} \Lambda_t \ . \quad (25)$$

Under P_0, (y_t) is standard Brownian motion, (x_t) and (y_t) are independent, and (x_t) has the same distribution under P_0 and P. Now,

$$\pi_t(\varphi)E[\varphi(x_t)|\mathcal{F}_t^y] = \frac{E_0[\varphi(x_t)\Lambda_t|\mathcal{F}_t^y]}{E_0(\Lambda_t|\mathcal{F}_t^y)} \stackrel{\triangle}{=} \frac{\rho_t(\varphi)}{\rho_t(1)} \ . \quad (26)$$

Furthermore. we can prove that

$$\tilde{\Lambda}_t \stackrel{\triangle}{=} E_0(\Lambda_t|\mathcal{F}_t^y) = \exp\left(\int_0^t \pi_s(h)dy_s - \frac{1}{2}\int_0^t \pi_s(h)^2 dy_s\right) \ .$$

From (26), $\rho_t(\varphi) = \pi_t(\varphi)\tilde{\Lambda}_t$. Then using the Itô differential rule, we get (24). This derivation is due to Davis and Marcus[†], where the full details can be found. The measure transformation idea in stochastic differential equations is due to Girsanov (cf. Liptser and Shiryayev[*] and the references cited there).

Eq. (24) is an Itô stochastic partial differential equation. There is a calculus, the so-called Stratanovich calculus, which in many ways is like ordinary calculus. The conditional density equation for nonlinear filtering was derived using this calculus by Stratanovich[#]. For the relation between the two calculi see Wong[$]. This is an important modeling question.

[†] M.H.A. Davis and S.I. Marcus, An Introduction to Nonlinear Filtering, in *Stochastic Systems: The Mathematics of Nonlinear Filtering and Identification and Applications* eds. M. Hazewinkel and J. C. Willems, Reidel, Dordrecht, 1981.

[*] R.S. Liptser and A.N. Shiryayer, *Statistics of Random Processes I: General Theory*, Springer-Verlag, New York, 1977.

[#] R.L. Stratonovich, *Theory Prob. Appl. (USSR)*, Vol. 5, pp. 156–178, 1960.

[$] E. Wong, *Stochastic Processes in Information & Dynamical Systems*, McGraw Hill, New York, 1971.

The Stratanovich form of (24) is

$$\rho_1(\varphi) = \rho_0(\varphi) + \int_0^t \rho_s \left(L\varphi - \frac{1}{2} h^2 \varphi \right) ds + \int_0^t \rho_s (h\varphi) \circ dy_s \tag{27}$$

where the last integral is a (symmetric) Stratanovich integral. It should be noted that geometry is preserved when we work with the Stratanovich form of the equation. The relation between (24) and (27) involves the Wong–Zakai correction [note that the generator L in (27) has been replaced by $L - \frac{1}{2}h^2$]. If ρ_t has a density $q(t,x)$ then $q(t,x)$ satisfies a linear stochastic partial differential equation

$$dq(t,x) = \left(L^* - \frac{1}{2} h^2 \right) q(t,x) + h(x) q(t,x) \circ dy_t \ . \tag{28}$$

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

$$\bar{\mathbf{Y}} = \left\{ y \in \mathbf{Y} : 0 < \int_{\mathbf{X}} Q(x, y) P_X(dx) < \infty \right\}; \tag{29}$$

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

$$\begin{array}{rcll} H(x,y) & = & -\log(Q(x,y)) & \text{if } y \in \bar{\mathbf{Y}} \\ & & 0 & \text{otherwise :} \end{array} \tag{30}$$

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

$$P_{X|Y}(A,y) = \frac{\int_A \exp(-H(x,y)) P_X(dx)}{\int_{\mathbf{X}} \exp(-H(x,y)) P_X(dx)}, \tag{31}$$

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. (29)–(31) 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

$$h(\tilde{P}_X \,|\, \hat{P}_X) = \begin{cases} \displaystyle\int_{\mathbf{X}} \log\left(\frac{d\tilde{P}_X}{d\hat{P}_X}\right) d\tilde{P}_X & \text{if } \tilde{P}_X \ll \hat{P}_X \text{ and the integral exists} \\ +\infty & \text{otherwise,} \end{cases} \tag{32}$$

$$i(\tilde{H}) = \begin{cases} -\log\left(\displaystyle\int_{\mathbf{X}} \exp(-\tilde{H}) dP_X\right) & \text{if } 0 < \displaystyle\int_{\mathbf{X}} \exp(-\tilde{H}) dP_X < \infty \\ -\infty & \text{otherwise,} \end{cases} \tag{33}$$

$$\langle \tilde{H}, \tilde{P}_X \rangle = \begin{cases} \displaystyle\int_{\mathbf{X}} \tilde{H} d\tilde{P}_X & \text{if the integral exists} \\ +\infty & \text{otherwise.} \end{cases} \tag{34}$$

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

(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

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

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

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

3. Path estimation and the Stochastic Control View

Note: the change in notation for the SDE Model and the Observation Model

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:

$$\begin{aligned} X_t &= X_0 + \int_0^t b(X_s, s)\, ds + \int_0^t \sigma(X_s, s)\, dV_s, \quad \text{for } 0 \leq t \leq T, \\ X_0 &\sim \mu, \end{aligned} \tag{35}$$

$$Y_t = \int_0^t g(X_s)\, ds + W_t \quad \text{for } 0 \leq t \leq T, \tag{36}$$

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 (35) and (36) 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

$$M_{y,t} = M_{y,0} \exp\left(\int_0^t U'_{y,s}(dX_s - b(X_s, s)\, ds) - \frac{1}{2} \int_0^t |\sigma(X_s, s)' U_{y,s}|^2 \, ds \right) \tag{37}$$

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 (35), 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

$$\tilde{X}_t = \theta + \int_0^t \left(b(\tilde{X}_s, s) + a(\tilde{X}_s, s) u(\tilde{X}_s, s) \right) ds + \int_0^t \sigma(\tilde{X}_s, s)\, d\tilde{V}_s, \quad (38)$$

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

$$\Gamma^u = \exp\left(\int_0^T u' \sigma(X_t^{\theta,0}, t)\, dV_t - \frac{1}{2} \int_0^T |\sigma' u(X_t^{\theta,0}, t)|^2\, dt \right), \quad (39)$$

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* (38) *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 (38) 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].

$$
\begin{aligned}
J(u,\theta,y) &= h(\tilde{P}_X \,|\, P_X^{\theta,0}) + \langle H_p(0,T,\theta,\,\cdot\,,y), \tilde{P}_X\rangle \\
&= \frac{1}{2}\tilde{\mathbf{E}}\int_0^T |\sigma' u(\tilde{X}_t,t)|^2\,dt - y_T' g(\theta) + \frac{1}{2}\tilde{\mathbf{E}}\int_0^T |g(\tilde{X}_t)|^2\,dt \\
&\quad - \tilde{\mathbf{E}}\int_0^T (y_T - y_t)'(\mathsf{cl} g + \mathcal{D}g)(\tilde{X}_t,t)\,dt \\
&\qquad \text{if the integrals exist} \\
&+\infty \qquad \text{otherwise,}
\end{aligned}
\tag{40}
$$

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:

$$
J(u,\theta,y) = \frac{1}{2}\tilde{\mathbf{E}}\int_0^T \left(|\sigma' u(\tilde{X}_t,t)|^2 + |\dot{y}_t - g(\tilde{X}_t)|^2\right)\,dt - \frac{1}{2}\int_0^T |\dot{y}_t|^2\,dt. \tag{41}
$$

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 (38), (41) 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*

$$u_* = -(\mathcal{D}v)', \tag{42}$$

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),*

$$J(u_*(\,\cdot\,,\,\cdot\,,y),\theta,y) \leq h(\tilde{P}_X \,|\, P_X^{\theta,0}) + \langle H_p(0,T,\theta,\,\cdot\,,y), \tilde{P}_X \rangle. \tag{43}$$

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

$$\tilde{z}' a(z) \tilde{z} \geq \epsilon |\tilde{z}|^2 \qquad \text{for all } z, \tilde{z} \in \mathbb{R}^n. \tag{44}$$

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.

$$\frac{\partial \rho}{\partial s} + \mathcal{L}\rho + \left(\dot{y} - \frac{1}{2} g\right)' g\rho = 0 \quad \text{on } \mathbb{R}^n \times (0, T), \quad \rho(\,\cdot\,, T, y) = 1. \tag{45}$$

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

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

$$\text{on } \mathbb{R}^n \times (0,T), \quad v(\,\cdot\,, T, y) = 0. \tag{46}$$

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

$$\begin{aligned} \tilde{H}(X) &= -\log\left(\frac{d\tilde{P}_X}{dP_X}(X)\right) + K \\ &= -\log\left(\frac{d\tilde{\mu}}{d\mu}(X_0)\right) - \int_0^T u'\sigma(X_t, t)\, dV_t \\ &\quad + \frac{1}{2}\int_0^T |\sigma' u(X_t, t)|^2\, dt + K, \end{aligned} \tag{47}$$

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

Integral log-likelihood functions of the form (47) 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:

$$\begin{aligned} G(\tilde{H}, y) &= K - \langle \tilde{H}\,,\, P_{X|Y}(\,\cdot\,, y)\rangle \\ &= K + h(\mu_Y(\,\cdot\,, y)\,|\,\mu) - h(\mu_Y(\,\cdot\,, y)\,|\,\tilde{\mu}) \\ &\qquad + \int_0^T \int_{\mathbb{R}^n} \left(u_* - \frac{1}{2}u\right)' a u(z,t,y) \\ &\qquad\quad \cdot\, P_{X|Y}(\chi_t^{-1}(dz), y)\, dt. \end{aligned} \tag{48}$$

Log-likelihood functions of the form (47) 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 (36), 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:

$$-\frac{\partial \tilde{v}}{\partial t} = \mathcal{L}\tilde{v} - \frac{1}{2}\mathcal{D}\tilde{v}a(\mathcal{D}\tilde{v})' - \left(\dot{\tilde{y}}_t - \frac{1}{2}\tilde{g}\right)'\tilde{g}; \quad \tilde{v}(\,\cdot\,, T) = 0. \tag{49}$$

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

$$\begin{aligned}\Pi(\tilde{g}, \tilde{y}) \;=\; & \int_0^T \int_{\mathbb{R}^n} \mathcal{D}\tilde{v}a\left(u_* - \frac{1}{2}(\mathcal{D}\tilde{v})'\right)(z, t, y) \\ & \cdot\; P_{X|Y}(\chi_t^{-1}(dz), y)\, dt.\end{aligned} \tag{50}$$

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.

Variational Bayes and a Problem of Reliable Communication: Part II: Infinite Systems (*)

Ensemble Filtering

Extensions to State Process described by Partial Differential Equation

Infinite-time Behavior of Filter

Controlled Filtering

Filtering Extensions of Majda's work

Recent work of Alexandre Chorin on Implicit Sampling for Particle Filtering

(*) Submitted: *Journal of Statistical Mechanics*