Simultaneous Input and State Smoothing for Linear Discrete-time Stochastic Systems with Unknown Inputs

Sze Zheng Yong[a] Minghui Zhu[b] Emilio Frazzoli[a]

Abstract— **This paper considers the problem of simultaneously estimating the states and unknown inputs of linear discrete-time systems in the presence of additive Gaussian noise based on observations from the entire time interval. A fixed-interval input and state smoothing algorithm is proposed for this problem and the input and state estimates are shown to be unbiased and to achieve minimum mean squared error and maximum likelihood. A numerical example is included to demonstrate the performance of the smoother.**

I. Introduction

State and input estimation of stochastic systems is important to a number of applications found across a wide range of disciplines. For example, in the state estimation problem of vehicles at an intersection, the input of the other vehicle is not available, and is not well modeled by a zero-mean Gaussian white noise process, thus, the standard Kalman filter and smoother cannot be applied. Other examples in which the estimates of both the state and input are desirable include real-time estimation of mean areal precipitation during a storm [1], fault detection and diagnosis [2] and input estimation in physiological systems [3].

Literature review. Filter algorithms estimate current values of the variables of interest in real-time, whereas smoothing algorithms post-process all measurements to also estimate past values. Kalman filter [4] is the most widely used linear filtering algorithm, while the most common fixed-interval smoothing algorithm (a.k.a. Kalman smoother) is probably the Rauch-Tung-Striebel (RTS) smoother [5], which has the advantage of being computationally cheaper than the forward-backward smoother [6]. However, these algorithms are not applicable to systems with unknown inputs that cannot be modeled as a zero mean Gaussian white noise.

Optimal filter algorithms for linear systems with unknown inputs can be classified into those that only estimate the system states [1], [7], [8] and those that simultaneously estimate the unknown inputs and states [9]–[14]. In fact, most state only estimators are shown to yield the same estimates as the simultaneous input and state estimators [10], [11], [14], which seems to make sense in hindsight.

On the other hand, to the best knowledge of the authors, the only available input and state smoothing algorithm is designed for nonlinear systems with unknown inputs [15]. However, this algorithm makes an implicit assumption that is equivalent to imposing a full rank condition on the system direct feedthrough matrix, which is somewhat restrictive (cf. [14]). Furthermore, neither an analytical solution nor any claims of optimality were presented. Thus, an optimal smoothing algorithm for linear stochastic systems with unknown inputs is still lacking at present.

Contributions. This paper proposes a smoothing algorithm for simultaneous estimation of input and states for linear discrete-time stochastic systems with unknown inputs. The proposed algorithm is an extension of the RTS smoother [5] to allow for unknown inputs and consists similarly of a forward and a backward pass. The forward pass is computed with the optimal filter from a previous work [14], and in this paper, we present an algorithm for the backward pass. Two commonly used inference techniques are used to show that the algorithm has many desirable properties. To be precise, the estimates are shown to be unbiased and to achieve minimum mean squared error and maximum likelihood.

II. Problem Formulation

In this paper, we consider the linear time-varying discrete-time stochastic system with unknown inputs

$$\begin{aligned} x_{k+1} &= A_k x_k + B_k u_k + G_k d_k + w_k \\ y_k &= C_k x_k + D_k u_k + H_k d_k + v_k \end{aligned} \tag{1}$$

where $x_k \in \mathbb{R}^n$ is the state vector at time k, $u_k \in \mathbb{R}^m$ is a known input vector, $d_k \in \mathbb{R}^p$ is an unknown input vector, and $y_k \in \mathbb{R}^l$ is the measurement vector. The process noise $w_k \in \mathbb{R}^n$ and the measurement noise $v_k \in \mathbb{R}^l$ are assumed to be mutually uncorrelated, zero-mean, Gaussian white random signals with known covariance matrices, $Q_k = \mathbb{E}[w_k w_k^\top] \succeq 0$ and $R_k = \mathbb{E}[v_k v_k^\top] \succ 0$, respectively. We assume that d_k is uncorrelated with $\{w_k\}$ and $\{v_k\}$ for all k, as well as with $\{d_j\}$ for all $j \neq k$, i.e., d_k is completely unknown and cannot be predicted from the knowledge of d_j for all $j \neq k$. In addition, x_k and d_k are assumed to be jointly normally distributed with non-informative priors, and as is observed in [14], with means and covariances given by their unbiased estimates and error covariance matrices. x_0 is also assumed to be independent of v_k and w_k for all k and is a Gaussian vector with a mean $\pi_0 = \mathbb{E}[x_0]$ and covariance matrix $\mathcal{P}_0^x$.

Without loss of generality, we assume throughout the paper that $n \geq l \geq 1$, $l \geq p \geq 0$ and $m \geq 0$, and that the current time variable is strictly nonnegative. We also assume that the matrices A_k, B_k, C_k, D_k, G_k and $H_k = \begin{bmatrix} U_{1,k} & U_{2,k} \end{bmatrix} \begin{bmatrix} \Sigma_k & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} V_{1,k}^\top \\ V_{2,k}^\top \end{bmatrix}$ are known, and that the sys-

[a] S.Z Yong and E. Frazzoli are with the Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, MA, USA (e-mail: szyong@mit.edu, frazzoli@mit.edu).

[b] M. Zhu is with the Department of Electrical Engineering, Pennsylvania State University, 201 Old Main, University Park, PA 16802, USA (e-mail: muz16@psu.edu).

tem is strongly detectable and $\mathrm{rank}(U_{2,k}^\top C_k G_{k-1} V_{2,k-1}) = p - \mathrm{rank}(H_{k-1})$ (such that there exists a stable minimum-variance unbiased filter [14, Theorem 9]). This linear system can also been viewed as a hidden Markov model (HMM) with an infinite number of states that is "parameterized" by the known and unknown inputs, as depicted in Fig. 1. Note that only $d_{1,N} = V_{1,N}^\top d_N$, which is the projection of d_N that is observable from y_N, is included at the end of the Markov chain (The reader is referred to [14] for its justification).

The problem we address is the simultaneous estimation of x_k and d_k from the observations given by $y_{0:N}$ and $u_{0:N}$, where we denote the set of consecutive signals $\{s_{t_0}, s_{t_0+1}, \dots, s_{t_f}\}$ as $s_{t_0:t_f}$. The estimation problem is commonly known as (i) *filtering* if $k = N$, (ii) *smoothing* if $k < N$ and (iii) *prediction* if $k > N$. While it may be possible to consider all three cases above when there are no unknown inputs, it is clear that without any knowledge of the future d_j for $k+1 \le j \le N$, the prediction problem is not possible (except when $l = p = \mathrm{rank}(H_k)$ for all k). A previous work is dedicated to the optimal filtering problem [14], but the optimal smoothing problem has not been solved; thus, the objective of this paper is to *develop an optimal fixed-interval smoothing algorithm for linear discrete-time stochastic systems with unknown input for which the state and input estimates are unbiased and achieve minimum mean squared error and maximum likelihood.*

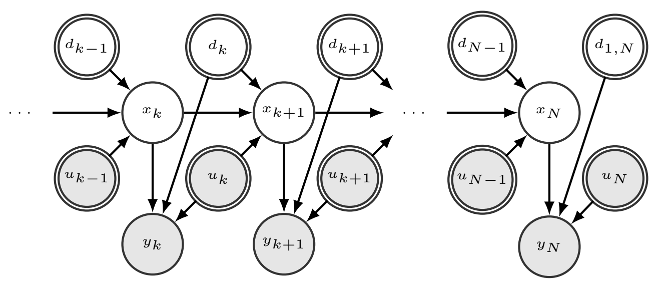

Fig. 1: A hidden Markov model (HMM) "parameterized" by u_k and d_k, with state transition A_k and emission probability C_k. Shaded nodes are observed, while unshaded nodes are hidden and to be estimated.

III. Optimal Smoothing

We cast the problem of optimal smoothing (more specifically, fixed-interval smoothing) as a problem of directly or indirectly computing the joint distribution of the state x_k and (completely) unknown input d_k, $p(x_k, d_k | y_{0:N}, u_{0:N})$, at time step k using measurements y_k and known inputs u_k up to time N, where $N > k$. The result is a direct extension of the Rauch-Tung-Striebel (RTS) smoother [5] to systems with unknown inputs. Note the contrast to the joint distribution of the optimal filtering problem given by $p(x_k, d_k | y_{0:k}, u_{0:k})$, which itself is an extension of the Kalman filter [4]. As with the RTS smoother, our proposed optimal smoothing algorithm, ULISS (Updated Linear Input & State Smoother), consists of two passes—*forward* and *backward* passes:

(i) *Forward pass:* This pass involves the filtering problem which can be solved with the optimal filtering algorithm,

Algorithm 1 ULISS Algorithm

Initialize: $\hat{x}_{0|0} = \mathbb{E}[x_0]$; $P^x_{0|0} = \mathcal{P}^x_0$; $\hat{A}_0 = A_0 - G_{1,0}\Sigma_0^{-1}C_{1,0}$; $\hat{Q}_0 = G_{1,0}\Sigma_0^{-1}R_{1,0}\Sigma_0^{-1}G_{1,0}^\top + Q_0$; $\hat{d}_{1,0} = \Sigma_0^{-1}(z_{1,0} - C_{1,0}\hat{x}_{0|0} - D_{1,0}u_0)$; $P^d_{1,0} = \Sigma_0^{-1}(C_{1,0}P^x_{0|0}C_{1,0}^\top + R_{1,0})\Sigma_0^{-1}$;

▷ **Forward Pass**

for $k = 1$ to N **do**

▷ Estimation of $d_{2,k-1}$ and d_{k-1}

$\hat{A}_{k-1} = A_{k-1} - G_{1,k-1}M_{1,k-1}C_{1,k-1}$;

$\hat{Q}_{k-1} = G_{1,k-1}M_{1,k-1}R_{1,k-1}M_{1,k-1}^\top G_{1,k-1}^\top + Q_{k-1}$;

$\tilde{P}_k = \hat{A}_{k-1}P^x_{k-1|k-1}\hat{A}_{k-1}^\top + \hat{Q}_{k-1}$;

$\tilde{R}_{2,k} = C_{2,k}\tilde{P}_k C_{2,k}^\top + R_{2,k}$;

$P^d_{2,k-1} = (G_{2,k-1}^\top C_{2,k}^\top \tilde{R}_{2,k}^{-1} C_{2,k} G_{2,k-1})^{-1}$;

$M_{2,k} = P^d_{2,k-1} G_{2,k-1}^\top C_{2,k}^\top \tilde{R}_{2,k}^{-1}$;

$\hat{x}_{k|k-1} = A_{k-1}\hat{x}_{k-1|k-1} + B_{k-1}u_{k-1} + G_{1,k-1}\hat{d}_{1,k-1}$;

$\hat{d}_{2,k-1} = M_{2,k}(z_{2,k} - C_{2,k}\hat{x}_{k|k-1} - D_{2,k}u_k)$;

$\hat{d}_{k-1} = V_{1,k-1}\hat{d}_{1,k-1} + V_{2,k-1}\hat{d}_{2,k-1}$;

$P^d_{12,k-1} = M_{1,k-1}C_{1,k-1}P^x_{k-1|k-1}A_{k-1}^\top C_{2,k}^\top M_{2,k}^\top - P^d_{1,k-1}G_{1,k-1}^\top C_{2,k}^\top M_{2,k}^\top$;

$P^d_{k-1} = V_{k-1}\begin{bmatrix} P^d_{1,k-1} & P^d_{12,k-1} \\ P^{d\top}_{12,k-1} & P^d_{2,k-1}\end{bmatrix} V_{k-1}^\top$;

$P^{xd}_{1,k-1} = -P^x_{k-1|k-1}C_{1,k-1}^\top M_{1,k-1}^\top$;

$P^{xd}_{2,k-1} = -P^x_{k-1|k-1}A_{k-1}^\top C_{2,k}^\top M_{2,k}^\top - P^{xd}_{1,k-1}G_{1.k-1}^\top C_{2,k}^\top M_{2,k}^\top$;

$P^{xd}_{k-1} = P^{xd}_{1,k-1}V_{1,k-1}^\top + P^{xd}_{2,k-1}V_{2,k-1}^\top$

▷ Time update

$\hat{x}^\star_{k|k} = \hat{x}_{k|k-1} + G_{2,k-1}\hat{d}_{2,k-1}$;

$P^{\star x}_{k|k} = G_{2,k-1}M_{2,k}R_{2,k}M_{2,k}^\top G_{2,k}^\top + (I - G_{2,k-1}M_{2,k}C_{2,k})\tilde{P}_k(I - G_{2,k-1}M_{2,k}C_{2,k})^\top$;

$\tilde{R}^\star_{2,k} = C_{2,k}P^{\star x}_{k|k}C_{2,k}^\top + R_{2,k} - C_{2,k}G_{2,k-1}M_{2,k}R_{2,k} - R_{2,k}M_{2,k}^\top G_{2,k-1}^\top C_{2,k}$;

▷ Measurement update

$\tilde{L}_k = (P^{\star x}_{k|k}C_{2,k}^\top - G_{2,k-1}M_{2,k}R_{2,k})\tilde{R}^{\star\dagger}_{2,k}$;

$\hat{x}_{k|k} = \hat{x}^\star_{k|k} + \tilde{L}_k(z_{2,k} - C_{2,k}\hat{x}^\star_{k|k} - D_{2,k}u_k)$;

$P^x_{k|k} = (I - \tilde{L}_k C_{2,k})G_{2,k-1}M_{2,k}R_{2,k}\tilde{L}_k^\top + \tilde{L}_k R_{2,k}M_{2,k}^\top G_{2,k-1}^\top (I - \tilde{L}_k C_{2,k})^\top + (I - \tilde{L}_k C_{2,k})P^{\star x}_{k|k}(I - \tilde{L}_k C_{2,k})^\top + \tilde{L}_k R_{2,k}\tilde{L}_k^\top$;

▷ Estimation of $d_{1,k}$

$\tilde{R}_{1,k} = C_{1,k}P^x_{k|k}C_{1,k}^\top + R_{1,k}$;

$M_{1,k} = \Sigma_k^{-1}$;

$P^d_{1,k} = M_{1,k}\tilde{R}_{1,k}M_{1,k}$;

$\hat{d}_{1,k} = M_{1,k}(z_{1,k} - C_{1,k}\hat{x}_{k|k} - D_{1,k}u_k)$;

end for

▷ **Backward Pass**

for $k = N-1$ to 1 **do**

$J_k = \begin{bmatrix} P^x_{k|k}A_k^\top + P^{xd}_k G_k^\top \\ P^{dx}_k A_k^\top + P^d_k G_k^\top \end{bmatrix}(P^{\star x}_{k+1|k+1})^{-1}$;

$\begin{bmatrix} \hat{x}_{k|N} \\ \hat{d}_{k|N} \end{bmatrix} = \begin{bmatrix} \hat{x}_{k|k} \\ \hat{d}_k \end{bmatrix} + J_k(\hat{x}_{k+1|N} - \hat{x}^\star_{k+1|k+1})$;

$\begin{bmatrix} P^x_{k|N} & P^{xd}_{k|N} \\ P^{dx}_{k|N} & P^d_{k|N} \end{bmatrix} = \begin{bmatrix} P^x_{k|k} & P^{xd}_k \\ P^{dx}_k & P^d_k \end{bmatrix} + J_k(P^x_{k+1|N} - P^{\star x}_{k+1|k+1})J_k^\top$;

end for

ULISE, in [14], from which the current name of the proposed smoother is derived. From a single pass of ULISE, we can obtain the forward pass estimates $\hat{x}_{k|k}$ and $\hat{d}_k$, as well as covariances $P^x_{k|k}$, $P^{dx}_k = (P^{xd}_k)^\top$ and P^d_k for all $k = \{1, 2, \dots, N-1\}$, whereas for $k = N$, we only have $\hat{x}_{N|N}$ and $P^x_{N|N}$, which we shall see are the only two quantities needed to start the backward recursion.

(ii) *Backward pass:* The backward pass[1] essentially uses the output measurements of the future to further improve the filtered state and input estimates. This can be seen in the backward pass algorithm presented below, where the information obtained from the difference between the smoothed and predicted *future* state is used to improve the state and input estimates from the forward pass. The backward pass can be computed with the following:

$$
\begin{aligned}
J_k := \begin{bmatrix} J_{1,k} \\ J_{2,k} \end{bmatrix} &= \begin{bmatrix} P^x_{k|k} A_k^\top + P_k^{xd} G_k^\top \\ P_k^{dx} A_k^\top + P_k^d G_k^\top \end{bmatrix} (P^{\star\, x}_{k+1|k+1})^{-1} \\
&:= \begin{bmatrix} J^-_{1,k} \\ J^-_{2,k} \end{bmatrix} (P^{\star\, x}_{k+1|k+1})^{-1} \\
\begin{bmatrix} \hat{x}_{k|N} \\ \hat{d}_{k|N} \end{bmatrix} &= \begin{bmatrix} \hat{x}_{k|k} \\ \hat{d}_k \end{bmatrix} + J_k(\hat{x}_{k+1|N} - \hat{x}^\star_{k+1|k+1}) \\
\begin{bmatrix} P^x_{k|N} & P^{xd}_{k|N} \\ P^{dx}_{k|N} & P^d_{k|N} \end{bmatrix} &= J_k(P^x_{k+1|N} - P^{\star\, x}_{k+1|k+1}) J_k^\top + \begin{bmatrix} P^x_{k|k} & P^{xd}_k \\ P^{dx}_k & P^d_k \end{bmatrix},
\end{aligned}
\tag{2}
$$

where $\hat{x}_{k|k}$, $\hat{x}^\star_{k+1|k+1}$, $\hat{d}_k$, $P^x_{k|k}$, $P^{\star\, x}_{k|k}$, $P^{dx}_k = (P^{xd}_k)^\top$ and P^d_k are computed by the forward pass with the ULISE algorithm [14], while $\hat{x}_{k|N}$ and $\hat{d}_{k|N}$ are the unbiased smoothed state and input estimates, and $P^x_{k|N}$, $P^{dx}_{k|N} = (P^{xd}_{k|N})^\top$ and $P^d_{k|N}$ are covariances of the smoothed estimates. The backward in time recursion is started from the time step $N-1$ with $\hat{x}_{N|N}$ and $P^x_{N|N}$ that are computed in the forward pass.

In the following theorem, we present some properties of the smoothed input and state estimates:

Theorem 1. *The discrete-time fixed interval smoother consisting of the forward pass [14] and the backward pass given in* (2) *minimizes the mean squared error and maximizes the posterior likelihood of x_k and d_k for all $k < N$ given all observations up to time N. Thus, the smoothed estimates are unbiased and achieve minimum mean squared error (i.e., $\begin{bmatrix} \hat{x}_{k|N} \\ \hat{d}_{k|N} \end{bmatrix} = \mathbb{E}\left[\begin{bmatrix} x_k \\ d_k \end{bmatrix} | y_{0:N}, u_{0:N}\right]$) and also maximum likelihood (i.e., $\begin{bmatrix} \hat{x}_{k|N} \\ \hat{d}_{k|N} \end{bmatrix} = \arg\max_{x_k, d_k} \; p\left(\begin{bmatrix} x_k \\ d_k \end{bmatrix} | y_{0:N}, u_{0:N}\right)$).*

Proof. The recursive backward pass equations in (2) can be independently derived using two methods—minimum mean squared error estimation and by maximum likelihood estimation. Then, we will show that minimum mean squared error estimates are unbiased, from which we conclude that the stated properties hold. Both derivations will be presented in detail in Section IV. ∎

Remark 1. *From* (2)*, we observe that if $\hat{x}_{k+1|N} = \hat{x}^\star_{k+1|k+1}$, then the smoothed estimates cannot be further improved from the filtered estimate. It can be shown by induction that this is indeed the case when $l = p = \mathrm{rank}(H_k)$ for all k. In fact, this is the (only) case when prediction is possible in real-time and the filtered estimate is exactly the predicted estimate; and with* (2)*, we now establish that the filtered estimate is also the smoothed estimate in this case.*

Remark 2. *A stable smoother exists whenever the corresponding filter (forward pass) is stable, for which necessary and sufficient conditions are given in [14].*

The lag-one covariance smoother[2] follows directly from the derivation of the optimal smoother ((5) in Section IV).

Corollary 1 (Lag-One Covariance Smoother)**.** *The lag-one covariance of the optimal smoother is given by*

$$
\begin{aligned}
P^x_{k+1,k|N} &:= \mathbb{E}[(\hat{x}_{k+1|N} - x_{k+1})(\hat{x}_{k|N} - x_k)^\top] \\
&= P^x_{k+1|N} J_{1,k}^\top \\
P^{xd}_{k+1,k|N} &:= \mathbb{E}[(\hat{x}_{k+1|N} - x_{k+1})(\hat{d}_{k|N} - d_k)^\top] \\
&= P^{x,s}_{k+1|k+1} J_{2,k}^\top,
\end{aligned}
\tag{3}
$$

where $J^-_{1,k}$ and $J^-_{2,k}$ are as defined in (2)*.*

Remark 3. *In the special case when there are no unknown inputs (as is the case with the RTS smoother), the lag-one covariance smoother is given by*

$$P^x_{k+1,k|N} = P^x_{k+1|N}(P^x_{k+1|k})^{-1} A_k P^x_{k|k},$$

which is a simpler expression when compared to the formulation in [16, pp. 334–335] which involves recursions.

IV. Fixed-Interval Smoother Analysis

In this section, we derive the backward pass equations (2) using two different methods—via minimum mean squared error estimation and maximum likelihood estimation. In the first proof via minimum mean squared error estimation (a.k.a. Bayesian least squares), the optimal estimate is given by the expected value of x_k and d_k given all observations, i.e., $\begin{bmatrix} \hat{x}_k^{MMSE} \\ \hat{d}_k^{MMSE} \end{bmatrix} = \mathbb{E}\left[\begin{bmatrix} x_k \\ d_k \end{bmatrix} | y_{0:N}, u_{0:N}\right]$ (see, e.g. [17, Section 4.6]). To compute this, we seek the likelihood of the estimates of x_k and d_k and evaluate the expected value of the likelihood function. In the second proof via maximum likelihood estimation, as with the derivation of the RTS smoother, we define a loss function for the problem and find a smoothing solution that minimizes the posterior loss, which in turn guarantees that the estimates attain maximum likelihood $\begin{bmatrix} \hat{x}_k^{MLE} \\ \hat{d}_k^{MLE} \end{bmatrix} = \arg\max_{x_k, d_k} \; p\left(\begin{bmatrix} x_k \\ d_k \end{bmatrix} | y_{0:N}, u_{0:N}\right)$. By inspection, it can be seen that both derivation methods produce the same estimates as given in (2). Moreover, by law of iterated expectations, the estimates are unbiased since $\mathbb{E}\left[\begin{bmatrix} \hat{x}_k^{MMSE} \\ \hat{d}_k^{MMSE} \end{bmatrix}\right] = \mathbb{E}\left[\mathbb{E}\left[\begin{bmatrix} x_k \\ d_k \end{bmatrix} | y_{0:N}, u_{0:N}\right]\right] = \begin{bmatrix} x_k \\ d_k \end{bmatrix}$. Thus, Theorem 1 holds.

Proof 1. *(Minimum Mean-Squared Error Estimation)* This proof relies extensively on the properties of Gaussian distributions given in Appendix A (Lemmas 1 and 2).

From the forward pass algorithm, ULISE, we first find the joint probability of x_{k+1}, x_k and d_k given $y_{0:k'}$ and $u_{0:k'}$, where k' is an intermediate step between k and $k+1$ such

[1]The backward pass is to be differentiated from the backward filter, as the former refers the use of future observations for updating the filtered estimate from the forward pass, i.e., finding the joint distribution, $p(x_k, d_k | y_{0:N}, u_{0:N})$, whereas the latter refers to the application of forward filter backward in time.

[2]This is a useful result that, e.g., enables the use of the EM-algorithm for parameter estimation (similar to [16]). This is part of an ongoing work.

that d_k can be entirely estimated, but before the estimate of x_{k+1} can be updated (right after the filtering step in Line 14 of Algorithm 1). Note that this corresponds to the propagation/prediction step, which is before the update step in ULISE, as described in [14, Section 4], in line with the approach of the RTS smoother [5]. This joint probability is computed in the forward pass (filtering) and is given by [14, Sections 4 and 5]:

$$\begin{aligned}&p(x_k, d_k, x_{k+1}|y_{0:k'}, u_{0:k'})\\ &= \mathcal{N}\left(\begin{bmatrix} x_k \\ d_k \\ x_{k+1}\end{bmatrix}; \begin{bmatrix} \hat{x}_{k|k} \\ \hat{d}_k \\ \hat{x}^\star_{k+1|k+1}\end{bmatrix}, \begin{bmatrix} P^x_{k|k} & P^{xd}_k & J^-_{1,k} \\ P^{dx}_k & P^d_k & J^-_{2,k} \\ J^{-\top}_{1,k} & J^{-\top}_{2,k} & P^{\star x}_{k+1|k+1}\end{bmatrix}\right),\end{aligned} \tag{4}$$

where $J^-_{1,k}$ and $J^-_{2,k}$ are as defined in (2), and $\hat{x}_{k|k}$, $\hat{x}^\star_{k+1|k+1}$, $\hat{d}_k$, $P^x_{k|k}$, $P^{\star x}_{k|k}$, $P^{dx}_k = (P^{xd}_k)^\top$ and P^d_k are computed by the forward pass. From the Markov property of the state transition in (1) and by Lemma 2, we obtain

$$\begin{aligned} p(x_k, d_k|x_{k+1}, y_{0:k'}, u_{0:k'}) &= p(x_k, d_k|x_{k+1}, y_{0:k}, u_{0:k}) \\ &= \mathcal{N}\left([x_k^\top \; d_k^\top]^\top; \mu', P'\right),\end{aligned}$$

where $\mu' = \begin{bmatrix} \hat{x}_{k|k} \\ \hat{d}_k\end{bmatrix} + J_k(x_{k+1} - \hat{x}^\star_{k+1|k+1})$ and $P' = \begin{bmatrix} P^x_{k|k} & P^{xd}_k \\ P^{dx}_k & P^d_k\end{bmatrix} - J_k P^{\star x}_{k+1|k+1} J_k^\top$ with J_k defined in (2).

Similarly, we obtain from the Markov property of the state transition in (1) the following marginal distribution

$$\begin{aligned} p(x_k, d_k|x_{k+1}, y_{0:N}, u_{0:N}) &= p(x_k, d_k|x_{k+1}, y_{0:k}, u_{0:k}) \\ &= \mathcal{N}\left([x_k^\top \; d_k^\top]^\top; \mu', P'\right).\end{aligned}$$

Next, assuming that we are given $p(x_{k+1}|y_{0:N}, u_{0:N}) = \mathcal{N}(x_{k+1}; \hat{x}_{k+1|N}, P^x_{k+1|N})$, then by Lemma 1, we can find the joint distribution of x_k, d_k and x_{k+1} given all the data

$$\begin{aligned}&p(x_k, d_k, x_{k+1}|y_{0:N}, u_{0:N}) \\ &= p(x_k, d_k|x_{k+1}, y_{0:N}, u_{0:N}) p(x_{k+1}|y_{0:N}, u_{0:N}) \\ &= \mathcal{N}\left([x_{k+1}^\top \; x_k^\top \; d_k^\top]^\top; \mu'', P''\right),\end{aligned}$$

where by the property of joint density of Gaussian distributions given in Lemma 1, we find μ'' and P'' as follows

$$\begin{aligned}\mu'' &= \begin{bmatrix} \hat{x}_{k+1|N} \\ \begin{bmatrix} \hat{x}_{k|k} \\ \hat{d}_k\end{bmatrix} + J_k(\hat{x}_{k+1|N} - \hat{x}^\star_{k+1|k+1})\end{bmatrix} \\ P'' &= \begin{bmatrix} P^x_{k+1|N} & P^x_{k+1|N} J_k^\top \\ J_k P^x_{k+1|N} & J_k P^x_{k+1|N} J_k^\top + P'\end{bmatrix}.\end{aligned} \tag{5}$$

Finally, we find the joint marginal posterior distribution of x_k and d_k using the property of marginal density of partitioned Gaussians (with $x = x_{k+1}$, $y = [x_k^\top \; d_k^\top]^\top$ in Lemma 2) as

$$p(x_k, d_k|y_{0:N}, u_{0:N}) = \mathcal{N}\left(\begin{bmatrix} x_k \\ d_k\end{bmatrix}; \begin{bmatrix} \hat{x}_{k|N} \\ \hat{d}_{k|N}\end{bmatrix}, \begin{bmatrix} P^x_{k|N} & P^{xd}_{k|N} \\ P^{dx}_{k|N} & P^d_{k|N}\end{bmatrix}\right)$$

with $\begin{bmatrix} \hat{x}_{k|N} \\ \hat{d}_{k|N}\end{bmatrix}$ and $\begin{bmatrix} P^x_{k|N} & P^{xd}_{k|N} \\ P^{dx}_{k|N} & P^d_{k|N}\end{bmatrix}$ given by (2). Then, we perform the backward recursion starting from the last time step $N-1$, where $\hat{x}_{N|N}$, $\hat{x}^\star_{N|N}$ $P^x_{N|N}$ and $P^{\star x}_{N|N}$. This concludes the proof. ∎

Proof 2. *(Maximum Likelihood Estimation)* This proof is based on the maximization of the log-likelihood for all $k < N$ (with non-informative priors for x_k, x_{k+1}, d_k):

$$\begin{aligned}&\mathcal{L}(x_k, d_k, x_{k+1}, y_{0:N}, u_{0:N}) = \log P(x_k, d_k, x_{k+1}|y_{0:N}, u_{0:N}) \\ &= \log P(x_k, d_k, x_{k+1}, y_{k'+1:N}|y_{0:k'}, u_{0:N}) \\ &\quad - \log P(y_{k'+1:N}|y_{0:k'}, u_{0:N}).\end{aligned} \tag{6}$$

Then, using conditional probabilities, we observe that

$$\begin{aligned}&P(x_k, d_k, x_{k+1}, y_{k'+1:N}|y_{0:k'}, u_{0:N}) \\ &= P(y_{k'+1:N}|x_k, d_k, x_{k+1}, y_{0:k'}, u_{0:N}) \\ &\quad P(x_k, d_k, x_{k+1}|y_{0:k'}, u_{0:N}) \\ &= P(y_{k'+1:N}|x_{k+1}, u_{k+1:N}) P(x_k, d_k, x_{k+1}|y_{0:k'}, u_{0:k'}),\end{aligned} \tag{7}$$

where the final equality is the result of the Markov property of the system (cf. Figure 1) and $P(x_k, d_k, x_{k+1}|y_{0:k'}, u_{0:k'})$ can be found using (4), derived in [14, Sections 4 and 5].

Substituting (4) and (7) into (6), we have

$$\begin{aligned}&\mathcal{L}(x_k, d_k, x_{k+1}, y_{0:N}, u_{0:N}) \\ &= -\left(\begin{bmatrix} x_k \\ d_k \\ x_{k+1}\end{bmatrix} - \begin{bmatrix} \hat{x}_{k|k} \\ \hat{d}_k \\ \hat{x}^\star_{k+1|k+1}\end{bmatrix}\right)^\top \\ &\begin{bmatrix} P^x_{k|k} & P^{xd}_k & J^-_{1,k} \\ P^{dx}_k & P^d_k & J^-_{2,k} \\ J^{-\top}_{1,k} & J^{-\top}_{2,k} & P^{\star x}_{k+1|k+1}\end{bmatrix}^{-1} \left(\begin{bmatrix} x_k \\ d_k \\ x_{k+1}\end{bmatrix} - \begin{bmatrix} \hat{x}_{k|k} \\ \hat{d}_k \\ \hat{x}^\star_{k+1|k+1}\end{bmatrix}\right) \\ &+ \text{terms which do not involve } x_k \text{ and } d_k.\end{aligned} \tag{8}$$

Next, our goal is to obtain a backward recursion to find the maximum likelihood estimates of $\hat{x}_{k|N}$ and $\hat{d}_{k|N}$ given the estimate $\hat{x}_{k+1|N}$. It follows from (8) that $\hat{x}_{k|N}$ and $\hat{d}_{k|N}$ are the solution that minimizes the following *loss function*:

$$\begin{aligned}\mathcal{J} &= [(x_k - \hat{x}_{k|k})^\top \;\; (d_k - \hat{d}_k)^\top \;\; (\hat{x}_{k+1|N} - \hat{x}^\star_{k+1|k+1})^\top] \\ &\begin{bmatrix} P^x_{k|k} & P^{xd}_k & J^-_{1,k} \\ P^{dx}_k & P^d_k & J^-_{2,k} \\ J^{-\top}_{1,k} & J^{-\top}_{2,k} & P^{\star x}_{k+1|k+1}\end{bmatrix}^{-1} \begin{bmatrix} x_k - \hat{x}_{k|k} \\ d_k - \hat{d}_k \\ \hat{x}_{k+1|N} - \hat{x}^\star_{k+1|k+1}\end{bmatrix} \\ &= \begin{bmatrix} \begin{bmatrix} x_k - \hat{x}_{k|k} \\ d_k - \hat{d}_k\end{bmatrix} \\ \hat{x}_{k+1|N} - \hat{x}^\star_{k+1|k+1}\end{bmatrix}^\top \Lambda \begin{bmatrix} \begin{bmatrix} x_k - \hat{x}_{k|k} \\ d_k - \hat{d}_k\end{bmatrix} \\ \hat{x}_{k+1|N} - \hat{x}^\star_{k+1|k+1}\end{bmatrix} \\ &= \begin{bmatrix} x_k - \hat{x}_{k|k} \\ d_k - \hat{d}_k\end{bmatrix}^\top \Lambda_{11}^{-1} \begin{bmatrix} x_k - \hat{x}_{k|k} \\ d_k - \hat{d}_k\end{bmatrix} \\ &\quad + (\hat{x}_{k+1|N} - \hat{x}^\star_{k+1|k+1})^\top \Lambda_{22}^{-1} (\hat{x}_{k+1|N} - \hat{x}^\star_{k+1|k+1}) \\ &\quad + \begin{bmatrix} x_k - \hat{x}_{k|k} \\ d_k - \hat{d}_k\end{bmatrix}^\top \Lambda_{12}^{-1} (\hat{x}_{k+1|N} - \hat{x}^\star_{k+1|k+1}) \\ &\quad + (\hat{x}_{k+1|N} - \hat{x}^\star_{k+1|k+1})^\top (\Lambda_{12}^\top)^{-1} \begin{bmatrix} x_k - \hat{x}_{k|k} \\ d_k - \hat{d}_k\end{bmatrix},\end{aligned} \tag{9}$$

where $\Lambda := \begin{bmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{12}^\top & \Lambda_{22}\end{bmatrix}$ can be obtained using the block-wise matrix inversion formula given in Appendix B:

$$\begin{aligned}\Lambda_{11} &= \left(\begin{bmatrix} P^x_{k|k} & P^{xd}_k \\ P^{dx}_k & P^d_k\end{bmatrix} - \begin{bmatrix} J^-_{1,k} \\ J^-_{2,k}\end{bmatrix} P^{\star x}_{k+1|k+1} \begin{bmatrix} J^-_{1,k} \\ J^-_{2,k}\end{bmatrix}^\top\right)^{-1} \\ \Lambda_{12} &= -\Lambda_{11} \begin{bmatrix} J^-_{1,k} \\ J^-_{2,k}\end{bmatrix} P^{\star x}_{k+1|k+1} \\ \Lambda_{22} &= \left(P^{\star x}_{k+1|k+1} - \begin{bmatrix} J^-_{1,k} \\ J^-_{2,k}\end{bmatrix}^\top \begin{bmatrix} P^x_{k|k} & P^{xd}_k \\ P^{dx}_k & P^d_k\end{bmatrix} \begin{bmatrix} J^-_{1,k} \\ J^-_{2,k}\end{bmatrix}\right)^{-1}.\end{aligned}$$

Setting the gradient of $\mathcal{J}$ in (9) to zero, we obtain the estimates of $\hat{x}_{k|N}$ and $\hat{d}_{k|N}$ as is given in (2). Next, subtracting $\begin{bmatrix} x_k \\ d_k \end{bmatrix}$ from both sides of (2) and rearranging, we have

$$\begin{bmatrix} \tilde{x}_{k|N} \\ \tilde{d}_{k|N} \end{bmatrix} + J_k \hat{x}_{k+1|N} = \begin{bmatrix} \tilde{x}_{k|k} \\ \tilde{d}_{k|k} \end{bmatrix} + J_k \hat{x}^{\star}_{k+1|k+1} \tag{10}$$

Using the following facts:

$$\begin{aligned} \mathbb{E}[\hat{x}_{k+1|N}\hat{x}^{\top}_{k+1|N}] &= \mathbb{E}[x_{k+1}x^{\top}_{k+1}] - P^{x}_{k+1|N} \\ \mathbb{E}[\hat{x}^{\star}_{k+1|k+1}\hat{x}^{\star\top}_{k+1|k+1}] &= \mathbb{E}[x_{k+1}x^{\top}_{k+1}] - P^{\star x}_{k+1|k+1} \\ \mathbb{E}\left[\begin{bmatrix} \tilde{x}_{k|N} \\ \tilde{d}_{k|N} \end{bmatrix} \hat{x}^{\top}_{k+1|N}\right] &= \mathbb{E}\left[\begin{bmatrix} \tilde{x}_{k|k} \\ \tilde{d}_{k|k} \end{bmatrix} \hat{x}^{\star\top}_{k+1|k+1}\right] = 0, \end{aligned}$$

where the last fact is a result of the orthogonality principle of minimum mean square error estimation (see, e.g., [18, Section 4.2]), we obtain the recursive equation for the error covariance of the smoothed estimates given in (2); hence, this completes the proof. ∎

V. Illustrative Example

In this example, we consider the fault identification and state estimation problem when the system dynamics is afflicted by disturbances or faults, d_k, that can either affect the system dynamics through the input matrix G_k or the outputs through the direct feedthrough matrix H_k. Thus, the objective is to estimate the states of the system and to identify the faults that the system is experiencing for continued operation or self-repair. Specifically, the linear discrete-time problems we consider are based on the system given in [14]:

$$A = \begin{bmatrix} 0.5 & 2 & 0 & 0 & 0 \\ 0 & 0.2 & 1 & 0 & 1 \\ 0 & 0 & 0.3 & 0 & 1 \\ 0 & 0 & 0 & 0.7 & 1 \\ 0 & 0 & 0 & 0 & 0.1 \end{bmatrix}; \quad \begin{aligned} B &= 0_{5\times 1}; \\ C &= I_5; \\ D &= 0_{5\times 1}; \end{aligned}$$

$$G = \begin{bmatrix} 1 & 0 & -0.3 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}; \; Q = 10^{-4} \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0.5 & 0 & 0 \\ 0 & 0.5 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix};$$

$$R = 10^{-2} \begin{bmatrix} 1 & 0 & 0 & 0.5 & 0 \\ 0 & 1 & 0 & 0 & 0.3 \\ 0 & 0 & 1 & 0 & 0 \\ 0.5 & 0 & 0 & 1 & 0 \\ 0 & 0.3 & 0 & 0 & 1 \end{bmatrix};$$

with six different H matrices

$$H^1 = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}; H^2 = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix}; H^3 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix};$$

$$H^4 = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}; H^5 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}; H^6 = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix};$$

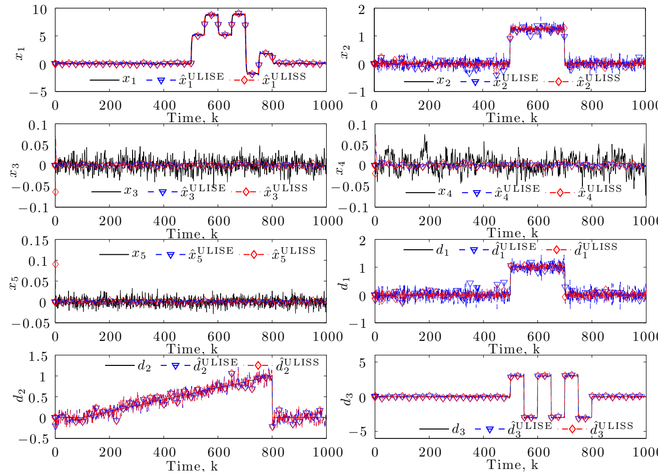

Fig. 2: Actual states x_1, x_2, x_3, x_4, x_5 and their estimates, as well as unknown inputs d_1, d_2 and d_3 and their estimates.

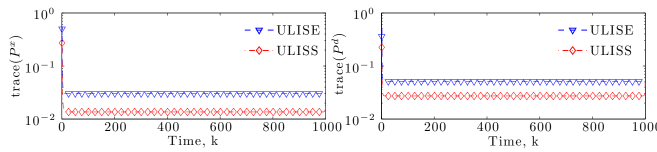

Fig. 3: Trace of estimate error covariance of states, trace(P^x), and unknown inputs, trace(P^d).

to illustrate the effect of parameter changes on the performance of the backward and forward passes of ULISS. With the above H matrices, the invariant zeros of the systems are respectively $\{0.3, 0.8\}$, $\{0.1, 0.3, 0.5, 0.7, 0.8\}$, $\emptyset$, $\{0.3, -0.8\}$, $\emptyset$ and $\{0.1, 0.7, 0.3, -0.8, 0.35\}$. Thus, all six systems are verified to be strongly detectable.

The unknown inputs used in this example are

$$d_{k,1} = \begin{cases} 1, & 500 \le k \le 700 \\ 0, & \text{otherwise} \end{cases}$$

$$d_{k,2} = \begin{cases} \frac{1}{700}(k-100), & 100 \le k \le 800 \\ 0, & \text{otherwise} \end{cases}$$

$$d_{k,3} = \begin{cases} 3, & 500 \le k \le 549, 600 \le k \le 649, 700 \le k \le 749 \\ -3, & 550 \le k \le 599, 650 \le k \le 699, 750 \le k \le 799 \\ 0, & \text{otherwise.} \end{cases}$$

The simulations were implemented in MATLAB on a 2.2 GHz Intel Core i7 CPU. Figure 2 shows a comparison of the input and state estimates of the forward and backward passes, i.e., of the filtered and smoothed estimates for the last system with H^6. As expected, the smoothed estimates are better than the filtered estimates, given that the smoothing algorithm is designed to improve on the filter estimates by considering the observations of the entire time interval. This improvement is more apparent from Figure 3, when the traces of the estimate error covariance of the forward and backward passes are compared. Furthermore, the same trend is seen in the results of all six systems in Table I, regardless of the rank of the direct feedthrough matrix, H_k. Note that even in the fourth case with H^4, there is an improvement of the smoothed estimates over the filtered ones, although it is not apparent due to the process of rounding to four decimal places. However, it can also be observed that the amount of improvement of the smoothed estimates over the filtered estimates does depend on the system at hand.

TABLE I: Minimum estimate variance over the entire time interval of forward (ULISE) and backward (ULISS) passes.

		P^x_{11}	P^x_{22}	P^x_{33}	P^x_{44}	P^x_{55}	P^d_{11}	P^d_{22}	P^d_{33}
H^1	ULISE	0.1843	0.0091	0.0002	0.0004	0.0001	0.0099	0.0102	0.1923
	ULISS	0.1843	0.0091	0.0002	0.0004	0.0001	0.0099	0.0102	0.1922
H^2	ULISE	0.1494	0.0052	0.0002	0.0004	0.0001	0.0097	0.0102	0.1574
	ULISS	0.1485	0.0048	0.0002	0.0004	0.0001	0.0047	0.0102	0.1565
H^3	ULISE	0.0076	0.0052	0.0002	0.0004	0.0001	0.0097	0.0102	0.3906
	ULISS	0.0076	0.0048	0.0002	0.0004	0.0001	0.0047	0.0102	0.3836
H^4	ULISE	0.0076	0.0257	0.0002	0.0004	0.0001	0.0348	0.0102	0.4925
	ULISS	0.0076	0.0257	0.0002	0.0004	0.0001	0.0348	0.0102	0.4925
H^5	ULISE	0.0079	0.0074	0.0002	0.0004	0.0001	0.0089	0.0102	0.0099
	ULISS	0.0070	0.0030	0.0002	0.0004	0.0001	0.0039	0.0102	0.0099
H^6	ULISE	0.0076	0.0218	0.0002	0.0004	0.0001	0.0309	0.0102	0.0097
	ULISS	0.0075	0.0054	0.0002	0.0004	0.0001	0.0074	0.0102	0.0096

VI. Conclusion

This paper presented an optimal smoothing algorithm for simultaneously estimating the states and unknown inputs of linear discrete-time stochastic systems based on all observations in a fixed interval. The input and state smoother is derived using widely used methods of minimizing mean squared error and maximizing likelihood. The resulting estimates are unbiased and achieve minimum mean squared error and maximum likelihood. Simulation results have shown a clear improvement of the input and states estimates when compared to the filtered estimates in all test cases.

Acknowledgments

The work presented in this paper was supported by the National Science Foundation, grant #1239182. M. Zhu is partially supported by ARO W911NF-13-1-0421 (MURI).

References

[1] P.K. Kitanidis. Unbiased minimum-variance linear state estimation. *Automatica*, 23(6):775–778, November 1987.

[2] R. Patton, R. Clark, and P.M. Frank. *Fault diagnosis in dynamic systems: theory and applications*. Prentice-Hall international series in systems and control engineering. Prentice Hall, 1989.

[3] G. De Nicolao, G. Sparacino, and C. Cobelli. Nonparametric input estimation in physiological systems: Problems, methods, and case studies. *Automatica*, 33(5):851–870, 1997.

[4] R.E. Kalman. A new approach to linear filtering and prediction problems. *Transactions of the ASME–Journal of Basic Engineering*, 82(Series D):35–45, 1960.

[5] H.E. Rauch, C.T. Striebel, and F. Tung. Maximum likelihood estimates of linear dynamic systems. *Journal of the American Institute of Aeronautics and Astronautics*, 3(8):1445–1450, August 1965.

[6] D. Fraser and J. Potter. The optimum linear smoother as a combination of two optimum linear filters. *IEEE Transactions on Automatic Control*, 14(4):387–390, August 1969.

[7] M. Darouach, M. Zasadzinski, and M. Boutayeb. Extension of minimum variance estimation for systems with unknown inputs. *Automatica*, 39(5):867–876, 2003.

[8] Y. Cheng, H. Ye, Y. Wang, and D. Zhou. Unbiased minimum-variance state estimation for linear systems with unknown input. *Automatica*, 45(2):485–491, 2009.

[9] C.S. Hsieh. Robust two-stage Kalman filters for systems with unknown inputs. *IEEE Transactions on Automatic Control*, 45(12):2374–2378, December 2000.

[10] S. Gillijns and B. De Moor. Unbiased minimum-variance input and state estimation for linear discrete-time systems. *Automatica*, 43(1):111–116, January 2007.

[11] S. Gillijns and B. De Moor. Unbiased minimum-variance input and state estimation for linear discrete-time systems with direct feedthrough. *Automatica*, 43(5):934–937, 2007.

[12] H. Fang, Y. Shi, and J. Yi. On stable simultaneous input and state estimation for discrete-time linear systems. *International Journal of Adaptive Control and Signal Processing*, 25(8):671–686, 2011.

[13] S.Z. Yong, M. Zhu, and E. Frazzoli. Simultaneous input and state estimation for linear discrete-time stochastic systems with direct feedthrough. In *Conference on Decision and Control (CDC)*, pages 7034–7039, 2013.

[14] S.Z. Yong, M. Zhu, and E. Frazzoli. A unified filter for simultaneous input and state estimation of linear discrete-time stochastic systems. *Automatica*, 2014. Provisionally accepted. Available from: `http://arxiv.org/abs/1309.6627`.

[15] H. Fang and R.A. de Callafon. Simultaneous input and state smoothing and its application to oceanographic flow field reconstruction. In *American Control Conference (ACC)*, pages 4705–4710, June 2013.

[16] R.H. Shumway and D.S. Stoffer. *Time Series Analysis and Its Applications: With R Examples*. Springer Texts in Statistics. Springer, 2010.

[17] D.P. Bertsekas and J.N. Tsitsiklis. *Introduction To Probability*. Athena Scientific books. Athena Scientific, 2002.

[18] A.H. Sayed. *Adaptive Filters*. Wiley-IEEE Press, 2008.

[19] C.M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., 2006.

Appendix

A. Properties of Gaussian Distribution

The following are properties of Gaussian distributions (for proofs, see e.g., [19, pp.85-93]:

Lemma 1 (Joint density of Gaussian variables)**.** *Given Gaussian random variables x and y described by $x \sim \mathcal{N}(x;\mu,P)$ and $y|x \sim \mathcal{N}(y;Fx+g,R)$, the joint density of x and y is*

$$\begin{bmatrix} x \\ y \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} x \\ y \end{bmatrix}; \begin{bmatrix} \mu \\ F\mu+g \end{bmatrix}, \begin{bmatrix} P & PF^\top \\ FP & FPF^\top + R \end{bmatrix}\right).$$

Lemma 2 (Conditional and marginal densities of partitioned Gaussians)**.** *Given a joint Gaussian distribution with density*

$$\begin{bmatrix} x \\ y \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} x \\ y \end{bmatrix}; \begin{bmatrix} a \\ b \end{bmatrix}, \begin{bmatrix} \mathcal{A} & \mathcal{C} \\ \mathcal{C}^\top & \mathcal{B} \end{bmatrix}\right),$$

then, the conditional and marginal densities of x and y are:

$$\begin{aligned} x|y &\sim \mathcal{N}(p; a + \mathcal{C}\mathcal{B}^{-1}(y-b), \mathcal{A} - \mathcal{C}\mathcal{B}^{-1}\mathcal{C}^\top) \\ y|x &\sim \mathcal{N}(q; b + \mathcal{C}^\top\mathcal{A}^{-1}(x-a), \mathcal{B} - \mathcal{C}^\top\mathcal{A}^{-1}\mathcal{C}) \\ x &\sim \mathcal{N}(x; a, \mathcal{A}),\ y \sim \mathcal{N}(y; b, \mathcal{B}). \end{aligned}$$

B. Block Matrix Inversion

A well known analytical block matrix inversion formula is given in the following lemma:

Lemma 3 (Block Matrix Inversion)**.** *Given a matrix $\mathcal{M} = \begin{bmatrix} \mathcal{A} & \mathcal{B} \\ \mathcal{C} & \mathcal{D} \end{bmatrix}$ with invertible $\mathcal{A}$ and $\mathcal{D}$, its inverse $\mathcal{M}^{-1} := \begin{bmatrix} \hat{\mathcal{M}}_{11} & \hat{\mathcal{M}}_{12} \\ \hat{\mathcal{M}}_{21} & \hat{\mathcal{M}}_{22} \end{bmatrix}$ can be found with*

$$\begin{aligned} \hat{\mathcal{M}}_{11} &= (\mathcal{A} - \mathcal{B}\mathcal{D}^{-1}\mathcal{C})^{-1} \\ &= \mathcal{A}^{-1} + \mathcal{A}^{-1}\mathcal{B}(\mathcal{D} - \mathcal{C}\mathcal{A}^{-1}\mathcal{B})^{-1}\mathcal{C}\mathcal{A}^{-1} \\ \hat{\mathcal{M}}_{12} &= -(\mathcal{A} - \mathcal{B}\mathcal{D}^{-1}\mathcal{C})^{-1}\mathcal{B}\mathcal{D}^{-1} \\ &= -\mathcal{A}^{-1}\mathcal{B}(\mathcal{D} - \mathcal{C}\mathcal{A}^{-1}\mathcal{B})^{-1} \\ \hat{\mathcal{M}}_{21} &= -\mathcal{D}^{-1}\mathcal{C}(\mathcal{A} - \mathcal{B}\mathcal{D}^{-1}\mathcal{C})^{-1} \\ &= -(\mathcal{D} - \mathcal{C}\mathcal{A}^{-1}\mathcal{B})^{-1}\mathcal{C}\mathcal{A}^{-1} \\ \hat{\mathcal{M}}_{22} &= \mathcal{D}^{-1} + \mathcal{D}^{-1}\mathcal{C}(\mathcal{A} - \mathcal{B}\mathcal{D}^{-1}\mathcal{C})^{-1}\mathcal{B}\mathcal{D}^{-1} \\ &= (\mathcal{D} - \mathcal{C}\mathcal{A}^{-1}\mathcal{B})^{-1}. \end{aligned}$$