ISR TECHNICAL REPORT 2009-19

Certifying the Optimality of a Distributed State Estimation System via Majorization Theory

Gabriel Lipsa and Nuno Martins

Certifying the Optimality of a Distributed State Estimation System via Majorization Theory

Gabriel M. Lipsa, *Member, IEEE,* and Nuno C. Martins, *Fellow, OSA,*

Abstract

Consider a first order linear time-invariant discrete time system driven by process noise, a pre-processor that accepts causal measurements of the state of the system, and a state estimator. The pre-processor and the state estimator are not co-located, and, at every time-step, the pre-processor transmits either a real number or an erasure symbol to the estimator. We seek the pre-processor and the estimator that jointly minimize a cost that combines two terms; the expected squared state estimation error and a communication cost. In our formulation, the transmission of a real number from the pre-processor to the estimator incurs a positive cost while erasures induce zero cost. This paper is the first to prove analytically that a symmetric threshold policy at the pre-processor and a Kalman-like filter at the estimator, which updates its estimate linearly in the presence of erasures, are jointly optimal for our problem.

I. Introduction

We address the design of a finite horizon optimal state estimation system featuring two causal operators; a pre-processor $\mathcal{P}_{0,T}$ and a remote estimator $\mathcal{E}$, where T denotes the time-horizon. At each time instant, the pre-processor outputs either an erasure symbol or a real number, based on causal measurements of the state of a first order linear time-invariant system driven by process noise. The estimator has causal access to the output of the pre-processor and its output is denoted as state estimate. We consider an optimization problem characterized by cost functions that combine the state estimation error and a communication cost. In our formulation, the communication cost depends on the output of the pre-processor, where we ascribe zero cost to the erasure symbol and a pre-specified positive constant otherwise. The state process, denoted

G. Lipsa and N. Martins are with the Department of Electrical and Computer Engineering, University of Maryland College Park, College Park, MD, 20742 USA e-mail: glipsa@umd.edu, nmartins@isr.umd.edu.

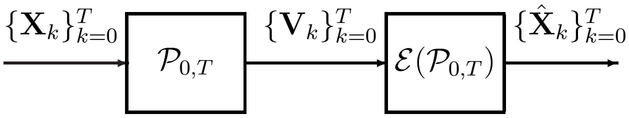
$\{\mathbf{X}_k\}_{k=0}^T$
$\mathcal{P}_{0,T}$
$\{\mathbf{V}_k\}_{k=0}^T$
$\mathcal{E}(\mathcal{P}_{0,T})$
$\{\hat{\mathbf{X}}_k\}_{k=0}^T$

Fig. 1. Schematic representation of the distributed estimation system considered in this paper. It depicts the pre-processor $\mathcal{P}_{0,T}$ and the corresponding optimal estimator $\mathcal{E}(\mathcal{P}_{0,T})$, which produces the minimum mean squared error estimate of the process $\{\mathbf{X}_k\}_{k=0}^T$ given in (5).

as $\mathbf{X}_k$, is given and the two causal operators $\mathcal{P}_{0,T}$ and $\mathcal{E}$ are to be jointly designed so as to minimize the given cost function.

Most of this Section is dedicated to precisely formulating such an optimal estimation problem. In subsection I-A we give a description of the information structure of our framework, followed by subsections I-B and I-C, where we give the problem formulation and a comparison with existing work, respectively. In Section II, we describe a particular solution, while in Section V we prove its optimality. Towards this goal, Section III presents auxiliary optimality results and Section IV is dedicated to introducing concepts from majorization theory and preliminary results, notation and definitions. Section VI presents conclusions and ideas for future work, while in Appendices I and II we state and prove lemmas that are supporting results used throughout the paper.

Notation: In this paper, we use lower case letters for constants, such as a, c and d. For random variables we will use bold upper case letters, such as $\mathbf{X}$, while a particular realization is represented as a constant x. The lower case letters f, g and h are used mainly for probability density functions, with the exception of h, which can also be used to indicate a general function. We denote sets by double bared upper case font, such as $\mathbb{A}$ and $\mathbb{B}$. For sets, we make use of standard operations such as union ($\mathbb{A}\cup\mathbb{B}$), intersection ($\mathbb{A}\cap\mathbb{B}$) and set difference ($\mathbb{A}\backslash\mathbb{B}$). If $\mathbb{A}$ and $\mathbb{B}$ are two subsets of the real line $\mathbb{R}$, we express set difference as $\mathbb{A}\backslash\mathbb{B} = \{x \in \mathbb{R} : x \in \mathbb{A}, x \notin \mathbb{B}\}$. General functions are denoted using calligraphic, upper case font, such as $\mathcal{V}$ and $\mathcal{J}$. Further notation is described throughout the paper on a need basis.

A. Preliminary definitions and information pattern description

We start by describing the three stochastic processes and the two classes of causal operators (pre-processor and estimator) that constitute our problem formulation.

Definition 1: (**State Process**) Given a real constant a, and a positive real constant σ_W^2, consider the state of the following first order, linear time-invariant discrete-time system driven by process noise:

$$\mathbf{X}_0 \overset{def}{=} x_0 \tag{1}$$

$$\mathbf{X}_{k+1} \overset{def}{=} a\mathbf{X}_k + \mathbf{W}_k,\ k \geq 0 \tag{2}$$

where $\{\mathbf{W}_k\}_{k=0}^T$ is an independent identically distributed (i.i.d.) Gaussian zero mean stochastic process with variance σ_W^2 and x_0 is a real number. The filtration generated by $\{\mathbf{X}_k\}_{k=0}^T$ is denoted as:

$$\mathcal{X}_k \overset{def}{=} \sigma\left(\mathbf{X}_t; 0 \leq t \leq k\right) \tag{3}$$

where $\sigma\left(\mathbf{X}_t; 0 \leq t \leq k\right)$ is the smallest sigma algebra generated by $\{\mathbf{X}_t, 0 \leq t \leq k\}$, for all integers k.

Definition 2: (**Pre-processor and remote link process**) Consider an erasure symbol denoted as $\mathfrak{E}$ and a causal pre-processor $\mathcal{P}_{0,T} : (x_0, \ldots, x_k) \mapsto v_k$, defined for $k \in \{0, \ldots, T\}$ and $v_k \in \mathbb{R} \cup \{\mathfrak{E}\}$. Hence, at each time instant k, the preprocessor outputs a real number or the erasure symbol, based on past observations of the state process. Notice that a pre-processor generates a stochastic process $\{\mathbf{V}_k\}_{k=0}^T$ via the application of the operator $\mathcal{P}_{0,T}$ to the process $\{\mathbf{X}_k\}_{k=0}^T$ (See Figure 1). The map $\mathcal{P}_{0,T}$ is a valid pre-processor if the following two conditions hold: (1) The pre-processor transmits the initial state x_0 at time zero, i.e., $v_0 = x_0$. (2) The pre-processor is measurable in the sense that the process $\{\mathbf{V}_k\}_{k=0}^T$ is adapted to $\mathcal{X}_k$.

The filtration generated by $\{\mathbf{V}_k\}_{k=0}^T$ is denoted as $\{\mathcal{B}_k\}_{k=0}^T$ and it is obtained as:

$$\mathcal{B}_k \overset{def}{=} \sigma\left(\mathbf{V}_t; 0 \leq t \leq k\right) \tag{4}$$

where $\sigma\left(\mathbf{V}_t; 0 \leq t \leq k\right)$ is the smallest sigma algebra generated by $\{\mathbf{V}_t, 0 \leq t \leq k\}$, for all non-negative integers k.

Remark 1: Notice that any finite vector of reals can be encoded into a single real number via a suitable invertible transformation. Hence, without loss of generality, we can also assume that the pre-processor can transmit either a vector of real numbers or the erasure symbol.

Definition 3: (**Optimal estimate and optimal estimator**) Given a pre-processor $\mathcal{P}_{0,T}$, we consider optimal estimators in the expected squared sense whose optimal estimate at time k is denoted as $\hat{\mathbf{X}}_k$ and is expressed as follows:

$$\hat{x}_k \overset{def}{=} \begin{cases} E\left[\mathbf{X}_k | \{v_t\}_{t=0}^{k}\right] & \text{if } k \geq 1 \\ x_0 & \text{if } k = 0 \end{cases} \tag{5}$$

where $E\left[\mathbf{X}_k | \{v_t\}_{t=0}^{k}\right]$ represents the expectation of the state $\mathbf{X}_k$ conditioned on the observed current and past outputs of the pre-processor $\{v_t\}_{t=0}^{k}$ (see Figure 1). We use $\mathcal{E}(\mathcal{P}_{0,T})$ to denote the **optimal estimator** associated with a given pre-processor policy $\mathcal{P}_{0,T}$.

Notice that from Definition 2 we assume that the pre-processor always transmits the initial state x_0. Hence, the initial estimate is set to satisfy $\hat{x}_0 = v_0 = x_0$. Such an assumption is a key element that will allow us to prove the optimality of a certain scheme, via an inductive method. This will be discussed later on in Section V.

Remark 2: It is important to note that all the information available at the estimator $\mathcal{E}(\mathcal{P}_{0,T})$ is also available at the pre-processor $\mathcal{P}_{0,T}$. Hence, the pre-processor $\mathcal{P}_{0,T}$ can construct the state estimate $\hat{\mathbf{X}}_k$ by reproducing the estimation algorithm executed at the optimal estimator.

B. Problem statement

In this subsection, we define the optimal estimation paradigm that is central to this paper. We start by specifying the cost, which is used as a merit criterion throughout the paper, followed by the problem definition.

Definition 4: **(Finite time horizon cost function)** Given a measurable pre-processor $\mathcal{P}_{0,T}$ (Definition 2), a real constant a, a positive integer T, a positive real number d less than one and positive real constants σ_W^2 and c, consider the following cost:

$$\mathcal{J}_{0,T}\left(a, \sigma_W^2, c, \mathcal{P}_{0,T}\right) \overset{def}{=} \sum_{k=1}^{T} d^{k-1} E\left[\left(\mathbf{X}_k - \hat{\mathbf{X}}_k\right)^2 + \underbrace{c\mathbf{R}_k}_{\text{communication cost}}\right] \tag{6}$$

where $\mathbf{X}_k$ is the state of the system defined in (1)-(2), $\hat{\mathbf{X}}_k$ is the optimal estimate specified in Definition 3, and $\mathbf{R}_k$ is the following indicator function:

$$\mathbf{R}_k \overset{def}{=} \begin{cases} 0 & \text{if } \mathbf{V}_k = \mathfrak{E} \\ 1 & \text{otherwise} \end{cases}, \qquad k \geq 1 \tag{7}$$

Remark 3: (**Cost does not depend on $\mathbf{X}_0$**) Notice that because the plant (1)-(2) is linear, the fact that $\hat{x}_0 = x_0$ holds (see Definition 3) implies that the homogenous part of the state can be reproduced at the estimator. Hence, the optimal estimator will incorporate such an homogeneous term, thus subtracting it out from the estimation error $\mathbf{X}_k - \hat{\mathbf{X}}_k$, for $k \geq 0$. This also implies that the cost (6) does not depend on the homogeneous term nor on the initial condition $\mathbf{X}_0$.

The following is the main problem addressed in this paper.

Problem 1: Let a real constant a, the variance of the process noise σ_W^2 and the initial condition x_0 be given. In addition, consider that a positive real c, a positive real number d less then one and a positive integer T are given, specifying the cost as in Definition 4. We want to find an optimal solution $\mathcal{P}_{0,T}^*$ to the following optimization problem:

$$\mathcal{P}_{0,T}^* = \arg\min_{\mathcal{P}_{0,T}} \mathcal{J}_{0,T}(a, \sigma_W^2, c, \mathcal{P}_{0,T}) \tag{8}$$

C. Comparison with the state of the art

There is a significant body of work in distributed estimation and in filtering in multiple areas. Of particular interest to this paper is the work in [1], which explores the optimization of paging and registration policies in mobile cellular networks. In [1], motion is modeled as a discrete-time Markov process, and the optimization is carried out for a discounted cost evaluated over an infinite horizon.

The authors of [1] use majorization theory and Riesz's rearrangement inequality to show that, for Gaussian random walk models, nearest-location-first paging and distance threshold registration are jointly optimal. In comparison with the work in [1], which considers random walks and indicator-type costs, our work addresses the optimal estimation in the expected square error sense for scalar linear time invariant systems (stable or unstable).

In [7], the authors consider a sequential estimation problem with two decision makers, where the first observes the state of a stochastic process and decides whether to transmit information to the second agent, which will act as a state estimator. These agents have the common objective of minimizing a performance criterion, with the constraint that the first agent can transmit information to the estimator only a pre-specified finite number of times. In contrast with [7], where the authors assume that the decision policies at the estimator are constrained a-priory to be of the threshold type, here we prove the optimality of symmetric threshold policies. Yet another difference between this paper and [7] is that we adopt a communication cost, instead of constraining the number of transmissions. The problem of obtaining optimal estimates subject to a finite number of sampling actions, in continuous time, is addressed in [14], [15] and related work by the same authors cited therein. Notice that neither the work [7] nor [14], [15] can be used for Problem 1 because there is no explicit relationship between the cost for communication in Problem 1 and the constraint on the number of sampling actions, as adopted in [14], [7]. A general framework for a distinct, yet related, class of problems in continuous time is studied in [9], which is conducive to establishing existence of solutions and optimality results via quasi-variational inequalities. The formulation in [9] is stated in terms of the optimal scheduling of sensors to achieve an optimal estimate of a function of the state at the end of a finite horizon.

The work in [10] is motivated by large-scale sensor networks where simultaneous data transfer to a fusion center is not feasible. In [10], the sensors are part of a networked control system in which a controller is collocated with the fusion center, who must decide which sensor to observe and each choice has a cost associated with it. The main paradigm in [10] is similar to our Problem 1, for which the authors of [10] illustrated numerically that the best policy is of the threshold type.

The author in [11] investigates an optimal control problem, where measurements can be collected one sensor at a time and each sensor has an associated cost. In [11] it is shown that the problem of selecting the optimal strategy can be formulated as a deterministic control problem. The computation of the measurement policy takes place offline and the optimal strategy is adopted. In contrast to our result, the policies adopted in [11] are off-line.

The authors of [8] adopt a formulation that is similar to ours. They consider a networked control problem with transmission costs, where they adopt a Kalman-like estimator and show, using dynamic programming, that, for such a pre-determined choice of estimator, the optimal pre-processor is a memoryless function of the state estimation error. In contrast to our paper, the problem analyzed in [8] deals also with the multidimensional case, while we handle the scalar case, but we prove analytically that there exist a Kalman-like filter at the estimator and a threshold policy at the pre-processor that are jointly optimal.

Notice that the communication link in our framework is not noisy, in the sense that the pre-processor can predict with certainty what the estimator receives after every transmission. A significant advance in the understanding of the problem of designing optimal causal pre-processors and estimators in the presence of noisy transmission, without communication cost, can be found in [3], [4].

II. Optimal Solution to Problem 1

In this section, we start by defining a particular choice of estimator (section II-A) and pre-processor (section II-C), which we denote as Kalman-like and symmetric threshold policy, respectively. As we argue later on, in Theorem 1, such estimator and pre-processor are optimal for Problem 1.

A. A Kalman-like estimator

Definition 5: (**Kalman-like estimator**) Given the process defined in (1)-(2) and a pre-processor $\mathcal{P}_{0,T}$, define the map $\mathcal{Z} : (v_0, \ldots, v_k) \mapsto z_k$, for k in the set $\{0, \ldots, T\}$, where z_k is computed as follows:

$$z_0 \overset{def}{=} x_0 \tag{9}$$

$$z_k \overset{def}{=} \begin{cases} az_{k-1} & \text{if } v_k = \mathfrak{E} \\ v_k & \text{otherwise} \end{cases}, \text{ with } k \geq 1 \tag{10}$$

Remark 4: The Kalman-like filter generates the process $\{\mathbf{Z}_k\}_{k=0}^{T}$ via the operator $\mathcal{Z}$ applied to the process $\{\mathbf{V}_k\}_{k=0}^{T}$. Notice that the pre-processor has access to the estimate $\mathbf{Z}_k$ because it has access and full control of the input applied to $\mathcal{Z}$.

B. The Set $\mathbb{P}_T$ - of Admissible Pre-Processors

We proceed by defining a class of pre-processors, which is amenable to the use of recursive methods for performance analysis. If a pre-processor belongs to such a class then we denote it as admissible, and we argue in Remark 6 that there always exist an admissible pre-processor that is an optimal solution to Problem 1. This implies that we incur no loss of generality in constraining our analysis to admissible pre-processors.

Definition 6: (**Admissible pre-processor**) Let a horizon T larger than zero and a pre-processor policy $\mathcal{P}_{0,T}$ be given. The pre-processor $\mathcal{P}_{0,T}$ is admissible if there exist maps $\mathcal{P}_{m,T} : (x_m, \ldots, x_k) \mapsto v_k$, with $0 \leq m \leq T$ and $k \geq m$, such that $\mathcal{P}_{0,T}$ can be specified recursively as follows:

Description of the Algorithm for $\mathcal{P}_{m,T}$

- (**Initial step**) Set $k = m$, $r_m = 1$ and transmit the current state, i.e., $v_m = x_m$.

- (**Step A**) Increase the counter k by one. If $k > T$ holds then terminate, otherwise execute Step B.
- (**Step B**) Obtain the pre-processor output at time k via $v_k = \mathcal{P}_{m,T}(x_m, \dots, x_k)$. If $v_k = \mathfrak{E}$ then set $r_k = 0$ and go back to Step A. If $v_k \neq \mathfrak{E}$ then execute algorithm $\mathcal{P}_{k,T}$.

End of the description of the Algorithm for $\mathcal{P}_{m,T}$

The class of all admissible pre-processors is denoted as $\mathbb{P}_T$.

The following Remark provides an equivalent characterization of the class of admissible pre-processors.

Remark 5: Let a horizon T larger than zero and a pre-processor policy $\mathcal{P}_{0,T}$ be given. The pre-processor $\mathcal{P}_{0,T}$ is admissible if and only if for each $m \in \{0, \dots, T\}$ there exists a map $\mathcal{P}_{m,T} : (x_m, \dots, x_k) \mapsto v_k$ such that the following holds:

$$r_m = 1 \implies \mathcal{P}_{q,T}(x_q, \dots, x_k) = \mathcal{P}_{m,T}(x_m, \dots, x_k), \qquad x_q, \dots, x_k \in \mathbb{R}, k \geq m \geq q \geq 0 \quad (11)$$

Given an admissible pre-processor $\mathcal{P}_{0,T}$, later on we will also refer to the time-restricted pre-processors $\{\mathcal{P}_{m,T}\}_{m=1}^{T}$ according to Definition 6, or equivalently as implied by (11).

Remark 6: Given a positive time-horizon T, there is no loss of generality in constraining our search - for optimal an pre-processor - to the set $\mathbb{P}_T$. In order to justify this assertion, consider that an optimal pre-processor policy $\mathcal{P}^*_{0,T}$ is given. If a transmission takes place at some time m ($r_m = 1$ holds) then the optimal output at the pre-processor is $v_k = x_k$. In fact, given that a real number is transmitted, the choice $v_k = x_k$ must be optimal because it leads to a perfect estimate $\hat{x}_m = x_m$. Hence, given that $r_m = 1$, by Markovianity we conclude that the current and future output produced by the pre-processor $\{\mathbf{V}_k\}_{k=m}^{T}$ will not depend on the state $\mathbf{X}_k$ for times k prior to m. Consequently, $\mathcal{P}^*_{0,T}$ satisfies (11), and hence it is admissible.

C. Symmetric threshold pre-processor

Definition 7: In order to simplify our notation, we define the following process:

$$\mathbf{Y}_k \overset{def}{=} \mathbf{X}_k - a\mathbf{Z}_{k-1} \quad (12)$$

Using Definitions 1 and 5, we find that $\{\mathbf{Y}_k\}_{k=0}^T$ can be rewritten as:

$$\mathbf{Y}_0 = 0 \tag{13}$$

$$\mathbf{Y}_{k+1} = \begin{cases} a\mathbf{Y}_k + \mathbf{W}_k & \text{if } \mathbf{R}_k = 0 \\ \mathbf{W}_k & \text{if } \mathbf{R}_k = 1 \end{cases} \tag{14}$$

Remark 7: We notice that $\mathbf{Y}_k$ has an even probability density function. This fact makes $\{\mathbf{Y}_k\}_{k=0}^T$ a more convenient process to work with, in comparison to $\{\mathbf{X}_k\}_{k=0}^T$, which motivates its use in our analysis hereon, whenever possible. This decision incurs no loss of generality because $\{\mathbf{Y}_k\}_{k=0}^T$ can be recovered from $\{\mathbf{X}_k\}_{k=0}^T$, and vice-versa, via the use of $\{\mathbf{Z}_k\}_{k=0}^T$, which is common information at the pre-processor and estimator (See Remark 4). In addition, notice that the cost (6) can be re-written in terms of $\{\mathbf{Y}_k\}_{k=0}^T$ as follows:

$$\mathcal{J}_{0,T}\left(a, \sigma_W^2, c, \mathcal{P}_{0,T}\right) \overset{def}{=} \sum_{k=1}^{T} d^{k-1} E\left[\left(\mathbf{Y}_k - \hat{\mathbf{Y}}_k\right)^2 + c\mathbf{R}_k\right] \tag{15}$$

where $\hat{\mathbf{Y}}_k \overset{def}{=} E\left[\mathbf{Y}_k | \{\mathbf{V}_t\}_{t=0}^k\right]$. A key fact here is that $\hat{\mathbf{Y}}_k = \hat{\mathbf{X}}_k - a\mathbf{Z}_{k-1}$ holds, leading to the validity of the identity $\mathbf{Y}_k - \hat{\mathbf{Y}}_k = \mathbf{X}_k - \hat{\mathbf{X}}_k$.

Definition 8: Given a positive integer horizon T and an arbitrary sequence of positive real numbers (thresholds) $\tau = \{\tau_k\}_{k=1}^T$, for each m in the set $\{0, \ldots, T\}$, we define the following algorithm for $k \geq m$, which we denote as $\mathcal{S}_{m,T}$:

Description of Algorithm $\mathcal{S}_{m,T}$

- **(Initial step)** Set $k = m$, $r_m = 1$ and transmit the current state, i.e., $v_m = x_m$ or equivalently set $y_m = 0$.
- **(Step A)** Increase the time counter k by one. If $k > T$ holds then terminate, otherwise execute Step B.
- **(Step B)** If $|y_k| < \tau_k$ holds then set $r_k = 0$, transmit the erasure symbol, i.e., $v_k = \mathfrak{E}$, and return to Step A. If $|y_k| \geq \tau_k$ holds then set $m = k$ and execute $\mathcal{S}_{m,T}$.

End of description of Algorithm $\mathcal{S}_{m,T}$

Definition 9: (**Symmetric threshold policy**) The algorithm $\mathcal{S}_{0,T}$, as in Definition 8, is denoted as symmetric threshold pre-processor. The pre-processor $\mathcal{S}_{0,T}$ is admissible and the class of all symmetric threshold policies is denoted as $\mathbb{S}_T$.

The following is the main result of this paper.

Theorem 1: Let the parameters specifying Problem 1 be given, i.e., the variance of the process noise σ_W^2, the system's dynamic constant a, the communication cost c, the discount factor d and the time horizon T are pre-selected. There exists a sequence of positive real numbers $\tau^* = \{\tau_k^*\}_{k=1}^T$, such that the corresponding symmetric threshold policy $\mathcal{S}_{0,T}^*$ is an optimal solution to (8) and the corresponding optimal estimator $\mathcal{E}(\mathcal{S}_{0,T}^*)$ is $\mathcal{Z}$. Here $\mathcal{S}_{0,T}^*$ and $\mathcal{Z}$ follow Definitions 9 and 5, respectively.

Note: The proof of Theorem 1 is given in Section V.

III. Auxiliary optimality results

We start by defining the following class of path-dependent pre-processor policies, which is an extension of Definition 9 so as to allow time-varying thresholds that depend on past decisions. Such a class of admissible pre-processors will be used later in Section V, where we provide a proof for Theorem 1.

Definition 10: (**Algorithm** $\mathcal{D}_{m,T}$) Given a horizon T, consider that a sequence of (threshold) functions $\mathcal{T} \overset{def}{=} \{\mathcal{T}_{m,k} | m < k \leq T, 1 \leq m \leq T\}$, with $\mathcal{T}_{m,k} : \{0,1\}^{m-k} \rightarrow \mathbb{R}$, is given. For every m in the set $\{1, \dots, T\}$, we define the following algorithm, which we denote as $\mathcal{D}_{m,T}$:

Description of Algorithm $\mathcal{D}_{m,T}$

- **(Initial step)** Set $k = m$, $r_m = 1$ and transmit the current state, i.e., $v_m = x_m$ or equivalently set $y_m = 0$.
- **(Step A)** Increase the time counter k by one. If $k > T$ holds then terminate, otherwise execute Step B.
- **(Step B)** If $|y_k| < \mathcal{T}_{m,k}(r_m, \dots, r_{k-1})$ holds then set $r_k = 0$, transmit the erasure symbol, i.e., $v_k = \mathfrak{E}$, and return to Step A. If $|y_k| \geq \mathcal{T}_{m,k}(r_m, \dots, r_{k-1})$ holds then execute $\mathcal{D}_{k,T}$.

End of description of Algorithm $\mathcal{D}_{m,T}$

Recall that r_0 through r_{k-1} represent past decisions by the pre-processor, where $r_k = 1$ indicates that the state is transmitted to the estimator at time k, while $r_k = 0$ implies that an erasure was sent.

Definition 11: (**Path-dependent symmetric threshold policy**) Given a horizon T, consider that a sequence of (threshold) functions $\mathcal{T} \overset{def}{=} \{\mathcal{T}_{m,k} | m < k \leq T, 1 \leq m \leq T\}$, with $\mathcal{T}_{m,k} : \{0,1\}^{m-k} \rightarrow \mathbb{R}$, is given. The path-dependent symmetric threshold pre-processor associated with $\mathcal{T}$ is implemented via the $\underline{\text{execution of the algorithm } \mathcal{D}_{0,T}}$, as specified in Definition 10.

Typically, we denote such an admissible pre-processor as $\mathcal{D}_{0,T}$. We use $\mathbb{D}_{0,T}$ to denote the entire class of path-dependent symmetric threshold pre-processors with time horizon T.

The goal of this section is to provide the following two results that are crucial in the proof of Theorem 1: In Proposition 1, we prove that if $\mathcal{D}_{0,T}$ is any given path-dependent symmetric threshold pre-processor policy then the associated optimal estimator $\mathcal{E}(\mathcal{D}_{0,T})$ is $\mathcal{Z}$. In Lemma 1 we prove that if we optimize within the class of path-dependent policies then the optimum is of the path-independent type, as specified in Definition 9. This fact might raise the question of whether Definition 11 is needed. The answer is *yes* because we adopt a constructive argument in the proof of Theorem 1 in Section V, which uses Definition 11.

Proposition 1: Let $\mathcal{D}_{0,T}$ be a pre-selected path-dependent symmetric threshold policy (Definition 11), it holds that the optimal estimator $\mathcal{E}(\mathcal{D}_{0,T})$ is $\mathcal{Z}$, as described in Definition 5.

Remark 8: Proposition 1 could be recast by stating that $\hat{\mathbf{X}}_k = \mathbf{Z}_k$ holds in the presence of path-dependent symmetric threshold pre-processors.

Proof: (of Proposition 1) In order to simplify the proof, we define $\{\tilde{\mathbf{X}}_k\}_{k=0}^{T}$ as the process quantifying the error incurred by adopting a Kalman-like estimator $\mathcal{Z}$ (See Definition 5), i.e., $\tilde{\mathbf{X}}_k \overset{def}{=} \mathbf{X}_k - \mathbf{Z}_k$. More specifically, $\{\tilde{\mathbf{X}}_k\}_{k=0}^{T}$ can be equivalently expressed as follows:

$$\tilde{\mathbf{X}}_0 = 0 \tag{16}$$

$$\tilde{\mathbf{X}}_{k+1} = \begin{cases} a\tilde{\mathbf{X}}_k + \mathbf{W}_k & \text{if } \mathbf{R}_k = 0 \\ 0 & \text{if } \mathbf{R}_k = 1 \end{cases}, \qquad 0 \leq k \leq T-1 \tag{17}$$

The proof follows from the symmetry of all probability density functions involving $\tilde{\mathbf{X}}_k$ and $\mathbf{V}_k$. More specifically, under symmetric path-dependent threshold policies the probability density function of $\tilde{\mathbf{X}}_k$, given the past and current observations $\{\mathbf{V}_t\}_{t=0}^{k}$, is even. Hence, we conclude that $E[\tilde{\mathbf{X}}_k|\{\mathbf{V}_t\}_{t=0}^{k}] = 0$, which implies that $\hat{\mathbf{X}}_k \overset{def}{=} E[\mathbf{X}_k|\{\mathbf{V}_t\}_{t=0}^{k}] = \mathbf{Z}_k$. ■

A. Optimizing within the class $\mathbb{D}_T$

Remark 9: If $\mathcal{D}_{0,T}$ is a symmetric path-dependent threshold pre-processor (see Definition 11) then $\hat{\mathbf{Y}}_k = 0$ holds, leading to the following equality:

$$\mathcal{J}_{0,T}\left(a, \sigma_W^2, c, \mathcal{D}_{0,T}\right) = \sum_{k=1}^{T} d^{k-1} E\left[\mathbf{Y}_k^2 + c\mathbf{R}_k\right], \qquad \mathcal{D}_{0,T} \in \mathbb{D}_T \tag{18}$$

The process defined in (14) is a Markov Decision Process (MDP) whose state and control are $\mathbf{Y}_k$ and $\mathbf{R}_k$, respectively. Hence the minimization of (18) with respect to pre-processor policies $\mathcal{D}_{0,T}$ in the class $\mathbb{D}_T$ can be cast as a dynamic program [13]. To do so, we define the sequence of functions $\mathcal{V}_{t,T} : \mathbb{R} \to \mathbb{R}, t \in \{1, \dots, T+1\}$ which represent the cost-to-go as observed by the pre-processor. Here T represents the horizon, while t denotes the time at which the decision was taken, and the argument of the function is the MDP state $\mathbf{Y}_t$. In order to simplify our notation, we adopt the convention that $\mathcal{V}_{T+1,T}(y_{T+1}) \overset{def}{=} 0,\ y_{T+1} \in \mathbb{R}$. Using dynamic programming, we can find the following recursive equations for $\mathcal{V}_{t,T}(y_t),\ t \in \{1, \dots, T\}$:

$$\mathcal{V}_{t,T}(y_t) \overset{def}{=} \min_{r_t \in \{0,1\}} \mathcal{C}_{t,T}(y_t, r_t),\ t \in \{1, \dots, T\} \tag{19}$$

where $\mathcal{C}_{t,T} : \mathbb{R} \times \{0,1\} \to \mathbb{R}$ is defined as:

$$\mathcal{C}_{t,T}(y_t, r_t) \overset{def}{=} \begin{cases} c + dE\left[\mathcal{V}_{t+1,T}(\mathbf{W}_t)\right] & \text{if } r_t = 1 \\ y_t^2 + dE\left[\mathcal{V}_{t+1,T}\left(ay_t + \mathbf{W}_t\right)\right] & \text{if } r_t = 0 \end{cases} \tag{20}$$

From (20) it immediately follows that an optimal decision policy r_t^* at any time t is given by:

$$r_t^* = \begin{cases} 1 & \text{if } \mathcal{C}_{t,T}(y_t, 1) \leq \mathcal{C}_{t,T}(y_t, 0) \\ 0 & \text{if } \mathcal{C}_{t,T}(y_t, 0) < \mathcal{C}_{t,T}(y_t, 1) \end{cases} \tag{21}$$

Using the MDP given in Definition 7 and the value functions from equation (19), we prove the following Lemma, which states that, *within the class of symmetric path-dependent pre-processors* $\mathbb{D}_T$ *(Definition 11)*, there exists an optimal path-*independent* symmetric threshold policy $\mathcal{S}_{0,T}^*$ (Definition 9) for Problem 1.

Lemma 1: Let the parameters specifying Problem 1 be given, i.e., the variance of the process noise σ_W^2, the system's dynamic constant a, the communication cost c, the discount factor d and the time horizon T are pre-selected. Consider Problem 1 with the additional constraint that the pre-processor must be of the symmetric path-dependent type $\mathbb{D}_T$ specified in Definition 11. There exists an optimal path-*independent* symmetric threshold policy $\mathcal{S}_{0,T}^*$, as given in Definition 9, whose associated threshold selection $\{\tau_k^*\}_{k=1}^T$ is given by a solution to the following equations:

$$\mathcal{C}_{t,T}(\tau_t^*, 0) = \mathcal{C}_{t,T}(\tau_t^*, 1),\ t \in \{1, \dots, T\} \tag{22}$$

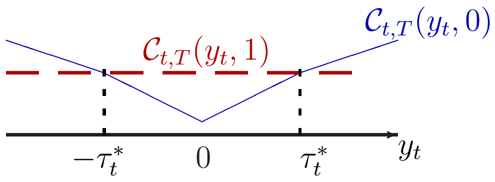

Fig. 2. Illustration suggesting that Facts A.1 through A.4. imply the existence of thresholds for which (23) holds.

Proof: From (21), we conclude that in order to prove this Lemma we only need to show that there exist thresholds $\{\tau_k^*\}_{k=1}^T$ for which the following equivalences hold:

$$|y_t| \geq \tau_t^* \Longleftrightarrow \mathcal{C}_{t,T}(y_t,1) \leq \mathcal{C}_{t,T}(y_t,0), \qquad t \in \{1,\dots,T\} \tag{23}$$

Indeed, if (23) holds then the optimal strategy in (21) can be implemented via a threshold policy. In order to prove that there exist thresholds $\{\tau_k^*\}_{k=1}^T$ such that (23) holds, we will use the following facts (A.1 thorugh A.4):

- **(Fact A.1)**: For every t in the set $\{1,\dots,T\}$, $\mathcal{C}_{t,T}(y_t,1)$ depends only on t, i.e., it is a time-dependent constant independent of y_t.
- **(Fact A.2)**: It holds that $\mathcal{C}_{t,T}(0,0) < \mathcal{C}_{t,T}(y_t,1)$ for $y_t \in \mathbb{R}$.
- **(Fact A.3)**: For every t in the set $\{1,\dots,T\}$ there exists a positive constant u_t such that $\mathcal{C}_{t,T}(y_t,0) > \mathcal{C}_{t,T}(y_t,1)$ and $\mathcal{C}_{t,T}(-y_t,0) > \mathcal{C}_{t,T}(-y_t,1)$ hold for every y_t satisfying $|y_t| > u_t$.
- **(Fact A.4)**: It holds that $\mathcal{C}_{t,T}(y_t,0)$ is a continuous, even, quasi-convex and unbounded function of y_t, for every t in the set $\{1,\dots,T\}$.

Facts A.1 and A.2 follow directly from (20), while Fact A.3 follows from Fact A.4, which requires a proof that we defer to a later stage. At this point we assume that Fact A.4 is valid, and we proceed by noticing that continuity of $\mathcal{C}_{t,T}(y_t,0)$ with respect to y_t, as well as Facts A.2 and A.3, imply that the equations in (22) have at least one solution $\{\tau_k^*\}_{k=1}^T$. Moreover, from Facts A.1 through A.4 we can conclude that such a solution $\{\tau_k^*\}_{k=1}^T$ guarantees that (23) is true (See Figure 2).

(Proof of Fact 4) Since y_t^2 is an even, convex, unbounded and continuous function of y_t, from (20) we conclude that it suffices to prove by induction that $\mathcal{V}_{t,T}(y_t)$ is even, quasiconvex, bounded and continuous for each t in the set $\{1,\dots,T\}$.

Since $\mathcal{V}_{T+1,T}(y_{T+1}) = 0$ holds by convention, the following is true:

$$\mathcal{V}_{T,T}(y_T) = \min\left(c, y_T^2\right), \qquad y_T \in \mathbb{R}$$

Hence $\mathcal{V}_{T,T}(y_T)$ is an even, quasiconvex, bounded and continuous function of y_T. Using Lemma 14 in Appendix II, we conclude that $E\left[\mathcal{V}_{T,T}(ay_{T-1} + \mathbf{W}_{T-1})\right]$ is also an even, quasiconvex, bounded and continuous function of y_{T-1}, which implies that so is $\mathcal{V}_{T-1,T}(y_{T-1})$. By induction it follows that $\mathcal{V}_{t,T}(y_t)$ is an even, quasiconvex, bounded and continuous of y_t, for each t in the set $\{1, \ldots, T\}$. ■

IV. Notation, Definitions and Basic Results for the Proof of Theorem 1

This section is dedicated to introducing notation, definitions and basic results in majorization theory that will streamline our proof of Theorem 1. The proof of Theorem 1 is given in Section V.

In Subsection IV-A, we introduce basic majorization theory and state a few Lemmas, which are supporting results for the proof of Theorem 1. In Subsection IV-B, we introduce notation and we derive recursive equations for the time update of certain conditional probability density functions of interest.

A. Basic Results, Notation and Definitions from Theory of Majorization

In [1], the authors define what a neat probability mass functions is. We will adapt this definition for probability density functions on $\mathbb{R}$.

Definition 12: (**Neat pdf**) Let $f : \mathbb{R} \to \mathbb{R}$ be a probability density function. We say that f is neat if f is quasiconcave and there exists a real number b such that f is non-decreasing on the interval $(-\infty, b]$ and non-increasing on $[b, \infty)$.

Remark 10: Throughout the paper, we will use the useful fact that the convolution of two neat and even probability density functions is also neat and even. The complete proof of this fact is given in Lemma 5 in Appendix I.

Hajek gives in [1] the definition of symmetric non-increasing function on $\mathbb{R}^n$. Since we work only on the real line, it suffices to notice that a probability density function $f : \mathbb{R} \to \mathbb{R}$ is symmetric non-increasing if and only if it is neat and even. Hence, without loss of generality, in this paper only use *symmetric non-increasing* to qualify certain probability density functions throughout the paper.

Let $\mathbb{A}$ be a given Borel measurable subset of $\mathbb{R}$, we denote its Lebesgue measure by $\mathcal{L}(\mathbb{A})$. If the Lebesgue measure of $\mathbb{A}$ is finite then the symmetric rearrangement of $\mathbb{A}$, denoted by $\mathbb{A}^\sigma$, is a symmetric closed interval centered around the origin with Lebesgue measure $\mathcal{L}(\mathbb{A})$:

$$\mathbb{A}^\sigma = \left\{ x \in \mathbb{R} : |x| \leq \frac{\mathcal{L}(\mathbb{A})}{2} \right\}$$

Let $f : \mathbb{R} \to \mathbb{R}$ be a given non-negative function, we define f^σ, the symmetric non-decreasing rearrangement of f, as follows:

$$f^\sigma(x) \overset{def}{=} \int_0^\infty \mathcal{I}_{\{z\in\mathbb{R}:f(z)>\rho\}^\sigma}(x) d\rho \tag{24}$$

where $\mathcal{I}_{\{z\in\mathbb{R}:f(z)>\rho\}^\sigma} : \mathbb{R} \to \{0,1\}$ is the following indicator function:

$$\mathcal{I}_{\{z\in\mathbb{R}:f(z)>\rho\}^\sigma}(x) \overset{def}{=} \begin{cases} 1 & \text{if } x \in \{z \in \mathbb{R} : f(z) > \rho\}^\sigma \\ 0 & \text{otherwise} \end{cases}, \qquad x \in \mathbb{R}$$

If f and g are two probability density functions on $\mathbb{R}$, then we say that f majorizes g, which we denote as $f \succ g$, provided that the following holds:

$$\int_{|x|\leq\rho} g^\sigma(x)dx \leq \int_{|x|\leq\rho} f^\sigma(x)dx, \text{ for all } \rho \geq 0 \tag{25}$$

One interpretation of the inequality in (25) is that, f majorizes g, if and only if for any Borel set $\mathbb{F}' \subset \mathbb{R}$ with finite Lebesgue measure, there exists another Borel set $\mathbb{F} \subset \mathbb{R}$ satisfying $\mathcal{L}(\mathbb{F}') = \mathcal{L}(\mathbb{F})$ and such that the following holds:

$$\int_{\mathbb{F}'} g(x)dx \leq \int_{\mathbb{F}} f(x)dx$$

Given a probability density function $f : \mathbb{R} \to \mathbb{R}$ and a Borel set $\mathbb{K}$, such that $\int_{\mathbb{K}} f(x)dx > 0$, we define the restriction of f to $\mathbb{K}$ as follows:

$$f_{\mathbb{K}}(x) \overset{def}{=} \begin{cases} \frac{f(x)}{\int_{\mathbb{K}} f(x)dx} & \text{if } x \in \mathbb{K} \\ 0 & \text{otherwise} \end{cases} \tag{26}$$

It is clear that $f_{\mathbb{K}}$ is also a probability density function.

The following Lemma is a supporting result for the proof of Theorem 1 given in Section V.

Lemma 2: Let $f, g : \mathbb{R} \to \mathbb{R}$ be two probability density functions, such that f is neat and even and $f \succ g$. Let κ be a real number in the interval $\kappa \in (0,1)$, and let $\mathbb{A} = [-\tau, \tau]$ be the symmetric closed interval such that $\int_{-\tau}^{\tau} f(x)dx = 1-\kappa$. For any function $h : \mathbb{R} \rightarrow [0,1]$ satisfying $\int_{\mathbb{R}} g(x)h(x)dx = 1-\kappa$, the following holds:

$$f_{\mathbb{A}} \succ \frac{g \cdot h}{1-\kappa} \tag{27}$$

where $g \cdot h : \mathbb{R} \rightarrow \mathbb{R}$ is defined as $g \cdot h(x) \overset{def}{=} g(x)h(x)$, for $x \in \mathbb{R}$.

Proof: From Lemma 10 given in Appendix I, we know that for any function $h : \mathbb{R} \rightarrow [0,1]$ satisfying $\int_{\mathbb{R}} g(x)h(x)dx = 1-\kappa$, there exists a set $\mathbb{A}' \subset \mathbb{R}$, satisfying $\int_{\mathbb{A}'} g(x)dx = 1-\kappa$, such that the following holds:

$$g_{\mathbb{A}'} \succ \frac{g \cdot h}{1-\kappa} \tag{28}$$

From Lemma 9 given in Appendix I, we know that $f_{\mathbb{A}} \succ g_{\mathbb{A}'}$. From equation (28) and the fact that $f_{\mathbb{A}} \succ g_{\mathbb{A}'}$ holds, it follows that:

$$f_{\mathbb{A}} \succ \frac{g \cdot h}{1-\kappa}$$

■

The following Lemma, which we state without proof, can be found in [1]:

Lemma 3: [1, Lemma 6.7] Let f and g be two probability density functions on $\mathbb{R}$, with f symmetric non-increasing and $f \succ g$. For a symmetric non-increasing probability density function h the following holds:

$$f * h \succ g * h \tag{29}$$

Lemma 4: Let f be a neat and even probability density function on the real line. Let g be a probability density function on the real line satisfying $g \prec f$. The following holds:

$$\int_{\mathbb{R}} x^2 f(x)dx \leq \int_{\mathbb{R}} (x-y)^2 g(x)dx, \qquad y \in \mathbb{R} \tag{30}$$

Proof: The result follows by selecting $h(x) = x^2$ in Lemma 13 found in Appendix A. ■

Remark 11: Consider the conditions of Lemma 4. The fact that the probability density function f is even implies that $\int_{\mathbb{R}} xf(x)dx = 0$. Hence, if we select $y = \int_{\mathbb{R}} xg(x)dx$ then it follows from equation (30) that the variance of f is less than or equal to the variance of g.

B. Conditional probabilities and conditional probability density functions

Before proving Theorem 1, in this subsection we need to make a few remarks and introduce more notation, which will streamline our proof. This subsection contains two parts: We start by introducing the notation for certain conditional probability density functions of interest, while in the second part we will derive recursive equations for the time update of the conditional densities, and we will also obtain a recursive expansion for the cost associated with any given admissible pre-processor policy $\mathcal{P}_{0,T}$.

Definition 13: Let a pre-processor $\mathcal{P}_{0,T}$, implementing a decision policy as in Definition 2, be given. We define the following notation for conditional probability densities, which will streamline our proof of Theorem 1:

1) Define the conditional probability density function of $\mathbf{Y}_k$ given that only erasure symbols were transmitted up until time k as follows:

$$\gamma_{k|k}\left(y\right) \overset{def}{=} f_{\mathbf{Y}_k|\mathbf{R}_1=0,\ldots,\mathbf{R}_k=0}\left(y\right), \qquad y \in \mathbb{R}$$

2) Define the conditional probability density function of $\mathbf{Y}_k$ given that only erasure symbols were transmitted up until time $k-1$ as follows:

$$\gamma_{k|k-1}\left(y\right) \overset{def}{=} f_{\mathbf{Y}_k|\mathbf{R}_1=0,\ldots,\mathbf{R}_{k-1}=0}\left(y\right), \qquad y \in \mathbb{R}$$

Definition 14: We define the following streamlined notation for certain conditional probabilities of interest:

1) Define the probability that, under policy $\mathcal{P}_{0,T}$, only erasure symbols have been transmitted up until time k:

$$\varsigma_k \overset{def}{=} \begin{cases} P\left(\mathbf{R}_1 = 0, \ldots, \mathbf{R}_k = 0\right) & \text{if } k \geq 1 \\ 1 & \text{if } k = 0 \end{cases}$$

2) Define the conditional probability that, under policy $\mathcal{P}_{0,T}$, the pre-processor transmits the erasure symbol at time k, given that only erasure symbols have been transmitted up until time $k-1$.

$$\varsigma_{k|k-1} \overset{def}{=} \begin{cases} P\left(\mathbf{R}_k = 0|\mathbf{R}_1 = 0, \ldots, \mathbf{R}_{k-1} = 0\right) & \text{if } k > 1 \\ \varsigma_1 & \text{if } k = 1 \end{cases}$$

Definition 15: Let $\mathcal{P}_{0,T}$ be a decision policy given as in Definition 2. Let k be a positive integer and y be a real number. For a positive integer k, define the function $\rho_k : \mathbb{R} \to [0,1]$ as follows:

$$\rho_k\left(y\right) \overset{def}{=} P\left(\mathbf{R}_k = 0 | \mathbf{Y}_k = y, \mathbf{R}_1 = 0, \dots, \mathbf{R}_{k-1} = 0\right), \qquad x \in \mathbb{R} \tag{31}$$

which is the probability that, at time k, the erasure symbol is transmitted, given that $\mathbf{Y}_k = y$, where y is any real number, and the fact that only erasure symbols have been transmitted up until time $k-1$.

Notation: For a random variable $\mathbf{Y}$ described by a probability density function f and a real function h, we denote by $E_f[h(\mathbf{Y})]$, the expected value of the random variable $h(\mathbf{Y})$ under the probability density function f.

C. Time Evolution

Now, we describe how the conditional probability density functions presented in subsection IV-B evolve in time, for a given policy $\mathcal{P}_{0,T}$. For a real number a, below we define the conditional probability density function of $a\mathbf{Y}_k$ given that no observation was received up until time k:

$$\gamma^a_{k|k}(y) \overset{def}{=} f_{a\mathbf{Y}_k|\mathbf{R}_1=0,\dots,\mathbf{R}_k=0}\left(y\right)$$

We denote by $\mathcal{N}_{\sigma^2_W}$ the probability density function of $\mathbf{W}_k$, for all k, i.e., the Gaussian zero mean probability density with variance σ^2_W, or more concretely $\mathcal{N}_{\sigma^2_W}(x) = \frac{1}{\sqrt{2\pi\sigma^2_W}} e^{-\frac{x^2}{2\sigma^2_W}}$. Since the sequence $\{\mathbf{W}_k\}^T_{k=0}$ is i.i.d., $\mathbf{W}_{k-1}$ is also independent of $\{Y_l\}^{k-1}_{l=0}$, which implies that the following holds:

$$\gamma_{k|k-1} = \gamma^a_{k-1|k-1} * \mathcal{N}_{\sigma^2_w} \tag{32}$$

Proposition 2: The conditional densities $\gamma_{k|k-1}$ and $\gamma_{k|k}$ are related via the following time-recursion:

$$\gamma_{k|k}(y) = \frac{\gamma_{k|k-1}(y)\rho_k\left(y\right)}{\varsigma_{k|k-1}}, \qquad \varsigma_{k|k-1} \neq 0, k \geq 1 \tag{33}$$

Proof: In order to arrive at (33), we use Baye's rule to write:

$$f_{\mathbf{Y}_k|\mathbf{R}_1=\mathfrak{E},\dots,\mathbf{R}_k=\mathfrak{E}}\left(y\right) = \frac{P\left(\mathbf{R}_k = 0|\mathbf{Y}_k = y, \mathbf{R}_1 = 0, \dots, \mathbf{R}_{k-1} = 0\right)}{P\left(\mathbf{R}_k = 0|\mathbf{R}_1 = 0, \dots, \mathbf{R}_{k-1} = 0\right)} f_{\mathbf{Y}_k|\mathbf{R}_1=0,\dots,\mathbf{R}_{k-1}=0}\left(y\right) \tag{34}$$

The recursion (33) follows from (34) by rewriting it according to Definitions 13, 14 and 15. Equation (34) holds only if $P\left(\mathbf{R}_k = 0 | \mathbf{R}_1 = 0, \dots, \mathbf{R}_{k-1} = 0\right) = \varsigma_{k|k-1} \neq 0$. If $\varsigma_{k|k-1} = 0$ then the conditional density function $f_{\mathbf{Y}_k|\mathbf{R}_1=0,\dots,\mathbf{R}_k=0}\left(y\right)$ is no longer defined. ■

Definition 16: Given an admissible pre-processor $\mathcal{P}_{0,T}$ and an integer $m \in \{0, \dots, T\}$, we adopt the following definition for the partial cost computed for the horizon $\{m+1, \dots, T\}$ under the assumption that $r_m = 1$:

$$\mathcal{J}_{m,T}\left(a, \sigma_W^2, c, \mathcal{P}_{m,T}\right) \overset{def}{=} \begin{cases} \sum_{k=m+1}^{T} d^{k-m-1} E\left[\left(\mathbf{Y}_k - \hat{\mathbf{Y}}_k\right)^2 + c\mathbf{R}_k\right] & \text{if } 0 \leq m < T \\ 0 & \text{if } m = T \end{cases} \quad (35)$$

Remark 12: Given an integer m, we notice that the cost in (35) will not depend on the value of the state at time m. This is so because, according to Definition 6, since $\mathcal{P}_{0,T}$ is admissible it holds that the current and future *output* of $\mathcal{P}_{m,T}$ will not depend on the current and past state observations. This Remark is an extension of Remark 3, which considered the case for $m = 0$.

Proposition 3: Given an arbitrarily selected admissible pre-processor $\mathcal{P}_{0,T}$, the finite horizon cost (6) can be expanded as:

$$\begin{aligned} &\mathcal{J}_{0,T}\left(a, \sigma_W^2, c, \mathcal{P}_{0,T}\right) \\ &\quad = \sum_{k=1}^{T} d^{k-1} \left(E_{\gamma_{k|k}}\left[\left(\mathbf{Y}_k - \hat{\mathbf{Y}}_k\right)^2\right] \varsigma_k + \left(c + \mathcal{J}_{k,T}\left(a, \sigma_W^2, c, \mathcal{P}_{k,T}\right)\right) \varsigma_{k-1}(1 - \varsigma_{k|k-1}) \right) \end{aligned} \quad (36)$$

Here we use the notation $E_{\gamma_{k|k}}\left[\left(\mathbf{Y}_k - \hat{\mathbf{Y}}_k\right)^2\right] \overset{def}{=} E\left[\left(\mathbf{Y}_k - \hat{\mathbf{Y}}_k\right)^2 | \mathbf{R}_1 = 0, \dots, \mathbf{R}_k = 0\right]$, where $\gamma_{k|k}$ is given in Definition 13.

Proof: We start by noticing that, by the total probability law, we can expand the cost as:

$$\begin{aligned} &\mathcal{J}_{0,T}\left(a, \sigma_W^2, c, \mathcal{P}_{0,T}\right) \\ &\quad = \sum_{k=1}^{T} d^{k-1} \Bigg(E\left[\left(\mathbf{Y}_k - \hat{\mathbf{Y}}_k\right)^2 | \mathbf{R}_1 = 0, \dots, \mathbf{R}_k = 0\right] P\left(\mathbf{R}_1 = 0, \dots, \mathbf{R}_k = 0\right) + \\ &\qquad + \left(c + E\left[\mathcal{J}_{k,T}\left(a, \sigma_W^2, c, \mathcal{P}_{k,T}\right) | \mathbf{R}_k = 1, \mathbf{R}_1 = 0, \dots, \mathbf{R}_{k-1} = 0\right]\right) \times \\ &\qquad\qquad P\left(\mathbf{R}_k = 1, \mathbf{R}_1 = 0, \dots, \mathbf{R}_{k-1} = 0\right) \Bigg) \end{aligned} \quad (37)$$

We proceed by obtaining the following identities:

$$\begin{aligned}&P\left(\mathbf{R}_k=1,\mathbf{R}_1=0,\ldots,\mathbf{R}_{k-1}=0\right)=P\left(\mathbf{R}_1=0,\ldots,\mathbf{R}_{k-1}=0\right)-\\&-P\left(\mathbf{R}_1=0,\ldots,\mathbf{R}_k=0\right)=P\left(\mathbf{R}_1=0,\ldots,\mathbf{R}_{k-1}=0\right)-\\&-P\left(\mathbf{R}_k=0|\mathbf{R}_1=0,\ldots,\mathbf{R}_{k-1}=0\right)P\left(\mathbf{R}_1=0,\ldots,\mathbf{R}_{k-1}=0\right)=\\&=\varsigma_{k-1}(1-\varsigma_{k|k-1}),\qquad k\geq 1\end{aligned}\tag{38}$$

Notice that, using standard probability theory, from $\{\varsigma_k\}_{k=1}^T$ we can compute $\left\{\varsigma_{k|k-1}\right\}_{k=1}^T$ and vice versa. Here, equation (38) is still valid for $k=1$, since we defined $\varsigma_0=1$ and $\varsigma_{1|0}=\varsigma_1$. Finally, notice that from Remark 12, we conclude the following:

$$E\left[\mathcal{J}_{k,T}\left(a,\sigma_W^2,c,\mathcal{P}_{k,T}\right)|\mathbf{R}_k=1,\mathbf{R}_1=0,\ldots,\mathbf{R}_{k-1}=0\right]=\mathcal{J}_{k,T}\left(a,\sigma_W^2,c,\mathcal{P}_{k,T}\right)\tag{39}$$

The proof of this Proposition is complete once we substitute (38) and (39) into (37). ■

Definition 17: The following is a convenient definition for the optimal cost:

$$\mathcal{J}^*_{m,T}\left(a,\sigma_W^2,c\right)\overset{def}{=}\begin{cases}\min_{\mathcal{P}_{m,T}\in\mathbb{P}_{T-m}}\mathcal{J}_{m,T}\left(a,\sigma_W^2,c,\mathcal{P}_{m,T}\right), & T\geq 1\\ 0, & T=0\end{cases}\tag{40}$$

From Proposition 3, we can immediately state the following Corollary:

Corollary 1: The following inequality holds for every admissible pre-processor $\mathcal{P}_{0,T}$:

$$\begin{aligned}&\mathcal{J}_{0,T}\left(a,\sigma_W^2,c,\mathcal{P}_{0,T}\right)\geq\\&\sum_{k=1}^T d^{k-1}\left(E_{\gamma_{k|k}}\left[\left(\mathbf{Y}_k-\hat{\mathbf{Y}}_k\right)^2\right]\varsigma_k+\left(c+\mathcal{J}^*_{k,T}\left(a,\sigma_W^2,c\right)\right)(1-\varsigma_{k|k-1})\varsigma_{k-1}\right)\end{aligned}\tag{41}$$

V. Proof of Theorem 1

Our **strategy** to prove Theorem 1 is to show that for every admissible pre-processor policy $\mathcal{P}_{0,T}$, there exists a path-dependent symmetric threshold policy $\mathcal{D}^o_{0,T}$ which does not underperform $\mathcal{P}_{0,T}$. This fact, which we denote as **Fact B.1**, leads to the following conclusions:

- (**Fact B.2**): Lemma 1 (Section III-A), in conjunction with Fact B.1, implies that an optimum $\mathcal{S}^*_{0,T}$ for Problem 1 exists and that it is of the symmetric threshold type $\mathbb{S}_T$ (Definition 9).
- (**Fact B.3**): From Fact B.2 and Proposition 1 (Section III), we conclude there exists a symmetric threshold policy $\mathcal{S}^*_{0,T}$ and a Kalman-like estimator $\mathcal{Z}$ (Definition 5) that are jointly optimal for Problem 1.

Proof: (of Theorem 1): Facts B.2 and B.3 constitute a proof for Theorem 1. It remains to prove the validity of Fact B.1.

(**Proof of Fact B.1**): Here we will use an inductive approach that is analogous to the one used in [1, Lemma 6.5]. Our proof for Fact B.1 is organized in two parts. In **Part I**, we will prove Fact B.1 for the case when the time-horizon T is one, while in **Part II**, we prove the general induction step.

Notation: According to the definitions of Section IV-B , any given pre-processor has associated with it conditional probability density functions $\left\{\gamma_{k|k}\right\}_{k=1}^{T}$ and $\left\{\gamma_{k|k-1}\right\}_{k=1}^{T}$, as well as conditional probabilities $\{\varsigma_k\}_{k=1}^{T}$ and $\left\{\varsigma_{k|k-1}\right\}_{k=1}^{T}$. Hence, we assume that the path-dependent symmetric threshold policy $\mathcal{D}_{0,T}^{o}$ - to be constructed as part of this proof - defines conditional probability density functions $\left\{\gamma_{k|k}^{o}\right\}_{k=1}^{T}$ and $\left\{\gamma_{k|k-1}^{o}\right\}_{k=1}^{T}$ as well as conditional probabilities $\{\varsigma_k^o\}_{k=1}^{T}$ and $\left\{\varsigma_{k|k-1}^{o}\right\}_{k=1}^{T}$.

Part I: Here we will prove Fact B.1 for $T=1$. We will do so by constructing a policy $\mathcal{D}_{0,1}^{o}$ as follows:

$$r_1^o \overset{def}{=} \begin{cases} 1 & \text{if } |y_1| > \tau_1 \\ 0 & \text{otherwise} \end{cases} \tag{42}$$

where τ_1 is a threshold that we will select appropriately. Hence, if the absolute value of y_1 is less than or equal to τ_1 then the pre-processor transmits the erasure symbol, otherwise it sends x_1. Consider that a policy $\mathcal{P}_{0,1}$ is given. We start by noticing that for $\mathcal{P}_{0,1}$ and $\mathcal{D}_{0,1}^{o}$ it holds that $\gamma_{1|0} = \gamma_{1|0}^{o} = N_{\sigma_W^2}$, while the cost associated with policy $\mathcal{P}_{0,1}$ is:

$$\mathcal{J}_{0,1}\left(a, \sigma_W^2, c, \mathcal{P}_{0,1}\right) = E_{\gamma_{1|1}}\left[\left(\mathbf{Y}_1 - \hat{\mathbf{Y}}_1\right)^2\right]\varsigma_1 + c(1-\varsigma_1) \tag{43}$$

where $\hat{\mathbf{Y}}_1 = E_{\gamma_{1|1}}\left[\mathbf{Y}_1\right]$. We construct a desirable $\mathcal{D}_{0,1}^{o}$ by selecting τ_1 such that $\varsigma_1^o = \varsigma_1$, which from (42) leads to a probability density function $\gamma_{1|1}^{o}$ that is neat and even. Furthermore, Lemma 2 implies that $\gamma_{1|1} \prec \gamma_{1|1}^{o}$ holds. From Lemma 4 we arrive at the following inequality:

$$E_{\gamma_{1|1}^{o}}\left[\left(\mathbf{Y}_1 - \hat{\mathbf{Y}}_1^o\right)^2\right] \leq E_{\gamma_{1|1}}\left[\left(\mathbf{Y}_1 - \hat{\mathbf{Y}}_1\right)^2\right] \tag{44}$$

The cost associated with the policy $\mathcal{D}_{0,1}^{o}$ is given by:

$$\mathcal{J}_{0,1}\left(a, \sigma_W^2, c, \mathcal{D}_{0,1}^{o}\right) = E_{\gamma_{1|1}^{o}}\left[\left(\mathbf{Y}_1 - \hat{\mathbf{Y}}_1^o\right)^2\right]\varsigma_1 + c(1-\varsigma_1) \tag{45}$$

Finally, we conclude from (43), (44) and (45) that:

$$\mathcal{J}_{0,1}\left(a, \sigma_W^2, c, \mathcal{P}_{0,1}\right) \geq \mathcal{J}_{0,1}\left(a, \sigma_W^2, c, \mathcal{D}_{0,1}^o\right) \tag{46}$$

which leads to the desired conclusion that $\mathcal{D}_{0,1}^o$ does not underperform $\mathcal{P}_{0,1}$.

Part II: (General induction step) Let T^I be a given horizon that is strictly larger than one. Assume the **inductive hypothesis** that Fact B.1 is valid for any horizon T less than T^I.

We start by noticing that the validity of our inductive hypothesis implies the following facts:

- (**Fact B.4**): The inductive hypothesis in conjunction with Lemma 1 implies that Problem 1 has an optimum for every horizon T less than T^I.
- (**Fact B.5**): The inductive hypothesis also implies that Problem 1 admits an optimal preprocessor policy of the symmetric threshold type (Definition 9), for every horizon T less than T^I.

Hence, Fact B.5 implies that there exist $\mathcal{S}_{1,T^I}^*$ through $\mathcal{S}_{T^I,T^I}^*$ that satisfy the following:

$$\mathcal{J}_{m,T^I}(a, \sigma_W^2, c, \mathcal{S}_{m,T^I}^*) = \min_{\tilde{\mathcal{P}}_{m,T^I} \in \mathbb{P}_{T^I-m}} \mathcal{J}_{m,T^I}(a, \sigma_W^2, c, \tilde{\mathcal{P}}_{m,T^I}) \underset{(a)}{=} \mathcal{J}_{m,T^I}^*(a, \sigma_W^2, c) \qquad 1 \leq m \leq T^I \tag{47}$$

where $\mathcal{S}_{m,T^I}^*$ is of the symmetric threshold type $\mathbb{S}_{T^I-m}$ and (a) above follows by definition from (40).

Now we proceed to showing that the general induction step holds. In order to do so, we show that for any admissible policy $\mathcal{P}_{0,T^I}$, we can construct a path-dependent symmetric threshold policy $\mathcal{D}_{0,T^I}^o$ that does not underperform $\mathcal{P}_{0,T^I}$. Henceforth, assume that $\mathcal{P}_{0,T^I}$ is an arbitrarily chosen admissible policy.

The following is our algorithm for $\mathcal{D}_{0,T^I}^o$:

Description of Algorithm for $\mathcal{D}_{0,T^I}^o$

- **(Initial step)** Set $k = 0$ and transmit the current state, i.e., $v_0 = x_0$ or equivalently set $y_0 = 0$.
- **(Step A)** Increase the time counter k by one. If $k > T^I$ holds then terminate, otherwise execute Step B.
- **(Step B)** If $|y_k| < \tau_k^o$ holds then set $r_k = 0$, transmit the erasure symbol, i.e., $v_k = \mathfrak{E}$, and return to Step A. If $|y_k| \geq \tau_k^o$ holds then execute $\mathcal{S}_{k,T^I}^*$, as defined in (47).

where $\{\tau_k^o\}_{k=1}^{T^I}$ are appropriately chosen thresholds, as described next.

End of description of Algorithm for $\mathcal{D}_{0,T^I}^o$

Notice that $\mathcal{D}_{0,T^I}^o$ is a path-dependent symmetric threshold strategy (Definition 10), for which we can also conclude that $\mathcal{D}_{m,T^I}^o = \mathcal{S}_{m,T^I}^*$ holds for $1 \leq m \leq T^I$.

In order to complete the specification of $\mathcal{D}_{0,T^I}^o$ so that it does not underform $\mathcal{P}_{0,T^I}$, we proceed by appropriately selecting the thresholds $\{\tau_k^o\}_{k=1}^{T^I}$.

(**Selection of thresholds** $\{\tau_k^o\}_{k=1}^{T^I}$) We proceed to describing how to choose the threshold sequence $\{\tau_k^o\}_{k=1}^{T}$ and what this choice implies. Notice that $\gamma_{1|0}^o = \mathcal{N}_{\sigma_W^2}$ and that the Gaussian probability density function is neat and symmetric. Choose τ_1^o such that $\varsigma_1^o = \varsigma_1$, it follows that the probability density function $\gamma_{1|1}^o$ is neat and even. From equation (32), which describes how the conditional probability density functions evolve in time, it holds that $\gamma_{2|1}^o$ is neat and even. By further selecting τ_2^o such that $\varsigma_{2|1}^o = \varsigma_{2|1}$, it also follows that $\gamma_{2|2}^o$ and $\gamma_{3|2}^o$ are neat and even. By repeated execution of this selection process, we can choose all the thresholds τ_k^o such that $\varsigma_{k|k-1}^o = \varsigma_{k|k-1}$ for all k in $\left\{1, \dots, T^I\right\}$. These choices also imply that $\gamma_{k|k}^o$ and $\gamma_{k|k-1}^o$ are neat and even for all k in $\left\{1, \dots, T^I\right\}$. Since $\varsigma_{k|k-1}^o = \varsigma_{k|k-1}$ holds for all k in $\left\{1, \dots, T^I\right\}$, it follows that $\varsigma_k^o = \varsigma_k$ is satisfied for all k in $\left\{1, \dots, T^I\right\}$.

At this point, we know that $\gamma_{1|0} = \gamma_{1|0}^o = \mathcal{N}_{\sigma_W^2}$ and that the Gaussian probability density function $\mathcal{N}_{\sigma_W^2}$ is neat and even. Hence, then from Lemma 2, we conclude that $\gamma_{1|1} \prec \gamma_{1|1}^o$. It also follows from Lemma 11 in the Appendix I and Lemma 3 that $\gamma_{2|1} \prec \gamma_{2|1}^o$ holds. From the repeated application of this idea, it follows that $\gamma_{k|k} \prec \gamma_{k|k}^o$ for all k in $\left\{1, \dots, T^I\right\}$ and, in addition, since $\gamma_{k|k}^o$ is neat and even, it holds that $\hat{\mathbf{Y}}_k^o = E_{\gamma_{k|k}^o}[\mathbf{Y}_k] = 0$ for all k in $\left\{1, \dots, T^I\right\}$. Since $\gamma_{k|k} \prec \gamma_{k|k}^o$ holds and $\gamma_{k|k}^o$ is neat and even, Lemma 4 implies that the following is true:

$$E_{\gamma_{k|k}^o}\left[\left(\mathbf{Y}_k - \hat{\mathbf{Y}}_k^o\right)^2\right] \leq E_{\gamma_{k|k}}\left[\left(\mathbf{Y}_k - \hat{\mathbf{Y}}_k\right)^2\right], \qquad k \in \left\{1, \dots, T^I\right\} \tag{48}$$

The cost obtained by applying the pre-processor policy $\mathcal{P}^o$ can be expressed using (36) as follows:

$$\mathcal{J}_{0,T^I}\left(a, \sigma_W^2, c, \mathcal{D}_{0,T^I}^o\right) = \sum_{k=1}^{T^I} d^{k-1}\Bigg(E_{\gamma_{k|k}^o}\left[\left(\mathbf{Y}_k - \hat{\mathbf{Y}}_k^o\right)^2\right]\varsigma_k + \\ \left(c + \mathcal{J}_{k,T^I}\left(a, \sigma_W^2, c, \mathcal{D}_{0,T^I}^o\right)\right)(1 - \varsigma_{k|k-1})\varsigma_{k-1}\Bigg) \tag{49}$$

Using (47), we can re-write (49) as follows:

$$\mathcal{J}_{0,T^I}\left(a,\sigma_W^2,c,\mathcal{D}^o_{0,T^I}\right) = \sum_{k=1}^{T^I} d^{k-1}\Bigg(E_{\gamma^o_{k|k}}\left[\left(\mathbf{Y}_k - \hat{\mathbf{Y}}^o_k\right)^2\right]\varsigma_k + \left(c+\mathcal{J}^*_{k,T^I}\left(a,\sigma_W^2,c\right)\right)\left(1-\varsigma_{k|k-1}\right)\varsigma_{k-1}\Bigg) \quad (50)$$

From inequality (41), which lower bounds the cost associated with any pre-processor policy, equation (50) and equation (48), we conclude that:

$$\mathcal{J}_{0,T^I}\left(a,\sigma_W^2,c,\mathcal{D}^o_{0,T^I}\right) \leq \mathcal{J}_{0,T^I}\left(a,\sigma_W^2,c,\mathcal{P}_{0,T^I}\right) \quad (51)$$

That we were able to construct $\mathcal{D}^o_{0,T^I}$ satisfying (51) for an arbitrarily chosen admissible pre-processor $\mathcal{P}_{0,T^I}$ constitutes a proof for Fact B.1. ■

VI. CONCLUSIONS

This paper addresses the design of a distributed estimation system comprising of two blocks connected in series, via a link that conveys either a real number or an erasure symbol. Transmission of a real number incurs a positive communication cost, while the erasure symbol features zero cost. The first block is a pre-processor that accepts causal state measurements of a scalar linear and time invariant plant driven by process noise, while the second block must produce an optimal estimate of the state, according to a cost that combines the expected squared estimation error and the communication cost. This paper is the first to prove that threshold policies at the pre-processor and a class of kalman-like filters (previously proposed in the literature) at the estimator are jointly optimal. The problem addressed here is non-convex, implying that standard arguments based on symmetry will not hold. In order to circumvent this difficulty, we introduce the use of majorization theory to establish a convenient partial order among candidate solutions. The proof follows by appropriate use of the partial order via a constructive argument that exploits the structure of the cost function.

APPENDIX I
MAJORIZATION THEORY

Lemma 5: If f and h are neat and even probability density functions, then $f * h$ is also neat and even, where by $f * h$ we mean the convolution between f and h.

Proof: The proof adopted here is analogous to the one in [1, Lemma 6.2], which deals with probability mass functions. Since h is a probability density function, it implies that is also measurable. Let $g : \mathbb{R} \to \mathbb{R}$ be defined as:

$$g(x) = \begin{cases} 1, x \in [-\alpha, \alpha] \\ 0, x \notin [-\alpha, \alpha] \end{cases}$$

where α is a positive real number. We notice that g is an indicator function. We claim that $f * g$ is neat and even.

$$(f * g)(x) = \int_{-\infty}^{\infty} f(x-t)g(t)dt = \int_{-\alpha}^{\alpha} f(x-t)dt = \int_{-\alpha+x}^{\alpha+x} f(y)dy \tag{52}$$

Since the function f is neat and even, it is clear that $f * g$ is neat and even from equation (52). The function $f * g$ is neat and even also for the case when $g(x) = 1$ on a symmetric open interval $(-\alpha, \alpha)$ and $g(x) = 0$ for $x \in (-\infty, -\alpha] \cup [\alpha, \infty)$.

We need to prove the main claim of Lemma 5. We do this by approximating the function h with a sum of indicator functions like the function g. Since h is neat and even it follows that $h(0) \geq h(x)$, for any real number x. For a positive integer number n, define the function h_n as follows:

$$h_n(x) = h(0)\frac{k}{n}, \quad x \in \left\{ x \in \mathbb{R} : h(0)\frac{k}{n} \leq h(x) < h(0)\frac{k+1}{n} \right\}, k \in \{0, \ldots, n-1\} \tag{53}$$

It follows that $h_n(x) \leq h_{n+1}(x)$ for every real number x and that $h_n \to h$. Moreover, from the monotone convergence Theorem, it follows that $f * h_n \to f * h$.

Since h is neat and even it follows that for every integer n and integer $k \leq n$, there exists a positive α_k^n such that $h(x) \geq h(0)\frac{k}{n}$ for $x \in \mathbb{I}_k^n = [-\alpha_k^n, \alpha_k^n]$ or $x \in \mathbb{I}_k^n = (-\alpha_k^n, \alpha_k^n)$ and $h(x) < h(0)\frac{k}{n}$ outside $\mathbb{I}_k^n$, and moreover $\mathbb{I}_k^n \subset \mathbb{I}_{k+1}^n$ for all positive integers $k < n$. The function h_n can be written as follows:

$$h_n(x) = h(0)\frac{1}{n}\sum_{k=0}^{n} \mathcal{I}_{\mathbb{I}_k^n}(x)$$

where by $\mathcal{I}_{\mathbb{I}_k^n}$ we denote the indicator function of the interval $\mathbb{I}_k^n$.

$$f * h_n = h(0)\frac{1}{n}\sum_{k=0}^{n} f * \mathcal{I}_{\mathbb{I}_k^n}$$

It follows that $f * h_n$ is neat and even, hence taking the limit as n goes to infinity, it implies that $f * h$ is neat and even. ■

Remark 13: From the proof of Lemma 5, it follows that the claim of Lemma 5 holds if f and h are any non-negative, even, quasiconcave and integrable functions.

We will state now two important inequalities, which are useful for this paper. The first one is the Riesz's rearrangement inequality:

Lemma 6 (Riesz's Rearrangement inequality [2]): If f, g and h are non-negative functions on $\mathbb{R}^n$, then:

$$\int_{\mathbb{R}^n} f(x)\left(g * h\right)(x)dx \leq \int_{\mathbb{R}^n} f^{\sigma}(x)\left(g^{\sigma} * h^{\sigma}\right)(x)dx \tag{54}$$

The second important inequality, which we need is the Hardy-Littlewood inequality [5].

Lemma 7 (Hardy-Littlewood inequality [5]): If f and g are two non-negative measurable functions defined on the real line, which vanish at infinity, then the following holds:

$$\int_{\mathbb{R}} f(x)g(x)dx \leq \int_{\mathbb{R}} f^{\sigma}(x)g^{\sigma}(x)dx \tag{55}$$

We state and prove the following Lemmas, which are a supporting results for Lemma 2 in Subsection IV-A.

Lemma 8: Let $f : \mathbb{R} \to \mathbb{R}$ be a symmetric and non-increasing probability density function. For any positive $\kappa \leq 1$, there exists a symmetric interval $\mathbb{I}$ centered around zero such that the following holds [1]:

$$f_{\mathbb{I}} \succ f_{\mathbb{I}'} \tag{56}$$

$$\int_{\mathbb{I}} f(x)dx = 1 - \kappa \tag{57}$$

for any Borel set (not necessarily interval) $\mathbb{I}' \subset \mathbb{R}^n$, satisfying $\int_{\mathbb{I}'} f(x)dx = 1 - \kappa$.

Proof: **Case I:** Assume that there exists ρ such that $\int_{\{x\in\mathbb{R}: f(x)>\rho\}} f(x)dx = 1 - \kappa$, then let $\mathbb{I} = \{x \in \mathbb{R} : f(x) > \rho\}$. Since, f is symmetric and non-increasing, it follows that $\mathbb{I}$ is a symmetric interval. Let any other set $\mathbb{I}'$ such that $\int_{\mathbb{I}'} f(x)dx = 1 - \kappa$. Choose any set $\mathbb{F}' \subset \mathbb{I}'$, if $\mathcal{L}(\mathbb{F}') \geq \mathcal{L}\left(\mathbb{I}\right)$, let $\mathbb{F} \subset \mathbb{R}$ be any Borel set, such that $\mathcal{L}(\mathbb{F}) = \mathcal{L}(\mathbb{F}')$ and $\mathbb{I} \subset \mathbb{F}$, it follows that :

$$\int_{\mathbb{F}} f_{\mathbb{I}}(x)dx = 1 \geq \int_{\mathbb{F}'} f_{\mathbb{I}'}(x)dx$$

since both $f_{\mathbb{I}}$ and $f_{\mathbb{I}'}$ are probability density functions. If $\mathcal{L}(\mathbb{F}') \leq \mathcal{L}(\mathbb{I})$, then choose any set $\mathbb{F} \subset \mathbb{I}$, such that $\mathcal{L}(\mathbb{F}) = \mathcal{L}(\mathbb{F}')$. Let $\mathbb{F}_1 = \mathbb{F} \cap \mathbb{F}'$, then, for any real number $x \in \mathbb{F}' \setminus \mathbb{F}_1$ it holds

[1] Here $f_{\mathbb{I}}$ and $f_{\mathbb{I}'}$ follow the definition in (26)

that $f(x) \leq \rho$, while on the set $\mathbb{F} \setminus \mathbb{F}_1$, $f(x) \geq \rho$.

$$\begin{aligned}\int_{\mathbb{F}} f_{\mathbb{I}}(x)dx = \frac{1}{1-\kappa}\int_{\mathbb{F}} f(x)dx &= \frac{1}{1-\kappa}\left(\int_{\mathbb{F}_1} f(x)dx + \int_{\mathbb{F}\setminus\mathbb{F}_1} f(x)dx\right)\\ &\geq \frac{1}{1-\kappa}\left(\int_{\mathbb{F}_1} f(x)dx + \int_{\mathbb{F}\setminus\mathbb{F}_1} \rho dx\right)\\ &\geq \frac{1}{1-\kappa}\left(\int_{\mathbb{F}_1} f(x)dx + \int_{\mathbb{F}'\setminus\mathbb{F}_1} f(x)dx\right)\\ &= \frac{1}{1-\kappa}\int_{\mathbb{F}'} f(x)dx = \int_{\mathbb{F}'} f_{\mathbb{I}'}(x)dx\end{aligned}$$

The second inequality is due to the fact that $\mathbb{F} \setminus \mathbb{F}_1$ and $\mathbb{F}' \setminus \mathbb{F}_1$ have the same Lebegue measure.

Case II: Assume that, there is no such ρ, such that $\int_{\{x\in\mathbb{R}:f(x)>\rho\}} f(x)dx = 1-\kappa$. The integral $\int_{\{x\in\mathbb{R}:f(x)>\rho\}} f(x)dx$ is decreasing as a function of ρ and is also bounded. It follows than that, there exist a ρ such that $\int_{\{x\in\mathbb{R}:f(x)>\rho\}} f(x)dx < 1-\kappa$ and $\int_{\{x\in\mathbb{R}:f(x)\geq\rho\}} f(x)dx \geq 1-\kappa$. Both the sets $\{x \in \mathbb{R}^n : f(x) > \rho\}$ and $\{x \in \mathbb{R} : f(x) \geq \rho\}$ are symmetric intervals and $\{x \in \mathbb{R} : f(x) > \rho\} \subset \{x \in \mathbb{R} : f(x) \geq \rho\}$. Then we can find an interval $\mathbb{I} \subset \{f(x) \geq \rho\}$ symmetric around the origin such that $\int_{\mathbb{I}} f(x)dx = 1-\kappa$. Using the same type of arguments like in the first case we get that $f_{\mathbb{I}} \succ f_{\mathbb{I}'}$ for any $\mathbb{I}' \subset \mathbb{R}$ such that $\int_{\mathbb{I}'} f(x)dx = 1-\kappa$. ■

Lemma 9: Let $f, g : \mathbb{R} \to \mathbb{R}$ be two probability density functions, such that f is neat and even and $f \succ g$. Let κ be a real number such that $0 < \kappa < 1$. Let $\mathbb{I}$ be the symmetric interval given by Lemma 8 for the probability density function f and the number κ. If $\mathbb{I}' \subset \mathbb{R}$ is a set such that $\int_{\mathbb{I}'} g(x)dx = 1-\kappa$ is satisfied then $f_{\mathbb{I}} \succ g_{\mathbb{I}'}$ holds, where $f_{\mathbb{I}}$ and $g_{\mathbb{I}'}$ follow the definition in (26).

Proof: Fix a Borel set $\mathbb{I}' \in \mathbb{R}$ such that $\int_{\mathbb{I}'} g(x)dx = 1-\kappa$ and choose a Borel set $\mathbb{F}' \in \mathbb{I}'$ with strictly positive Lebesgue measure. If $\mathcal{L}(\mathbb{F}') \geq \mathcal{L}(\mathbb{I})$, choose $\mathbb{F}$ any Borel set with $\mathcal{L}(\mathbb{F}) = \mathcal{L}(\mathbb{F}')$, such that $\mathbb{I} \subset \mathbb{F}$. It is clear in this case that $\int_{\mathbb{F}} f_{\mathbb{I}}(x)dx = 1 \geq \int_{\mathbb{F}'} g_{\mathbb{I}'}(x)dx$. If $\mathcal{L}(\mathbb{F}') \leq \mathcal{L}(\mathbb{I})$, then because $f \succ g$, there exists a set $\mathbb{F}'' \in \mathbb{R}$, such that $\mathcal{L}(\mathbb{F}'') = \mathcal{L}(\mathbb{F}')$ and $\int_{\mathbb{F}''} f(x)dx \geq \int_{\mathbb{F}'} g(x)dx$. Choose $\mathbb{I}''$ a set which contains $\mathbb{F}''$ and $\int_{\mathbb{I}''} f(x)dx = 1-\kappa$. By Lemma 8, $f_{\mathbb{I}''} \prec f_{\mathbb{I}}$, so it follows that there exists a set $\mathbb{F} \subset \mathbb{I}$, with the same Lebesgue measure as $\mathbb{F}''$ such that $\int_{\mathbb{F}} f(x)dx \geq \int_{\mathbb{F}''} f(x)dx \geq \int_{\mathbb{F}'} g(x)dx$ ■

Lemma 10: Let $f : \mathbb{R} \to \mathbb{R}$ be a probability density function and let κ be a positive real number, less than one. Let $h : \mathbb{R} \to [0,1]$ be a measurable positive function such that $\int_{\mathbb{R}} h(x)f(x)dx = 1-\kappa$. There exists a Borel set $\mathbb{A}$ such that $\int_{\mathbb{A}} f(x)dx = 1-\kappa$ and $f_{\mathbb{A}} \succ \frac{h \cdot f}{1-\kappa}$.

Proof: If exists ρ such that $\int_{\{x\in\mathbb{R}:f(x)>\rho\}} f(x)dx = 1-\kappa$, then let $\mathbb{A} = \{x \in \mathbb{R} : f(x) > \rho\}$. If no such ρ exists, just like in the proof of Lemma 8, there exists a ρ such that:

$$\int_{\{x\in\mathbb{R}:f(x)>\rho\}} f(x)dx < 1-\kappa, \text{ and } \int_{\{x\in\mathbb{R}:f(x)\geq\rho\}} f(x)dx \geq 1-\kappa$$

i.e., there exists a set of Lebesgue measure strictly positive, such that $f(x) = \rho$. Choose a set $\mathbb{A}' = \{x \in \mathbb{R} : f(x) > \rho\}$. From the set $\{x \in \mathbb{R} : f(x) = \rho\}$, choose a subset $\mathbb{A}''$ of measure $\frac{1-\kappa-\int_{\{x\in\mathbb{R}:f(x)>\rho\}} f(x)dx}{\rho}$. Let $\mathbb{A} = \mathbb{A}' \cup \mathbb{A}''$, it follows then that $\int_{\mathbb{A}} f(x)dx = 1-\kappa$ and that $f(x) \geq \rho$, for all $x \in \mathbb{A}$.

Let $\mathbb{F}'$ be a Borel set in $\mathbb{R}$, if $\mathcal{L}(\mathbb{F}') \geq \mathcal{L}(\mathbb{A})$, choose $\mathbb{F}$ such that $\mathcal{L}(\mathbb{F}') = \mathcal{L}(\mathbb{F})$ and $\mathbb{A} \subset \mathbb{F}$. Then the following holds:

$$\int_{\mathbb{F}} f_{\mathbb{A}}(x)dx = 1 \geq \int_{\mathbb{F}'} \frac{f(x)}{1-\kappa} h(x)dx$$

If $\mathcal{L}(\mathbb{F}') \leq \mathcal{L}(\mathbb{A})$, let $\mathbb{F}_1 = \mathbb{F}' \cap \mathbb{A}$ and let $\mathbb{F}_2 \subset \mathbb{A} \setminus \mathbb{F}_1$ such that $\mathcal{L}(\mathbb{F}_1 \cup \mathbb{F}_2) = \mathcal{L}(\mathbb{F}')$. If $x \in \mathbb{F}_1$, $f(x) \geq h(x)f(x)$, and if $x \in \mathbb{F}_2$, $f(x) \geq \rho$, and if $x \in \mathbb{F}' \setminus \mathbb{F}_1$, $h(x)f(x) \leq f(x) \leq \rho$. It follows then:

$$\int_{\mathbb{F}_1\cup\mathbb{F}_2} f_{\mathbb{A}}(x)dx \geq \int_{\mathbb{F}'} h(x)\frac{f(x)}{1-\kappa}dx$$

■

Lemma 11: Let $f, g : \mathbb{R} \to \mathbb{R}$ be two probability density functions such that $f \succ g$. For any non- zero constant a, define the following probability density functions:

$$\tilde{f}(x) \overset{def}{=} \frac{1}{|a|} f\left(\frac{x}{a}\right)$$
$$\tilde{g}(x) \overset{def}{=} \frac{1}{|a|} g\left(\frac{x}{a}\right)$$

Under the definitions above, $\tilde{f} \succ \tilde{g}$ holds.

Remark 14: We notice that Lemma 11 is well posed since $\tilde{f}$ and $\tilde{g}$ are also probability density functions. If f is the probability density function of a random variable X, then $\tilde{f}$ is the probability density function of the random variable aX.

Proof: For a set $\mathbb{A} \subset \mathbb{R}$ and for a strictly positive constant α, define the set $\alpha\mathbb{A} = \left\{x \in \mathbb{R} : \frac{1}{\alpha}x \in \mathbb{A}\right\}$. Assume a to be positive and let $\mathbb{F}'$ be a set of positive and finite Lebesgue measure.

$$\int_{\mathbb{F}'} \tilde{g}(x)dx = \int_{\frac{1}{a}\mathbb{F}'} g(x)adx$$

since $f \succ g$, there exists a set $\mathbb{F}''$ with the same Lebesgue measure as $\frac{1}{a}\mathbb{F}'$ such that:

$$\int_{\frac{1}{a}\mathbb{F}'} g(x)adx \leq \int_{\mathbb{F}''} f(x)adx = \int_{a\mathbb{F}''} \tilde{f}(x)dx$$

Choose $\mathbb{F} = a\mathbb{F}''$, clearly, $\mathbb{F}$ and $\mathbb{F}'$ have the same Lebesgue measure, then it follows that:

$$\int_{\mathbb{F}'} \tilde{g}(x)dx \leq \int_{\mathbb{F}} \tilde{f}(x)dx$$

which implies that $\tilde{g} \prec \tilde{f}$. Similar arguments hold for a negative. ■

From the Riesz's rearrangement inequality, Hajek states and proves in [1] the following result:

Lemma 12: [1, Page 619] Let f and g be probability density functions defined on the real line, such that, f is neat and even, and $f \succ g$. Let h be a non-negative, symmetric and non-increasing function. The following holds:

$$\int_{\mathbb{R}} h(x)g(x)dx \leq \int_{\mathbb{R}} h(x)f(x)dx \tag{58}$$

In order to prove Lemma 4, we state the following Lemma.

Lemma 13: Let f be a neat and even probability density function on the real line, Let g, be a probability density function on the real line, such that $g \prec f$. Let h be a positive, even and quasiconvex function. Then the following holds:

$$\int_{\mathbb{R}} h(x)f(x)dx \leq \int_{\mathbb{R}} h(x-y)g(x)dx \tag{59}$$

where y is any real number.

Proof: Let c be a positive real number and define the functions:

$$h_c(x) = c - \min\left(c, h(x)\right)$$

$$h_c(x,y) = c - \min\left(c, h(x-y)\right)$$

for any real number y. We notice that the function h_c is symmetric and non-increasing, it is then immediate, that $h_c = h_c^\sigma$ and $h_c = h_c^\sigma(\cdot, y)$ for all real numbers y. The following inequalities are true:

$$\int_{\mathbb{R}} h_c(x,y)g(x)dx \leq \int_{\mathbb{R}} h_c(x)g^\sigma(x)dx \leq \int_{\mathbb{R}} h_c(x)f(x)dx$$

for any $y \in \mathbb{R}$. The first inequality follows from the Hardy-Littlewood inequality (7), while the second inequality follows from Lemma 12. It follows that:

$$\begin{gathered}\int_{\mathbb{R}} h_c(x,y)g(x)dx \leq \int_{\mathbb{R}} h_c(x)f(x)dx \Rightarrow \\ \int_{\mathbb{R}} \left(c - \min\left(c, h(x-y)\right)\right) g(x)dx \leq \int_{\mathbb{R}} \left(c - \min\left(c, h(x)\right)\right) f(x)dx \Rightarrow \\ \int_{\mathbb{R}} \min\left(c, h(x-y)\right) g(x)dx \geq \int_{\mathbb{R}} \min\left(c, h(x)\right) f(x)dx\end{gathered}$$

Taking the limit as c goes to infinity and using the monotone convergence theorem the result follows. ■

APPENDIX II
QUASICONVEX LEMMA

Lemma 14: Let $h : \mathbb{R} \to \mathbb{R}$, be a measurable, bounded, even and quasiconvex function. Let $\mathbf{W}$ be a random variable with an even and quasiconcave probability density function. Define $\bar{h} : \mathbb{R} \to \mathbb{R}$, such that $\bar{h} \overset{def}{=} E\left[h(x + \mathbf{W})\right]$, then $\bar{h}$ is a bounded, even and quasiconvex function. If the function h is also continuous then $\bar{h}$ is also continuous.

Proof: Define $g : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ as $g(x,c) \overset{def}{=} E\left[c - \min\left(c, h(x + \mathbf{W})\right)\right]$. We will show that the function $g(x,c)$ is continuous in c for every fixed real number x, and for every c the function $g(x,c)$ is even and quasiconcave in x. The function h is even and quasiconvex then, it follows that zero is a global minimizer of h. For any real number c and any real number x define the set $D(x,c) \overset{def}{=} \{w \in \mathbb{R} : h(x+w) \leq c\}$ Since h is even and quasiconvex then $D(0,c)$ is a symmetric interval around zero or the empty set. Note that for $h(0) \leq c < \sup_x h(x)$, the set $D(0,c)$ is a symmetric interval, which can be either closed or open. Hence it follows that:

$$D(x,c) = \begin{cases} \emptyset, & c < h(0) \\ [-\alpha(c) - x, \alpha(c) - x] \text{ or } (-\alpha(c) - x, \alpha(c) - x), & h(0) \leq c < \sup_x h(x) \\ (-\infty, \infty), & \sup_x h(x) \leq c \end{cases}$$

where by $\emptyset$ we denote the empty set and $\alpha(c)$ is the real number such that $h(x) \leq c$ if and only if $0 \leq x \leq \alpha(c)$ ($0 \leq x < \alpha(c)$). We will show that the function $g(x,c)$ is even and quasiconvex in x for any real number c. Let $f : \mathbb{R} \to \mathbb{R}$, be the probability density function of $\mathbf{W}$. We can write $g(x, c)$:

$$g(x,c) = E\left[c - \min(c, h(x + \mathbf{W}))\right] = c\int_{-\alpha(c)-x}^{\alpha(c)-x} f(w)dw - \int_{-\alpha(c)-x}^{\alpha(c)-x} h(x+w)f(w)dw$$

For any positive real number δ, any real numbers c and x, it holds that:

$$\begin{aligned}&\left|E\left[g(x+\mathbf{W}, c+\delta)\right] - E\left[g(x+\mathbf{W}, c)\right]\right| \leq \\ &E\left[\left|\delta + \min(c+\delta, h(x+\mathbf{W})) - \min(c, h(x+\mathbf{W}))\right|\right] \leq 2\delta\end{aligned}$$

It follows that for any real number x and any real number c, for any positive real number ϵ, choose $\delta = \frac{\epsilon}{2}$, then for any real number $\bar{c} \in (c - \delta, c + \delta)$, $|g(x, \bar{c}) - g(x, c)| < \epsilon$, hence the function $g(x, c)$ is a continuous function in c for every real number x.

Since the function h is even and quasiconvex, it follows that the function $c - \min(c, h(x))$ is even and quasiconcave, i.e. is neat and even. Moreover, from the definition of the set $D(0, c)$, we notice that the function $c - \min(c, h(x))$ is non-negative, bounded and takes the value zero outside the set $D(0, c)$. If $c < \sup_x h(x)$, then the set $D(0, c)$ is the empty set or a finite interval (open or closed), it follows that, if $c < \sup_x h(x)$ the function $c - \min(c, h(x))$ is integrable. Hence, it holds that:

$$\begin{aligned}g(x,c) = E\left[c - \min(h(x+\mathbf{W}, c)\right] &= \int_{-\infty}^{\infty}(c - \min(h(x+w, c))f(w)dw \\ = \int_{-\infty}^{\infty}(c - \min(h(x+w, c))f(-w)dw &= \int_{-\infty}^{\infty}(c - \min(h(x-\eta, c))f(\eta)d\eta\end{aligned}$$

The second equality comes from the fact that f is even, while the third equality comes from the change of variable $\eta = -w$. It follows from Lemma 5 and Remark 13 that $g(x, c)$ is a neat and even function for every $c < \sup_x h(x)$. Since $g(x, c)$ is continuous in c it implies that $g(x, c)$ is neat and even for every real c and moreover the function $E\left[\min(c, h(x + \mathbf{W}))\right]$ is even and quasiconvex. From the monotone convergence theorem, it holds that:

$$\bar{h}(x) = \lim_{c\to\infty} E\left[\min(h(x+\mathbf{W}), c)\right]$$

and the properties of $E\left[\min(h(x + \mathbf{W}), c)\right]$ in x are kept for $\bar{h}$, i.e. $\bar{h}$ is even and quasiconvex. Since h is bounded, it follows that $\bar{h}$ is bounded and we only need to prove the continuity of $\bar{h}$. We are given that h is even and quasiconvex, which implies that h is non-decreasing on $[0, \infty)$ and non-increasing on $(-\infty, 0]$. We are also given that h is bounded and continuous, which implies that h is uniform continuous on the interval $[0,\infty)$ and is also uniform continuous on the interval $(-\infty,0]$. It follows that the entire function h is uniform continuous, i.e. for any real number x, for any positive real number ϵ, there exists a positive real number δ, which does not depend on x, such that for any real number $y \in (x-\delta, x+\delta)$, it holds that $|h(x)-h(y)| < \epsilon$. It follows that, for any real number x and for any real number $y \in (x-\delta, x+\delta)$, it holds that:

$$\begin{aligned}|E\left[h(x+\mathbf{W})\right] - E\left[h(y+\mathbf{W})\right]| &= |\int_{-\infty}^{\infty} h(x+w)f(w)dw - \int_{-\infty}^{\infty} h(y+w)f(w)dw| \\ &\leq \int_{-\infty}^{\infty} |h(x+w)-h(y+w)|f(w)dw \leq \epsilon\end{aligned}$$

This implies that $\bar{h}$ is continuous. ■

References

[1] Bruce Hajek, Kevin Mitzel and Sichao Yang, "*Paging and Registration in Cellular Networks: Jointly Optimal Policies and an Iterative Algorithm,*" IEEE Transactions on Information Theory, vol. 54, no. 2, Feb 2008, pp 608-622

[2] F. Riesz, "*Sur une inegalite integrale,*" J. London Math. Soc., vol. 5, pp. 162168, 1930.

[3] A. Mahajan, D. Teneketzis, "*Optimal Design of Sequential Real-Time Communication Systems,*" IEEE Transactions on Information Theory, Vol 55, No 11, November 2009

[4] A. Mahajan, " *Sequential decomposition of sequential teams: applications to real-time communication and networked control systems,*" PhD Dissertation, University of Michigan, September 2008

[5] G. E. Hardy, J. E. Littlewood, and G. Polya, "*Inequalities.*" First / second edition, Cambridge University Press, London and New York, 1934 / 1952.

[6] Albert W. Marshall, Ingram Olkin, Barry Arnold, "*Inequalities: Theory of Majorization and Its Applications (Springer Series in Statistics),*" Academic Press, New York, 569 pp

[7] O. C. Imer and T. Basar. "*Optimal estimation with limited measurements,*" in Proc. IEEE CDC/ECC 2005, Seville, Spain submitted to IEEE Transactions on Automatic Control

[8] Y. Xu and J. Hespanha, "*Optimal Communication Logics for Networked Control Systems,*" Proc. of the 43rd Conf. on Decision and Contr., Dec. 2004

[9] J.S. Baras and A. Bensoussan, "*Optimal Sensor Scheduling in Nonlinear Filtering of Diffusion Processes,*" SIAM Journal on Control and Optimization, Vol. 27, No. 2, pp. 786-814, July 1989.

[10] Wei Wu and Ari Arapostathis, "*Optimal Sensor Querying: General Markovian and LQG Models With Controlled Observations,*" IEEE Transactions on Automatic Control, vol. 53, no. 6, July 2008, pp 1392-1405.

[11] Michael Athans, "*On the determination of optimal costly measurement strategies for linear stochastic systems,*" Automatica, vol. 8, 1972, pp 397-412.

[12] D. Bertsekas, "*Dynamic Programming and Optimal Control,*" Athena Scientific, 3rd edition, Vol. I 2005, Vol. II 2007

[13] D. Bertsekas and S. E. Shreve, "*Stochastic Optimal Control: The Discrete Time Case,*" Athena Scientific, 1996

[14] M. Rabi, G.V. Moustakides and J. S. Baras, "*Multiple Sampling for Real-time Estimation on a Finite Horizon,*" Proc. of the 45th Conf. on Decision and Contr., Dec. 2006, pp 1351-1357

[15] M. Rabi, "*Packet Based Inference and Control,*" Ph.D. Dissertation, University of Maryland, 2006