Optimal State Estimation for Stochastic Systems: An Information Theoretic Approach

Xiangbo Feng, Kenneth A. Loparo, *Senior Member, IEEE*, and Yuguang Fang, *Member, IEEE*

***Abstract*— In this paper, we examine the problem of optimal state estimation or filtering in stochastic systems using an approach based on information theoretic measures. In this setting, the traditional minimum mean-square measure is compared with information theoretic measures, Kalman filtering theory is reexamined, and some new interpretations are offered. We show that for a linear Gaussian system, the Kalman filter is the optimal filter not only for the mean-square error measure, but for several information theoretic measures which are introduced in this work. For nonlinear systems, these same measures generally are in conflict with each other, and the feedback control policy has a dual role with regard to regulation and estimation. For linear stochastic systems with general noise processes, a lower bound on the achievable mutual information between the estimation error and the observation are derived. The properties of an optimal (probing) control law and the associated optimal filter, which achieve this lower bound, and their relationships are investigated. It is shown that for a linear stochastic system with an affine linear filter for the homogeneous system, under some reachability and observability conditions, zero mutual information between estimation error and observations can be achieved only when the system is Gaussian.**

***Index Terms*— Dual control, entropy, Kalman filtering, state estimation.**

I. Introduction

A SIGNIFICANT research effort has been devoted to the problem of state estimation for stochastic systems. Following the classical work of Gauss on least squares estimation and the modern day approach introduced by Kalman [1], [2] and investigated by other researchers [3]–[7], there have been intensive studies on least squares estimation. When applied to stochastic control systems, Kalman filtering theory also provides a tool for solving control problems, especially the stochastic optimal control problem for linear Gaussian (LG) systems where a separation principle holds.

On the other hand, information theory developed by Shannon [29] laid down a concrete mathematical framework for communication systems. Shannon's entropy has found considerable applications in many other fields. Recognizing the many similarities between state estimation and communication systems, many researchers have attempted to make the connection between estimation/control theory and information theory [10]–[20]. Recently, Saridis [13] proposed a general conceptual framework for stochastic and adaptive control problems using Jaynes maximum entropy principle [21]; this work provides a nice interpretation for many performance criteria and control algorithms.

It is well known that except for LG systems (and/or linear systems with a quadratic cost), the stochastic optimal control problem is in general a dual control problem [23], [24], and an optimal feedback control has two different, usually conflicting, attributes known as *probing* (or *learning*) and *regulating*. A general solution to such dual control problems has yet to be found. The probing aspect of a feedback control is critical in the dual control problem. We expect that a better understanding of how the control will effect the way the system performs the learning or probing for uncertainties will provide more insight into the final solution of the dual control problem. In this paper, we study the state estimation problem and the probing effects of a feedback control on state estimation. We use an information theoretic approach and illustrate the distinquishing features of an LG system that enable solutions to certain control problems for these systems to be obtained. We also investigate some interesting properties for the information theoretic measures introduced. In particular, for linear stochastic systems, we shall show that the entropy measures introduced in this work enjoy the same invariance properties as the mean-square error. A lower bound on the attainable mutual information between the estimation error and observation processes among all admissible controls and filters is established for linear stochastic systems. We conjecture that under some weak conditions this bound is zero if and only if the system is Gaussian. We partially justify this conjecture in Section IV by showing that it is true for systems with a linear affine filter for the homogeneous system. An example is given to illustrate some of the main results of the paper. The paper is organized as follows. In Section II after a brief survey of the existing results, we formulate the general estimation problem for nonlinear stochastic systems in an information theoretic framework. In Section III, we establish some important properties of the optimal filter for linear systems. In Section IV, for linear systems with an affine linear filter for the homogeneous system, we prove that zero error/observation mutual information can be achieved only for an LG system, provided that some reachability and observability conditions are satisfied. Section V contains an example for the linear non-Gaussian case, and Section VI contains our concluding remarks.

Manuscript received July 29, 1993. Recommended by Associate Editor, W. S. Wong. This work was supported in part by the Ford Motor Company, Dearborn, Michigan and the Center for Automation and Intelligent Systems Research at Case Western Reserve University, Cleveland, OH.

The authors are with the Department of Systems and Control Engineering, Case Western Reserve University, Cleveland, OH 44106-7070 USA.

Publisher Item Identifier S 0018-9286(97)04275-X.

II. Entropy and the General Estimation Problem with Active Probing

Consider the nonlinear stochastic system

$$\begin{cases} x_{k+1} = a_k(x_k, u_k, w_k), & k \geq 0 \\ y_k = b_k(x_k, v_k) \end{cases} \tag{N}$$

where in (N), $x_k \in \mathcal{R}^n$ is the system state, $u_k \in U \subset \mathcal{R}^m$ is the control variable, and $y_k \in \mathcal{R}^p$ is the output. $U \subset \mathcal{R}^m$ denotes the admissible control set. The sequences of independent random vectors $\{w_k \in \mathcal{R}^q : k = 0, 1, \cdots\}$ and $\{v_k \in \mathcal{R}^s : k = 0, 1, \cdots\}$ model the driving and measurement noise processes, respectively. We assume that the basic random variables x_0, w_i, and v_j, defined on a probability space (Ω, F, P), are independent with finite covariance matrices Σ_0, Q_i, and R_j for all $i, j \geq 0$. For a random vector ξ defined on (Ω, F, P), $\sigma\{\xi\} \subset F$ is the σ-algebra generated by ξ. $\mathcal{Y}_k = \{y_0, y_1, \cdots, y_k\}$ is the observation (information) vector at time k, and a feedback control is a sequence $g = \{g_0, g_1, \cdots,\}$ of measurable functions with $u_k = g_k(\mathcal{Y}_k) \in U$. Let $\mathcal{G}$ denote the collection of all feedback controls. A filter is a sequence $f = \{f_0, f_1, \cdots\}$ of measurable functions with $\hat{x}_k = f_k(\mathcal{Y}_k) \in \mathcal{R}^n$ the estimate of the state x_k based on $\mathcal{Y}_k$. Let $\mathcal{F}$ denote the collection of all filters.

Our objective is to design a feedback control $g \in \mathcal{G}$ and a filter $f \in \mathcal{F}$ in a certain optimal fashion and to study the relationship between g and f. Although the least squares family of estimators has been extensively studied, most notably, the minimum mean-square estimator for LG systems and extensions to affine linear minimum mean-square estimators for general linear systems, few researchers have asked the questions: Is it possible to go beyond the traditional approaches? If so, what is an appropriate framework for estimation, a criteria for optimality, and what is the relationship to control system design? This paper proposes a framework for state estimation, develops performance indexes for state estimation using information theoretic concepts, and investigates the intimate relationship between the design of an optimal filter and a feedback control for the system.

Kalman filtering theory [1]–[8] revolutionized the classical least squares method and also provided a methodology for solving certain control problems, in particular, the stochastic optimal control problem for LG systems where a separation principle holds. For general classes of stochastic systems, the difficulty in determining an optimal control and filter is immediately apparent, even for a linear non-Gaussian system with a quadratic cost functional. Even though a separation principle holds for this case and the optimal filter is the (true) minimum mean-square filter, it is generally not that the Kalman filter and recursive finite-dimensional realizations of this optimal filter are unknown. Further, complications result from the dual concept of a feedback control law introduced by Feldbaum [23]. Researchers have devoted a significant effort to studying the dual optimal control problem, and even though many meaningful results have been obtained, an optimal solution to the problem is unknown. Consider, for example, an LG system with unknown parameters in the system matrix taking values from a finite discrete set and a quadratic cost functional. It is shown [9] that a finite-dimensional nonlinear filter exists for the joint estimation of state and the unknown parameter; however, the dynamic programming approach does not lead to a closed-form analytical solution for the optimal control. The cost-to-go involves three terms: one for regulation, one for probing (estimation), and one for equivocation. Due to the conflicting nature of each of these cost terms, it is expected that an inferior solution to this "multiobjective" optimization problem should provide a reasonable solution to the problem. In spite of the developments of control design using quadratic cost functionals for LG systems, the computational difficulties of dual control problems suggest that the traditional quadratic cost functionals may not be the best criteria to use in the synthesis of feedback control policies for general classes of stochastic systems.

Information theory has revolutionized communication and coding theory and has had significant applications in other fields like statistics, physics, economics, and computer science. A marriage between information and control theory may provide better insight and understanding of many complicated control problems. Several researchers have developed some important results [10]–[14], [17]–[20]. In particular, Saridis [13] has given an entropy formulation of optimal and adaptive control problems and has interpreted the dual effect in terms of information theoretic measures. In [11] and [12], Weidemann and Stear studied estimation and feedback control systems and analyzed various information quantities associated with such a system. Kalata and Priemer [17] and Tomita *et al.* [18], [19] subsequently investigated prediction, filtering, and smoothing problems for signals generated by an LG system and showed that the Kalman filter minimizes the mutual information between the estimation error and observation process as well as the error entropy. A study of the optimal dual control problem for linear stochastic systems with parameter uncertainty using Saridis' formulation is presented in [15] and [16]. Before presenting a generalization of the setup presented in [10] for the problem of synthesizing a probing feedback control associated with a filtering problem, some elementary results from information theory are discussed. A detailed account of information theory and its contribution to communication theory and applications are presented in [26] and [27].

An essential notion of Shannon's information theory is entropy, including conditional entropy and mutual information. For random variables ξ and η given, the *(differential) entropy* $h(\xi)$ of ξ and the *joint (differential) entropy* $h(\xi, \eta)$ of (ξ, η) are defined as

$$\begin{aligned} h(\xi) &= \int p_\xi(x) \log(1/p_\xi(x))\, dx \\ &= -\int p_\xi(x) \log(p_\xi(x))\, dx \\ h(\xi, \eta) &= \int p_{\xi\eta}(x, y) \log(1/p_{\xi\eta}(x, y))\, dx\, dy \\ &= -\int p_{\xi\eta}(x, y) \log(p_{\xi\eta}(x, y))\, dx\, dy \end{aligned} \tag{1}$$

where $p_\xi(x)$ and $p_{\xi\eta}(x, y)$ are the (joint) density functions of ξ and (ξ, η), respectively. Likewise, the *conditional (differential)* *entropy* of ξ given η is defined as

$$h(\xi \mid \eta) = \iint p_{\xi|\eta}(x \mid y) \log(1/p_{\xi|\eta}(x \mid y))\, dx\, dy \qquad (2)$$

where $p_{\xi|\eta}(x \mid y)$ is the conditional density function. Another important concept is the *mutual information* $I(\xi;\eta)$ of (ξ,η), which is defined as

$$I(\xi;\eta) = \int p_{\xi\eta}(x,y) \log \frac{p_{\xi\eta}(x,y)}{p_\xi(x)p_\eta(y)}\, dx\, dy. \qquad (3)$$

Roughly speaking, the differential entropy $h(\xi)$ measures the dispersion of the random variable ξ. The smaller the differential entropy $h(\xi)$, the more concentrated the density function of ξ. Because $h(\xi)$ is not invariant with respect to a bijective transformation, it is usually not a measure of information. However, the nonnegative quantity $I(\xi;\eta)$, which is invariant with respect to a bijective transformation, is a quantitative measure of the mutual information of ξ and η. The following properties [which we will refer to as Property (P)] can be easily proved.

(P1) $h(\xi,\eta) = h(\xi) + h(\eta \mid \xi) = h(\eta) + h(\xi \mid \eta) = h(\eta,\xi)$.
(P2) $I(\xi;\eta) = I(\eta;\xi) = h(\xi) + h(\eta) - h(\xi,\eta) \geq 0$ with equality iff ξ,η independent.
(P3) $h(\xi + f(\eta)|\eta) = h(\xi|\eta)$, for any measurable function f. (P)
(P4) $h(\xi \mid \eta) = h(\xi \mid \zeta)$, if $\sigma\{\eta\} = \sigma\{\zeta\}$. $h(\xi \mid \eta) = h(\xi)$ iff ξ,η independent.
(P5) $h(\xi + c) = h(\xi)$, $I(\xi + c;\eta) = I(\xi;\eta)$, $h(\xi \mid \eta + c) = h(\xi \mid \eta)$, for any constant c.
(P6) $h(\xi + \eta) \geq \max\{h(\xi), h(\eta)\}$, if ξ,η independent.

References [26] and [27] provide a detailed account of these properties and their interpretation in information theory.

A general framework for state estimation (filtering) in terms of the information theoretic measures defined in (1)–(3) is provided next. For a feedback control $g \in \mathcal{G}$ given, the closed-loop quantities, like the state x_k, the output y_k, and the control u_k, are random variables which depend on g. We shall use the notation x_k^g, y_k^g, and u_k^g to signify this fact. Similarly, the superscript f indicates that the estimate $\hat{x}_k^{g,f}$ also depends on f. Let $\tilde{x}_k^{g,f} = x_k^g - \hat{x}_k^{g,f}$ denote the filtering error. A diagram for the combined estimation/control can be represented as in Fig. 1.

In Fig. 1, $b(\cdot,\cdot)$ denotes the output (or sensor) channel with output function b_k and measurement noise v_k, and $f(\cdot)$ denotes the filter. The objective is to design f and g so that a certain optimal estimate $\hat{x}_k^{g,f}$ is obtained. Note that the estimate $\hat{x}_k^{g,f}$ and the error $\tilde{x}_k^{g,f}$ depend on the selections of both the filter f and the feedback controller g. This is a simple indication of the dual property of the control g. To make this relationship more precise, we define the following types of optimal estimators and optimal (probing) control laws in terms of some useful information theoretic measures.

Definition 2.1: Consider the stochastic system (N) and the estimation/control configuration given in Fig. 1.

1) For $g \in \mathcal{G}$ given, we call $\bar{f}_g \in \mathcal{F}$ the *minimum mean-square estimator* (of x_k^g based on $\mathcal{Y}_k^g$), if $\bar{f}_g \in \mathcal{F}$ minimizes the mean-square error $\sigma^{g,f}(\tilde{x}_k) \stackrel{\text{def}}{=} E\{\|\tilde{x}_k^{g,f}\|^2\}$ for all $k \geq 0$, i.e.,

$$\sigma^{g,\bar{f}_g}(\tilde{x}_k) = E\{\|\tilde{x}_k^{g,\bar{f}_g}\|^2\} = \inf_{f\in\mathcal{F}} \sigma^{g,f}(\tilde{x}_k) = \inf_{f\in\mathcal{F}} E\{\|\tilde{x}_k^{g,f}\|^2\}.$$

2) For $g \in \mathcal{G}$ given, we call $\tilde{f}_g \in \mathcal{F}$ the *minimum error/observation information estimator* (of x_k^g given $\mathcal{Y}_k^g$), if $\tilde{f}_g \in \mathcal{F}$ minimizes the mutual information $I(\tilde{x}_k^{g,f};\mathcal{Y}_k^g)$ between the information vector $\mathcal{Y}_k^g$ and the estimation error $\tilde{x}_k^{g,f}$ for all $k \geq 0$, i.e.,

$$I(x_k^{g,\tilde{f}_g};\mathcal{Y}_k^g) = \inf_{f\in\mathcal{F}} I(x_k^{g,f};\mathcal{Y}_k^g).$$

3) For $g \in \mathcal{G}$ given, we call $\hat{f}_g \in \mathcal{F}$ the *minimum error entropy estimator* (of x_k^g given $\mathcal{Y}_k^g$), if $\hat{f}_g \in \mathcal{F}$ minimizes the error (differential) entropy $h(\tilde{x}_k^{g,f})$ of the estimation error $\tilde{x}_k^{g,f}$ for all $k \geq 0$, i.e.,

$$h(\tilde{x}_k^{g,\hat{f}_g}) = \inf_{f\in\mathcal{F}} h(\tilde{x}_k^{g,f}).$$

4) We call $g^* \in \mathcal{G}$ the *optimal probing control in the mean-square sense* and $f^* \in \mathcal{F}$ the minimum mean-square estimator, if for all $k \geq 0$

$$E\{\|\tilde{x}_k^{g^*,f^*}\|^2\} = \inf_{g\in\mathcal{G},f\in\mathcal{F}} E\{\|\tilde{x}_k^{g,f}\|^2\}.$$

Clearly, f^* is the minimum mean-square estimator associated with g^*; refer to 1) above.

5) We call $g^\# \in \mathcal{G}$ the *optimal probing control in the error/observation information sense* and $f^\# \in \mathcal{F}$ the minimum error/observation information estimator, if for all $k \geq 0$

$$I(\tilde{x}_k^{g^\#,f^\#};\mathcal{Y}_k^{g^\#}) = \inf_{g\in\mathcal{G},f\in\mathcal{F}} I(\tilde{x}_k^{g,f};\mathcal{Y}_k^g).$$

Clearly, $f^\#$ is the minimum error/observation information estimator associated with $g^\#$; refer to 2) above.

6) We call $g^\% \in \mathcal{G}$ the *optimal probing control in the error entropy sense* and $f^\% \in \mathcal{F}$ the minimum error entropy estimator, if for all $k \geq 0$

$$h(\tilde{x}_k^{g^\%,f^\%}) = \inf_{g\in\mathcal{G},f\in\mathcal{F}} h(\tilde{x}_k^{g,f}).$$

Clearly, $f^\%$ is the minimum error entropy estimator associated with $g^\%$; refer to 3) above.

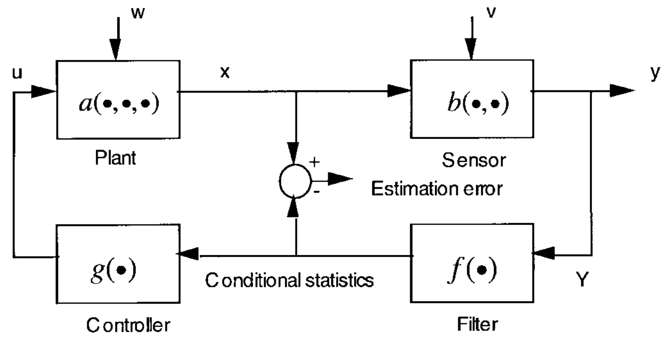

Fig. 1. Estimation/control system configuration.

7) We call $g^{\$} \in \mathcal{G}$ the *optimal probing control in the sensor channel transmittance sense*, if g maximizes the *sensor channel transmittance* $I(x_k^g; \mathcal{Y}_k^g)$ for all $k \geq 0$, i.e.,

$$I(x_k^{g^{\$}}; \mathcal{Y}_k^{g^{\$}}) = \sup_{g \in \mathcal{G}} I(x_k^g; \mathcal{Y}_k^g).$$

We shall call (g^*, f^*) the minimum mean-square (control and filter) pair, $(g^\#, f^\#)$ the minimum error/observation pair, and $(g^\%, f^\%)$ the minimum error entropy pair. In particular, when $I^{g^\#, f^\#}(\tilde{x}_k, \mathcal{Y}_k) = 0$, the minimum error/observation information pair $(g^\#, f^\#)$ is referred to as the zero error/observation information pair. We assume that all information theoretic quantities, like entropy and mutual information, are well defined and that the indicated infimum and supremum are achievable (the optimal performance measures can, however, take extended real values). Next, we comment on the probing effect of a feedback control g.

The probing effect of a control g can be defined as the influence that control actions have on estimation. This will be quantified by a certain performance measure defined for estimation. To be more precise, suppose that $P^{g,f}(\tilde{x}_k)$ is a performance measure for estimation, which we would like to minimize. We will say $g \in \mathcal{G}$ has no probing effect [with respect to $P^{g,f}(\tilde{x}_k)$], if for any $g \in \mathcal{G}$, there exists a filter $f_g \in \mathcal{F}$ such that P^{g,f_g} is the global minimum defined by $\inf_{g \in \mathcal{G}, f \in \mathcal{F}} P^{g,f}(\tilde{x}_k)$. This probing effect of g is the same probing or learning effect as in the dual control problem.

The concept of the minimum mean-square estimator $\bar{f}_g$ associated with a feedback control $g \in \mathcal{G}$ is well known. For (N), we have $\bar{f}_g(\mathcal{Y}_k^g) = E\{x_k^g \mid \mathcal{Y}_k^g\}$. In particular, for LG systems, i.e., for (N) with $a_k(x, u, w) = A_k x + B_k u + G_k w$, $b_k(x, v) = C_k x + H_k v$, and x_0, w_j, v_k Gaussian, $\bar{f}_g$ is realized by the Kalman filter. It is also well known that in this case the error covariance matrix $\Sigma_{k|k}$ is independent of the selection of g, and therefore g has no probing effect. As mentioned previously, the quantity $I(\tilde{x}_k^{g,f}; \mathcal{Y}_k^g) \geq 0$ measures the mutual information between the estimation error $\tilde{x}_k^{g,f}$ and the information vector $\mathcal{Y}_k^g$. $I(\tilde{x}_k^{g,f}; \mathcal{Y}_k^g) = 0$ if and only if $\tilde{x}_k^{g,f}$ is independent of $\mathcal{Y}_k^g$ and therefore can be regarded as a measure of the dependence between $\tilde{x}_k^{g,f}$ and $\mathcal{Y}_k^g$. Minimizing this measure can lead to better estimation. Again, for an LG system, it is well known that the minimum mean-square estimation error $\tilde{x}_k^{g,\bar{f}_g}$ is independent of $\mathcal{Y}_k^g$ for any $g \in \mathcal{G}$, and as a consequence $g \in \mathcal{G}$ has no probing effect in this sense as well. If the mutual information $I(\tilde{x}^{g,\bar{f}_g}, \mathcal{Y}_k^g) = 0$, then the minimum mean-square estimator $\bar{f}_g$ extracts all the relevant information about x_k^g from the observation $\mathcal{Y}_k^g$. The error (differential) entropy measure $h(\tilde{x}_k^{g,f})$ measures the dispersion of the estimation error $\tilde{x}_k^{g,f}$, and for an LG system it can be shown that $h(\tilde{x}_k^{g,\bar{f}_g}) = \alpha \log \det \Sigma_{k|k} + \beta$, where α and β are constants and $\Sigma_{k|k}$ is the error covariance matrix. Hence, the Kalman filter $\bar{f}_g$ also minimizes the error (differential) entropy measure $h(\tilde{x}_k^{g,f})$ for any $g \in \mathcal{G}$, and the control g has no probing effect in this sense as well. We conclude that for LG systems, the Kalman filter $\bar{f}_g$ is optimal in all senses 1)–3), and the optimal costs $E\{\|\tilde{x}_k^{g,\bar{f}_g}\|^2\}$, $I(\tilde{x}_k^{g,\bar{f}_g}; \mathcal{Y}_k^g) = 0$, and $h(\tilde{x}_k^{g,\bar{f}_g})$ are independent of the choice of the feedback law $g \in \mathcal{G}$. The pair $(g, \bar{f}_g)$ for any $g \in \mathcal{G}$ is also optimal in all senses 4)–6), and $g \in \mathcal{G}$ has no probing effect. This can be considered as an equivalent statement of a separation principle for the linear-quadratic-Gaussian (LQG) problem. In general, for nonlinear systems a feedback control law $g \in \mathcal{G}$ does have an influence on state estimation, and the solutions to any of the above optimal (probing) estimation problems defined by 1)–6) remain unknown. Before going into further analysis, we comment on the sensor channel transmittance $I(x_k^g; \mathcal{Y}_k^g)$ [11], which is a quantity that only depends on the sensor $b(\cdot, \cdot)$ and the feedback law $g \in \mathcal{G}$. For a fixed sensor $b(\cdot, \cdot)$, the data-processing inequality [32, p. 208] yields $\sup_{f \in \mathcal{F}} I(x_k^g; \hat{x}_k^{g,f}) \leq I(x_k^g; \mathcal{Y}_k^g)$, for any $g \in \mathcal{G}$. Therefore, $I(x_k^g; \mathcal{Y}_k^g)$ provides an upper bound on the amount of information that can be extracted from the observations by a filter f. Intuitively, the synthesis of a probing control $g \in \mathcal{G}$ should maximize $I(x_k^g; \mathcal{Y}_k^g)$ so that the maximum amount of information about x_k^g is obtained from $\mathcal{Y}_k^g$ through the design of the filter f. If we regard the sensor as a communication channel, $\sup_{g \in \mathcal{G}} I(x_k^g; \mathcal{Y}_k^g)$ can be interpreted as the channel capacity.

Remark: To simplify notation, we will use the superscripts "g" and "f" on variables and functionals of variables to denote the dependence on the choice of the feedback law $g \in \mathcal{G}$ and the filter $f \in \mathcal{F}$. For example, $E\{\|\tilde{x}_k^{g,f}\|^2\}$ and $E^{g,f}\{\|\tilde{x}_k\|^2\}$ both denote the mean-square error of the filter associated with the given $g \in \mathcal{G}$ and $f \in \mathcal{F}$.

The synthesis of a probing control g requires maximizing the sensor channel transmittance

$$\sup_{g \in \mathcal{G}} I^g(x_k; \mathcal{Y}_k) = \sup_{g \in \mathcal{G}} [h^g(x_k) - h^g(x_k \mid \mathcal{Y}_k)] \tag{4}$$

and minimizing the error/observation (mutual) information

$$\begin{aligned} &\inf_{g \in \mathcal{G}, f \in \mathcal{F}} I^{g,f}(\tilde{x}_k; \mathcal{Y}_k) \\ &\quad = \inf_{g \in \mathcal{G}, f \in \mathcal{F}} [h^{g,f}(\tilde{x}_k) - h^{g,f}(\tilde{x}_k | \mathcal{Y}_k)] \\ &\quad \overset{\text{(P3)}}{=} \inf_{g \in \mathcal{G}, f \in \mathcal{F}} [h^{g,f}(\tilde{x}_k) - h^g(x_k \mid \mathcal{Y}_k)]. \end{aligned} \tag{5}$$

Thus, from (4), select $g \in \mathcal{G}$ to minimize $h^g(x_k \mid \mathcal{Y}_k)$, while from (5), select $g \in \mathcal{G}$ to maximize $h^g(x_k \mid \mathcal{Y}_k)$. Therefore, (4) and (5) are in conflict, and one approach is to maximize the difference, i.e.,

$$\sup_{g \in \mathcal{G}, f \in \mathcal{F}} W^{g,f} \overset{\text{def}}{=} \max_{g \in \mathcal{G}, f \in \mathcal{F}} [I^g(x_k; \mathcal{Y}_k) - I^{g,f}(\tilde{x}_k; \mathcal{Y}_k)]. \tag{6}$$

From (4) and (5)

$$W^{g,f} = I^g(x_k; \mathcal{Y}_k) - I^{g,f}(\tilde{x}_k; \mathcal{Y}_k) = h^g(x_k) - h^{g,f}(\tilde{x}_k)$$

and (6) is equivalent to

$$\begin{aligned} \sup_{g \in \mathcal{G}, f \in \mathcal{F}} W^{g,f} &= \sup_{g \in \mathcal{G}, f \in \mathcal{F}} [I^g(x_k; \mathcal{Y}_k) - I^{g,f}(\tilde{x}_k; \mathcal{Y}_k)] \\ &= \sup_{g \in \mathcal{G}, f \in \mathcal{F}} [h^g(x_k) - h^{g,f}(\tilde{x}_k)]. \end{aligned} \tag{7}$$

The constant $W^{g,f}$ is referred to as the *entropy performance index* of the filter channel because it measures the reduction (or difference) between the differential entropy of the input to the channel x_k^g, $h^g(x_k)$, and the differential entropy of the output from the channel $\tilde{x}_k^{g,f}$, $h^{g,f}(\tilde{x}_k)$; see Fig. 1. Next, we show that $W^{g,f} \geq 0$ for all "nontrivial" filters f. Because the mutual information is nonnegative, $W^{g,f} \leq I^g(x_k; \mathcal{Y}_k)$ for all g and f if the filter $f = \{f_0, f_1 \cdots, f_k \cdots\}$ is a given deterministic sequence $f_k = \alpha_k$, which is referred to as a *constant filter*, then

$$h^g(x_k) = h^{g,f}(\tilde{x}_k + \alpha_k) = h^{g,f}(\tilde{x}_k). \tag{8}$$

It follows that $W^{g,f} = 0$. Thus, a reasonable choice of a filter should make the mutual information difference $W^{g,f} = I^g(x_k; \mathcal{Y}_k) - I^{g,f}(\tilde{x}_k; \mathcal{Y}_k)$ nonnegative. Indeed, in the LG case, the Kalman filter $\bar{f}$ gives $W^{g,\bar{f}} = I^g(x_k; \mathcal{Y}_k) > 0$. A filter $f \in \mathcal{F}$ is called a *nontrivial filter* associated with g, if $W^{g,f} \geq 0$. Because $\hat{x}_k^{g,f} = f_k(\mathcal{Y}_k^g)$ is a measurable function of $\mathcal{Y}_k^g$, for a nontrivial filter

$$h^g(x_k) \geq h^{g,f}(\tilde{x}_k) \geq h^{g,f}(\tilde{x}_k \mid \mathcal{Y}_k) = h^g(x_k \mid \mathcal{Y}_k). \tag{9}$$

This inequality is important because it relates the three differential entropies in (9) that are the quantities to be optimized in (4), (5), and (7).

An optimal probing control $g \in \mathcal{G}$ simultaneously optimizes (4) and (5) with appropriately defined filters. If a solution exists, we say that the optimal probing control is nonconflicting.

For an LG system

$$\begin{cases} x_{k+1} = A_k x_k + B_k u_k + G_k w_k \\ y_k = C_k x_k + H_k v_k \end{cases} \tag{LG}$$

with

$$x_0 \sim N(\bar{x}_0, \Sigma_0), \qquad w_i \sim N(0, Q_i), \qquad v_j \sim N(0, R_j)$$

independent basic random variables, because there is no dual effect, there is no conflict for any $g \in \mathcal{G}$, and the Kalman filter $\bar{f}_g$ simultaneously optimizes all criteria. However, for a general nonlinear stochastic system, this problem is extremely difficult, and no general answers have been obtained to date. Some preliminary results for this problem are presented in this paper. In particular, we shall show in the next section that for a linear (non-Gaussian) system, there exists a single probing control which is optimal for all the measures defined. This is not true for general nonlinear systems.

Before concluding this section, we derive some general results for the nonlinear system (N) using an assumption which is similar to the one adopted by several researchers [11], [12], [17]–[19]. This assumption makes the probing effect of g partially nonconflicting. In Section IV, we shall show that this restrictive assumption, under a reachability and observability condition, is equivalent to the Gaussian assumption on the basic random variables for a linear stochastic system, given a linear affine filter for the homogeneous system.

Theorem 2.2: Consider the nonlinear stochastic system (N) with the filtering configuration shown in Fig. 1.

a) Let $g \in \mathcal{G}$ be fixed. Then, the minimum mean-square estimator $\bar{f}_g = \{\bar{f}_0, \bar{f}_1, \cdots\}$ is given by $\hat{x}_k^{g,\bar{f}_g} = E^g\{x_k \mid \mathcal{Y}_k\}$. Furthermore, the minimum error/observation information filter $\tilde{f}_g$ is equivalent to the minimum error entropy filter $\hat{f}_g$.

b) *Given Assumption A)*: For any $g \in \mathcal{G}$ given, there is a filter $f_g \in \mathcal{F}$ such that $I^{g,f_g}(\tilde{x}_k; \mathcal{Y}_k) = 0$. That is, for any $g \in \mathcal{G}$, the zero error/observation information filter $f_g \in \mathcal{F}$ exists and the estimation error $\tilde{x}_k^{g,f_g}$ is independent of $\mathcal{Y}_k^g$.

Then, the minimum mean-square filter $\bar{f}_g$ associated with $g \in \mathcal{G}$ is such a filter f_g, i.e., $(g, \bar{f}_g)$, forms a zero error/observation information pair. Let $g^{\%} \in \mathcal{G}$ satisfy

$$h^{g^{\%}}(x_k \mid \mathcal{Y}_k) = \inf_{g \in \mathcal{G}} h^g(x_k \mid \mathcal{Y}_k) \tag{10}$$

then $(g^{\%}, \bar{f}_{g^{\%}})$ is the minimum error entropy pair. Let $g^{\$} \in \mathcal{G}$ be the optimal probing control in the sensor channel transmittance sense, then $(g^{\$}, \bar{f}_{g^{\$}})$ maximizes the entropy performance index $W^{g,f}$ with

$$W^{g^{\$}, \bar{f}_{g^{\$}}} = \sup_{g \in \mathcal{G}, f \in \mathcal{F}} W^{g,f} = \sup_{g \in \mathcal{G}} I^g(x_k; \mathcal{Y}_k). \tag{11}$$

Proof: We show a) first. Let $g \in \mathcal{G}$ be a fixed feedback control. The fact that $\hat{x}_k^{g,\bar{f}_g} = \bar{f}_k(\mathcal{Y}_k^g) = E^g\{x_k \mid \mathcal{Y}_k\}$ is the minimum mean-square estimator is well known. With g fixed, for any $f \in \mathcal{F}$, we have

$$\begin{aligned} I^{g,f}(\tilde{x}_k; \mathcal{Y}_k) &= h^{g,f}(\tilde{x}_k) - h^{g,f}(\tilde{x}_k \mid \mathcal{Y}_k) \\ &\overset{\text{(P3)}}{=} h^{g,f}(\tilde{x}_k) - h^g(x_k \mid \mathcal{Y}_k). \end{aligned}$$

The next statement of a) follows directly. For b), given that Assumption A) holds, we first show that the minimum mean-square filter $\bar{f}_g$ associated with a $g \in \mathcal{G}$ is such a filter f_g. Consider that

$$\begin{aligned} \hat{x}_k^{g,\bar{f}_g} &= E^g\{x_k \mid \mathcal{Y}_k\} = E\{\hat{x}_k^{g,f_g} + \tilde{x}_k^{g,f_g} \mid \mathcal{Y}_k^g\} \\ &= \hat{x}_k^{g,f_g} + E\{\tilde{x}_k^{g,f_g} \mid \mathcal{Y}_k^g\} = \hat{x}_k^{g,f_g} + E\{\tilde{x}_k^{g,f_g}\} \quad \text{(a.s.)}. \end{aligned} \tag{12}$$

It follows that $\tilde{x}_k^{g,\bar{f}_g} = \tilde{x}_k^{g,f_g} + c_k^{g,f_g}$ with c_k^{g,f_g} a constant. Thus, $\tilde{x}_k^{g,\bar{f}_g}$ is also independent of $\mathcal{Y}_k^g$ and $\bar{f}_g$ is a zero error/observation information filter associated with $g \in \mathcal{G}$. Now, with $I^{g,\bar{f}_g}(\tilde{x}_k; \mathcal{Y}_k) = h^{g,\bar{f}_g}(\tilde{x}_k) - h^g(x_k \mid \mathcal{Y}_k) = 0$ for any $g \in \mathcal{G}$, it is easy to see that $(g^{\%}, \bar{f}_{g^{\%}})$ is also the minimum error entropy pair. Similarly, with $W^{g,f} = I^g(x_k; \mathcal{Y}_k) - I^{g,f}(\tilde{x}_k; \mathcal{Y}_k)$ for any $g \in \mathcal{G}$ and $f \in \mathcal{F}$, it is easy to see that under Assumption A), $(g^{\$}, \bar{f}_{g^{\$}})$ maximizes the entropy performance index $W^{g,f}$ and (11) holds.

In Theorem 2.2, the probing control laws $g^{\%}$ and $g^{\$}$ are defined by

$$h^{g^{\%}}(x_k \mid \mathcal{Y}_k) = \inf_{g \in \mathcal{G}} h^g(x_k \mid \mathcal{Y}_k) \tag{13}$$

and

$$I^{g^{\$}}(x_k; \mathcal{Y}_k) = \sup_{g \in \mathcal{G}} I^g(x_k; \mathcal{Y}_k) = \sup_{g \in \mathcal{G}} [h^g(x_k) - h^g(x_k \mid \mathcal{Y}_k)] \tag{14}$$

respectively. Given Assumption A), these two optimal probing control laws are not necessarily conflicting. We shall show later that for a linear stochastic system, even without Assumption A), a single feedback control $g^* \in \mathcal{G}$ can be selected which is optimal in all senses defined earlier.

Although Assumption A) is restrictive, it is also intuitive from an information theory perspective since for any $g \in \mathcal{G}$ given, the filter f_g extracts all the necessary information about x_k^g from the observations $\mathcal{Y}_k^g$. In this case, the control $g^\$$ sends the largest amount of information about x_k to the observation $\mathcal{Y}_k$ for a given sensor configuration, and either $(g^\$, \bar{f}_{g^\$})$ or $(g^\$, f_{g^\$})$ is the optimal control and filter pair in $\mathcal{G} \times \mathcal{F}$ in terms of mutual information. For an LG system, Assumption A) is satisfied with $f_g = \bar{f}_g$, the Kalman filter. We show in the next section that for a general linear stochastic system, $h^g(x_k \mid \mathcal{Y}_k)$ is actually independent of $g \in \mathcal{G}$. Thus, $g^\$$ is the control which maximizes $h^g(x_k)$, i.e., the feedback control which maximizes the dispersion of the state. This is a very naive solution from a signal detection perspective. One may ask: Can Assumption A) hold for a linear system without the Gaussian assumption? We conjecture that in general the answer is no, although we do not have a definitive proof.

III. Filtering in Linear Stochastic Systems

A general framework for filtering and state estimation was proposed, and an appealing but restrictive assumption, satisfied by an LG system which guarantees that the probing control is nonconflicting for a general nonlinear system, was presented. In the next section it is shown that for a linear stochastic system with an affine linear filter for the homogeneous system, Assumption A) is equivalent to the Gaussian assumption, under some reachability and observability conditions. Therefore, we expect that in general there is no zero error/observation pair $(g^\$, f^\$)$ for linear non-Gaussian systems, i.e., there is no feedback control $g \in \mathcal{G}$ and filter $f \in \mathcal{F}$ such that $I^{g,f}(\tilde{x}_k; \mathcal{Y}_k) = 0$. The question then becomes: What is the lower bound $I^{g^\$, f^\$}(\tilde{x}_k; \mathcal{Y}_k) > 0$ for the error/observation information measure $I^{g,f}(\tilde{x}_k; \mathcal{Y}_k)$, and how do we characterize the pair $(g^\$, f^\$)$? In this section, we study linear stochastic systems defined by

$$\begin{cases} x_{k+1}^g = A_k x_k^g + B_k u_k^g + G_k w_k \\ y_k^g = C_k x_k^g + H_k v_k \end{cases} \tag{L}$$

where in (L), the independent basic random variables $x_0^g = x_0, w_i$, and v_j are assumed to have the first- and second-order statistics

$$\begin{aligned} E\{x_0\} &= \bar{x}_0, & 0 \le \mathrm{Cov}(x_0) &= \Sigma_0 < +\infty \\ E\{w_i\} &= 0, & 0 \le \mathrm{Cov}(w_i) &= Q_i < +\infty \\ E\{v_j\} &= 0, & 0 < \mathrm{Cov}(v_j) &= R_j < +\infty \end{aligned} \tag{15}$$

for all $i, j \ge 0$. In (L), the superscript g is used to signify the fact that all closed-loop variables depend on the selection of a feedback control law $u_k^g = g_k(\mathcal{Y}_k^g)$. Here, $\mathcal{Y}_k^g = \{y_0^g, y_1^g, \cdots\}$ is the measurement (information) vector. When x_0, w_i, and v_j are Gaussian, (L) is an LG system. We begin by examining the Kalman filter $\bar{f}_g$ for an LG system in the present framework.

Proposition 3.1: Consider an LG system. Let $\bar{f}_g = \{\bar{f}_0, \bar{f}_1 \cdots, \bar{f}_k \cdots\}$ be the Kalman filter, i.e., $\hat{x}_k^{g,\bar{f}_g} = \bar{f}_k(\mathcal{Y}_k^g) = E\{x_k^g \mid \mathcal{Y}_k^g\}$. Then we have the following.

i) $\bar{f}_g$ is the optimal filter in the senses 1)–3) given in Definition 2.1. Furthermore, the optimal costs associated with $\bar{f}_g$ are $E\{\|\tilde{x}_k^{g,\bar{f}_g}\|^2\} = \mathrm{tr}(\Sigma_{k|k})$, $I(\tilde{x}_k^{g,\bar{f}_g}; \mathcal{Y}_k^g) = 0$, and $h(\tilde{x}_k^{g,\bar{f}}) = \alpha \log\det(\Sigma_{k|k}) + \beta$, which are all independent of the selection of $g \in \mathcal{G}$. Here $\alpha > 0$ and β are constants and $\Sigma_{k|k}$ is the error covariance matrix of $\tilde{x}_k^{g,\bar{f}_g}$ (which is independent of $g \in \mathcal{G}$). Therefore, for any feedback control law $g \in \mathcal{G}$, the pair $(g, \bar{f}_g)$ is the optimal control and filter pair in the senses defined in 4)–6) in Definition 2.1.

ii) For any $g \in \mathcal{G}$

$$\begin{aligned} I(x_k^g; \mathcal{Y}_k^g) &= W^{g,\bar{f}_g}(\tilde{x}_k) = h(x_k^g) - h(\tilde{x}_k^{g,\bar{f}_g}) \\ &= h(x_k^g) - (\alpha \log\det(\Sigma_{k|k}) + \beta). \end{aligned}$$

Also, there exists a $g \in \mathcal{G}$ such that $h(x_k^g) = \gamma + \eta \log\det(\bar{\Sigma}_k)$ with $\eta > 0$ and γ constants and $\bar{\Sigma}_k$ the solution of the matrix equation

$$\bar{\Sigma}_{k+1} = A_k' \bar{\Sigma}_k A_k + G_k' Q_k G_k, \quad \bar{\Sigma}_0 = \Sigma_0. \tag{16}$$

Therefore, if Σ_0, A_k, G_k, and Q_k are such that $\det(\bar{\Sigma}_k)$ goes to $+\infty$ as k tends to $+\infty$, then, the sensor channel transmittance (or capacity) tends to $+\infty$ as k goes to $+\infty$.

Proof: For any feedback control law $g \in \mathcal{G}$, it is well known [22, pp. 93–104] that for the LG system, the Kalman filter $\bar{f}_g = \{\bar{f}_0, \bar{f}_1, \cdots\}$ is the minimum mean-square estimator, and the filter is given by the conditional mean $\bar{f}_k(\mathcal{Y}_k^g) = E\{x^g \mid \mathcal{Y}_k^g\}$. Also, the error covariance matrix $\Sigma_{k|k} = \mathrm{Cov}(\tilde{x}_k^{g,\bar{f}_g})$ can be computed off-line and is independent of the selection of g. This shows that $\bar{f}_g$ is optimal in the sense 1) of Definition 2.1. It is also known that the estimation error $\tilde{x}_k^{g,\bar{f}_g}$ is independent of $\mathcal{Y}_k^g$ and is distributed according to $N(0, \Sigma_{k|k})$. Then, it follows that $I(\tilde{x}_k^{g,\bar{f}_g}; \mathcal{Y}_k^g) = 0$. Because $I(\tilde{x}_k^{g,f}; \mathcal{Y}_k^g) \ge 0$ for any $f \in \mathcal{F}$, this shows that $\bar{f}_g$ is optimal in the sense of 2). For $g \in \mathcal{G}$ fixed, by 1) of Theorem 2.2, we know that $\bar{f}_g$ is also the minimum error entropy filter associated with g. Since $\tilde{x}_k^{g,\bar{f}_g} \sim N(0, \Sigma_{k|k})$, we have by direct computation

$$h(\tilde{x}_k^{g,\bar{f}_g}) = \alpha \log\det(\Sigma_{k|k}) + \beta \tag{17}$$

for some $\alpha > 0$ and β constants. The fact that for any $g \in \mathcal{G}$, the pair $(g, \bar{f}_g)$ is optimal for any measures of 4)–6) is a direct consequence of the fact that $\Sigma_{k|k}$ is independent of the selection of $g \in \mathcal{G}$. This proves i). For ii), by the independence of $\tilde{x}_k^{g,\bar{f}_g}$, and $\mathcal{Y}_k^g$, we have from (P3) and (P4) that for any $g \in \mathcal{G}$

$$\begin{aligned} I(x_k^g; \mathcal{Y}_k^g) &= h(x_k^g) - h(x_k^g \mid \mathcal{Y}_k^g) = h(x_k^g) - h(\tilde{x}_k^{g,\bar{f}_g} \mid \mathcal{Y}_k^g) \\ &= h(x_k^g) - h(\tilde{x}_k^{g,\bar{f}_g}) = W^{g,\bar{f}_g}(\tilde{x}_k) \\ &= h(x_k^g) - (\alpha \log\det(\Sigma_{k|k}) + \beta). \end{aligned} \tag{18}$$

Now, take g as an open-loop control, i.e., $g = \{u_0, u_1, \cdots\}$ a deterministic sequence. Then, the state process x_k^g is Gaussian with distribution $N(\bar{x}_k, \bar{\Sigma}_k)$. Here $\bar{\Sigma}_k$ satisfies (16). Therefore, we obtain

$$h(x_k^g) = \gamma + \eta \log\det(\bar{\Sigma}_k) \tag{19}$$

with $\gamma > 0$ and η constants. Thus, if $\det(\bar{\Sigma}_k)$ goes to $+\infty$, (18) and (19) imply that $I(x_k^g; \mathcal{Y}_k^g)$ will go to $+\infty$ as well. This completes the proof. □

Proposition 3.1 states that for an LG system, the control law $g \in \mathcal{G}$ has no probing effect in any of the above senses, and the zero error/observation information pair is defined via the Kalman filter. This is another explanation for the separation principle which holds for LG systems. For general linear stochastic systems, it is known that the (true) minimum mean-square filter $\bar{f}_g$ associated with $g \in \mathcal{G}$ (which may not be the Kalman filter for non-Gaussian systems) also makes the error covariance matrix independent of the selection of g. However, $\bar{f}_g$ may not be optimal in other senses. Another observation is that infinite sensor channel transmittance corresponds to infinite channel capacity of an ideal Gaussian channel without power limitation. From the state detection perspective, this result says that optimal state detection is realized when the dispersion of the state distribution is maximized, i.e., $h(x_k^g)$ tends to infinity.

In general, the feedback control law $g \in \mathcal{G}$ does have a probing effect on state estimation, and the answer to any of the above optimal (probing) estimation problems defined by 1)–6) remain as open questions, even for the simplest nonlinear/non-Gaussian systems. We address some of these issues in the remainder of the paper.

We begin by studying a general linear system (L). The following decomposition of (L) [22] is essential to our development. For a $g \in \mathcal{G}$, we can decompose x_k^g and y_k^g as

$$x_k^g = \bar{x}_k^g + x_k, \qquad y_k^g = \bar{y}_k^g + y_k. \tag{20}$$

In (20), the processes $\{\bar{x}_k^g\}$ and $\{\bar{y}_k^g\}$ are defined by the equations

$$\begin{cases} \bar{x}_{k+1}^g = A_k \bar{x}_k^g + B_k u_k^g \\ \bar{y}_k^g = C_k \bar{x}_k^g \end{cases} \tag{D}$$

with $u_k^g = g_k(\mathcal{Y}_k^g)$ and the initial state $\bar{x}_0^g = \bar{x}_0$. The processes $\{x_k\}$ and $\{y_k\}$ are defined by the equations

$$\begin{cases} x_{k+1} = A_k x_k + G_k w_k \\ y_k = C_k x_k + H_k v_k \end{cases} \tag{H}$$

where $x_0 = x_0^g - \bar{x}_0$, w_k, and v_j are given as before. Let $\mathcal{U}_k^g = \{u_0^g, u_1^g, \cdots, u_{k-1}^g\}$ be the past control inputs up to time k, and let $\mathcal{Z}_k^g = \{\mathcal{Y}_k^g, \mathcal{U}_k^g\}$ be the information vector. Let $\mathcal{Y}_k = \{y_0, y_1, \cdots, y_k\}$ be the information vector associated with the homogeneous system (H). Consider the following lemma, whose proof is given in [22].

Lemma 3.2: For the linear system (L), let $g \in \mathcal{G}$ be arbitrarily given. We have

$$\sigma\{\mathcal{Y}_k^g\} = \sigma\{\mathcal{Z}_k^g\} = \sigma\{\mathcal{Y}_k\} \tag{21}$$

where in (21), $\sigma\{\xi\}$ denotes the σ-algebra generated by the random vector ξ. Equivalently, $\mathcal{Z}_k^g, \mathcal{Y}_k^g$, and $\mathcal{Y}_k$ are (measurable) functions of each other.

Proposition 3.3: For the linear system (L), we have the following.

1) For any $g \in \mathcal{G}$ and $f \in \mathcal{F}$

$$h(x_k^g \mid \mathcal{Y}_k^g) = h(\tilde{x}_k^{g,f} \mid \mathcal{Y}_k^g) = h(x_k \mid \mathcal{Y}_k). \tag{22}$$

Note that the conditional differential entropy $h(x_k \mid \mathcal{Y}_k)$ is associated with the "homogeneous" system (H) and therefore is independent of the selection of g and f.

2) For any $g, \hat{g} \in \mathcal{G}$ and $f \in \mathcal{F}$ given, there is a $\hat{f} \in \mathcal{F}$ such that $h(\tilde{x}_k^{g,f}) = h(\tilde{x}_k^{\hat{g},\hat{f}})$.

Proof: For 1), let $g \in \mathcal{G}$ and $f \in \mathcal{F}$ be given. Since $\tilde{x}_k^{g,f} = x_k^g - f_k(\mathcal{Y}_k^g)$, the equality $h(x_k^g \mid \mathcal{Y}_k^g) = h(\tilde{x}_k^{g,f} \mid \mathcal{Y}_k^g)$ follows from (P3). Since from (20), we have $h(x_k^g \mid \mathcal{Y}_k^g) = h(\bar{x}_k^g + x_k \mid \mathcal{Y}_k^g)$ and $\bar{x}_k^g$ is a measurable function of $\mathcal{Y}_k^g$ from (D), we have $h(x_k^g \mid \mathcal{Y}_k^g) = h(x_k \mid \mathcal{Y}_k^g)$. However, by Lemma 3.2, $\sigma\{\mathcal{Y}_k^g\} = \sigma\{\mathcal{Y}_k\}$, and using (P4), we have $h(x_k \mid \mathcal{Y}_k^g) = h(x_k \mid \mathcal{Y}_k)$. This shows (22). For 2), let $g, \hat{g} \in \mathcal{G}$ and $f \in \mathcal{F}$ be given. By Lemma 3.2, there are measurable functions $T_k, T_k', \hat{T}_k$, and $\hat{T}_k'$ such that

$$\begin{aligned} \mathcal{Y}_k^g &= T_k(\mathcal{Y}_k), \qquad \mathcal{Y}_k = T_k'(\mathcal{Y}_k^g) \\ \mathcal{Y}_k^{\hat{g}} &= \hat{T}_k(\mathcal{Y}_k), \qquad \mathcal{Y}_k = \hat{T}_k'(\mathcal{Y}_k^{\hat{g}}.) \end{aligned} \tag{23}$$

Then, $\mathcal{Y}_k^g = T_k(\hat{T}_k'(\mathcal{Y}_k^{\hat{g}})) \stackrel{\text{def}}{=} S_k(\mathcal{Y}_k^{\hat{g}})$ and we have from (20) that

$$\begin{aligned} h(\tilde{x}_k^{g,f}) &= h\big(\bar{x}_k^g(\mathcal{Y}_k^g) + x_k - f_k(\mathcal{Y}_k^g)\big) \\ &= h\big(\bar{x}_k^g(S_k(\mathcal{Y}_k^{\hat{g}})) + x_k - f_k(S_k(\mathcal{Y}_k^{\hat{g}}))\big) \\ &= h\big(\bar{x}_k^{\hat{g}}(\mathcal{Y}_k^{\hat{g}}) + x_k - \big[f_k(S_k(\mathcal{Y}_k^{\hat{g}})) \\ &\quad - \bar{x}_k^g(S_k(\mathcal{Y}_k^{\hat{g}})) + \bar{x}_k^{\hat{g}}(\mathcal{Y}_k^{\hat{g}})\big]\big) \\ &= h\big(\bar{x}_k^{\hat{g}}(\mathcal{Y}_k^{\hat{g}}) + x_k - \hat{f}_k(\mathcal{Y}_k^{\hat{g}})\big) \\ &= h\big(\tilde{x}_k^{\hat{g},\hat{f}}\big). \end{aligned} \tag{24}$$

where in (23), the filter $\hat{f} \in \mathcal{F}$ is defined by

$$\hat{f}_k(\mathcal{Y}_k^{\hat{g}}) \stackrel{\text{def}}{=} \big[f_k(S_k(\mathcal{Y}_k^{\hat{g}})) - \bar{x}_k^g(S_k(\mathcal{Y}_k^{\hat{g}})) + \bar{x}_k^{\hat{g}}(\mathcal{Y}_k^{\hat{g}})\big]. \tag{25}$$

□

Proposition 3.3 is a very interesting result because 1) states that for a linear system, the conditional entropy $h(x_k^g \mid \mathcal{Y}_k^g)$, which is the equivocation of the sensor channel, is independent of the selection of $g \in \mathcal{G}$. From the previous discussion, given Assumption A) for the nonlinear system (N), the optimal probing controls $g^{\%}$ and $g^{\$}$ defined in Theorem 2.2 may not be conflicting. From Proposition 3.3, for the linear system (L) without Assumption A), a single feedback control $g^* \in \mathcal{G}$, which is optimal for all the performance measures defined for the probing estimation problem, can be determined. It is in this sense that the probing estimation problem for a linear systems is nonconflicting.

Theorem 3.4: For the linear stochastic system (L), there is a control and filter pair $(g^*, f^*) \in \mathcal{G} \times \mathcal{F}$ such that

$$\begin{aligned} h^{g^*}(x_k) &= \sup_{g \in \mathcal{G}} h^g(x_k) \\ h^{g^*,f^*}(\tilde{x}_k) &= \inf_{f \in \mathcal{F}} h^{g^*,f}(\tilde{x}_k) = \inf_{g \in \mathcal{G}, f \in \mathcal{F}} h^{g,f}(\tilde{x}_k) \\ I^{g^*}(x_k; \mathcal{Y}_k) &= \sup_{g \in \mathcal{G}} I^g(x_k; \mathcal{Y}_k) \\ I^{g^*,f^*}(\tilde{x}_k; \mathcal{Y}_k) &= \inf_{g \in \mathcal{G}, f \in \mathcal{F}} I^{g,f}(\tilde{x}_k; \mathcal{Y}_k) \\ W^{g^*,f^*} &= \sup_{g \in \mathcal{G}, f \in \mathcal{F}} W^{g,f}. \end{aligned} \tag{26}$$

Proof: Recall the assumption that the $\inf$ and $\sup$ operations given above are achievable. Let $g^* \in \mathcal{G}$ be such that the first equation in (26) holds. Take $f^* \in \mathcal{F}$ such that

$$h^{g^*,f^*}(\tilde{x}_k) = \inf_{f \in \mathcal{F}} h^{g^*,f}(\tilde{x}_k). \tag{27}$$

From Proposition 3.3 for any $g \in \mathcal{G}$ and $f \in \mathcal{F}$, a $f' \in \mathcal{F}$ (which depends on g, f, and g^*) exists such that $h^{g,f}(\tilde{x}_k) = h^{g^*,f'}(\tilde{x}_k)$. It follows that

$$\begin{aligned} \inf_{g \in \mathcal{G}, f \in \mathcal{F}} h^{g,f}(\tilde{x}_k) &= \inf_{g \in \mathcal{G}, f \in \mathcal{F}} h^{g^*,f'}(\tilde{x}_k) \geq \inf_{f'' \in \mathcal{F}} h^{g^*,f''}(\tilde{x}_k) \\ &\stackrel{(27)}{=} h^{g^*,f^*}(\tilde{x}_k). \end{aligned} \tag{28}$$

We conclude that

$$h^{g^*,f^*}(\tilde{x}_k) = \inf_{g \in \mathcal{G}, f \in \mathcal{F}} h^{g,f}(\tilde{x}_k). \tag{29}$$

We show the third equation in (26) next. From Proposition 3.3, for any $g \in \mathcal{G}$, we have

$$I^g(x_k; \mathcal{Y}_k) = h^g(x_k) - h(x_k \mid \mathcal{Y}_k).$$

Note that $h(x_k \mid \mathcal{Y}_k)$ is associated with the homogeneous system (H). Thus

$$\begin{aligned} I^{g^*}(x_k; \mathcal{Y}_k) &= h^{g^*}(x_k) - h(x_k \mid \mathcal{Y}_k) \\ &= \sup_{g \in \mathcal{G}} [h^g(x_k) - h(x_k \mid \mathcal{Y}_k)] = \sup_{g \in \mathcal{G}} I^g(x_k; \mathcal{Y}_k). \end{aligned}$$

Similarly, we have from the previous proposition that

$$\begin{aligned} \inf_{g \in \mathcal{G}, f \in \mathcal{F}} I^{g,f}(\tilde{x}_k; \mathcal{Y}_k) &= \inf_{g \in \mathcal{G}, f \in \mathcal{F}} [h^{g,f}(\tilde{x}_k) - h^{g,f}(\tilde{x}_k \mid \mathcal{Y}_k)] \\ &\stackrel{(P3)}{=} \inf_{g \in \mathcal{G}, f \in \mathcal{F}} [h^{g,f}(\tilde{x}_k) - h^g(x_k \mid \mathcal{Y}_k)] \\ &= \inf_{g \in \mathcal{G}, f \in \mathcal{F}} [h^{g,f}(\tilde{x}_k) - h(x_k \mid \mathcal{Y}_k)] \\ &\stackrel{(29)}{=} h^{g^*,f^*}(\tilde{x}_k) - h(x_k \mid \mathcal{Y}_k) \\ &= I^{g^*,f^*}(\tilde{x}_k; \mathcal{Y}_k). \end{aligned} \tag{30}$$

This establishes the fourth equation in (26). The last equation follows directly from the fact that $W^{g,f} = I^g(x_k; \mathcal{Y}_k) - I^{g,f}(\tilde{x}_k; \mathcal{Y}_k)$ from the third and the fourth equations in (26). □

Theorem 3.4 states that for the linear stochastic system (L), the pair (g^*, f^*) is simultaneously the optimal control and filter pair in all the senses defined in 2), 3), and 5)–7) given in Definition 2.1. Furthermore, g^* will maximize the entropy $h^g(x_k)$ and therefore will tend to increase the dispersion of the state x_k^g. This destabilizes the system so that the sensor channel transmittance $I^g(x_k; \mathcal{Y}_k)$ is maximized. f^* is the optimal filter which minimizes the error entropy $h^{g^*,f}(\tilde{x}_k)$, and therefore $I^{g,f}(\tilde{x}_k; \mathcal{Y}_k)$ is minimized. We know from Theorem 2.2 that under Assumption A), f^* can be selected as the minimum mean-square estimator, similar to the LG case. However, we will see later that Assumption A) does not hold for linear non-Gaussian systems in general. Thus, f^* may not in general be the minimum mean-square filter.

Now, suppose we ignore the probing effect of $g \in \mathcal{G}$ on the sensor channel transmittance $I^g(x_k; \mathcal{Y}_k)$ and only concern ourselves with the error/observation information $I^{g,f}(\tilde{x}_k; \mathcal{Y}_k)$ or the error entropy $h^{g,f}(\tilde{x}_k)$. Then, a direct consequence of Proposition 3.3 is that $g \in \mathcal{G}$ has no probing effect, and the attainable lower bound on $I^{g,f}(\tilde{x}_k; \mathcal{Y}_k)$ only depends on the homogeneous system (H). We study this in some detail next. In what follows, we shall use $\mathcal{F}_h$ to denote the family of filters associated with the homogeneous system (H).

Theorem 3.5: For the linear stochastic system (L), let $f^0 \in \mathcal{F}_h$ be such that

$$h^{f^0}(\tilde{x}_k) = \inf_{f \in \mathcal{F}_h} h^f(\tilde{x}_k). \tag{31}$$

Then

$$I^{f^0}(\tilde{x}_k; \mathcal{Y}_k) = \inf_{f \in \mathcal{F}_h} I^f(\tilde{x}_k; \mathcal{Y}_k) \tag{32}$$

and for any $g \in \mathcal{G}$ given, there is a $f_g \in \mathcal{F}$ such that

$$h^{g,f_g}(\tilde{x}_k) = \inf_{g \in \mathcal{G}, f \in \mathcal{F}} h^{g,f}(\tilde{x}_k) = h^{f^0}(\tilde{x}_k) \tag{33}$$

or equivalently

$$\begin{aligned} I^{g,f_g}(\tilde{x}_k; \mathcal{Y}_k) &= \inf_{g \in \mathcal{G}, f \in \mathcal{F}} I^{g,f}(\tilde{x}_k, \mathcal{Y}_k) \\ &= I^{f^0}(\tilde{x}_k; \mathcal{Y}_k). \end{aligned} \tag{34}$$

Proof: For the trivial control $0 \in \mathcal{G}$ given by $u_k = 0$ for all k, we have that $h^{0,f}(\tilde{x}_k)$, etc., reduces to $h^f(\tilde{x}_k)$, etc., for the homogeneous system (H), except for a constant bias induced by $\bar{x}_0$. Let $f^0 \in \mathcal{F}_h$ be defined by (31). Then, (32) follows from (22) immediately. Now, for the control/filter pair $(0, f^0)$ and $g \in \mathcal{G}$, by 2) of Proposition 3.3, there is an $f_g \in \mathcal{F}$ such that

$$h^{g,f_g}(\tilde{x}_k) = h^{0,f^0}(\tilde{x}_k) = h^{f^0}(\tilde{x}_k) \geq \inf_{g \in \mathcal{G}, f \in \mathcal{F}} h^{g,f}(\tilde{x}_k). \tag{35}$$

Again, from 2) of Proposition 3.3, for any pair $(g, f) \in \mathcal{G} \times \mathcal{F}$, there is a pair $(0, f_h(g, f))$ such that $h^{g,f}(\tilde{x}_k) = h^{0,f_h}(\tilde{x}_k)$. Thus, we have

$$\begin{aligned} \inf_{g \in \mathcal{G}, f \in \mathcal{F}} (\tilde{x}_k) &= \inf_{g \in \mathcal{G}, f \in \mathcal{F}} h^{0,f_h}(\tilde{x}_k) \leq \inf_{f \in \mathcal{F}_h} h^f(\tilde{x}_k) \\ &= h^{f^0}(\tilde{x}_k). \end{aligned} \tag{36}$$

From (35) and (36), we conclude (33). Equation (34) follows directly from (22) and (P3). □

It is important to note that if a realization of the “homogeneous” optimal filter $f^0 \in \mathcal{F}_h$ for the homogeneous system (H) can be determined, then the associated optimal filter $f_g \in \mathcal{G}$

which minimizes $I^{g,f}(\tilde{x}_k;\mathcal{Y}_k)$ can be realized for any given $g \in \mathcal{G}$ from (25), i.e.,

$$\begin{aligned}(f_g)_k(\mathcal{Y}_k^g) &= \left[f_k^0(S_k(\mathcal{Y}_k^g)) + \bar{x}_k^g(\mathcal{Y}_k^g) - \bar{x}_k^0(S_k(\mathcal{Y}_k^g))\right] \pmod{c_k} \\ &= \left[f_k^0(S_k(\mathcal{Y}_k^g)) + \bar{x}_k^g(\mathcal{Y}_k^g)\right] \pmod{c_k'} \end{aligned} \tag{37}$$

because $\bar{x}_k^0$ is deterministic. In (37), c_k and c_k' are sequences that can be computed directly. We also observe from Theorem 3.5 that the achievable lower bound on the error/observation information measure depends only on the homogeneous system (H). $g \in \mathcal{G}$ has no probing effect for this measure. Equation (37) gives an equivalent optimal filter for (L). A corollary to this theorem is given next.

Corollary 3.6: For the linear stochastic system (L), there exists a control and filter pair $(g,f) \in \mathcal{G}\times\mathcal{F}$ such that $I^{g,f}(\tilde{x}_k;\mathcal{Y}_k) = 0$ for all k, i.e., there is a zero error/observation pair (g,f) if and only if there exists a filter $f^0 \in \mathcal{F}_h$ for the homogeneous system (H) such that for all k

$$I^{f^0}(\tilde{x}_k;\mathcal{Y}_k) = \inf_{f\in\mathcal{F}_h} I^f(\tilde{x}_k;\mathcal{Y}_k) = 0. \tag{38}$$

This corollary provides conditions for which Assumption A) is satisfied. Furthermore, from our previous results, if such an f^0 exists, then the minimum mean-square estimator $\bar{f} \in \mathcal{F}_h$ for the homogeneous system given by $\bar{f}_k = E\{\tilde{x}_k \mid \mathcal{Y}_k\}$ is such a filter. We know that for LG systems, the Kalman filter is a realization of the minimum mean-square estimator. In general, however, the minimum mean-square estimator is a nonlinear function of $\mathcal{Y}_k$ and determining a finite-dimensional realization of the filter may not be possible. One may ask: What is the condition for the existence of a zero error/observation filter f^0? A partial answer to this question is presented in the next section, where we show that for the linear stochastic system (L) with an affine linear filter f for the homogeneous system, under some reachability and observability conditions, a necessary and sufficient condition for the existence of a zero error/observation control and filter pair is that the basic random variables are Gaussian.

IV. Zero Error/Observation Information and the Gaussian Assumption

In the previous sections we discussed the importance of a zero error/observation control and filter pair $(g^{\#}, f^{\#})$ such that

$$I^{g^{\#},f^{\#}}(\tilde{x}_k;\mathcal{Y}_k) = \inf_{g\in\mathcal{G},f\in\mathcal{F}} I^{g,f}(\tilde{x}_k;\mathcal{Y}_k) = 0. \tag{39}$$

Roughly speaking, this is the control and filter pair such that all the information about x_k contained in $\mathcal{Y}_k$ is extracted by the filter. For a linear stochastic system (L), from the previous section, we see that the feedback control g plays no role in this case, and the problem is equivalent to determining the existence of a filter $f^0 \in \mathcal{F}_h$ for the homogeneous system (H) such that the zero error/observation information is achieved by f^0. For an LG system, the Kalman filter is such a filter. Or in other words, the Gaussian assumption on the basic random variables is a sufficient condition for the existence of such an f^0. Is the Gaussian assumption also necessary? In this section, we partially answer this question for linear stochastic systems when restricting our attention to the family of affine linear filters $f^0 \in \mathcal{F}_h$ for the homogeneous system and assuming all basic random variables have independent components.

For the homogeneous system (H), the family of affine linear filters $\mathcal{A}_h \subset \mathcal{F}_h$ is defined as

$$\begin{aligned}\mathcal{A}_h \stackrel{\text{def}}{=} \{&f = \{f_0, f_1, \cdots, f_k, \cdots\} \in \mathcal{F}_h : f_k(\mathcal{Y}_k) \\ &= L_k\mathcal{Y}_k + d_k,\ L_k, d_k \text{ constant matrices}\}.\end{aligned}$$

An affine linear filter $f = \{f_0, f_1, \cdots\} \in \mathcal{A}_h$ with $f_k(\mathcal{Y}_k) = L_k\mathcal{Y}_k + d_k$ is called a linear filter, if $d_k = 0$. Let $\mathcal{L}_h \subset \mathcal{A}_h$ denote the family of linear filters for (H). For the error entropy $h^f(\tilde{x}_k)$ or the error/observation information $I^f(\tilde{x}_k;\mathcal{Y}_k)$, a shift $-d_k$ in the error

$$\tilde{x}_k^f = x_k - (L_k\mathcal{Y}_k + d_k) = (x_k - L_k\mathcal{Y}_k) - d_k = \tilde{x}_k^l - d_k \tag{40}$$

has no effect. Note, the superscript $l = \{L_k\mathcal{Y}_k : k \geq 0\} \in \mathcal{L}_h$ in (40) denotes a linear filter in $\mathcal{L}_h$, and there exists a linear zero error/observation filter $l^0 \in \mathcal{L}_h$ for the homogeneous system (H) if and only if there exists an affine linear zero error/observation information filter $f^0 \in \mathcal{A}_h$. For simplicity, we restrict our attention to time-invariant single-input/single-output (SISO) homogeneous systems of the form

$$\begin{cases} x_{k+1} = Ax_k + Gw_k \\ y_k = Cx_k + dv_k \end{cases} \tag{H$'$}$$

where $x_k \in \mathcal{R}^n$ and $y_k, w_i, v_j \in \mathcal{R}$. Let $x_0 = (x_{01}, \cdots, x_{0n})'$. Assume that x_{0i}, w_j, and v_k are zero mean independent random variables with finite and positive variances. The results can be easily modified to treat more general systems (H) with the assumption that all the basic random variables have independent components.

Definition 4.1: The triple (A, G, C) or the SISO discrete-time system

$$\begin{cases} x_{k+1} = Ax_k + Gu_k \\ y_k = Cx_k \end{cases} \tag{Σ'}$$

1) is said to be componentwise observable, if for any $1 \leq i \leq n$, the initial state x_0 in the form $x_0 = (0, \cdots, 0, x_{0i}, 0, \cdots, 0)'$ is observable;
2) is said to be output reachable from zero, if for $x_0 = 0$ and any $y^* \in \mathcal{R}$ given, there is a control such that $y_k = y^*$ for some k.

The following lemma, whose proof is omitted here, follows directly from well-known results in linear systems theory.

Lemma 4.2: The SISO linear discrete-time system (Σ') is componentwise observable if and only if no column of the observability matrix $\mathcal{O} = (C', (CA)', \cdots, (CA^{n-1})')'$ is identically zero. (Σ') is output reachable from zero if only if the row vector $C\mathcal{C} = C(G, AG, \cdots, A^{n-1}G)$ is not identically zero.

Note that the usual observability and reachability (from zero) condition for (Σ') is a sufficient condition (not necessary) for the componentwise observability and output reachability we just defined. The next theorem is the main result of this section.

Theorem 4.3: For the system (H′), a necessary and sufficient condition for Statement B) given below to be true is that $d \neq 0$ and the system (Σ') is both componentwise observable and output reachable from zero.

Statement (B): If a linear (or affine linear) zero error/observation information filter f exists so that $\tilde{x}_k^f$ is independent of $\mathcal{Y}_k$ for all $k \geq 0$, then the basic random variables x_0, w_j, and v_i must be Gaussian for all $j, i \geq 0$.

Remark: If all the basic random variables are Gaussian, then the Kalman filter is an affine linear zero error/observation information filter, and Theorem 4.3 says that under the conditions posed, the existence of a linear (affine linear) zero error/observation filter f is equivalent to the Gaussian assumption on the system.

Before we begin the proof of this important theorem, some technical results, stated as lemmas, are required. A key result from statistics [28], Lemma 4.4 is presented next.

Lemma 4.4: Let $\eta_1, \cdots, \eta_k \in \mathcal{R}$ be independent random variables with positive and finite variances. If there are constants $a_i, b_i \in \mathcal{R}$ for $i = 1, \cdots, k$ with $a_i b_i \neq 0$ for all i such that $\sum_{i=1}^k a_i \eta_i$ and $\sum_{i=1}^k b_i \eta_i$ are independent, then η_i is Gaussian for each $i \in \{1, \cdots, k\}$.

A direct consequence of this Lemma is as follows.

Lemma 4.5: Given independent random variables $\xi_1, \cdots, \xi_n$ with positive and finite variances, suppose that there are nonzero constants $\alpha_1, \cdots, \alpha_l$ and $\beta_1, \cdots, \beta_m$ with $0 < l < m \leq n$ such that

$$\begin{aligned} X &= \alpha_1 \xi_1 + \cdots + \alpha_l \xi_l \\ Y &= \beta_1 \xi_1 + \cdots, \beta_m \xi_m \end{aligned} \tag{41}$$

are independent, then $\xi_1, \cdots, \xi_l$ must be Gaussian. However, the distribution of $\xi_{l+1}, \cdots, \xi_n$ can be arbitrary (unspecified).

Proof: Define $\tilde{Y} = Y - (\beta_{l+1}\xi_{l+1} + \cdots + \beta_m \xi_m)$. Then, by the assumption of the lemma, we see that X and $\tilde{Y}$ are independent because X is independent of both Y and $(\beta_{l+1}\xi_{l+1} + \cdots + \beta_m \xi_m)$. From Lemma 4.4, we know that $\xi_1, \cdots, \xi_l$ must be Gaussian. Clearly, the distributions of $\xi_{l+1}, \cdots, \xi_n$ remain unspecified. □

Lemma 4.6: Let $A = (a_{ij})_{n \times m}$ be a $n \times m$ matrix, and let a_j be the jth column of A. There is a row vector $\alpha = (\alpha_1, \cdots, \alpha_n)$ such that $\alpha a_j \neq 0$ for all j if and only if $a_j \neq 0$ for all j.

Proof (Necessity): Assume that $a_j = 0$ for some j. Then $\alpha a_j = 0$ for all α. For sufficiency, assume that $a_j \neq 0$ for all j. Consider that for any j fixed, the relation $\alpha a_j = 0$ defines an $(n-1)$-dimensional subspace in $\mathcal{R}^n$. Then, with $S_j \stackrel{\text{def}}{=} \{\alpha \in \mathcal{R}^n : \alpha a_j = 0\}$, the set $D \stackrel{\text{def}}{=} \cup_j S_j$ has volume zero, and we can always select $\alpha \in \mathcal{R}^n \backslash D$ such that $\alpha a_j \neq 0$ for all j. □

Now, we begin to prove Theorem 4.3. We do so by showing a slightly more general result. With x_0, v_i, and w_j the basic random variables, and f_k a linear function of $\mathcal{Y}^k$, it is easy to see that we have the following functional relationships for all i, j, k:

$$\begin{cases} y_i = y_i(x_0, w_0, \cdots, w_{i-1}, v_0, \cdots, v_i) \\ x_k = x_k(x_0, w_0, \cdots, w_{k-1}) \\ \tilde{x}_j = \tilde{x}_j(x_0, w_0, \cdots, w_{j-1}, v_0, \cdots, v_j) \end{cases} \tag{42}$$

where all functions are linear in their arguments. It follows that (a Taylor series expansion and linearity) we have (43), as shown at the bottom of the page. If we represent the second equation in (43) in a matrix form, we have

$$\mathcal{Y}^k = \frac{\partial \mathcal{Y}^k}{\partial x_0} x_0 + (H_0, H_1, \cdots, H_{k-1}) \begin{pmatrix} w_0 \\ \vdots \\ w_{k-1} \end{pmatrix} + d \begin{pmatrix} v_0 \\ \vdots \\ v_k \end{pmatrix} \tag{44}$$

where

$$\frac{\partial \mathcal{Y}^k}{\partial x_0} = \begin{pmatrix} \frac{\partial y_0}{\partial x_0}' \\ \frac{\partial y_1}{\partial x_0}' \\ \vdots \\ \frac{\partial y_k}{\partial x_0}' \end{pmatrix} = \begin{pmatrix} C \\ CA \\ \vdots \\ CA^k \end{pmatrix} \in \mathcal{R}^{(k+1) \times n}$$

$$H_i = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \frac{\partial y_{i+1}}{\partial w_i} \\ \vdots \\ \frac{\partial y_k}{\partial w_i} \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ CG \\ \vdots \\ CA^{k-1-i}G \end{pmatrix} \in \mathcal{R}^{k+1} \tag{45}$$

for all $i = 0, 1, \cdots, (k-1)$.

Theorem 4.7: For any $k \geq n$ fixed, if there is a linear filter such that $\tilde{x}_k$ is independent of $\mathcal{Y}^k$, then

$$x_0, w_0, \cdots, w_{k-n}, v_0, \cdots, v_k$$

are Gaussian if and only if there exists $\alpha = (\alpha_0, \alpha_1, \cdots, \alpha_k) \in \mathcal{R}^{k+1}$ and $\beta = (\beta_1, \cdots, \beta_n) \in \mathcal{R}^n$ such that

$$\begin{aligned} &\alpha \frac{\partial \mathcal{Y}^k}{\partial x_0} \neq 0 \text{ Componentwise,} \\ &\alpha H_i \neq 0, \quad \forall i = 0, \cdots, (k-n); \\ &\alpha d \neq 0 \text{ Componentwise} \end{aligned} \tag{46}$$

and also

$$\begin{aligned} &\beta \frac{\partial \tilde{x}_k}{\partial x_0} \neq 0 \text{ Componentwise,} \\ &\beta \frac{\partial \tilde{x}_k}{\partial w_i} \neq 0, \quad \forall i = 0, \cdots, (k-n); \quad \beta \frac{\partial \tilde{x}_k}{\partial v_j} \neq 0, \\ &\qquad \forall j = 0, \cdots, k. \end{aligned} \tag{47}$$

$$\begin{cases} \tilde{x}_k = \frac{\partial \tilde{x}_k}{\partial x_0} x_0 + \frac{\partial \tilde{x}_k}{\partial w_0} w_0 + \cdots + \frac{\partial \tilde{x}_k}{\partial w_{k-1}} w_{k-1} + \frac{\partial \tilde{x}_k}{\partial v_0} v_0 + \cdots + \frac{\partial \tilde{x}_k}{\partial v_k} v_k \\ y_i = \frac{\partial y_i}{\partial x_0} x_0 + \frac{\partial y_i}{\partial w_0} w_0 + \cdots + \frac{\partial y_i}{\partial w_{i-1}} w_{i-1} + \frac{\partial y_i}{\partial v_0} v_0 + \cdots + \frac{\partial y_i}{\partial v_i} v_i, \\ \qquad i = 0, 1, \cdots, k \end{cases} \tag{43}$$

Proof: We show sufficiency first. Note that $\tilde{x}_k$ and $\mathcal{Y}^k$ are independent if and only if $\alpha\tilde{x}_k$ and $\beta\mathcal{Y}^k$ are independent for all α and β. For the α and β given in the theorem, since $\tilde{x}_k$ is independent of $\mathcal{Y}^k$, it follows that $\beta\tilde{x}_k$ and $\alpha\mathcal{Y}^k$ are independent. However, from (43) and (44), we have

$$\beta\tilde{x}_k = \left[\sum_{i=1}^{n}\left(\beta\frac{\partial\tilde{x}_k}{\partial x_0}\right)_i x_{0i} + \sum_{i=0}^{k-n}\left(\beta\frac{\partial\tilde{x}_k}{\partial w_i}\right)w_i + \sum_{i=0}^{k}\left(\beta\frac{\partial\tilde{x}_k}{\partial v_i}\right)v_i\right] + \sum_{i=k-n+1}^{k-1}\left(\beta\frac{\partial\tilde{x}_k}{\partial w_i}\right)w_i$$

$$\alpha\mathcal{Y}^k = \left[\sum_{i=1}^{n}\left(\alpha\frac{\partial\mathcal{Y}^k}{\partial x_0}\right)_i x_{0i} + \sum_{i=0}^{k-n}(\alpha H_i)w_i + \sum_{i=0}^{k} d\alpha_i v_i\right] + \sum_{i=k-n+1}^{k-1}(\alpha H_i)w_i. \tag{48}$$

Equations (46) and (47) are equivalent to the fact that all the coefficients inside $[\cdots]$ on the right-hand side of (48) (except those for $w_{k-n+1},\cdots,w_{k-1}$) are not zero. We conclude that $x_{01},\cdots,x_{0n},w_0,\cdots,w_{k-n}$, and $v_0,\cdots,v_k$ are Gaussian from Lemma 4.5.

For necessity, assume that for all α and β, (46) and (47) do not hold. That is, for any α and β constant vectors, at least one coefficient inside $[\cdots]$ on the right-hand side of (48) is zero. It is easy to see that this implies that there is a coefficient, without loss of generality, say $d\alpha_i$, that is zero for almost all α. This implies that $\mathcal{Y}^k$ is independent of v_i. Then, by Lemma 4.5, the distribution of v_i can be arbitrary (even though $\tilde{x}_k$ is independent of $\mathcal{Y}^k$). □

At this stage, from Lemma 4.6, the necessary and sufficient condition given in Theorem 4.7 is equivalent to the following two conditions:

$$\begin{cases}\frac{\partial\tilde{x}_k}{\partial x_{0i}} \neq 0, & \forall i=1,2,\cdots,n,\\ \frac{\partial\tilde{x}_k}{\partial w_i} \neq 0, & \forall i=0,1,\cdots,(k-n),\\ \frac{\partial\tilde{x}_k}{\partial v_i} \neq 0, & \forall i=1,2,\cdots,k.\end{cases} \tag{C1}$$

and

$$\begin{cases}\frac{\partial\mathcal{Y}^k}{\partial x_{0i}} = \mathcal{O}_k e_i \neq 0, & \forall i=1,2,\cdots,n,\\ \frac{\partial\mathcal{Y}^k}{\partial w_i} = H_i \neq 0, & \forall i=0,1,\cdots,(k-n),\\ \frac{\partial\mathcal{Y}^k}{\partial v_i} = (d,d,\cdots,d)' \neq 0, & \forall i=1,2,\cdots,k.\end{cases} \tag{C2}$$

where in (C2), $\mathcal{O}_k \stackrel{\text{def}}{=} (C',(CA)',\cdots,(CA^k)')'$ and e_i is the ith standard basis vector for $\mathcal{R}^n$. Note that Condition (C2) is equivalent to $d \neq 0$ and the system is componentwise observable and output reachable. Therefore, the necessity part of Theorem 4.3 is proved. To show sufficiency, we prove the following result.

Proposition 4.8: Assume that $d \neq 0$, and (Σ') is both componentwise observable and output reachable. Then, for the Kalman filter and i,j,l arbitrary but fixed, we have

$$\begin{cases}\frac{\partial\tilde{x}_k}{\partial x_{0i}} \neq 0 & \text{for some } k\geq 0,\\ \frac{\partial\tilde{x}_k}{\partial w_j} \neq 0 & \text{for some } k\geq 0,\\ \frac{\partial\tilde{x}_k}{\partial v_l} \neq 0 & \text{for some } k\geq 0.\end{cases} \tag{C3}$$

Proof: Without loss of generality, we can assume that all the basic random variables are Gaussian. (Otherwise, just regard the notation $E\{x_l \mid \mathcal{Y}_l\}$ as the orthogonal projection of x_l onto the Hilbert space $Y^k = \text{span}\{y_0,\cdots,y_l\}$, and change "independent" to "uncorrelated" in the following proof, etc. For a detailed account on the Hilbert space approach to the filtering problem, see [25].) Note also that all random variables involved are zero mean. We show the first equation in (C3) next. Suppose that for all $k\geq 0$, $\frac{\partial\tilde{x}_k}{\partial x_{0i}} = 0$ for some i, say, $i=1$. Since $\tilde{x}_k = \tilde{x}_k(x_{01},\cdots,x_{0n},w_0,\cdots,w_{k-1})$ is a linear function of its arguments, then $\tilde{x}_k$ is independent of x_{01}. Define a new system

$$\begin{cases}\bar{x}_{k+1} = A\bar{x}_k + Gw_k\\ \bar{y}_k = C\bar{x}_k + dv_k\end{cases} \tag{K}$$

where the initial state $\bar{x}_0 = x_0 + \zeta e_1$ with $\zeta\in\mathcal{R}$ a Gaussian $N(0,\sigma^2)$ random variable independent of x_0, w_i, and v_j. Let $\Sigma_{k|k}$ and $\bar{\Sigma}_{k|k}$ denote the error covariances of the Kalman filter for (H′) and (K), respectively. Since $\tilde{x}_k$ is not a function of x_{01}, we have $\bar{\Sigma}_{k|k} = \Sigma_{k|k}$ for any σ^2. We show next this will lead to $\mathcal{O}e_1 = 0$.

By linearity, we have

$$\begin{aligned}\bar{x}_k &= x_k + A^k e_1\zeta\\ \bar{\mathcal{Y}}_k &= \mathcal{Y}_k + \mathcal{O}_k e_1\zeta\end{aligned} \tag{49}$$

where $\bar{\mathcal{Y}}_k = (\bar{y}_0,\cdots,\bar{y}_k)'$ is the information vector for (K) and $\mathcal{O}_k = (C',(CA)',\cdots,(CA^k)')'$. It follows that

$$E\{\bar{x}_k \mid \mathcal{Y}_k,\zeta\} = E\{x_k \mid \mathcal{Y}_k,\zeta\} + (A^k e_1)\zeta.$$

Since ζ is independent of $(\mathcal{Y}_k, x_k)$, this implies that $E\{x_k \mid \mathcal{Y}_k,\zeta\} = E\{x_k \mid \mathcal{Y}_k\}$. Therefore, $E\{\bar{x}_k \mid \mathcal{Y}_k,\zeta\} = E\{x_k \mid \mathcal{Y}_k\} + (A^k e_1)\zeta$ and

$$\bar{x}_k - E\{\bar{x}_k \mid \mathcal{Y}_k,\zeta\} = x_k - E\{x_k \mid \mathcal{Y}_k\} = \tilde{x}_k. \tag{50}$$

Hence

$$\Sigma_{k|k} = \text{Cov}(\bar{x}_k - E\{\bar{x}_k \mid \mathcal{Y}_k,\zeta\}). \tag{51}$$

From (49), we have $\sigma\{\bar{\mathcal{Y}}_k,\zeta\} = \sigma\{\mathcal{Y}_k,\zeta\}$. It follows that

$$\bar{x}_k - E\{\bar{x}_k \mid \mathcal{Y}_k,\zeta\} = \bar{x}_k - E\{\bar{x}_k \mid \bar{\mathcal{Y}}_k,\zeta\}. \tag{52}$$

Let $\tilde{\zeta} = \zeta - E\{\zeta \mid \bar{\mathcal{Y}}_k\}$ be the innovation. Then

$$\bar{x}_k - E\{\bar{x}_k \mid \bar{\mathcal{Y}}_k,\zeta\} = \bar{x}_k - E\{\bar{x}_k \mid \bar{\mathcal{Y}}_k,\tilde{\zeta}\} \tag{53}$$

where $\bar{\mathcal{Y}}_k$ and $\tilde{\zeta}$ are independent. By a direct computation, we have from (51)–(53)

$$\begin{aligned}\Sigma_{k|k} &= \text{Cov}(\bar{x}_k - E\{\bar{x}_k \mid \bar{\mathcal{Y}}_k,\tilde{\zeta}\})\\ &= \bar{\Sigma}_{k|k} - (E\{\bar{x}_k\tilde{\zeta}'\})(E\{\bar{x}_k\tilde{\zeta}'\})'/(E\{\tilde{\zeta}^2\}).\end{aligned} \tag{54}$$

Thus, $\Sigma_{k|k} = \bar{\Sigma}_{k|k}$ if and only if

$$E\{\bar{x}_k\tilde{\zeta}'\} = 0. \tag{55}$$

We shall show (55) implies that $\mathcal{O}e_1 = 0$. Let us compute $E\{\bar{x}_k\tilde{\zeta}\} = E\{\bar{x}_k\tilde{\zeta}'\}$ (note that $\tilde{\zeta} = \tilde{\zeta}'$ is a scalar). The innovation can be computed as

$$\tilde{\zeta} = \zeta - E\{\zeta \mid \bar{\mathcal{Y}}_k\} = \zeta - E\{\zeta\bar{\mathcal{Y}}_k'\}(E\{\bar{\mathcal{Y}}_k\bar{\mathcal{Y}}_k'\})^{-1}\bar{\mathcal{Y}}_k. \tag{56}$$

Since ζ is independent of $\mathcal{Y}_k$, we have

$$E\{\zeta\bar{\mathcal{Y}}_k'\} = E\{\zeta(\mathcal{Y}_k + (\mathcal{O}_k e_1)\zeta)\} = \sigma^2(\mathcal{O}_k e_1)' \quad (57)$$

and

$$\begin{aligned} E\{\bar{\mathcal{Y}}_k\bar{\mathcal{Y}}_k'\} &= E\{(\mathcal{Y}_k + (\mathcal{O}_k e_1)\zeta)(\mathcal{Y}_k + (\mathcal{O}_k e_1)\zeta)'\} \\ &= (E\{\bar{\mathcal{Y}}_k\bar{\mathcal{Y}}_k'\} + \sigma^2\mathcal{O}_k e_1 e_1'\mathcal{O}_k') \stackrel{\text{def}}{=} (\Sigma_{\bar{\mathcal{Y}}}). \end{aligned} \quad (58)$$

Then, (56)–(58) yields

$$\tilde{\zeta} = \zeta - \sigma^2 e_1'\mathcal{O}_k'(\Sigma_{\bar{\mathcal{Y}}})^{-1}\bar{\mathcal{Y}}_k. \quad (59)$$

It follows that

$$E\{\bar{x}_k\tilde{\zeta}\} = E\{\bar{x}_k\zeta\} - \sigma^2 E\{\bar{x}_k\bar{\mathcal{Y}}_k'\}(\Sigma_{\bar{\mathcal{Y}}})^{-1}\mathcal{O}_k e_1. \quad (60)$$

However, because ζ is independent of x_k, we have

$$E\{\bar{x}_k\zeta\} = E\{(x_k + (A^k e_1)\zeta)\zeta\} = \sigma^2(A^k e_1). \quad (61)$$

Also, from the independence of ζ and $(x_k, \mathcal{Y}_k)$, we have

$$\begin{aligned} E\{\bar{x}_k\bar{\mathcal{Y}}_k'\} &= E\{(x_k + (A^k e_1)\zeta)(\mathcal{Y}_k + (\mathcal{O}_k e_1)\zeta)\} \\ &= E\{x_k\mathcal{Y}_k'\} + \sigma^2(A^k e_1)(\mathcal{O}_k e_1)'. \end{aligned} \quad (62)$$

Combining (60)–(62), we have

$$\begin{aligned} E\{\bar{x}_k\tilde{\zeta}\} = \sigma^2[&A^k e_1 - \sigma^2(A^k e_1)(\mathcal{O}_k e_1)'(\Sigma_{\bar{\mathcal{Y}}})^{-1}(\mathcal{O}_k e_1) \\ &- E\{x_k\mathcal{Y}_k'\}(\Sigma_{\bar{\mathcal{Y}}})^{-1}(\mathcal{O}_k e_1)]. \end{aligned} \quad (63)$$

Using the matrix inversion lemma, with $\Sigma_y = E\{\mathcal{Y}_k\mathcal{Y}_k'\}$ (the invertability of Σ_y is guaranteed by $d \neq 0$) we obtain

$$\begin{aligned} (\Sigma_{\bar{\mathcal{Y}}})^{-1} &= (\Sigma_y + \sigma^2(\mathcal{O}_k e_1)(\mathcal{O}_k e_1)')^{-1} \\ &= \left(\Sigma_y^{-1} - \frac{(\Sigma_y^{-1}\mathcal{O}_k e_1)(\Sigma_y^{-1}\mathcal{O}_k e_1)'}{\sigma^{-2} + (\Sigma_y^{-1}\mathcal{O}_k e_1)'\Sigma_y(\Sigma_y^{-1}\mathcal{O}_k e_1)}\right). \end{aligned} \quad (64)$$

From (63) and (64), we see that with σ^2 sufficiently large, the second term inside $[\cdots]$ in (63) dominates. Because (55) holds for all σ^2, this implies that this term must be zero for all σ^2 large, i.e.,

$$(A^k e_1)(\mathcal{O}_k e_1)'(\cdots)^{-1}(\mathcal{O}_k e_1) = 0, \qquad \forall k \geq 0, \qquad \sigma^2 \gg 1. \quad (65)$$

Since $(\Sigma_{\bar{\mathcal{Y}}})^{-1}$ is positive definite, (65) is equivalent to

$$\text{either } A^k e_1 = 0 \text{ or } \mathcal{O}_k e_1 = 0, \qquad \forall k \geq 0. \quad (66)$$

Using (66) for $k = 0, 1, \cdots, n-1$, we can easily conclude that

$$\mathcal{O}_{n-1}e_1 = \mathcal{O}e_1 = 0.$$

This contradicts the componentwise observability and proves the first statement in (C3).

For the second equation in (C3), suppose that for all $k \geq 0$, $\frac{\partial\tilde{x}_k}{\partial w_j} = 0$ for some j, say $j = 0$. Define a new system

$$\begin{cases} \bar{x}_{k+1} = A\bar{x}_k + G\bar{w}_k \\ \bar{y}_k = C\bar{x}_k + dv_k \end{cases} \quad (\text{K}')$$

where $\bar{w}_0 = w_0 + \zeta$ and ζ is defined as before, and $\bar{w}_j = w_j$ for $j \neq 0$. Then, by linearity, we have similar to (49) that for all $k \geq 0$ and with $A^{-1} \stackrel{\text{def}}{=} I$

$$\begin{aligned} \bar{x}_k &= x_k + (A^{k-1}G)\zeta \\ \bar{\mathcal{Y}}_k &= \mathcal{Y}_k + ((0', \mathcal{O}_{k-1}')'G)\zeta. \end{aligned} \quad (67)$$

Following the same arguments, with e_1 replaced by G and A^k by A^{k-1}, we obtain instead of (66)

$$\text{either } A^{k-1}G = 0 \quad \text{or} \quad (0', \mathcal{O}_{k-1}')'G = 0, \qquad \forall k \geq 0. \quad (68)$$

Using (68) for $k = 1, \cdots, n$, we can easily conclude that $C\mathcal{C} = C(G, AG, \cdots, A^{n-1}G) = 0$.

For the last equation in (C3), assume that $\frac{\partial\tilde{x}_k}{\partial v_l} = 0$ for all $k \geq 0$ and for some $l \geq 0$, i.e., $\tilde{x}_k$ is independent of v_l. Since $y_l = Cx_l + dv_l$ and v_l is independent of all other random variables and $d \neq 0$, this implies that $\tilde{x}_k$ is not a function of y_l. However, we have

$$y_l = CA^l x_0 + \sum_{i=0}^{l-1} CA^{l-1-i}Gw_i + dv_l$$

where $x_{01}, \cdots, x_{0n}, w_0, \cdots, w_{l-1}$, and v_l are independent. This implies that either

$$CA^l = 0, \qquad CG = CAG = \cdots = CA^{l-1}G = 0 \quad (69)$$

or

$$\frac{\partial\tilde{x}_k}{\partial x_{0i}} = 0, \qquad \frac{\partial\tilde{x}_k}{\partial w_j} = 0, \qquad 1 \leq i \leq n,\ 0 \leq j \leq (l-1). \quad (70)$$

Equation (70) contradicts the componentwise observability and output reachability, but (69) leads to $C\mathcal{C} = 0$, which contradicts the output reachability. This shows the last equation in (C3) and completes the proof. □

Now, we continue with the proof of Theorem 4.3.

Proof of Theorem 4.3: Necessity follows directly from Theorem 4.7 and (C2). For sufficiency, assume that $d \neq 0$ and (Σ') is both componentwise observable and output reachable. Note that for all $k \geq 0$, $\tilde{x}_k$ is independent of $\mathcal{Y}_k$ by assumption. For any i, j, l arbitrarily given, select k sufficiently large so that (C3) holds and (C2) holds with $k - n > \max\{j, l\}$. Then, it follows that there exist α and β such that (48) holds with

$$\left(\beta\frac{\partial\tilde{x}_k}{\partial x_0}\right)_i \neq 0, \qquad \beta\frac{\partial\tilde{x}_k}{\partial w_j} \neq 0, \qquad \beta\frac{\partial\tilde{x}_k}{\partial v_l} \neq 0$$

and

$$\left(\alpha\frac{\partial\mathcal{Y}_k}{\partial x_0}\right)_i \neq 0, \qquad \alpha H_l \neq 0, \qquad \alpha_l d \neq 0.$$

Now applying Lemma 4.5, we conclude that x_{0i}, w_j, and v_l must be Gaussian. Since i, j, l are arbitrary, this proves that x_0, w_i, and v_j are Gaussian for all $i, j \geq 0$. □

Next, we examine the linear inhomogeneous system. Define the family of *quasi-affine linear* filters f_g associated with a feedback control $g \in \mathcal{G}$ as

$$\begin{aligned} \mathcal{S} = \{f_g \in \mathcal{F} : (f_g)_k(\mathcal{Y}_k^g) = L_k\mathcal{Y}_k^g + P_k\mathcal{U}_k^g + d_k, \\ L_k, P_k, d_k \ \text{ are real-valued matrices.}\}. \end{aligned} \quad (71)$$

Clearly, the Kalman filter is a quasi-affine linear filter for any $g \in \mathcal{G}$. Suppose that f^0 is the affine linear minimum error entropy filter for the homogeneous system. Then, by Theorem 3.5, for any $g \in \mathcal{G}$ given, there is a filter f_g such that (33) and (34) hold. From (37), we know that this filter f_g can be taken as

$$(f_g)_k(\mathcal{Y}_k^g) = f_k^0(S_k(\mathcal{Y}_k^g)) + \bar{x}_k^g(\mathcal{Y}_k^g)$$

where S_k is the transformation such that $\mathcal{Y}_k = S_k(\mathcal{Y}_k^g)$. Since

$y_k = y_k^g - \bar{y}_k^g$ from (20), we can write $\mathcal{Y}_k = \mathcal{Y}_k^g - \bar{\mathcal{Y}}_k^g$ with $\bar{\mathcal{Y}}_k^g \stackrel{\text{def}}{=} (\bar{y}_0^g, \cdots, \bar{y}_k^g)'$. Then, we have

$$(f_g)_k(\mathcal{Y}_k^g) = f_k^0(\mathcal{Y}_k^g - \bar{\mathcal{Y}}_k^g) + \bar{x}_k^g(\mathcal{Y}_k^g). \tag{72}$$

From (D), we see that both $\bar{x}_k^g$ and $\bar{y}_k^g$ are linear functions of $\bar{x}_0$ and $\mathcal{U}_k^g$. Hence, from (72), we see that f_g is a quasi-affine linear filter associated with g if and only if f^0 is an affine linear filter for the homogeneous system. Therefore, as a consequence of Theorems 4.3 and 3.5 (or Corollary 3.6), we have the following result.

Theorem 4.9: For the linear stochastic system

$$\begin{cases} x_{k+1} = Ax_k + Bu_k + Gw_k \\ y_k = Cx_k + Dv_k \end{cases} \tag{L$'$}$$

where $x_0 \in \mathcal{R}^n, w_i \in \mathcal{R}$, and $v_j \in \mathcal{R}$ are independent random vectors with independent components and $y_k \in \mathcal{R}$. A necessary and sufficient condition for the Statement (C) given below to be true is that $D \neq 0$, and (A, G, C) is componentwise observable and output reachable.

Statement (C): If there is a feedback control $g \in \mathcal{G}$ and a quasi-affine linear filter f_g such that (g, f_g) is a zero error/observation information pair, then the basic random vectors x_0, w_j, and v_i must be Gaussian for all $j, i \geq 0$.

Remark: The key idea of the above proof is that the condition $D \neq 0$ guarantees that for any i fixed, there is a k such that y_k is a function of v_i so that v_i will affect y_k in a nontrivial way. Similarly, the output reachability condition guarantees that for any j fixed, there is a k such that y_k is a (nontrivial) function of w_j, and the componentwise observability condition guarantees that each component x_{0l} of the initial state x_0 will affect y_k in a nontrivial way, for some time instant k. This idea can be directly applied to the proof of the multivariable case where in (L$'$), y_k, v_i, and w_j are all vectors, say, $y_k \in \mathcal{R}^m, v_i \in \mathcal{R}^q$, and $w_j \in \mathcal{R}^s$ such that v_i and w_j have independent components. In this case, $C = (c_1', \cdots, c_m')' \in \mathcal{R}^{m \times n}$, $G = (g_1, \cdots g_s) \in \mathcal{R}^{n \times s}$, and $D = (d_1, \cdots, d_q) \in \mathcal{R}^{m \times q}$ are matrices. Without giving a detailed proof, we claim that the necessary and sufficient condition for Statement (C) is: $d_j \neq 0$ for each j, (A, c_i) is componentwise observable for some i, and for any l given, there is a k such that (A, g_l, c_k) is output reachable. (Note that the componentwise observability is well defined for (A, c_i) in an obvious way.)

V. Example

In this section, we present an example to illustrate some of the results presented in the previous sections.

Example: Consider the scalar linear stochastic system

$$\begin{cases} x_{k+1}^g = u_k^g + w_k \\ y_k^g = x_k^g + v_k \end{cases} \tag{73}$$

where x_0, w_j, and v_k are independent random variables with the uniform distributions

$$x_0, w_i \sim \text{uni}[-\alpha, \alpha], \qquad v_j \sim \text{uni}[-\beta, \beta]. \tag{74}$$

Assume that $\alpha \geq \beta > 0$ are constants. By the results of Section IV, to derive the optimal filter(s), we consider the homogeneous system for (73)

$$\begin{cases} x_{k+1} = w_k \\ y_k = x_k + v_k. \end{cases} \tag{75}$$

For this simple system, x_0 and w_0 have the same distribution, then we can consider the output equation

$$y_k = x_k + v_k, \qquad \forall k \geq 0. \tag{76}$$

Here x_k is an i.i.d. sequence of uni$[-\alpha, \alpha]$ random variables which are independent of v_j. We derive different filters for (75) next.

The Kalman Filter: Since all the random variables have zero mean, the Kalman filter is the linear minimum mean-square filter $k_j(\mathcal{Y}_j) = \hat{x}_j^k(\mathcal{Y}_j)$ given by

$$\hat{x}_j^k(\mathcal{Y}_j) = \sum_{i=0}^{j} m_i y_i$$

where the m_i are chosen so that the error covariance $\sigma_j^k(\tilde{x}_j) = \text{Cov}\{x_j - \hat{x}_j^k(\mathcal{Y}_j)\}$ is minimized. Since $x_j = w_j$ is independent of $\mathcal{Y}_{j-1}$, (76) yields

$$\begin{aligned} \text{Cov}\{\tilde{x}_j^k\} &= \text{Cov}\{x_j - m_0 y_0 - \cdots - m_j y_j\} \\ &= (1 - m_j)^2 \text{Cov}(x_j) + m_j^2 \text{Cov}(v_j) \\ &\quad + \sum_{i=0}^{j-1} m_j^2 \text{Cov}(y_j). \end{aligned}$$

Thus, we see that $m_l = 0$ for $l \leq j-1$ and $m_j = \alpha^2/(\alpha^2+\beta^2)$, i.e.,

$$\hat{x}_j^k(\mathcal{Y}_j) = \hat{x}_j^k(y_j) = \frac{\alpha^2}{\alpha^2 + \beta^2} y_j, \quad j = 0, 1, \cdots. \tag{77}$$

The Minimum Mean-Square Filter: The minimum mean-square filter is given by

$$\hat{x}_k^m(\mathcal{Y}_k) = E\{x_k \mid \mathcal{Y}_k\}.$$

By the independence, it is easy to see that $\hat{x}_k^m(\mathcal{Y}_k) = \hat{x}_k^m(y_k) = E\{x_k \mid y_k\}$. To determine $\hat{x}_k^m(y_k)$, we have to compute the conditional density function $p(x_k \mid y_k)$ and then carry out the integration. A direct computation yields, with $\delta = \alpha - \beta$ and $\epsilon = \alpha + \beta$

$$\hat{x}_k^m(y_k) = \begin{cases} (y_k - \delta)/2, & \text{if } -\epsilon \leq y_k \leq -\delta, \\ y_k, & \text{if } -\delta \leq y_k \leq \delta, \\ (y_k + \delta)/2, & \text{if } \delta \leq y_k \leq \epsilon. \end{cases} \tag{78}$$

The Affine Linear Minimum Error Entropy Filter: Let $\hat{x}_k^L = L_k \mathcal{Y}_k + d_k = \sum_{i=0}^{k} l_i^{(k)} y_i + d_k$ denote an affine linear filter. Then, the error entropy is given by

$$h^L(\tilde{x}_k) = h\left(x_k - \sum_{i=0}^{k} l_i^{(k)} y_i - d_k\right) \stackrel{\text{(P5)}}{=} h\left(x_k - \sum_{i=0}^{k} l_i^{(k)} y_i\right).$$

Since $x_k - l_k^{(k)} y_k$ is independent of y_j for $j \leq k-1$, then by (P6), we have

$$h^L(\tilde{x}_k) = h\left(x_k - \sum_{i=0}^{k} l_i^{(k)} y_i\right) \geq h\left(x_k - l_k^{(k)} y_k\right).$$

Hence, the affine linear minimum error entropy filter is given by $\hat{x}_k^l(\mathcal{Y}_k) = \hat{x}_k^l(y_k) = l_k y_k$, where $l_k \in \mathcal{R}$ is chosen to minimize the error entropy

$$h(x_k - l_k y_k) = \min_{l\in\mathcal{R}} h(x_k - l y_k) = \min_{l\in\mathcal{R}} h((1-l)x_k - l v_k) = \min_{l\in\mathcal{R}} h((1-l)x_k + l v_k). \quad (79)$$

The last equality follows the fact that x_k and v_k are independent and v_k has a symmetric distribution about zero. After some manipulation with densities and entropy, it can be shown that a global minimum is attained at $l_k = l^* = \alpha/(\alpha+\beta)$ and the affine linear minimum mean-square filter is given by

$$\hat{x}_k^l(\mathcal{Y}_k) = \hat{x}_k^l(y_k) = \frac{\alpha}{\alpha+\beta} y_k. \quad (80)$$

From the above, we see that the (true) minimum mean-square filter $\hat{x}_k^m$ is a piecewise linear function of y_k. In the case when $\alpha = \beta$, we have $\hat{x}_k^k = \hat{x}_k^m = \hat{x}_k^l = y_k/2$. That is, the Kalman filter, the minimum mean-square filter, and the affine linear minimum entropy filter are the same linear function of y_k. To compare the performance of these (optimal) filters according to different measures, we have to compute the error covariance, the error entropy, and the error/observation (mutual) information associated with each of them. The computation is straightforward but tedious and is omitted here. We present the results next. (Note that the superscripts k, m, and l indicate the Kalman filter, the minimum mean-square filter, and the minimum error entropy filter, respectively.)

For the error covariances, we have

$$\begin{aligned} \mathrm{Cov}(\tilde{x}_k^k) &= \frac{\alpha^2\beta^2}{3(\alpha^2+\beta^2)} \\ \mathrm{Cov}(\tilde{x}_k^m) &= \frac{1}{3}\beta^2\left(1 - \frac{\beta}{2\alpha}\right) \\ \mathrm{Cov}(\tilde{x}_k^l) &= \frac{2\alpha^2\beta^2}{3(\alpha+\beta)^2}. \end{aligned} \quad (81)$$

For the error entropy, we have

$$\begin{aligned} h(\tilde{x}_k^k) &= \frac{\alpha}{2\beta} - \log\frac{\alpha^2+\beta^2}{2\alpha\beta^2} \\ h(\tilde{x}_k^m) &= \frac{(\alpha-\beta)^2}{4\alpha\beta}\log(\alpha-\beta) - \frac{(\alpha+\beta)^2}{4\alpha\beta}\log(\alpha+\beta) \\ &\quad + \log(2\alpha\beta) + \frac{1}{2} \\ h(\tilde{x}_k^l) &= \frac{1}{2} - \log\frac{\alpha+\beta}{2\alpha\beta}. \end{aligned} \quad (82)$$

It can be shown by direct computation that $h(x_k \mid \mathcal{Y}_k) = \log(2\beta) - \beta/(2\alpha)$. Then, from $I^d(\tilde{x}_k; \mathcal{Y}_k) = h^d(\tilde{x}_k) - h(x_k \mid \mathcal{Y}_k)$ for $d = k, m, l$, the error/observation information measures are given by

$$\begin{aligned} I(\tilde{x}_k^k; \mathcal{Y}_k) &= \frac{\alpha^2+\beta^2}{2\alpha\beta} - \log\frac{\alpha^2+\beta^2}{\alpha\beta} \\ I(\tilde{x}_k^m; \mathcal{Y}_k) &= \frac{\alpha+\beta}{2\alpha} - \log\frac{\alpha+\beta}{\alpha} \\ &\quad + \frac{(\alpha-\beta)^2}{4\alpha\beta}\log\frac{\alpha-\beta}{\alpha+\beta} \\ I(\tilde{x}_k^l; \mathcal{Y}_k) &= \frac{\alpha+\beta}{2\alpha} - \log\frac{\alpha+\beta}{\alpha}. \end{aligned} \quad (83)$$

It can be shown that for any $\beta > 0$ given, and $\alpha \in [\beta, +\infty)$, we have the relations

$$\mathrm{Cov}(\tilde{x}_k^m) \le \mathrm{Cov}(\tilde{x}_k^k) \le \mathrm{Cov}(\tilde{x}_k^l) \quad (84)$$

with equality if and only if $\alpha = \beta$, and

$$\begin{aligned} &h(\tilde{x}_k^m) \le h(\tilde{x}_k^l) \le h(\tilde{x}_k^k) \\ &0 < I(\tilde{x}_k^m; \mathcal{Y}_k) \le I(\tilde{x}_k^l; \mathcal{Y}_k) \le I(\tilde{x}_k^k; \mathcal{Y}_k) \end{aligned} \quad (85)$$

with equality if and only if $\alpha = \beta$.

From (84) and (85), we see that there is no affine linear zero error/observation information filter. System (73) apparently satisfies the condition of Theorem 4.9, and the basic random variables are uniform rather than Gaussian. Also, we observe that in terms of all performance measures, the (true) minimum mean-square filter always has the best performance among these three filters. For the error covariance, the Kalman filter (the affine linear minimum mean-square filter) is superior to the affine linear minimum error entropy filter. But, for the error entropy (and error/observation information measures), the affine linear minimum error entropy filter is superior to the Kalman filter. Hence, we expect that there will be conflict or competition among these measures. One may ask whether or not there is a nonlinear filter f for the homogeneous system which will have a lower error entropy (or error/observation) information measure when compared to the (true) minimum mean-square filter, or what is the true minimum error/observation information filter? We are unable to answer these questions so far. Some simulation evidence suggests that for piecewise linear filters with one corner, no filter will perform better than the minimum mean-square filter $\hat{x}_k^m$.

Finally, from (72) and (20), we see that for any feedback control $g \in \mathcal{G}$, the optimal filter(s) for the inhomogeneous system (73), defined according to 1)–6) of Definition 2.1, is given by

$$\hat{x}_{k+1}^g = u_k^g + \hat{x}_k^d(y_{k+1}^g - u_k^g) \quad (86)$$

where $\hat{x}_k^d(\cdot)$ is the (piecewise) linear function defined in (77), (78), and (80) for $d = k, m, l$, respectively.

For a detailed discussion on systems with quantization and the application of the framework developed in this work to this class of systems, the readers are referred to [10].

VI. Concluding Remarks

A general framework for filtering and state estimation in stochastic systems from an information theoretic point of view is proposed in this paper. Linear stochastic systems are studied in some detail, and it is shown that under some reachability and observability conditions, the existence of an affine linear zero error/observation filter for the homogeneous system or a quasi-affine linear filter for the inhomogeneous system is equivalent to having all the basic random variables Gaussian. An example is given to illustrate some of the results for linear stochastic systems. Reference [10] illustrates the application of the proposed framework to quantized systems. We observe that, in general, there is conflict and competition between the different performance criteria that can be used to design an estimator. For linear stochastic systems, we showed that a feedback control strategy has no probing effect in the sense of the mean-square error, the error entropy, and the error/observation information measures. However, for a nonlinear system (e.g., the quantized system studied in [10]), the choice of the feedback control can have a significant effect on estimation in all of the senses considered in this paper.

For future research, there are many problems that need to be studied. The first and most important one is constructing a realization of the minimum error entropy filter, or a realization of a suboptimal filter for a general linear stochastic system, as a finite-dimensional dynamic system. We know that this realization should reduce to the Kalman filter for LG systems. Apparently, this will involve an entropy optimization problem. Current results available in the literature may not be sufficient to accomplish this task. One problem is that the orthogonality condition which results in the Kalman filter is lacking. The second interesting problem is to determine how far the results developed in Section IV can be extended. In particular, under what conditions does there exist a zero error/observation filter for the general class of stochastic systems? A third problem is to determine under what circumstances the (true) minimum mean-square filter is actually the minimum error entropy filter. We see that from the example given in Section V, the (true) minimum mean-square filter associated with a feedback control is actually the minimum error entropy and minimum error/observation information filter associated with the same control. Furthermore, the filtering error associated with the minimum mean-square filter has a density function $p(\tilde{x})$, which is a piecewise linear, even, nonincreasing function centered at $\tilde{x} = 0$ when $\tilde{x} \geq 0$. We wonder if this is the reason why the filters are equivalent.

REFERENCES

[1] R. E. Kalman, "A new approach to linear filtering and prediction problems," *Trans. ASME, J. Basic Eng.*, Series 82D, pp. 35–45, Mar. 1960; reprinted in [7].

[2] R. E. Kalman and R. S. Bucy, "New results in linear filtering and prediction theory," *Trans. ASME, J. Basic Eng.*, Series 83D, pp. 95–108, Mar. 1961; reprinted in [7].

[3] Y. C. Ho and R. C. K. Lee, "A Bayesian approach to problems in stochastic estimation and control," *IEEE Trans. Automat. Contr.*, vol. AC-9, pp. 333–339, Oct. 1964; reprinted in [7].

[4] T. Kailath, "An innovation approach to least-sqaure estimation—Part I: Linear filtering in additive white noise," *IEEE Trans. Automat. Contr.*, vol. AC-13, pp. 646–655, Dec. 1968; reprinted in [7].

[5] ——, "An innovation approach to least square estimation—Part II: Linear smoothing in additive white noise," *IEEE Trans. Automat. Contr.*, vol. AC-13, pp. 655–660, Dec. 1968; reprinted in [7].

[6] H. E. Rauch, F. Tung, and C. T. Striebel, "Maximum likelihood estimation of linear dynamic systems," *AIAA J.*, vol. 3, pp. 1445–1450, Aug. 1965; reprinted in [7].

[7] H. W. Sorenson, Ed., *Kalman Filtering: Theory and Application*. New York: IEEE Press, 1985.

[8] ——, "Least square estimation: From Gauss to Kalman," *IEEE Spectrum*, vol. 7, pp. 63–68, July 1970; reprinted in [7].

[9] B. E. Griffiths and K. A. Loparo, "Optimal control of jump-linear Gaussian systems," *Int. J. Contr.*, vol. 43, no. 4, pp. 792–819.

[10] X. Feng and K. A. Loparo, "Active probing for information in control systems with quantized state measurements: A minimum entropy approach," *IEEE Trans. Automat. Contr.*, vol. 41, pp. 216–238, Feb. 1997.

[11] H. L. Weidemann, "Entropy analysis of feedback control systems," in *Adv. Contr. Syst.*, C. Leonodes, Ed. New York: Academic, 1969.

[12] H. L. Weidemann and E. B. Stear, "Entropy analysis of estimating systems," *IEEE Trans. Inf. Theory*, vol. 16, no. 3, pp. 264–270, 1970.

[13] G. N. Saridis, "Entropy formulation of optimal and adaptive control," *IEEE Trans. Automat. Contr.*, vol. 33, no. 8, pp. 713–721, 1988.

[14] J. Zaborszky, "An information theory viewpoint for the general identification problem," *IEEE Trans. Automat. Contr.*, vol. 11, no. 1, pp. 130–131, 1966.

[15] Y. A Tsai, F. A. Casiello, and K. A. Loparo, "Discrete time entropy formulation of optimal and adaptive control problem," *IEEE Trans. Automat. Contr.*, vol. 37, no. 7, pp. 1083–1087, 1992.

[16] F. A. Casiello and K. A. Loparo, *Entropy Interpretation of Active and Passive Learning Controllers*, submitted for publication.

[17] P. Kalata and R. Priemer, "Linear prediction, filtering and smoothing: An information theoretic approach," *Inf. Sci.*, vol. 17, pp. 1–14, 1979.

[18] Y. Tomita, S. Omatu, and T. Soeda, "An application of the information theory to filtering problem," *Inf. Sci.*, vol. 11, pp. 13–27, 1976.

[19] ——, "Information theoretical optimal smoothing estimation," *Inf. Sci.*, vol. 22, pp. 201–215, 1980.

[20] R. S. Bucy, "Information and filtering," *Inf. Sci.*, vol. 18, pp. 179–187, 1979.

[21] E. T. Jaynes, "Information theory and statistical mechanics," *Phys. Rev.*, vol. 106, no. 4, pp. 620–630, 1957.

[22] P. R. Kumar and P. Varaiya, *Stochastic Systems*. Englewood Cliffs, NJ: Prentice Hall, 1986.

[23] A. A. Feldbaum, *Optimal Control Systems*. New York: Academic, 1965.

[24] G. N. Saridis, *Self-Organizing Control of Stochastic Systems*. New York: Marcel Dekker, 1977.

[25] D. G. Luenberger, *Optimization by Vector Space Methods*. New York: Wiley, 1969.

[26] R. G. Gallager, *Information Theory and Reliable Communication*. New York: Wiley, 1968.

[27] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.

[28] S. G. Ghurye and I. Olkin, "A characterization of the multivariable normal distribution," *Ann. Math. Statist.*, vol. 33, pp. 533–541, 1962.

[29] C. E. Shannon, "Extracting state information from a quantized output record," *Syst. Contr. Lett.*, vol. 13, pp. 365–372, 1989.

[30] D. F. Delchamps, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, pp. 379–423, 623–656, 1948.

[31] ——, "Controlling the flow of information in feedback systems with measurement quantization," in *Proc. 28th IEEE Conf. Decision Contr.*, Tampa, FL, Dec. 1989, pp. 2355–2360.

[32] R. M. Gray, *Entropy and Information Theory*. New York: Springer-Verlag, 1990.

Xiangbo Feng, for a photograph and biography, see p. 238 of the February 1997 issue of this TRANSACTIONS.

Kenneth A. Loparo (S'75–M'77–SM'89), for a photograph and biography, see p. 238 of the February 1997 issue of this TRANSACTIONS.

Yuguang Fang (S'92–M'94) received the B.S. and M.S. degrees in mathematics from Qufu Normal University, Shandong, The People's Republic of China, in 1984 and 1987, respectively, and the Ph.D. degree in systems, control and industrial engineering from Case Western Reserve University, Cleveland, OH, in 1994.

From 1987 to 1988, he held research and teaching positions in both the Mathematics Department and the Institute of Automation at Qufu Normal University. From 1989 to 1993, he was a Teaching and Research Assistant in the Department of Systems, Control and Industrial Engineering at Case Western Reserve University, where he became a Research Associate from January 1994 to May 1994. He was a Post-Doctoral Research Associate in the Department of Electrical and Computer Engineering at Boston University from 1994 to 1996. He is now completing his Ph.D. requirements in wireless and mobile communications networks. His research interests include wireless and mobile communications and networking, personal communications services (PCS), stochastic and adaptive systems, hybrid systems in integrated communications and controls, robust stability and control, nonlinear dynamical systems, and neural networks.