A VARIATIONAL APPROACH TO NONLINEAR ESTIMATION

SANJOY K. MITTER
DEPARTMENT OF ELECTRICAL ENGINEERING AND COMPUTER SCIENCE,
AND LABORATORY FOR INFORMATION AND DECISION SYSTEMS,
MASSACHUSETTS INSTITUTE OF TECHNOLOGY,
CAMBRIDGE, MA 02139, USA
AND
NIGEL J. NEWTON
DEPARTMENT OF ELECTRONIC SYSTEMS ENGINEERING,
UNIVERSITY OF ESSEX,
WIVENHOE PARK, COLCHESTER, CO4 3SQ, UK. *

Abstract. We consider estimation problems, in which the estimand, X, and observation, Y, take values in measurable spaces. Regular conditional versions of the forward and inverse Bayes formula are shown to have dual variational characterisations involving the minimisation of an *apparent information*, and the maximisation of a *compatible information*. These both have natural information theoretic interpretations, according to which Bayes' formula and its inverse are optimal information processors. The variational characterisation of the forward formula has the same form as that of Gibbs measures in statistical mechanics. The special case in which X and Y are diffusion processes governed by stochastic differential equations is examined in detail. The minimisation of apparent information can then be formulated as a stochastic optimal control problem, with cost that is quadratic in both the control and observation fit. The dual problem can be formulated in terms of infinite-dimensional deterministic optimal control. Local versions of the variational characterisations are developed, which quantify information *flow* in the estimators. In this context, the information conserving property of Bayesian estimators coincides with the Davis-Varaiya martingale stochastic dynamic programming principle.

Key words. Bayesian Inference, Information Theory, Legendre-type Transforms, Nonlinear Filtering, Stochastic Optimal Control.

AMS subject classifications. 93E11 93E20 94A15 62F15 60E10 60G35

1. Introduction.

This article investigates a variational formulation of Bayesian estimation with a natural information theoretic interpretation. The two 'directions' of an abstract Bayes formula (likelihood function to posterior distribution and vice-versa) are given variational representations. The forward representation involves the minimisation of an *apparent information* of probability measures on the space of the estimand. This apparent information is made up of two parts: the information gain of the measure over the prior distribution for the estimand, and a *residual* term representing the information value of the observation, complementary to this. The apparent information of probability measures is greater than or equal to the total information in the observation, with equality if and only if the measure is the posterior distribution of the estimand. Thus the (forward) Bayes formula can be thought of as an optimal 'information processor', in that it balances input and output information.

[*] This work was started while the second author was on a period of Study Leave visiting LIDS, supported by EPSRC grant GR/M97039. It continued during his visits to the INRIA research groups Omega and Sigma2. The second author would like to thank Denis Talay and Francois LeGland for their hospitality during these visits. Further support was provided by the Army Research Office under the MURI Grant: Data Fusion in Large Arrays of Microsensors DAAD19-00-1-0466; under Center for Imaging Science subcontract, Grant DAAD19-99-1-0012 (Johns Hopkins University), the MURI Grant: Vision Strategies and ATR Performance subcontract no. 654-21256 (Brown University) and an INTEL Grant. This publication is also an output from a research project funded by the Cambridge-MIT Institute ("CMI"). CMI is funded in part by the U.K. Government. The research was carried out for CMI by Massachusetts Institute of Technology. CMI can accept no responsibility for any information provided or views expressed.

Sub-optimal processors appear to have access to more information than that in the observation. The variational representation of the inverse Bayes formula involves the maximisation of a *compatible information* of likelihood functions on the space of the estimand. This is defined to be the difference between the information in an unspecified observation associated with the likelihood function, and that part of this information complementary to the (given) posterior distribution. The compatible information of likelihood functions is less than or equal to the information gain of the posterior distribution over the prior, with equality if and only if the likelihood function is equivalent to that provided by the inverse Bayes formula. Once again, the inverse Bayes formula can be thought of as an optimal processor, balancing input and output information. However, in this case, rather than appearing to have an additional source of information, sub-optimal processors lose (or fail to make use of) part of the input information.

In Section 2, the estimand, X, and the observation, Y, of the Bayesian problem are supposed to take values in Borel spaces $(\mathbf{X}, \mathcal{X})$ and $(\mathbf{Y}, \mathcal{Y})$, respectively. The starting point is a 'regular conditional' version of the Bayes formula. In Section 3, the results are specialised to the estimation of diffusion processes with partial observations. In that context, the regular conditional probability distribution can be chosen to be continuous in the observations. It also has the key property of being Markovian. This means that the family of measures over which apparent information is minimised can be restricted to the distributions of the process X when a 'finite energy', feedback control is applied through the drift coefficient. Thus, in this case, the minimisation of apparent information can be interpreted in terms of a problem in stochastic optimal control. This is explored in Section 4.

The dual variational problem for diffusion processes is developed in Section 5. One interpretation of it is as a problem in infinite-dimensional deterministic optimal control. The optimal trajectory of the dual problem is a 'likelihood filter' for the process X in reversed time, from which the corresponding nonlinear filter can be found. This gives new interpretation to a connection between an optimal control problem in one time direction and a nonlinear filter in the other, which was made for non-degenerate diffusions in [6] via the Hopf transformation, and used to give existence and uniqueness results for the unnormalised conditional density equation with unbounded observations. The results of Sections 3 to 5 are established under fairly weak conditions. In particular, they include the case of degenerate diffusions.

In the context of estimators for diffusion processes, there is a 'local' version of the variational formulations, which characterises flow rates of information, and shows that Bayesian processors are conservative in the sense that they balance input and output flow rates. This is the subject of Section 6.

A variational representation of the Fokker-Planck equation for diffusion processes is discussed in [10]. This involves the minimisation of the 'energy' of drift coefficients over those that give rise to a particular set of marginal densities. There, as here, the modification of the drift coefficient can be interpreted as the application of a *control* term, which re-expresses the variational problem as one in optimal control. The two problems are somewhat different though. In particular, the controls admitted in [10] give rise to mutually singular transition probabilities, which are certainly not permitted in the present context.

A preliminary account of some of the results herein was reported in [11].

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

$$\bar{\mathbf{Y}} = \left\{ y \in \mathbf{Y} : 0 < \int_{\mathbf{X}} Q(x,y) P_X(dx) < \infty \right\}; \tag{2.1}$$

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

$$\begin{aligned} H(x,y) = -\log(Q(x,y)) \qquad & \text{if } y \in \bar{\mathbf{Y}} \\ 0 \qquad & \text{otherwise}: \end{aligned} \tag{2.2}$$

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

$$P_{X|Y}(A,y) = \frac{\int_A \exp(-H(x,y)) P_X(dx)}{\int_{\mathbf{X}} \exp(-H(x,y)) P_X(dx)}, \tag{2.3}$$

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.

Equations (2.1)–(2.3) constitute an 'outcome-by-outcome' abstract Bayes formula, yielding a posterior probability distribution for X for each outcome of Y. Of course, for any y belonging to a set of P_Y-measure zero, $P_{X|Y}(\,\cdot\,,y)$ depends on the choice of version of the Radon-Nikodym derivative Q. However, in particular examples, we can often find a version such that $P_{X|Y}(A,\,\cdot\,)$ is continuous for each $A \in \mathcal{X}$.

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

$$\begin{aligned} h(\tilde{P}_X \,|\, \hat{P}_X) = \int_{\mathbf{X}} \log\left(\frac{d\tilde{P}_X}{d\hat{P}_X}\right) d\tilde{P}_X \quad & \text{if } \tilde{P}_X \ll \hat{P}_X \text{ and the integral exists} \\ +\infty \qquad & \text{otherwise}, \end{aligned} \tag{2.4}$$

$$\begin{aligned} i(\tilde{H}) = -\log\left(\int_{\mathbf{X}} \exp(-\tilde{H}) dP_X\right) \quad & \text{if } 0 < \int_{\mathbf{X}} \exp(-\tilde{H}) dP_X < \infty \\ -\infty \qquad & \text{otherwise}, \end{aligned} \tag{2.5}$$

$$\begin{aligned} \langle \tilde{H}, \tilde{P}_X \rangle = \int_{\mathbf{X}} \tilde{H} d\tilde{P}_X \qquad & \text{if the integral exists} \\ +\infty \qquad & \text{otherwise}. \end{aligned} \tag{2.6}$$

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 $Y = y$.)

PROPOSITION 2.1. *Suppose that (H1) is satisfied, and H and $P_{X|Y}$ are as defined above. Then for any y such that*

$$(2.7)\quad -\int_{\mathbf{X}} H(x,y)\exp(-H(x,y))P_X(dx) < \infty, \quad (\text{where } +\infty\exp(-\infty) = 0):$$

(i)

$$(2.8)\qquad i(H(\,\cdot\,,y)) = \min_{\tilde{P}_X \in \mathcal{P}(\mathcal{X})} \left\{ h(\tilde{P}_X \,|\, P_X) + \langle H(\,\cdot\,,y), \tilde{P}_X\rangle \right\};$$

(ii)

$$(2.9)\qquad h(P_{X|Y}(\,\cdot\,,y)\,|\,P_X) = \max_{\tilde{H}\in\mathcal{H}(\mathbf{X})} \left\{ i(\tilde{H}) - \langle \tilde{H}, P_{X|Y}(\,\cdot\,,y)\rangle \right\};$$

(iii) $P_{X|Y}(\,\cdot\,,y)$ is the unique minimiser in (2.8);
(iv) if H^* is a maximiser in (2.9), then there exists a real constant K such that

$$H^*(X) = H(X,y) + K \quad \text{a.s.}$$

Proof. If $y \in \bar{\mathbf{Y}}$ and (2.7) holds then $h(P_{X|Y}(\,\cdot\,,y)\,|\,P_X) < \infty$, $i(H(\,\cdot\,,y)) > -\infty$ and $H(\,\cdot\,,y) \in L_1(P_{X|Y}(\,\cdot\,,y))$. This is also true if $y \notin \bar{\mathbf{Y}}$ since, in that case, $H(\,\cdot\,,y) = 0$ and $P_{X|Y}(\,\cdot\,,y) = P_X$. Thus, it is clear that the minimum in (2.8) is less than $+\infty$, and the maximum in (2.9) is greater than $-\infty$.

Suppose that, for $\tilde{P}_X \in \mathcal{P}(\mathcal{X})$, $h(\tilde{P}_X\,|\,P_X) < \infty$ and $H(\,\cdot\,,y) \in L_1(\tilde{P}_X)$. It readily follows that $\tilde{P}_X \ll P_{X|Y}(\,\cdot\,,y)$, so that

$$h(\tilde{P}_X\,|\,P_X) = \int_{\mathbf{X}} \left(\log\left(\frac{d\tilde{P}_X}{dP_{X|Y}}(x,y)\right) + \log\left(\frac{dP_{X|Y}}{dP_X}(x,y)\right)\right)\tilde{P}_X(dx),$$

and

$$(2.10)\qquad h(\tilde{P}_X\,|\,P_X) + \langle H(\,\cdot\,,y), \tilde{P}_X\rangle = i(H(\,\cdot\,,y)) + h(\tilde{P}_X\,|\,P_{X|Y}(\,\cdot\,,y)).$$

It is easy to show that, for any $\tilde{P}_X \in \mathcal{P}(\mathcal{X})$, the relative entropy functional $h(\,\cdot\,|\,\tilde{P}_X)$ is non-negative, evaluates to zero at $\tilde{P}_X$, and is strictly convex on the subset of $\mathcal{P}(\mathcal{X})$ for which it is finite. This establishes parts (i) and (iii).

Suppose now that, for $\tilde{H} \in \mathcal{H}(\mathbf{X})$, $i(\tilde{H}) > -\infty$ and $\tilde{H} \in L_1(P_{X|Y}(\,\cdot\,,y))$. Let $\tilde{P}_X$ be defined by (2.3) with $\tilde{H}$ replacing $H(\,\cdot\,,y)$. It readily follows that $P_{X|Y}(\,\cdot\,,y) \ll \tilde{P}_X$, and so

$$
\begin{aligned}
i(\tilde{H}) - \tilde{H}(X) &= \log\left(\frac{d\tilde{P}_X}{dP_X}(X)\right) \\
&= \log\left(\frac{dP_{X|Y}}{dP_X}(X,y)\right) - \log\left(\frac{dP_{X|Y}}{d\tilde{P}_X}(X,y)\right).
\end{aligned}
$$

Thus

$$
(2.11) \qquad i(\tilde{H}) - \langle \tilde{H}, P_{X|Y}(\,\cdot\,,y)\rangle = h(P_{X|Y}(\,\cdot\,,y)\,|\,P_X) - h(P_{X|Y}(\,\cdot\,,y)\,|\,\tilde{P}_X).
$$

Suppose that there is a set $A \in \mathcal{X}$, for which $P_{X|Y}(A,y) = 0$ but $\tilde{P}_X(A) > 0$. Let $\tilde{P}'_X$ be defined by

$$
\tilde{P}'_X(B) = \left(\tilde{P}_X(A^C)\right)^{-1} \tilde{P}_X(A^C \cap B) \qquad \text{for all } B \in \mathcal{X}.
$$

Then $h(P_{X|Y}(\,\cdot\,,y)\,|\,\tilde{P}'_X) < h(P_{X|Y}(\,\cdot\,,y)\,|\,\tilde{P}_X)$, and so any maximiser in (2.11) must be absolutely continuous with respect to $P_{X|Y}(\,\cdot\,,y)$. It is easy to show that, for any $\tilde{P}_X \in \mathcal{P}(\mathcal{X})$, the relative entropy functional $h(\tilde{P}_X\,|\,\cdot\,)$ is non-negative, evaluates to zero at $\tilde{P}_X$, and is strictly convex on the subset of $\mathcal{P}(\mathcal{X})$ consisting of measures that are absolutely continuous with respect to $\tilde{P}_X$. This establishes parts (ii) and (iv). □

Remark 1. If the *mutual information* between X and Y is finite,

$$
(2.12) \qquad \int_{\mathbf{X}\times\mathbf{Y}} \log\left(\frac{dP_{XY}}{d(P_X \otimes P_Y)}\right) dP_{XY} < \infty,
$$

then there exists a version of Q for which (2.7) is satisfied for all y.

Remark 2. Proposition 2.1 is a special case of an energy-entropy duality that plays a major role in statistical physics and in the theory of large deviations. More general results of this nature are widely available in the literature. (See, for example, [5].) Our aim in this section is to provide an information-theoretic interpretation of the result in the Bayesian context. The simple proof we provide here makes use of the special nature of that context.

Parts (i) and (ii) of Proposition 2.1 both concern the processing of information over and above that in the prior P_X. In part (i), the source of additional information is the observation that $Y = y$. The abstract Bayes formula extracts the part of this information pertinent to X, $h(P_{X|Y}(\,\cdot\,,y)\,|\,P_X)$, and leaves the residual information, $\langle H(\,\cdot\,,y), P_{X|Y}(\,\cdot\,,y)\rangle$. One can think of the input information as being held in the likelihood function, $\exp(-H(\,\cdot\,,y))$, and the extracted information as being held in the distribution, $P_{X|Y}(\,\cdot\,,y)$. An arbitrary estimation procedure that postulates $\tilde{P}_X$ as a 'post-observation' distribution for X, appears to have access to additional information, in that it yields an information gain on X of $h(\tilde{P}_X\,|\,P_X)$, and a residual information of $\langle H(\,\cdot\,,y), \tilde{P}_X\rangle$. The sum of these two terms (the term in brackets on the right-hand side of (2.8)) is strictly greater than the actual information available, $i(H(\,\cdot\,,y))$, unless $\tilde{P}_X = P_{X|Y}(\,\cdot\,,y)$. We shall call it the *apparent information* of the estimator $\tilde{P}_X$. (Implicit in the interpretation of $h(\tilde{P}_X\,|\,P_X)$ as an information gain, is the assumption that $\tilde{P}_X$ represents a rational belief about X given the prior and some additional knowledge, such as an observation.)

In part (ii), the source of additional information is the posterior distribution, $P_{X|Y}(\,\cdot\,,y)$. The aim now is to postulate an observation (with likelihood function

$\exp(-\tilde{H}))$, which would give rise to this distribution. The input information here, $h(P_{X|Y}(\,\cdot\,,y)\,|\,P_X)$, is merged with the residual information of the postulated observation, $\langle \tilde{H}, P_{X|Y}(\,\cdot\,,y)\rangle$, and the result is greater than or equal to the total information in the postulated observation, $i(\tilde{H})$, with equality if and only if the observation is compatible with $P_{X|Y}(\,\cdot\,,y)$ in the sense of part (iv) of the proposition. The term in brackets on the right-hand side of (2.9) can be thought of as that part of the information in the postulated observation compatible with $P_{X|Y}(\,\cdot\,,y)$. We shall call it the *compatible information* of the likelihood function $\exp(-\tilde{H})$. Another interpretation is that the input information, $h(P_{X|Y}(\,\cdot\,,y)\,|\,P_X)$, is processed to produce compatible information resulting in a net loss of information except when the processor is optimal.

Throughout the rest of the paper, the apparent information and compatible information will be denoted by $F(\tilde{P}_X,y)$ and $G(\tilde{H},y)$, i.e.

$$F(\tilde{P}_X,y) = h(\tilde{P}_X\,|\,P_X) + \langle H(\,\cdot\,,y),\tilde{P}_X\rangle, \tag{2.13}$$

$$G(\tilde{H},y) = i(\tilde{H}) - \langle \tilde{H}, P_{X|Y}(\,\cdot\,,y)\rangle. \tag{2.14}$$

As equations (2.10) and (2.11) show, the minimisation of F is equivalent to the minimisation of the *information excess* of the estimator $\tilde{P}_X$, $h(\tilde{P}_X\,|\,P_{X|Y}(\,\cdot\,,y))$, and the maximisation of G is equivalent to the minimisation of the *information deficit* of the likelihood function $\exp(-\tilde{H})$, $h(P_{X|Y}(\,\cdot\,,y)\,|\,\tilde{P}_X)$. In fact (as was pointed out by an anonymous referee), these interpretations still hold in the absence of (2.7). However, in not identifying the source information or the extracted information, they do not show the information processing aspects of Bayesian estimation in quite the same way as the quantities F and G. Moreover, F and G make clear the compromises involved in Bayesian estimation. Part (i) of the proposition shows how $P_{X|Y}(\,\cdot\,,y)$ compromises between being close to the prior P_X and fitting with the observation $Y=y$, whereas part (ii) shows how $H(\,\cdot\,,y)$ (or its equivalents) compromise between holding a lot of information but not too much residual information.

Of course it is possible to give other variational characterisations of $P_{X|Y}(\,\cdot\,,y)$. For example, one could consider it as the minimiser of the total variation norm of the difference measure $\tilde{P}_X - P_{X|Y}(\,\cdot\,,y)$. However, such characterisations lack the information theoretic interpretation discussed above: F and G are natural error measures for sub-optimal estimation procedures. The characterisation (2.8) could be used as a basis for approximations. For example, we may wish to approximate a posterior distribution by a discrete law on a finite partition of $\mathbf{X}$. The size of the partition may be fixed, but we may be able to choose the law and the details of the partition by means of a finite number of parameters. The characterisation (2.8) could form the basis of an optimisation with respect to this set of parameters. Similarly, the characterisation (2.9) could be used as a basis for the study of modelling errors, in that it shows the information loss arising from the use of an incorrect likelihood function. Since the use of an incorrect prior, P_X^e (with $P_X^e \ll P_X$), with a Bayesian procedure is equivalent to the use of the incorrect likelihood function

$$\exp(-H^e(\,\cdot\,,y)) = \exp(-H(\,\cdot\,,y))\frac{dP_X^e}{dP_X},$$

(2.9), with $\tilde{H} = H^e(\,\cdot\,,y)$, also shows the information loss arising through the use of an incorrect prior. Furthermore, if there were any uncertainty in the likelihood function or the prior, the resulting information loss could be studied by means of game theoretic methods.

Proposition 2.1 is an instance of a Legendre-type transform between the relative entropy of probability measures and the logarithm of the exponential moment of real-valued random variables. A similar transform occurs in the characterisation of Gibbs measures in statistical mechanics, [8]. In that context, $(\mathbf{X}, \mathcal{X})$ is the *configuration space* of a physical system (the cartesian product of a number, N, of identical spaces), H is a *Hamiltonian* representing the energies of the configurations, and F is the *free energy* of the probability measure $\tilde{P}_X$ with respect to the reference measure P_X and H. A Gibbs measure represents a thermodynamic state of the system in thermodynamic equilibrium. If N is finite then there is only one Gibbs measure, and it takes the form (2.3). Gibbs theory comes into its full richness only when N is infinite, in which case there may be multiple Gibbs measures and formulae such as (2.3) are no longer appropriate. However, variational characterisations are. We note that the Bayesian estimator can be seen to compromise between being close to the prior and fitting with the observation in exactly the same way that a thermodynamic system in equilibrium compromises between maximising entropy and minimising average energy.

3. Path Estimators. The techniques of Section 2 are specialised here for the case in which the estimand, X, and observation, Y, are, respectively, continuous $I\!R^n$- and $I\!R^d$-valued processes governed by the following Itô integral equations:

$$\begin{aligned} X_t &= X_0 + \int_0^t b(X_s, s)\, ds + \int_0^t \sigma(X_s, s)\, dV_s, \quad \text{for } 0 \le t \le T, \\ X_0 &\sim \mu, \end{aligned} \tag{3.1}$$

$$Y_t = \int_0^t g(X_s)\, ds + W_t \quad \text{for } 0 \le t \le T, \tag{3.2}$$

where $X_t, V_t \in I\!R^n$, μ is a law on $(I\!R^n, \mathcal{B}^n)$, $Y_t, W_t \in I\!R^d$, 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 (3.1) and (3.2) are satisfied for all t. The abstract spaces $(\mathbf{X}, \mathcal{X})$ and $(\mathbf{Y}, \mathcal{Y})$ of Section 2 now become the spaces $(C([0,T]; I\!R^n), \mathcal{B}_T)$ and $(C([0,T]; I\!R^d), \mathcal{B}_T)$ of continuous functions, topologised 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 (H1) 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, $P_{X|Y}$, such that (2.7) 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

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

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

$$M_{y,t} = M_{y,0} \exp\left(\int_0^t U_{y,s}' \,(dX_s - b(X_s, s)\, ds) - \frac{1}{2} \int_0^t |\sigma(X_s, s)' U_{y,s}|^2\, ds \right) \tag{3.3}$$

for some progressively measurable, $I\!R^n$-valued process U_y. $P_{X|Y}(\,\cdot\,, y)$ will then be the distribution of a *controlled* process, X_y, satisfying an equation like (3.1), 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 characterisation of Section 2 will become a problem in stochastic optimal control.

We might also expect $P_{X|Y}(\,\cdot\,, y)$ to be Markov (at least for almost all y), in which case it will be appropriate to restrict admissible controls, $\tilde{U}$ to *feedback* controls of the form $u(\tilde{X}_t, t)$. It should also then be possible to define regular conditional *transition* probabilities for $P_{X|Y}$. With this in mind, let $(\chi_t, 0 \le t \le T)$ be the co-ordinate process on $\mathbf{X}$, and

$$\mathcal{X}_s^t = \sigma(\chi_r, s \le r \le t) \quad \text{for } 0 \le s \le t \le T. \tag{3.4}$$

We should be able to construct regular conditional probabilities

$$P_{X|Y}^{s+} : \mathcal{X}_s^T \times I\!R^n \times C([s,T];\, I\!R^d) \to [0,1]$$

such that, for all $A \in \mathcal{X}_s^T$,

$$P_{X|Y}(A, y) = \int_{R^n} P_{X|Y}^{s+}(A, z, (y_t - y_s, s \le t \le T))\, P_{X|Y}(\chi_s^{-1}(dz), y). \tag{3.5}$$

These will have variational characterisations in terms of the corresponding regular conditional probabilities for the prior, P_X, and appropriately constructed likelihood functions. This will lead towards a *localised* version of the results of Section 2.

In what follows, we develop the above ideas in a rigorous manner. We do this by placing constraints on b and σ such that (3.1) has a *strong* solution, and then use the techniques of stochastic flows. This has the advantage that we are able to include problems with degenerate diffusion coefficients, which are important in many areas of application. (In fact our approach also applies to some problems not satisfying a hypoellipticity condition.)

The constraints we place on μ, b, σ and g also fit well with Clark's *robustness* ideas (see [2]). These lead to an explicit function H and corresponding regular conditional probability, $P_{X|Y}$, that is Markov for every y. They also admit unbounded observation functions g, which are needed in the linear case.

We suppose that μ, b, σ and g satisfy the following technical conditions:

(H2) there exists an $\epsilon > 0$ such that

$$\int_{I\!R^n} \exp\left(\epsilon |z|^2\right) \mu(dz) < \infty;$$

(H3) σ is bounded, and b and σ are uniformly Lipschitz continuous on compact sets and differentiable with respect to the components of z, the derivatives being continuous and bounded;

(H4) g has continuous first, second and third derivatives, and there exist $C < \infty$ and $\alpha < \infty$ such that for all $z \in I\!R^n$

$$\sum_i \left| \frac{\partial g}{\partial z_i}(z) \right| \le C$$

$$\sum_{i,j} \left| \frac{\partial^2 g}{\partial z_i \partial z_j}(z) \right| \leq C(1 + |z|)$$

$$\text{and} \quad \sum_{i,j,k} \left| \frac{\partial^3 g}{\partial z_i \partial z_j \partial z_k}(z) \right| \leq C(1 + |z|^\alpha).$$

It follows from (H3) that (3.1) has a *strong* solution $\Phi : I\!R^n \times \mathbf{Y} \to \mathbf{X}$, so that on the probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t), P, X_0, (V, W))$ supporting an $I\!R^n$-valued random variable X_0 with distribution μ, and $(n + d)$-dimensional vector Brownian motion (V, W), independent of X_0, $(X_t = \Phi_t(X_0, V), \mathcal{F}_t; 0 \leq t \leq T)$ is a continuous semimartingale satisfying (3.1). (See, for example, [15].)

It follows from (H2)–(H4) that $\mathbf{E} \int_0^T |g(X_t)|^2 \, dt < \infty$, and from this and the independence of X and W it follows by standard results (see, for example, [9]) that (H1) is satisfied when the reference measure λ_Y is Wiener measure, and the Radon-Nikodym derivative takes the form:

$$\text{(3.6)} \qquad \frac{dP_{XY}}{d(P_X \otimes \lambda_Y)}(X, Y) = \exp\left(\int_0^T g(X_t)' \, dY_t - \frac{1}{2} \int_0^T |g(X_t)|^2 \, dt \right).$$

In order to develop the representations of Proposition 2.1 we first need a version of this that is well defined for all y. Under (H2)–(H4) the process $(g(X_t), \mathcal{F}_t, 0 \leq t \leq T)$ is a semimartingale, and so it is possible to 'integrate by parts' in (3.6) and define Q as any measurable function such that, for each y,

$$\text{(3.7)} \qquad Q(X, y) = \exp\left(y_T' g(X_T) - \int_0^T y_t' \, dg(X_t) - \frac{1}{2} \int_0^T |g(X_t)|^2 \, dt \right).$$

(See [2] and [3].) It can also be shown (see, for example, [13], [14]) that the resulting regular conditional probability, $P_{X|Y}$, is continuous in y in the sense of the topology associated with the convergence of means of bounded, measurable functions, that

$$\text{(3.8)} \qquad 0 < \mathbf{E} Q(X, y) < \infty \qquad \text{for all } y$$

and that

$$\text{(3.9)} \qquad \mathbf{E} Q(X, y) \log(Q(X, y)) \leq \mathbf{E} Q(X, y)^2 < \infty.$$

Thus the set $\bar{\mathbf{Y}}$ of (2.1) can be taken to be the entire space $\mathbf{Y}$ in this case, and (2.7) is satisfied for all y. Proposition 2.1 can thus be applied for each y, and $H = -\log(Q)$.

We can now split the path estimation problem as suggested by (3.5). For any $z \in I\!R^n$ and any $0 \leq s \leq T$, let $(X_t^{z,s}; s \leq t \leq T)$ be the solution of (3.1) on the interval $s \leq t \leq T$ with 'initial condition' $X_s^{z,s} = z$, and let

$$H_p : [0, T] \times [0, T] \times I\!R^n \times \mathbf{X} \times \mathbf{Y} \to I\!R$$

be a measurable function such that

$$\text{(3.10)} \qquad \begin{aligned} H_p(s, t, z, X^{z,s}, y) = &-y_t' g(X_t^{z,s}) + y_s' g(z) + \int_s^t y_r' \, dg(X_r^{z,s}) \\ &+ \frac{1}{2} \int_s^t |g(X_r^{z,s})|^2 \, dr \quad \text{for } 0 \leq s \leq t \leq T. \end{aligned}$$

The fact that such a function exists follows from the 'strong solution' hypothesis (H3), as does the decomposition

$$(3.11)\qquad H(X,y) = H_p(0,s,X_0,X,y) + H_p(s,T,X_s,(X_t, s \le t \le T),y).$$

$H_p(s,t,z,\,\cdot\,,\,\cdot\,)$ is the equivalent of H for the problem of estimating the path $(X_r^{z,s}, s \le r \le t)$ given the observation $(Y_r^{z,s}, s \le r \le t)$, where

$$Y_t^{z,s} = \int_s^t g(X_r^{z,s})\,dr + W_t - W_s \quad \text{for } s \le t \le T.$$

In particular, $H_p(s,T,z,\,\cdot\,,\,\cdot\,)$ is the equivalent of H for the problem of estimating $X^{z,s}$ given $Y^{z,s}$. Let $v(z,s,y)$ be the minimum apparent information for this problem; then, according to Proposition 2.1(i),

$$(3.12)\qquad v(z,s,y) = -\log\left(\mathbf{E}\exp(-H_p(s,T,z,X^{z,s},y))\right).$$

It now follows that, for any $A \in \mathcal{X}_0^s$,

$$(3.13)\qquad P_{X|Y}(A,y) = \frac{\mathbf{E}\mathbf{1}_A(X)\exp\left(-H_p(0,s,X_0,X,y) - v(X_s,s,y)\right)}{\mathbf{E}\exp\left(-H_p(0,s,X_0,X,y) - v(X_s,s,y)\right)},$$

and from Jensen's inequality and (3.9) it follows that $H_p(0,s,\chi_0,\,\cdot\,,y) + v(\chi_s(\,\cdot\,),s,y)$ satisfies (2.7) for all s. So, from Proposition 2.1, the path measure $P_{X|Y}$ restricted to $\mathcal{X}_0^s$ is the unique probability measure on $\mathcal{X}_0^s$ that minimises the apparent information

$$(3.14) F_s(\tilde{P}_{X,s},y) = h(\tilde{P}_{X,s}\,|\,P_{X,s}) + \langle H_p(0,s,\chi_0,\,\cdot\,,y)\,,\,\tilde{P}_{X,s}\rangle + \langle v(\chi_s,s,y)\,,\,\tilde{P}_{X,s}\rangle,$$

where $P_{X,s}$ is the restriction of P_X to $\mathcal{X}_0^s$. It also easily follows that the minimum apparent information in (3.14) does not depend on s.

These arguments show that the variational form of the path estimation problem (3.1), (3.2) can be interpreted in terms of dynamic programming, with value function v. For each s we can split the problem into two sub-problems: the estimation of $X^{z,s}$ for each z (resulting in a minimum apparent information of $v(z,s,y)$), followed by the estimation of $(X_t, 0 \le t \le s)$, where $v(X_s,s,y)$ plays a part in the likelihood function. $v(X_s,s,y)$ summarises that part of the likelihood function associated with increments of Y after time s. The first sub-problem can be interpreted in terms of stochastic optimal control, where the cost is the apparent information of the controlled process. This is developed in the next section.

4. A Stochastic Control Formulation. We consider the first variational sub-problem discussed above with $s = 0$. In keeping with the comments above on dynamic programming, it turns out that we need consider only feedback controls. Also, because controls are intended to produce a change in measure of the form (3.3), it is appropriate to let the control enter the drift through the map $z \mapsto az$, where $a = \sigma\sigma'$.

Consider the following controlled equation

$$(4.1)\qquad \tilde{X}_t = \theta + \int_0^t \left(b(\tilde{X}_s,s) + a(\tilde{X}_s,s)u(\tilde{X}_s,s)\right)\,ds + \int_0^t \sigma(\tilde{X}_s,s)\,d\tilde{V}_s,$$

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

$$\Gamma^u = \exp\left(\int_0^T u'\sigma(X_t^{\theta,0},t)\,dV_t - \tfrac{1}{2}\int_0^T |\sigma' u(X_t^{\theta,0},t)|^2\,dt\right), \tag{4.2}$$

and $(\Omega, \mathcal{F}, P)$, V and $X^{z,s}$ are as defined in Section 3.

LEMMA 4.1. *If b and σ satisfy (H3), and $u \in \mathbf{U}$ then equation (4.1) has a weak solution and is unique in law.*

Proof. From (H3) and (U1) it follows that

$$P\left(\int_0^T \left|\sigma' u(X_t^{\theta,0},t)\right|^2 dt < \infty\right) = 1.$$

This, together with (U2) and Girsanov's theorem, shows that V^u, defined by

$$V_t^u = V_t - \int_0^t \sigma' u(X_s^{\theta,0},s)\,ds, \tag{4.3}$$

is a standard Brownian motion under the probability measure P^u, defined by

$$\frac{dP^u}{dP} = \Gamma^u. \tag{4.4}$$

This shows that $(\Omega, \mathcal{F}, (\mathcal{F}_t), P^u, X^{\theta,0}, V^u)$ is a weak solution of (4.1).

Next, suppose that $(\tilde\Omega, \tilde{\mathcal{F}}, (\tilde{\mathcal{F}}_t), \tilde P, \tilde X, \tilde V)$ is a weak solution of (4.1), and, for each natural number N, let $\tau_N : \mathbf{X} \to [0,T]$ be defined by

$$\tau_N(x) = \inf\{t \geq 0 : |x_t| \geq N\} \wedge T.$$

Since $\tilde X$ is continuous $\tilde P\left(\tau_N(\tilde X) \to T\right) = 1$. Also, since u satisfies (U1),

$$\tilde{\mathbf{E}}\exp\left(\tfrac{1}{2}\int_0^{\tau_N(\tilde X)} \left|\sigma' u(\tilde X_s,s)\right|^2 ds\right) < \infty,$$

and so, from a standard variation of Novikov's theorem (see, for example, Theorem 6.1 in [9]), it follows that $(M_t, \tilde{\mathcal{F}}_t, 0 \leq t \leq T)$, where

$$M_t = \exp\left(-\int_0^t u'\sigma(\tilde X_s,s)\,d\tilde V_s - \tfrac{1}{2}\int_0^t \left|\sigma' u(\tilde X_s,s)\right|^2 ds\right), \tag{4.5}$$

is a local martingale with respect to the sequence of stopping times $(\tau_N(\tilde X);\ N = 1,2,\ldots)$. Let

$$\tilde V_t^N = \tilde V_t + \int_0^{t\wedge\tau_N(\tilde X)} \sigma' u(\tilde X_s,s)\,ds,$$

then, by Girsanov's theorem, $\tilde V^N$ is a standard Brownian motion under the probability measure $\tilde P^N$, defined by $d\tilde P^N = M_{\tau_N(\tilde X)} d\tilde P$. Let $(\mathcal{X}_t;\ 0 \leq t \leq T)$ be the filtration on $(\mathbf{X}, \mathcal{X})$ generated by the co-ordinate process (χ_t). Since

$$\tilde X_{t\wedge\tau_N(\tilde X)} = \Phi_{t\wedge\tau_N(\tilde X)}(\theta, \tilde V^N) \qquad \text{for } 0 \leq t \leq T,$$

where Φ is the strong solution to (3.1), the law of $\tilde{X}$ restricted to $\mathcal{X}_{\tau_N}$ is identical to that of $X^{\theta,0}$ under P^u, restricted to the same sigma-field. Finally, for any $A \in \mathcal{X}$,

$$\begin{aligned}\tilde{P}(\tilde{X} \in A, \tau_N(\tilde{X}) = T) &= \tilde{P}(\tilde{X} \in A) - \tilde{P}(\tilde{X} \in A, \tau_N(\tilde{X}) < T) \\ &\to \tilde{P}(\tilde{X} \in A),\end{aligned}$$

and so, since the events on the left-hand side each belong to one of $(\mathcal{X}_{\tau_N};\, N = 1, 2, \ldots)$, the law of $\tilde{X}$ on $\mathcal{X}$ is identical to that of $X^{\theta,0}$ under P^u. □

Let $(\tilde{\Omega}, \tilde{\mathcal{F}}, (\tilde{\mathcal{F}}_t), \tilde{P}, \tilde{X}, \tilde{V})$ be a weak solution of (4.1) 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 $H_p(0, T, \theta, \,\cdot\,, y)$ (as defined in (3.10).

$$\begin{aligned} J(u,\theta,y) &= h(\tilde{P}_X \mid P_X^{\theta,0}) + \langle H_p(0,T,\theta,\,\cdot\,,y), \tilde{P}_X\rangle \\ &= \tfrac{1}{2}\tilde{\mathbf{E}}\int_0^T |\sigma' u(\tilde{X}_t,t)|^2\,dt - y_T' g(\theta) + \tfrac{1}{2}\tilde{\mathbf{E}}\int_0^T |g(\tilde{X}_t)|^2\,dt \\ &\qquad -\tilde{\mathbf{E}}\int_0^T (y_T - y_t)'(\mathcal{L}g + \mathcal{D}gau)(\tilde{X}_t,t)\,dt \qquad \text{if the integrals exist} \\ &\qquad +\infty \qquad \text{otherwise,} \end{aligned} \tag{4.6}$$

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

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

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:

$$J(u,\theta,y) = \tfrac{1}{2}\tilde{\mathbf{E}}\int_0^T \left(|\sigma' u(\tilde{X}_t,t)|^2 + |\dot{y}_t - g(\tilde{X}_t)|^2\right)dt - \tfrac{1}{2}\int_0^T |\dot{y}_t|^2\,dt. \tag{4.7}$$

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 (4.1), (4.7) 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 4.2. *Suppose that b, σ and g satisfy (H3) and (H4), and let the function $u_* : I\!R^n \times [0,T] \times \mathbf{Y} \to I\!R^n$ be defined by*

$$u_* = -(\mathcal{D}v)', \tag{4.8}$$

where v is as defined in (3.12). Then, for each $y \in \mathbf{Y}$, $u_*(\,\cdot\,,\,\cdot\,,y)$ belongs to $\mathbf{U}$, and for all $\theta \in I\!R^n$, $y \in \mathbf{Y}$ and $\tilde{P}_X \in \mathcal{P}(\mathcal{X})$ (not necessarily the distribution of a controlled process),

$$J(u_*(\,\cdot\,,\,\cdot\,,y),\theta,y) \le h(\tilde{P}_X \mid P_X^{\theta,0}) + \langle H_p(0,T,\theta,\,\cdot\,,y), \tilde{P}_X\rangle. \tag{4.9}$$

Proof. The proof is in three parts. The first uses the methods of stochastic flows to establish a stochastic representation formula for u_*, (4.20). The second proves the statement of the theorem for non-degenerate systems with bounded coefficients. Finally, a truncation argument is used to extend this result to the general case. Only the time-homogeneous case (b and σ not dependent on t) is treated in order to avoid excessive notation. The arguments extend in an obvious way to the general case.

Standard moment bounding arguments (see, for example, Theorem 4.6 in [9]) show that for each natural number m there exists a $C_m < \infty$, not depending on z or s, such that

$$\sup_{s\le t\le T} \mathbf{E}|X_t^{z,s}|^{2m} \le C_m \left(1 + |z|^{2m}\right) \tag{4.10}$$

$$\text{and} \quad \sup_{s\le t\le T} \mathbf{E}\,\|\Psi_t^{z,s}\|^{2m} \le C_m, \tag{4.11}$$

where $(\Psi_t^{z,s} \in I\!\!R^{n\times n}; s \le t \le T)$ is the solution of the equation of first-order variation associated with $X^{z,s}$,

$$\Psi_t^{z,s} = I + \int_s^t \mathcal{D}b(X_r^{z,s})\Psi_r^{z,s}\,dr + \sum_i \int_s^t \mathcal{D}\sigma_i(X_r^{z,s})\Psi_r^{z,s}\,dV_{i,r}. \tag{4.12}$$

Here, and in what follows, σ_i is the i'th column of σ, and $V_{i,t}$ is the i'th component of V_t. For any $z, \tilde{z} \in I\!\!R^n$ and any $0 \le s \le t \le T$

$$X_t^{z,s} - X_t^{\tilde{z},s} = (z - \tilde{z}) + \int_s^t (b(X_r^{z,s}) - b(X_r^{\tilde{z},s}))\,dr + \int_s^t (\sigma(X_r^{z,s}) - \sigma(X_r^{\tilde{z},s}))\,dV_r,$$

and so for any natural number m there exists a $C_m < \infty$, not depending on s, t, z or $\tilde{z}$, such that

$$\begin{aligned}\mathbf{E}\sup_{s\le r\le t}\left|X_r^{z,s} - X_r^{\tilde{z},s}\right|^{2m} &\le 3^{2m-1}\Bigg(|z-\tilde{z}|^{2m} + \mathbf{E}\sup_{s\le r\le t}\left|\int_s^r (b(X_q^{z,s}) - b(X_q^{\tilde{z},s}))\,dq\right|^{2m} \\ &\qquad + \mathbf{E}\sup_{s\le r\le t}\left|\int_s^r (\sigma(X_q^{z,s}) - \sigma(X_q^{\tilde{z},s}))\,dV_q\right|^{2m}\Bigg) \\ &\le C_m\left(|z-\tilde{z}|^{2m} + \int_s^t \mathbf{E}\sup_{s\le q\le r}\left|X_q^{z,s} - X_q^{\tilde{z},s}\right|^{2m}\,dr\right),\end{aligned}$$

where we have used Doob's submartingale inequality, (4.10), (H3) and standard bounds for the moments of stochastic integrals. It thus follows from the Gronwall lemma that

$$\mathbf{E}\sup_{s\le t\le T}|X_t^{z,s} - X_t^{\tilde{z},s}|^{2m} \le C_m\exp(C_mT)|z-\tilde{z}|^{2m} \qquad \text{for all } (z,\tilde{z},s). \tag{4.13}$$

Similarly,

$$\mathbf{E}\sup_{s\le t\le T}|X_t^{z,s}|^{2m} \le C_m(1+|z|^{2m}) \qquad \text{for all } (z,s), \tag{4.14}$$

and so for any $\epsilon > 0$ and any bounded set $A \subset I\!\!R^n$ there exists a $C < \infty$ such that

$$P\left(\sup_{s\le t\le T}|X_t^{z,s}| > C\right) < \epsilon/4 \qquad \text{for all } (z,s) \in A\times[0,T].$$

From (H3) and (H4) it follows that $\mathcal{D}(\mathcal{L}g)$ is uniformly continuous on compacts, and so for any $\eta > 0$ there exists a $\delta > 0$ such that if $z, \tilde{z} \in A$ and $|z - \tilde{z}| < \delta$

$$P\left(\sup_{s\le t\le T}\left\|\mathcal{D}(\mathcal{L}g)(X_t^{z,s}) - \mathcal{D}(\mathcal{L}g)(X_t^{\tilde{z},s})\right\| > \eta, \sup_{s\le t\le T}(|X_t^{z,s}| \vee |X_t^{\tilde{z},s}|) \le C\right) < \epsilon/2,$$

so that

$$P\left(\sup_{s\le t\le T}\left\|\mathcal{D}(\mathcal{L}g)(X_t^{z,s}) - \mathcal{D}(\mathcal{L}g)(X_t^{\tilde{z},s})\right\| > \eta\right) < \epsilon. \tag{4.15}$$

The polynomial growth of $\mathcal{D}(\mathcal{L}g)$ together with (4.14) and the Vallée-Poussin theorem shows that, for any $0 < p < \infty$, the family

$$\left\{\sup_{s\le t\le T}\|\mathcal{D}(\mathcal{L}g)(X_t^{z,s})\|^p\,; z \in A, 0 \le s \le T\right\},$$

is uniformly integrable. This and (4.15) show that for any $0 < p < \infty$

$$\mathbf{E}\sup_{s\le t\le T}\left\|\mathcal{D}(\mathcal{L}g)(X_t^{z,s}) - \mathcal{D}(\mathcal{L}g)(X_t^{\tilde{z},s})\right\|^p = o(|z-\tilde{z}|^0) \tag{4.16}$$

uniformly on $A \times [0,T]$. Similar arguments show that $\mathcal{D}g(X_t^{z,s})$, $\mathcal{D}b(X_t^{z,s})$, $\mathcal{D}\sigma_i(X_t^{z,s})$ and $\mathcal{D}(\mathcal{D}g\sigma_i)(X_t^{z,s})$, for $i = 1, 2, \ldots, n$, have the same property.

It follows from the mean-value theorem that

$$\begin{aligned} X_t^{z,s} - X_t^{\tilde{z},s} = (z - \tilde{z}) &+ \int_s^t \mathcal{D}b\left(\alpha_{0,r}X_r^{z,s} + (1-\alpha_{0,r})X_r^{\tilde{z},s}\right)(X_r^{z,s} - X_r^{\tilde{z},s})\,dr \\ &+ \sum_i \int_s^t \mathcal{D}\sigma_i\left(\alpha_{i,r}X_r^{z,s} + (1-\alpha_{i,r})X_r^{\tilde{z},s}\right)(X_r^{z,s} - X_r^{\tilde{z},s})\,dV_{i,r}, \end{aligned}$$

where $0 < \alpha_{i,r} < 1$ and $\alpha_{i,r}$ is $\mathcal{F}_r$-measurable for each i. The above continuity properties, Hölder's inequality and techniques similar to those used to prove (4.13) now show that for any $0 < p < \infty$

$$\mathbf{E}\sup_{s\le t\le T}\left|X_t^{z,s} - X_t^{\tilde{z},s} - \Psi_t^{z,s}(z-\tilde{z})\right|^p = o(|z-\tilde{z}|^p), \tag{4.17}$$

and

$$\mathbf{E}\left|\Theta(z,s,y) - \Theta(\tilde{z},s,y) - \xi(z,s,y)\Theta(z,s,y)(z-\tilde{z})\right|^p = o(|z-\tilde{z}|^p), \tag{4.18}$$

both uniformly on $A \times [0,T]$, where

$$\Theta(z,s,y) = \exp\left(-H_p(s,T,z,X^{z,s},y)\right)$$

and

$$\begin{aligned} \xi(z,s,y) = (y_T - y_s)'\mathcal{D}g(z) &+ \sum_i \int_s^T (y_T - y_t)'\mathcal{D}(\mathcal{D}g\sigma_i)(X_t^{z,s})\Psi_t^{z,s}\,dV_{i,t} \\ &+ \int_s^T (y_T - y_t)'\mathcal{D}(\mathcal{L}g)(X_t^{z,s})\Psi_t^{z,s}\,dt - \int_s^T g'(X_t^{z,s})\mathcal{D}g(X_t^{z,s})\Psi_t^{z,s}\,dt. \end{aligned}$$

Thus $\mathcal{D}\rho = \mathbf{E}\xi\Theta$, where $\rho = \mathbf{E}\Theta$. Now, Jensen's inequality shows that

$$\inf_{z\in A, 0\le s\le T} \rho(z,s,y) \ge \inf_{z\in A, 0\le s\le T} \exp(\mathbf{E}\log(\Theta(z,s,y))) > 0, \tag{4.19}$$

and so

$$u_*(z,s,y) = \frac{\mathbf{E}\xi(z,s,y)\Theta(z,s,y)}{\mathbf{E}\Theta(z,s,y)}. \tag{4.20}$$

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

$$\tilde{z}' a(z)\tilde{z} \ge \epsilon|\tilde{z}|^2 \qquad \text{for all } z, \tilde{z} \in I\!R^n. \tag{4.21}$$

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. (see, for example, [7])

$$\frac{\partial \rho}{\partial s} + \mathcal{L}\rho + \left(\dot{y} - \tfrac{1}{2}g\right)' g\rho = 0 \quad \text{on } I\!R^n \times (0,T), \quad \rho(\,\cdot\,,T,y) = 1. \tag{4.22}$$

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

$$\frac{\partial v}{\partial s} + \mathcal{L}v - \tfrac{1}{2}\mathcal{D}va(\mathcal{D}v)' - \left(\dot{y} - \tfrac{1}{2}g\right)' g = 0 \quad \text{on } I\!R^n \times (0,T), \quad v(\,\cdot\,,T,y) = 0. \tag{4.23}$$

Now, because of (4.10), (4.11) and the boundedness of g and $\mathcal{D}g$, $u_*(\,\cdot\,,\,\cdot\,,y)$ is also bounded and, by Novikov's theorem, satisfies (U2). We have thus shown that in this special case $u_*(\,\cdot\,,\,\cdot\,,y) \in \mathbf{U}$. Let V^* and P^* be abbreviations for $V^{u_*(\,\cdot\,,\,\cdot\,,y)}$ and $P^{u_*(\,\cdot\,,\,\cdot\,,y)}$, respectively, where, for $u \in \mathbf{U}$, V^u and P^u are as defined by (4.3) and (4.4). Then Itô's rule and (4.23) show that

$$\begin{aligned} 0 = v(X_T^{\theta,0}, T, y) = v(\theta,0,y) &+ \int_0^T \left(\left(\dot{y}_t - \tfrac{1}{2}g\right)' g - \tfrac{1}{2}|\sigma' u_*|^2\right)(X_t^{\theta,0}, t, y)\,dt \\ &- \int_0^T (u_*'\sigma)(X_t^{\theta,0}, t, y)\,dV_t^*. \end{aligned}$$

As was pointed out in the proof of Lemma 4.1, $(\Omega, \mathcal{F}, (\mathcal{F}_t), P^*, X^{\theta,0}, V^*)$ is a weak solution of (4.1) and so, since g, $u_*(\,\cdot\,,\,\cdot\,,y)$ and σ are bounded,

$$\begin{aligned} v(\theta,0,y) &= \mathbf{E}^* \int_0^T \left(\tfrac{1}{2}|\sigma' u_*| - \left(\dot{y}_t - \tfrac{1}{2}g\right)' g\right)(X_t^{\theta,0}, t, y)\,dt \\ &= J(u_*(\,\cdot\,,\,\cdot\,,y), \theta, y). \end{aligned}$$

By definition, $v(\theta,0,y)$ is the minimum apparent information, and so we have established (4.9) in this special case. A consequence of (4.9), and the uniqueness of the measure minimising apparent information, is that the distribution of $\tilde{X}$ when $u = u_*(\,\cdot\,,\,\cdot\,,y)$ is the regular conditional distribution of $X^{\theta,0}$ given that $Y = y$. Thus, in this special case,

$$\Gamma^{u_*(\,\cdot\,,\,\cdot\,,y)} = \frac{\Theta(\theta,0,y)}{\rho(\theta,0,y)} \quad \text{a.s.}$$

Next, suppose that the additional constraints placed on y, b, g and σ are removed. For any natural number N, let

$$\begin{aligned}
&b_N(z) = b(z)\exp(-|z|^2/N),\\
&g_N(z) = g(z)\exp(-|z|^2/N),\\
&\sigma_N(z) = \big[\sigma\ N^{-1}I\big] \qquad \text{(an } n \times 2n \text{ matrix)},
\end{aligned}$$

and let y^N be a sequence of differentiable elements of $\mathbf{Y}$ with Hölder continuous derivatives such that $\|y - y^N\| \to 0$. Then b_N and g_N are bounded and σ_N satisfies (4.21), b_N, σ_N and g_N satisfy (H3) and (H4) uniformly in N, and b_N, σ_N, g_N, $\mathcal{D}b_N$, $\partial\sigma_N/\partial z_i$ and $\mathcal{D}g_N$ converge to b, $[\sigma\, 0]$, g, $\mathcal{D}b$, $[\partial\sigma/\partial z_i\, 0]$ and $\mathcal{D}g$ (respectively) uniformly on compacts. We add the subscript (or superscript) N to X, Ψ, Θ etc. to indicate that y, b, g and σ have been replaced by y^N, b_N, g_N and σ_N in the various definitions, and that V has been replaced by the $2n$-dimensional Brownian motion, (V_t, B_t). Now

$$\begin{aligned}
X_t^{z,s} - X_t^{N,z,s} = &\int_s^t \big(b_N(X_r^{z,s}) - b_N(X_r^{N,z,s})\big)\, dr + \int_s^t \big(\sigma(X_r^{z,s}) - \sigma(X_r^{N,z,s})\big)\, dV_r \\
&+ \int_s^t (b(X_r^{z,s}) - b_N(X_r^{z,s}))\, dr - N^{-1}(B_t - B_s).
\end{aligned}$$

Arguments similar to those used to prove (4.13), (4.17) and (4.18) show that, for any natural number m and any bounded set $A \subset I\!R^n$,

$$\mathbf{E} \sup_{s\le t\le T} \left| X_t^{z,s} - X_t^{N,z,s} \right|^{2m} \to 0, \tag{4.24}$$

$$\mathbf{E} \sup_{s\le t\le T} \left\| \Psi_t^{z,s} - \Psi_t^{N,z,s} \right\|^{2m} \to 0,$$

$$\mathbf{E} \left| \Theta(z,s,y) - \Theta_N(z,s,y^N) \right|^{2m} \to 0 \tag{4.25}$$

$$\text{and} \quad \mathbf{E} \left| \xi(z,s,y) - \xi_N(z,s,y^N) \right|^{2m} \to 0,$$

all uniformly on $A \times [0,T]$. This, Hölder's inequality and (4.19) show that

$$u_{*N}(\,\cdot\,,\,\cdot\,,y^N) \to u_*(\,\cdot\,,\,\cdot\,,y) \qquad \text{uniformly on } A \times [0,T]. \tag{4.26}$$

Thus $u_*(\,\cdot\,,\,\cdot\,,y)$ satisfies (U1). It follows from (4.24) and (4.26) that

$$\sup_{0\le t\le T} \left| u_*(X_t^{\theta,0},t,y) - u_{*N}(X_t^{N,\theta,0},t,y^N) \right| \to 0 \qquad \text{in probability,}$$

so that

$$\Gamma_N^{u_{*N}(\,\cdot\,,\,\cdot\,,y^N)} \to \Gamma^{u_*(\,\cdot\,,\,\cdot\,,y)} \qquad \text{in probability.} \tag{4.27}$$

It also follows from (4.25) and (4.19) that

$$\Gamma_N^{u_{*N}(\,\cdot\,,\,\cdot\,,y^N)} = \frac{\Theta_N(\theta,0,y^N)}{\rho_N(\theta,0,y^N)} \to \frac{\Theta(\theta,0,y)}{\rho(\theta,0,y)} \qquad \text{in probability,} \tag{4.28}$$

and so $u_*(\,\cdot\,,\,\cdot\,,y)$ satisfies (U2), and the unique distribution of $\tilde{X}$ under this control coincides with the regular conditional distribution of X given that $Y = y$. This establishes (4.9) in the general case. □

We return now to the path estimator with initial distribution μ. The minimisation of apparent information can be expressed in terms of the following controlled process with random initial condition:

$$(4.29)\qquad \begin{aligned} &\tilde{X}_t = \tilde{X}_0 + \int_0^t \left(b(\tilde{X}_s, s) + a(\tilde{X}_s, s)u(\tilde{X}_s, s) \right) ds + \int_0^t \sigma(\tilde{X}_s, s)\, d\tilde{V}_s, \\ &\tilde{X}_0 \sim \tilde{\mu}. \end{aligned}$$

A simple variant of Lemma 4.1 shows that, if u is continuous and satisfies (U2) for all $\theta \in \mathbb{R}^n$, then this equation is unique in law and has a weak solution for any initial law, $\tilde{\mu}$. Let $\tilde{P}_X$ be the distribution of $\tilde{X}$ corresponding to the pair $(\tilde{\mu}, u)$; it follows from (3.14) and the subsequent discussion that

$$(4.30)\qquad F(\tilde{P}_X, y) = F_0(\tilde{\mu}, y) = h(\tilde{\mu} \,|\, \mu) + \langle J(u, \cdot, y), \tilde{\mu} \rangle,$$

and this is minimised by the choice $u = u_*(\cdot, \cdot, y)$ and $\tilde{\mu} = \mu_Y(\cdot, y)$, where for $B \in \mathcal{B}^n$

$$(4.31)\qquad \mu_Y(B, y) = P_{X|Y}(\chi_0^{-1}(B), y).$$

Thus, for each y, the regular conditional probability distribution $P_{X|Y}(\cdot, y)$ is Markovian with 'initial' marginal $\mu_Y(\cdot, y)$ and differential operator

$$(4.32)\qquad \mathcal{L}_y = \sum_i (b + au_*(\cdot, \cdot, y))_i \frac{\partial}{\partial z_i} + \tfrac{1}{2} \sum_{i,j} a_{i,j} \frac{\partial^2}{\partial z_i \partial z_j}.$$

Of course, the nonlinear filter and interpolator for the process X can be found from the marginals of this path space measure.

5. The Inverse Problem.

The variational characterisation of the inverse problem (parts (ii) and (iv) of Proposition 2.1) can also be applied to the path estimator. This involves choosing a likelihood function to be compatible with the (given) regular conditional probability distribution, $P_{X|Y}(\cdot, y)$. In Section 4, we minimised apparent information over probability measures corresponding to weak solutions of (4.29). Here, we maximise 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 in Section 3. 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

$$(5.1)\qquad \begin{aligned} \tilde{H}(X) &= -\log\left(\frac{d\tilde{P}_X}{dP_X}(X) \right) + K \\ &= -\log\left(\frac{d\tilde{\mu}}{d\mu}(X_0) \right) - \int_0^T u'\sigma(X_t, t)\, dV_t + \tfrac{1}{2} \int_0^T |\sigma' u(X_t, t)|^2\, dt + K, \end{aligned}$$

where $K \in \mathbb{R}$ and $\tilde{P}_X$ is as defined following (4.29). We shall assume that $\mu_Y(\cdot, y) \ll \tilde{\mu}$. If this is not so, then, as shown in the proof of Proposition 2.1, we can always choose another $\tilde{\mu}$ resulting in more compatible information, for which it is. The term K in (5.1) is the information in the associated (unspecified) observation.

Integral log-likelihood functions of the form (5.1) 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 characterisation of $P_{X|Y}$ in terms if stochastic control can be used to express the compatible information corresponding to $\tilde{H}$, as follows:

$$\begin{aligned} G(\tilde{H}, y) &= K - \langle \tilde{H}\,,\, P_{X|Y}(\,\cdot\,, y)\rangle \\ &= K + h(\mu_Y(\,\cdot\,, y)\,|\,\mu) - h(\mu_Y(\,\cdot\,, y)\,|\,\tilde{\mu}) \\ &\quad + \int_0^T \int_{I\!R^n} \left(u_* - \frac{1}{2}u\right)' au(z,t,y)P_{X|Y}(\chi_t^{-1}(dz), y)\,dt. \end{aligned} \tag{5.2}$$

Log-likelihood functions of the form (5.1) 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 (3.2), 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:

$$-\frac{\partial \tilde{v}}{\partial t} = \mathcal{L}\tilde{v} - \frac{1}{2}\mathcal{D}\tilde{v}a(\mathcal{D}\tilde{v})' - \left(\dot{\tilde{y}}_t - \frac{1}{2}\tilde{g}\right)'\tilde{g}; \quad \tilde{v}(\,\cdot\,, T) = 0. \tag{5.3}$$

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

$$\Pi(\tilde{g}, \tilde{y}) = \int_0^T \int_{I\!R^n} \mathcal{D}\tilde{v}a\left(u_* - \frac{1}{2}(\mathcal{D}\tilde{v})'\right)(z,t,y)P_{X|Y}(\chi_t^{-1}(dz), y)\,dt. \tag{5.4}$$

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-normalised regular conditional probability distribution for X_s given observations $(Y_t - Y_s, s \le t \le T)$, 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 [6], 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 [12]. For a somewhat different problem involving optimisation over observation functions, see [16].

6. Information Flow and Localisation. The results of Section 2 concerning the information conserving properties of Bayesian estimators can be localised in the context of the diffusion problem (3.1), (3.2). Proposition 2.1 can be applied to provide variational characterisations of various conditional probabilities of the path measure $P_{X|Y}$, including transition probabilities, and these can be used to characterise the *flow* of information at a given time and in a given state.

For any initial law $\tilde{\mu} \ll \mu$, and any control u satisfying (U1) and (U2) for all θ, let $(\tilde{\Omega}, \tilde{\mathcal{F}}, (\tilde{\mathcal{F}}_t), \tilde{P}, \tilde{X}, \tilde{V})$ be a weak solution of (4.29), let $\tilde{P}_X$ be the distribution of

$\tilde{X}$, and let $P_{X,s}$, $\tilde{P}_{X,s}$ and $P_{X,s|Y}(\,\cdot\,,y)$ be the restrictions of P_X, $\tilde{P}_X$ and $P_{X|Y}(\,\cdot\,,y)$ to $\mathcal{X}_0^s$ (as defined in (3.4)). It follows from the results of Section 4 that $P_{X,s|Y}(\,\cdot\,,y)$ coincides with $\tilde{P}_{X,s}$ when $\tilde{\mu} = \mu_Y(\,\cdot\,,y)$ and $u(\,\cdot\,,t) = u_*(\,\cdot\,,t,y)$ for $0 \le t \le s$. As shown in the discussion following (3.13), this is the unique probability measure on $\mathcal{X}_0^s$ minimising the apparent information (3.14). The sum of the first two terms on the right-hand side of (3.14) is the apparent information of $\tilde{P}_{X,s}$ in the context of estimators of $(X_t, 0 \le t \le s)$ given observations $(Y_t, 0 \le t \le s)$, which we can think of as being the apparent information *up to* time s. The third term on the right-hand side of (3.14) is the information in the observations $(Y_t - Y_s, s \le t \le T)$, which we can think of as being the information *remaining* in the observations Y at time s. As s increases, the estimator corresponding to $(\tilde{\mu}, u)$ progressively converts observation information into apparent information. If $u = u_*(\,\cdot\,,\,\cdot\,,y)$ then this process is conservative, in that $F_s(\tilde{P}_{X,s}, y)$ does not change with s. However, if u is not optimal then the apparent information can increase faster than the observation information decreases.

We can refine this argument as follows. Let

$$\tilde{I}_s = \log\left(\frac{d\tilde{P}_{X,s}}{dP_{X,s}}(\tilde{X})\right) + H_p(0,s,\tilde{X}_0,\tilde{X},y) + v(\tilde{X}_s,s,y) \quad \text{for } 0 \le s \le T, \tag{6.1}$$

where H_p is defined in (3.10). Then it follows from (3.11) that, for all $0 \le s \le t \le T$,

$$\begin{aligned}\tilde{I}_t = \tilde{I}_s &+ \log\left(\frac{d\tilde{P}_{X,t}}{dP_{X,t}} \times \frac{dP_{X,s}}{d\tilde{P}_{X,s}}(\tilde{X})\right) + H_p(s,t,\tilde{X}_s,(\tilde{X}_r, s \le r \le T),y) \\ &+ v(\tilde{X}_t,t,y) - v(\tilde{X}_s,s,y).\end{aligned} \tag{6.2}$$

Let $\tilde{Q}_X$ and Q_X be, respectively, the distributions of $(X_r^{z,s}, s \le r \le t)$ (as defined in Section 3) with and without the application of the control $(u(X_r^{z,s},r), s \le r \le t)$. The apparent information of $\tilde{Q}_X$ in the context of estimators for $(X_r^{z,s}, s \le r \le t)$ given $Y^{z,s}$ is

$$\begin{aligned}F_{s,t}(z,\tilde{Q}_X,y) &= h(\tilde{Q}_X\,|\,Q_X) + \langle H_p(s,t,z,\,\cdot\,,y),\tilde{Q}_X\rangle + \langle v(\chi_t,t,y),\tilde{Q}_X\rangle, \\ &= v(z,s,y) + \frac{1}{2}\int_s^t \int_{I\!R^n} |\sigma'(u-u_*(\tilde{z},r,y))|^2\,\tilde{Q}_X(\chi_r^{-1}(d\tilde{z}))\,dr,\end{aligned} \tag{6.3}$$

where we have used (2.10). It now follows that

$$\tilde{\mathbf{E}}(\tilde{I}_t\,|\,\tilde{\mathcal{F}}_s) = \tilde{I}_s + \frac{1}{2}\int_s^t \tilde{\mathbf{E}}\left(|\sigma'(u-u_*)(\tilde{X}_r,r,y)|^2 \,\middle|\, \tilde{\mathcal{F}}_s\right) dr.$$

Thus $(\tilde{I},\tilde{\mathcal{F}}_t)$ is a sub-martingale, and a martingale if $u = u_*(\,\cdot\,,\,\cdot\,,y)$. This is the Davis-Varaiya characterisation of the optimal control for the problem of Section 4, [4].

Setting $t = s + \delta s$ in (6.3) we obtain the following local information quantities.

$$h(\tilde{Q}_X|Q_X) = \frac{1}{2}|\sigma' u(z,s)|^2\delta s + o(\delta s), \tag{6.4}$$

$$\langle H_p(s,s+\delta s,z,\,\cdot\,,y),\tilde{Q}_X\rangle = -g(z)'\delta y + \frac{1}{2}|g(z)|^2\delta s + o(\delta s), \tag{6.5}$$

$$\begin{aligned}\langle v(\chi_{s+\delta s},s+\delta s,y),\tilde{Q}_X\rangle &= v(z,s,y) + g(z)'\delta y \\ &- \left(\left(u - \frac{1}{2}u_*\right)' a u_* + \frac{1}{2}|g|^2\right)(z,s,y)\delta s + o(\delta s)\end{aligned} \tag{6.6}$$

Equation (6.4) shows the local increase in information gain of the distribution of the process (4.29) over P_X, equation (6.5) shows the local increase in the residual information of the estimator $\tilde{P}_X$, and equation (6.6) shows the local decrease in the average information remaining in the observation after time s. If y is differentiable at s, then there is a local rate of increase of apparent information of $|\sigma' u(z,s)|^2/2 - (\dot{y}_s - g/2)'g(z)$, and a local rate of decrease of remaining observation information of $(u - u_*/2)'au^*(z,s,y) - (\dot{y}_s - g/2)'g(z)$. The former exceeds the latter unless the control is optimal.

The dual problem can also be localised in this way. For u as above, let $\tilde{H}_p$ be a measurable function such that

$$\begin{aligned}\tilde{H}_p(s,t,z,X^{z,s}) = -\int_s^t u'\sigma(X_r^{z,s},r)\,dV_r + \frac{1}{2}\int_s^t |\sigma' u(X_r^{z,s},r)|^2\,dr \\ +(K_s - K_t).\end{aligned} \tag{6.7}$$

where K is differentiable and $K_T = 0$. This can be thought of as being the equivalent of $H_p(s,t,z,X^{z,s},y)$ for an unspecified time-distributed observation such that at time s the remaining information in the observation is K_s. (This corresponds to $\tilde{H}(X)$ of (5.1) with $K = K_0$.) Let Q_X^* be the distribution of $(X_r^{z,s}, s \le r \le t)$ when it is controlled by the optimal control. Taking expectation with respect to Q_X^* in (6.7), and taking the limit as $t \downarrow s$, we obtain a local rate of decrease of compatible information of $(u_* - u/2)'au(z,s,y)$. The local rate of increase of the information gain of $P_{X|Y}(\,\cdot\,,y)$ is, of course, $|\sigma' u_*(z,s,y)|^2/2$. The latter exceeds the former unless u is optimal.

In the global dual problem (5.1), the regular conditional probability $P_{X|Y}(\,\cdot\,,y)$ is the source of information. At time s the information in this source is

$$S_s = h(\tilde{\mu}|\mu) + \frac{1}{2}\int_0^s \int_{\mathbb{R}^n} |\sigma' u_*(z,t,y)|^2 P_{X|Y}(\chi_t^{-1}(dz),y)\,dt.$$

At time T there is no information in the observation and no residual information—all the information is still in the source. As s decreases, information flows out of the source at a rate $\dot{S}_s$; it is merged with residual information and flows into the observation at a rate $\dot{K}_s$. If u is optimal, then the flow is conservative, whereas more generally information is lost.

Let $\mathcal{H}_{z,s}$ be the Hilbert space of n-vectors of reals with inner product

$$\langle \alpha, \beta \rangle_{z,s} = \alpha' a(z,s)\beta.$$

The developments above show that the regular conditional probability $P_{X|Y}(\,\cdot\,,y)$ is locally characterised at the point (z,s) by the diffusion coefficients $a(z,s)$ and $(b(z,s) + a(z,s)\alpha_*)$, where α_* minimises

$$\frac{1}{2}\|\alpha\|_{z,s}^2 - \langle \alpha, u_*(z,s,y)\rangle_{z,s}; \tag{6.8}$$

whereas the optimal trajectory in the dual problem (5.3) is locally characterised in that its negative gradient at the point (z,s), β_*, maximises

$$\langle \beta, u_*(z,s,y)\rangle_{z,s} - \frac{1}{2}\|\beta\|_{z,s}^2. \tag{6.9}$$

The local balance of the Bayesian path estimator is thus characterised by the Legendre transform pair (6.8), (6.9). Of course, this is the characterisation of the optimal control problem of Section 4 provided by the stochastic maximum principle, the adjoint process being the gradient of the optimal dual state, $v(\,\cdot\,,\,\cdot\,,y)$, evaluated at $(\tilde{X}_t, t)$.

7. Conclusions. This article has developed dual variational characterisations of Bayesian estimation, in which the 'cost' functionals have particular information theoretic meaning. These characterisations provide a natural framework for the study of modelling and approximation errors in estimators such as nonlinear filters. They also link such issues with a broader theory of 'stochastic dissipativeness' (see [1]), on which the ideas and techniques of statistical physics can be brought to bear. We believe that this will have a number of advantages, for example in the study of the long-term behaviour of stochastic systems. The characterisations also provide a framework for the representation of estimators, in a broader context, as apparent information minimisers and compatible information maximisers. These issues will be explored elsewhere.

8. Acknowledgements. The authors would like to thank one of the referees and the Associate Editor for their comments and suggestions which have led to a substantial improvement in the paper.

REFERENCES

[1] V. S. Borkar and S. K. Mitter, *A note on stochastic dissipativeness*, in: Directions in Mathematical Systems Theory and Optimization, A. Rantzer and C.I. Byrnes (eds.), Springer-Verlag, 2002, pp. 41–49; in honor of Professor Anders Lindquist's 60th birthday.

[2] J. M. C. Clark, *The design of robust approximations to the stochastic differential equations of nonlinear filtering*, in: Communication Systems and Random Process Theory, J. K. Skwirzynski (ed.), NATO Advanced Study Institute Series, Sijthoff and Noordhoff, Alphen aan den Rijn, 1978, pp. 721–734.

[3] M. H. A. Davis, *A pathwise solution of the equations of nonlinear filtering*, Th. Pob. Appl., 27 (1983), pp. 167–175.

[4] M. H. A. Davis and P. P. Varaiya, *Dynamic programming conditions for partially observable stochastic systems*, SIAM J. Control Optim., 11 (1973), pp. 226–261.

[5] P. Dupuis and R. S. Ellis, *A Weak Convergence Approach to the Theory of Large Deviations*, Wiley, New York, 1997.

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

[7] A. Friedman, *Stochastic Differential Equations and Applications, Vol. 1*, Academic Press, 1975.

[8] H-O. Georgii, *Gibbs Measures and Phase Transitions*, de Gruyter, 1988.

[9] R. S. Liptser and A. N. Shiryayev, *Statistics of Random Processes 1—General Theory*, Springer-Verlag, 1977.

[10] T. Mikami, *Dynamical systems in the variational formulation of the Fokker-Planck equation by the Wasserstein metric*, Appl. Math. Optim., 42 (2000), pp. 203–227.

[11] S. K. Mitter and N. J. Newton, *The duality between estimation and control*, in Optimal Control and Partial Differential Equations, J. L. Menaldi, E. Rofman and A. Sulem (eds.), In honour of Professor Alain Bensoussan's 60th anniversary, IOS Press, Amsterdam, 2000.

[12] S. K. Mitter and N. J. Newton, *Information flow in nonlinear filters*, in preparation.

[13] N. J. Newton, *Observation sampling and quantisation for continuous-time estimators*, Stochastic Proc. Appl., 87 (2000), pp. 311–337.

[14] J. Picard, *Robustesse de la solution des problemes de filtrage avec bruit blanc independant*, Stochastics, 13 (1984), pp. 229–245.

[15] L. C. G. Rogers and D. Williams, *Diffusions, Markov Processes and Martingales: Part 2—Itô Calculus*, Wiley, New York, 1986.

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