DIMENSIONALITY REDUCTION, COMPRESSION AND QUANTIZATION FOR DISTRIBUTED ESTIMATION WITH WIRELESS SENSOR NETWORKS*

IOANNIS D. SCHIZAS†, ALEJANDRO RIBEIRO†,
AND GEORGIOS B. GIANNAKIS†

Abstract. The distributed nature of observations collected by inexpensive wireless sensors necessitates transmission of the individual sensor data under stringent bandwidth and power constraints. These constraints motivate: i) a means of reducing the dimensionality of local sensor observations; ii) quantization of sensor observations prior to digital transmission; and iii) estimators based on the quantized digital messages. These three problems are addressed in the present paper. We start deriving linear estimators of stationary random signals based on reduced-dimensionality observations. For uncorrelated sensor data, we develop mean-square error (MSE) optimal estimators in closed-form; while for correlated sensor data, we derive sub-optimal iterative estimators which guarantee convergence at least to a stationary point. We then determine lower and upper bounds for the Distortion-Rate (D-R) function and a novel alternating scheme that numerically determines an achievable upper bound of the D-R function for general distributed estimation using multiple sensors. We finally derive distributed estimators based on binary observations along with their fundamental error-variance limits for pragmatic signal models including: i) known univariate but generally non-Gaussian noise probability density functions (pdfs); ii) known noise pdfs with a finite number of unknown parameters; and iii) practical generalizations to multivariate and possibly correlated pdfs. Estimators utilizing either independent or colored binary observations are developed, analyzed and tested with numerical examples.

Key words. Wireless Sensor Networks, Distributed Parameter Estimation, Distributed Compression, Canonical Correlation Analysis, Distortion-Rate Analysis, Quantization, Estimation.

AMS(MOS) subject classifications. Primary 68W15, 62G05, 68P30, 90B15.

1. Introduction. Wireless sensor networks (WSNs) consist of low-cost energy limited transceiver nodes spatially deployed in large numbers to accomplish monitoring, surveillance and control tasks through cooperative actions [14]. The potential of WSNs for surveillance has by now been well appreciated especially in the context of data fusion and distributed detection; e.g., [32, 33] and references therein. However, except for recent works where spatial correlation is exploited to reduce the amount of information exchanged among nodes [2, 5, 9, 11, 15, 21, 25, 26], use of WSNs for

*Prepared through collaborative participation in the Communications and Networks Consortium sponsored by the U. S. Army Research Laboratory under the Collaborative Technology Alliance Program, Cooperative Agreement DAAD19-01-2-0011. The U. S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation thereon.

†Department of Electrical and Computer Engineering, University of Minnesota, 200 Union Street SE, Minneapolis, MN 55455. Tel/fax: 612-626-7781/612-625-4583 ({schizas,aribeiro,georgios}@ece.umn.edu.)

the equally important problem of distributed parameter estimation remains a largely uncharted territory.

While a number of statistical and information theoretic tools have been developed over the years, the unique characteristics of WSNs require rethinking of many algorithms traditionally designed for centralized estimation. Indeed, the distributed nature of the observations necessitates transmission of the individual sensor data; moreover, the power/bandwidth available for transmission and signal processing is severely limited. To complicate matters even more the parametric data models used and the knowledge of sensor noise distributions are not easy to characterize; observations taken by (small and inexpensive) sensors are very noisy; and the WSN size and topology may change dynamically.

To appreciate the challenges implied by these constraints, consider a mean-location parameter estimation problem with sensors collecting observations in order to estimate a parameter in additive zero-mean noise. The distributed nature of the observations necessitates transmission of the individual sensor data under stringent bandwidth and power constraints thus requiring: i) a means of combining local sensor observations in order to reduce their dimensionality while keeping the estimation MSE as small as possible; ii) quantization of the combined observations prior to digital transmission; and iii) estimators based on the quantized digital messages, certainly different from estimators based on the original analog-amplitude observations.

Overcoming the limitations of nonlinear/nonGaussian data models and non-ideal channel links, one of the major goals in this paper is to form estimates at the fusion center (FC) of a *random* stationary vector based on *analog-amplitude* multi-sensor observations. To enable estimation under the stringent power and computing limitations of WSNs, we seek linear dimensionality reducing operators (data compressing matrices) per sensor along with linear operators at the FC, in order to minimize the mean-square error (MSE) in estimation. If sufficiently strong error-control codes are used, we can treat links as ideal and formulate this intertwined compression-estimation task as a canonical correlation analysis problem [31]. Here, we explicitly account for non-ideal links and develop distributed estimators generally applicable to nonlinear and non-Gaussian setups (Section 2). We start by deriving in closed-form the MSE optimal matrices for compression and estimation when the sensor data are uncorrelated (Section 2.1), and we prove that the optimal solution amounts to optimally compressing the linear minimum mean-square error (LMMSE) signal estimate formed at each sensor. With correlated (coupled) sensor observations, globally optimal distributed estimation has been shown to be NP-hard when reduced-dimensionality sensor data are concatenated at the FC [18]. For this case, we develop a block coordinate descent iterative estimator (Section 2.2) which always converges to a stationary point and subsumes a recent distributed reconstruction algorithm in [10].

When the sensors are allowed to transmit only digital-amplitude data (due to encoding rate constraints), an issue of paramount importance is to determine bounds on the minimum achievable distortion between the signal of interest and its estimate formed at the FC using the encoded information transmitted by the sensors subject to rate constraints (Distortion-Rate function). In the *reconstruction* scenario, the FC wishes to accurately estimate the sensor observations. In the *estimation* scenario, the FC is interested in accurately estimating an underlying random vector which is correlated with, but not equal to, the sensor observations. In the single sensor setting, single-letter characterizations of the Distortion-Rate (D-R) function for both scenarios are known [8, p. 336], and the estimation problem, which is also referred to as rate-distortion with a remote source, has also been determined [3, p. 78]. In the distributed scenario, where there are multiple sensors with correlated observations, neither problem is well understood. The best analytical inner and outer bounds for the D-R function for reconstruction can be found in [4]. An iterative scheme has been developed in [10], which numerically determines an achievable upper bound for distributed reconstruction but not for signal estimation.

We present this D-R analysis in Section 3. We first determine the D-R function for estimating a *vector* parameter when applying rate-constrained encoding to the observation data, in closed form for the single-sensor case (Section 3.1). Without assuming that the number of parameters equals the number of observations, we prove that the optimal scheme achieving the D-R function amounts to first computing the minimum mean square error (MMSE) estimate at the sensor, and then optimally compressing the estimate via reverse water-filling (rwf). The D-R function for the single-sensor setup serves as a non-achievable lower D-R bound for rate constrained estimation in the multi-sensor setup. Next, we develop an alternating scheme that numerically determines an achievable D-R upper bound for the multi-sensor scenario (Section 3.2). Different from [10], which deals with WSN-based distributed reconstruction, our approach aims for general estimation problems.

Returning to the issue of estimation once the actual observations have been collected at the FC, we study the intertwining between quantization and estimation (Section IV). We begin with mean-location parameter estimation in the presence of known univariate but generally non-Gaussian noise pdfs (Section 4.1.1). We next develop mean-location parameter estimators based on binary observations and benchmark their performance when the noise variance is unknown; however, the same approach in principle applies to any noise pdf that is known except for a finite number of unknown parameters (Section 4.1.2). Subsequently, we move to the most challenging case where the noise pdf is completely unknown (Section 4.2). Finally, we consider vector generalizations where each sensor observes a given (possibly nonlinear) function of the unknown parameter vector in the presence of multivariate and possibly colored noise (Section 4.3). While

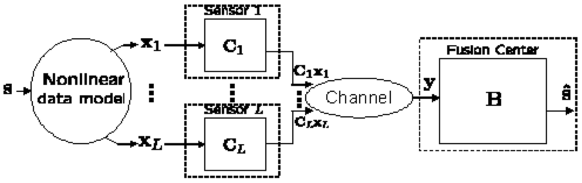

FIG. 1. *Distributed setup for estimating a random signal* $\mathbf{s}$.

challenging in general, it will turn out that under relaxed conditions, the resultant Maximum Likelihood Estimator (MLE) is the maximum of a concave function, thus ensuring convergence of Newton-type iterative algorithms. Moreover, in the presence of colored Gaussian noise, we show that judiciously quantizing each sensor's data renders the estimators' variance stunningly close to the variance of the clairvoyant estimator that is based on the unquantized observations; thus, nicely generalizing the results of Sections 4.1.1, 4.1.2, and [27] to the more realistic vector parameter estimation problem (Section 4.3.1).

Numerical examples corroborate our theoretical findings in Section 5, where we also test them on a motivating application involving distributed parameter estimation with a WSN for measuring a vector flow (Section 5.4). We conclude the paper in Section 6.

2. Dimensionality reduction for distributed estimation.

In this section we develop linear distributed estimators in a setup where the sensors observe and transmit analog-amplitude data.

Consider the WSN depicted in Fig. 1, comprising L sensors linked with an FC. Each sensor, say the ith one, observes an $N_i \times 1$ vector $\mathbf{x}_i$ that is correlated with a $p \times 1$ random signal of interest $\mathbf{s}$. Through a $k_i \times N_i$ fat matrix $\mathbf{C}_i$ each sensor transmits a compressed $k_i \times 1$ vector $\mathbf{C}_i\mathbf{x}_i$, using e.g., multicarrier modulation with one entry riding per subcarrier. Low-power and bandwidth constraints at the sensors encourage transmissions with $k_i \ll N_i$, while linearity in compression and estimation are well motivated by low-complexity requirements. Furthermore, we assume that:

(a1) No information is exchanged among sensors, and each sensor-FC link comprises a $k_i \times k_i$ full rank fading multiplicative channel matrix $\mathbf{D}_i$ along with zero-mean additive FC noise $\mathbf{z}_i$, which is uncorrelated with $\mathbf{x}_i$, $\mathbf{D}_i$, and across channels; i.e., noise covariance matrices satisfy $\mathbf{\Sigma}_{z_i z_j} = \mathbf{0}$ for $i \neq j$. Matrices $\{\mathbf{D}_i, \mathbf{\Sigma}_{z_i z_i}\}_{i=1}^{L}$ are available at the FC.

(a2) Data $\mathbf{x}_i$ and the signal of interest $\mathbf{s}$ are zero-mean with full rank auto- and cross-covariance matrices $\mathbf{\Sigma}_{ss}$, $\mathbf{\Sigma}_{sx_i}$ and $\mathbf{\Sigma}_{x_i x_j}$ $\forall\, i, j \in [1, L]$, all of which are available at the FC.

In multicarrier links, full rank of the channel matrices $\{\mathbf{D}_i\}_{i=1}^{L}$ is ensured if sensors do not transmit over subcarriers with zero channel gain.

Matrices $\{\mathbf{D}_i\}_{i=1}^L$ can be acquired via training, and likewise the signal and noise covariances in (a1) and (a2) can be estimated via sample averaging as usual. With multicarrier (and generally any orthogonal) sensor access, the noise uncorrelatedness across channels is also well justified. Notice that unlike [10, 18, 37, 38], we neither confine ourselves to a linear signal-plus-noise model $\mathbf{x}_i = \mathbf{H}\mathbf{s} + \mathbf{n}_i$, nor we invoke any assumption on the distribution (e.g., Gaussianity) of $\{\mathbf{x}_i\}_{i=1}^L$ and $\mathbf{s}$. Equally important, we do not assume ideal channel links.

Sensors transmit over orthogonal channels so that the FC separates and concatenates the received vectors $\{\mathbf{y}_i(\mathbf{C}_i) = \mathbf{D}_i\mathbf{C}_i\mathbf{x}_i + \mathbf{z}_i\}_{i=1}^L$, to obtain the $\sum_{i=1}^L k_i \times 1$ vector:

$$\mathbf{y}(\mathbf{C}_1, \ldots, \mathbf{C}_L) = \text{diag}(\mathbf{D}_1\mathbf{C}_1, \ldots, \mathbf{D}_L\mathbf{C}_L)\mathbf{x} + \mathbf{z}, \tag{2.1}$$

Left multiplying $\mathbf{y}$ by a $p \times (\sum_{i=1}^L k_i)$ matrix $\mathbf{B}$, we form the linear estimate $\hat{\mathbf{s}}$ of $\mathbf{s}$. For a prescribed power P_i per sensor, our problem is to obtain under (a1)-(a2) MSE optimal matrices $\{\mathbf{C}_i^o\}_{i=1}^L$ and $\mathbf{B}^o$; i.e., we seek:

$$\begin{gathered}(\mathbf{B}^o, \{\mathbf{C}_i^o\}_{i=1}^L) = \arg \min_{\mathbf{B}, \{\mathbf{C}_i\}_{i=1}^L} E[\|\mathbf{s} - \mathbf{B}\mathbf{y}(\mathbf{C}_1, \ldots, \mathbf{C}_L)\|^2], \\ \text{s. to } \text{tr}(\mathbf{C}_i\mathbf{\Sigma}_{x_ix_i}\mathbf{C}_i^T) \leq P_i, \quad i \in \{1, \ldots, L\}.\end{gathered} \tag{2.2}$$

2.1. Decoupled distributed estimation. We consider first the case where $\mathbf{\Sigma}_{x_ix_j} \equiv \mathbf{0}$, $\forall i \neq j$, which shows up e.g., when matrices $\{\mathbf{H}_i\}_{i=1}^L$ in the linear model $\mathbf{x}_i = \mathbf{H}_i\mathbf{s} + \mathbf{n}_i$ are mutually uncorrelated and also uncorrelated with the noise vectors $\mathbf{n}_i$. Then, the multi-sensor optimization task in (2.2) reduces to a set of L decoupled problems. Specifically, it is easy to show that the cost function in (2.2) can be written as [31]:

$$J(\mathbf{B}, \{\mathbf{C}_i\}_{i=1}^L) = \textstyle\sum_{i=1}^L E[\|\mathbf{s} - \mathbf{B}_i(\mathbf{D}_i\mathbf{C}_i\mathbf{x}_i + \mathbf{z}_i)\|^2] - (L-1)\text{tr}(\mathbf{\Sigma}_{ss}), \tag{2.3}$$

where $\mathbf{B}_i$ is the $p \times k_i$ submatrix of $\mathbf{B} := [\mathbf{B}_1 \ldots \mathbf{B}_L]$. As the ith nonnegative summand depends only on $\mathbf{B}_i, \mathbf{C}_i$ the MSE optimal matrices are given by

$$\begin{gathered}(\mathbf{B}_i^o, \mathbf{C}_i^o) = \arg \min_{\mathbf{B}_i, \mathbf{C}_i} E[\|\mathbf{s} - \mathbf{B}_i(\mathbf{D}_i\mathbf{C}_i\mathbf{x}_i + \mathbf{z}_i)\|^2], \\ \text{s. to } \ \text{tr}(\mathbf{C}_i\mathbf{\Sigma}_{x_ix_i}\mathbf{C}_i^T) \leq P_i, \ \ i \in \{1, \ldots, L\}.\end{gathered} \tag{2.4}$$

Since the cost function in (2.4) corresponds to a single-sensor setup ($L = 1$), we will drop the subscript i for notational brevity and write $\mathbf{B}_i = \mathbf{B}, \mathbf{C}_i = \mathbf{C}, \mathbf{x}_i = \mathbf{x}, \mathbf{z}_i = \mathbf{z}, P = P_i$ and $k = k_i$. The Lagrangian for minimizing (2.3) can be easily written as:

$$\begin{aligned}J(\mathbf{B}, \mathbf{C}, \mu) = J_o &+ \text{tr}(\mathbf{B}\mathbf{\Sigma}_{zz}\mathbf{B}^T) + \mu[\text{tr}(\mathbf{C}\mathbf{\Sigma}_{xx}\mathbf{C}^T) - P] \\ &+ \text{tr}[(\mathbf{\Sigma}_{sx} - \mathbf{B}\mathbf{D}\mathbf{C}\mathbf{\Sigma}_{xx})\mathbf{\Sigma}_{xx}^{-1}(\mathbf{\Sigma}_{xs} - \mathbf{\Sigma}_{xx}\mathbf{C}^T\mathbf{D}^T\mathbf{B}^T)],\end{aligned} \tag{2.5}$$

where $J_o := \text{tr}(\mathbf{\Sigma}_{ss} - \mathbf{\Sigma}_{sx}\mathbf{\Sigma}_{xx}^{-1}\mathbf{\Sigma}_{xs})$ is the minimum attainable MMSE for linear estimation of $\mathbf{s}$ based on $\mathbf{x}$. Continuing, we derive a simplified form of (2.5) the minimization of which will provide closed-form solutions for the MSE optimal matrices $\mathbf{B}^o$ and $\mathbf{C}^o$.

Aiming at this simplification, consider the SVD $\mathbf{\Sigma}_{sx} = \mathbf{U}_{sx}\mathbf{S}_{sx}\mathbf{V}_{sx}^T$, and the eigen-decompositions $\mathbf{\Sigma}_{zz} = \mathbf{Q}_z\mathbf{\Lambda}_z\mathbf{Q}_z^T$ and $\mathbf{D}^T\mathbf{\Sigma}_{zz}^{-1}\mathbf{D} = \mathbf{Q}_{zd}\mathbf{\Lambda}_{zd}\mathbf{Q}_{zd}^T$, where $\mathbf{\Lambda}_{zd} := \text{diag}(\lambda_{zd,1}\cdots\ \lambda_{zd,k})$ and $\lambda_{zd,1} \geq \cdots \geq \lambda_{zd,k} > 0$. Notice, that $\lambda_{zd,i}$ captures the SNR of the ith entry in the received signal vector at the FC. Further, define $\mathbf{A} := \mathbf{Q}_x^T\mathbf{V}_{sx}\mathbf{S}_{sx}^T\ \mathbf{S}_{sx}\mathbf{V}_{sx}^T\mathbf{Q}_x$ with $\rho_a := \text{rank}(\mathbf{A}) = \text{rank}(\mathbf{\Sigma}_{sx})$, and $\mathbf{A}_x := \mathbf{\Lambda}_x^{-1/2}\mathbf{A}\mathbf{\Lambda}_x^{-1/2}$ with corresponding eigen-decomposition $\mathbf{A}_x = \mathbf{Q}_{ax}\mathbf{\Lambda}_{ax}\mathbf{Q}_{ax}$, where $\mathbf{\Lambda}_{ax} = \text{diag}(\lambda_{ax,1},\cdots,\lambda_{ax,\rho_a}, 0,\cdots,0)$ and $\lambda_{ax,1} \geq \ldots \geq \lambda_{ax,\rho_a} > 0$. Moreover, let $\mathbf{V}_a := \mathbf{\Lambda}_x^{-1/2}\mathbf{Q}_{ax}$ denote the invertible matrix which simultaneously diagonalizes the matrices $\mathbf{A}$ and $\mathbf{\Lambda}_x$. Since matrices $(\mathbf{Q}_{zd}, \mathbf{Q}_x, \mathbf{V}_a, \mathbf{U}_{sx}, \mathbf{\Lambda}_{zd}, \mathbf{Q}_{zd}, \mathbf{D}, \mathbf{\Sigma}_{zz})$ are all invertible, for every matrix $\mathbf{C}$ (or $\mathbf{B}$) we can clearly find a unique matrix $\mathbf{\Phi}_C$ (correspondingly $\mathbf{\Phi}_B$) that satisfies:

$$\mathbf{C} = \mathbf{Q}_{zd}\mathbf{\Phi}_C\mathbf{V}_a^T\mathbf{Q}_x^T, \quad \mathbf{B} = \mathbf{U}_{sx}\mathbf{\Phi}_B\mathbf{\Lambda}_{zd}^{-1}\mathbf{Q}_{zd}^T\mathbf{D}^T\mathbf{\Sigma}_{zz}^{-1}, \tag{2.6}$$

where $\mathbf{\Phi}_C := [\phi_{c,ij}]$ and $\mathbf{\Phi}_B$ have sizes $k \times N$ and $p \times k$, respectively. Using (2.6), the Lagrangian in (2.5) becomes:

$$\begin{aligned} J(\mathbf{\Phi}_C, \mu) = J_o + \text{tr}(\mathbf{\Lambda}_{ax}) + \mu(\text{tr}(\mathbf{\Phi}_C\mathbf{\Phi}_C^T) - P) \\ -\text{tr}\left((\mathbf{\Lambda}_{zd}^{-1} + \mathbf{\Phi}_C\mathbf{\Phi}_C^T)^{-1}\mathbf{\Phi}_C\mathbf{\Lambda}_{ax}\mathbf{\Phi}_C^T\right). \end{aligned} \tag{2.7}$$

Applying the well known Karush-Kuhn-Tucker (KKT) conditions (e.g., [6, Ch. 5]) that must be satisfied at the minimum of (2.7), we prove in [31] that the matrix $\mathbf{\Phi}_C^o$ minimizing (2.7), is diagonal with diagonal entries:

$$\phi_{c,ii}^o = \begin{cases} \pm\sqrt{\left(\frac{\lambda_{ax,i}}{\mu^o\lambda_{zd,i}}\right)^{1/2} - \frac{1}{\lambda_{zd,i}}}, & 1 \leq i \leq \kappa \\ 0, & \kappa + 1 \leq i \leq k \end{cases}, \tag{2.8}$$

where κ is the maximum integer in $[1, k]$ for which $\{\phi_{c,ii}^o\}_{i=1}^{\kappa}$ are strictly positive, or, $\text{rank}(\mathbf{\Phi}_C^o) = \kappa$; and μ^o is chosen to satisfy the power constraint $\sum_{i=1}^{\kappa}(\phi_{c,ii}^o)^2 = P$ as:

$$\mu^o = \frac{(\sum_{i=1}^{\kappa}(\lambda_{ax,i}\lambda_{zd,i}^{-1})^{1/2})^2}{(P + \sum_{i=1}^{\kappa}\lambda_{zd,i}^{-1})^2}. \tag{2.9}$$

When $k > \rho_a$, the MMSE remains invariant [31]; thus, it suffices to consider $k \in [1, \rho_a]$. Summarizing, we have established that:

Proposition 2.1. *Under (a1), (a2), and for $k \leq \rho_a$, the matrices minimizing $J(\mathbf{B}_{p\times k}, \mathbf{C}_{k\times N}) = E[\|\mathbf{s} - \mathbf{B}_{p\times k}(\mathbf{D}\mathbf{C}_{k\times N}\mathbf{x} + \mathbf{z})\|^2]$, subject to tr$(\mathbf{C}_{k\times N}\mathbf{\Sigma}_{xx}\ \mathbf{C}_{k\times N}^T) \leq P$, are:*

$$\begin{aligned}\mathbf{C}^o &= \mathbf{Q}_{zd}\mathbf{\Phi}_C^o\mathbf{V}_a^T\mathbf{Q}_x^T,\\ \mathbf{B}^o &= \mathbf{\Sigma}_{sx}\mathbf{Q}_x\mathbf{V}_a\mathbf{\Phi}_C^{o\,T}\left(\mathbf{\Phi}_C^o\mathbf{\Phi}_C^{o\,T}+\mathbf{\Lambda}_{zd}^{-1}\right)^{-1}\mathbf{\Lambda}_{zd}^{-1}\mathbf{Q}_{zd}^T\mathbf{D}^T\mathbf{\Sigma}_{zz}^{-1},\end{aligned} \tag{2.10}$$

where $\mathbf{\Phi}_C^o$ is given by (2.8), and the corresponding Lagrange multiplier μ^o is specified by (2.9). The MMSE is

$$J_{\min}(k) = J_o + \sum_{i=1}^{\rho_a}\lambda_{ax,i} - \sum_{i=1}^{k}\frac{\lambda_{ax,i}(\phi_{c,ii}^o)^2}{\lambda_{zd,i}^{-1}+(\phi_{c,ii}^o)^2}. \tag{2.11}$$

According to Proposition 1, the optimal weight matrix $\mathbf{\Phi}_C^o$ in $\mathbf{C}^o$ distributes the given power across the entries of the pre-whitened vector $\mathbf{V}_a^T\mathbf{Q}_x\mathbf{x}$ at the sensor in a waterfilling-like manner so as to balance channel strength and additive noise variance at the FC with the degree of dimensionality reduction that can be afforded. It is worth mentioning that (2.8) dictates a minimum power per sensor. Specifically, in order to ensure that $\text{rank}(\mathbf{\Phi}_C^o) = \kappa$ the power must satisfy:

$$P > \frac{\sum_{i=1}^{\kappa}(\lambda_{ax,i}\lambda_{zd,i}^{-1})^{1/2}}{\sqrt{\lambda_{ax,\kappa}\lambda_{zd,\kappa}}} - \sum_{i=1}^{\kappa}\lambda_{zd,i}^{-1}\ . \tag{2.12}$$

The optimal matrices in Proposition 1 can be viewed as implementing a two-step scheme, where: i) we estimate $\mathbf{s}$ based on $\mathbf{x}$ at the sensor using the LMMSE estimate $\hat{\mathbf{s}}_{LM} = \mathbf{\Sigma}_{sx}\mathbf{\Sigma}_{xx}^{-1}\mathbf{x}$; and ii) compress and reconstruct $\hat{\mathbf{s}}_{LM}$ using the optimal matrices $\mathbf{C}^o$ and $\mathbf{B}^o$ implied by Proposition 1 after replacing $\mathbf{x}$ with $\hat{\mathbf{s}}_{LM}$. For this estimate-first compress-afterwards (EC) interpretation, we prove in [31] that:

Corollary 2.1. *For $k \in [1, \rho_a]$, the $k \times N$ matrix in (2.10) can be written as $\mathbf{C}^o = \hat{\mathbf{C}}^o\mathbf{\Sigma}_{sx}\mathbf{\Sigma}_{xx}^{-1}$, where $\hat{\mathbf{C}}^o$ is the $k \times p$ optimal matrix obtained by Proposition 1 when $\mathbf{x} = \hat{\mathbf{s}}_{LM}$. Thus, the EC scheme is MSE optimal in the sense of minimizing (2.3).*

Another interesting feature of the EC scheme implied by Proposition 1 is that the MMSE $J_{\min}(k)$ is non-increasing with respect to the reduced dimensionality k, given a limited power budget per sensor. Specifically, we establish in [31] that:

Corollary 2.2. *If $\mathbf{C}_{k_1\times N}^o$ and $\mathbf{C}_{k_2\times N}^o$ are the optimal matrices determined by Proposition 1 with $k_1 < k_2$, under the same channel parameters $\lambda_{zd,i}$ for $i = 1, \ldots, k_1$, and common power P, the MMSE in (2.11) is non-increasing; i.e., $J_{\min}(k_1) \geq J_{\min}(k_2)$ for $k_1 < k_2$.*

Notice that Corollary 2 advocates the efficient power allocation that the EC-n scheme performs among the compressed components.

2.2. Coupled distributed estimation. In this section, we allow the sensor observations to be correlated. Because $\mathbf{\Sigma}_{xx}$ is no longer block diagonal, decoupling of the multi-sensor optimization problem cannot be effected in this case. The pertinent MSE cost is [c.f. (2.2)]:

$$J(\{\mathbf{B}_i,\mathbf{C}_i\}_{i=1}^L) = E[\|\mathbf{s} - \textstyle\sum_{i=1}^L \mathbf{B}_i(\mathbf{D}_i\mathbf{C}_i\mathbf{x}_i + \mathbf{z}_i)\|^2]. \tag{2.13}$$

Minimizing (2.13) does not lead to a closed-form solution and incurs complexity that grows exponentially with L [18]. For this reason, we resort to iterative alternatives which converge at least to a stationary point of the cost in (2.13). To this end, let us suppose temporarily that matrices $\{\mathbf{B}_l\}_{l=1,l\neq i}^L$ and $\{\mathbf{C}_l\}_{l=1,l\neq i}^L$ are fixed and satisfy the power constraints $\text{tr}(\mathbf{C}_l\mathbf{\Sigma}_{x_lx_l}\mathbf{C}_l^T) = P_l$, for $l = 1,\dots,L$ and $l \neq i$. Upon defining the vector $\bar{\mathbf{s}}_i := \mathbf{s} - \sum_{l=1,l\neq i}^L(\mathbf{B}_l\mathbf{D}_l\mathbf{C}_l\mathbf{x}_l + \mathbf{B}_l\mathbf{z}_l)$ the cost in (2.13) becomes:

$$J(\mathbf{B}_i,\mathbf{C}_i) = E[\|\bar{\mathbf{s}}_i - \mathbf{B}_i\mathbf{D}_i\mathbf{C}_i\mathbf{x}_i - \mathbf{B}_i\mathbf{z}_i\|^2]\ , \tag{2.14}$$

which being a function of $\mathbf{C}_i$ and $\mathbf{B}_i$ only, falls under the realm of Proposition 1. This means that when $\{\mathbf{B}_l\}_{l=1,l\neq i}^L$ and $\{\mathbf{C}_l\}_{l=1,l\neq i}^L$ are given, the matrices $\mathbf{B}_i$ and $\mathbf{C}_i$ minimizing (2.14) under the power constraint $\text{tr}(\mathbf{C}_i\mathbf{\Sigma}_{x_ix_i}\mathbf{C}_i^T) \leq P_i$ can be directly obtained from (2.10), after setting $\mathbf{s} = \bar{\mathbf{s}}_i$, $\mathbf{x} = \mathbf{x}_i$, $\mathbf{z} = \mathbf{z}_i$ and $\rho_a = \text{rank}(\mathbf{\Sigma}_{\bar{s}_ix_i})$ in Proposition 1. The corresponding auto- and cross- covariance matrices needed must also be modified appropriately, namely $\mathbf{\Sigma}_{ss} = \mathbf{\Sigma}_{\bar{s}_i\bar{s}_i}$ and $\mathbf{\Sigma}_{sx_i} = \mathbf{\Sigma}_{\bar{s}_ix_i}$. We have thus established the following result for coupled sensor observations:

Proposition 2.2. *If (a1) and (a2) are satisfied, and* $k_i \leq$ *rank*$(\mathbf{\Sigma}_{\bar{s}_ix_i})$*, then for given matrices* $\{\mathbf{B}_l\}_{l=1,l\neq i}^L$ *and* $\{\mathbf{C}_l\}_{l=1,l\neq i}^L$ *satisfying tr*$(\mathbf{C}_l\mathbf{\Sigma}_{x_lx_l}\mathbf{C}_l^T) = P_l$*, the optimal* $\mathbf{B}_i^o$ *and* $\mathbf{C}_i^o$ *matrices minimizing* $E[\|\mathbf{s} - \sum_{l=1}^L \mathbf{B}_l(\mathbf{D}_l\mathbf{C}_l\mathbf{x}_l + \mathbf{z}_l)\|^2]$ *are provided by Proposition 1, after setting* $\mathbf{x} = \mathbf{x}_i$, $\mathbf{s} = \bar{\mathbf{s}}_i$ *and applying the corresponding covariance modifications.*

Proposition 2 suggests the following alternating algorithm for distributed estimation in the presence of fading and FC noise:

Algorithm 1 :

```
Initialize randomly the matrices $\{\mathbf{C}_i^{(0)}\}_{i=1}^L$ and $\{\mathbf{B}_i^{(0)}\}_{i=1}^L$, such that
$\text{tr}(\mathbf{C}_i^{(0)}\mathbf{\Sigma}_{x_ix_i}\mathbf{C}_i^{(0)^T}) = P_i$.
$n = 0$
repeat
  $n = n + 1$
  for $i = 1,L$ do
    Given the matrices $\mathbf{C}_1^{(n)},\mathbf{B}_1^{(n)},\dots,\mathbf{C}_{i-1}^{(n)},\mathbf{B}_{i-1}^{(n)},\mathbf{C}_{i+1}^{(n-1)},\mathbf{B}_{i+1}^{(n-1)}$,
    $\dots,\mathbf{C}_L^{(n-1)},\mathbf{B}_L^{(n-1)}$, determine $\mathbf{C}_i^{(n)},\mathbf{B}_i^{(n)}$ via Th. 2
  end for
until $|\text{MSE}^{(n)} - \text{MSE}^{(n-1)}| < \epsilon$ for given tolerance $\epsilon$
```

Notice that Algorithm 1 belongs to the class of block coordinate descent iterative schemes. At every step i during the nth iteration, it yields the optimal pair of matrices $\mathbf{C}_i^o, \mathbf{B}_i^o$, treating the rest as given. Thus, the $\text{MSE}^{(n)}$ cost per iteration is non-increasing and the algorithm always converges

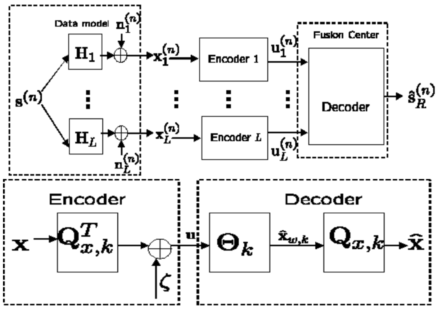

FIG. 2. *(Left): Distributed setup.; (Right): Test channel for* $\mathbf{x}$ *Gaussian in a point-to-point link.*

to a stationary point of (2.13). Beyond its applicability to possibly non-Gaussian and nonlinear model settings, it is the only available algorithm for handling fading and generally colored FC noise effects in distributed estimation.

3. Distortion-rate analysis for distributed estimation. In contrast to the previous section, here we consider digital-amplitude data transmission (bits) from the sensors to the FC. In such a setup, all the sensors are characterized by a rate constraint. In order to determine the minimum possible distortion (MSE) between the signal of interest and the estimate at the FC, under encoding rate constraints, we perform D-R analysis and determine bounds for the D-R function.

With reference to Fig. 2 (Left), consider a WSN comprising L sensors that communicate with an FC. Each sensor, say the ith, observes an $N_i \times 1$ vector $\mathbf{x}_i(t)$ which is correlated with a $p \times 1$ random signal (parameter vector) of interest $\mathbf{s}(t)$, where t denotes discrete time. Similar to [22,23,34], we assume that:

- **(a3)** No information is exchanged among sensors and the links with the FC are noise-free.
- **(a4)** The random vector $\mathbf{s}(t)$ is generated by a stationary Gaussian vector memoryless source with $\mathbf{s}(t) \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_{ss})$; the sensor data $\{\mathbf{x}_i(t)\}_{i=1}^{L}$ adhere to the linear-Gaussian model $\mathbf{x}_i(t) = \mathbf{H}_i\mathbf{s}(t) + \mathbf{n}_i(t)$, where $\mathbf{n}_i(t)$ denotes additive white Gaussian noise (AWGN); i.e., $\mathbf{n}_i(t) \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$; noise $\mathbf{n}_i(t)$ is uncorrelated across sensors, time and with $\mathbf{s}$; and $\mathbf{H}_i$ as well as (cross-) covariance matrices $\mathbf{\Sigma}_{ss}$, $\mathbf{\Sigma}_{sx_i}$ and $\mathbf{\Sigma}_{x_ix_j}$ are known $\forall i,j \in \{1,\ldots,L\}$.

Notice that (a3) assumes that sufficiently strong channel codes are used; while whiteness of $\mathbf{n}_i(t)$ and the zero-mean assumptions in (a4) are made without loss of generality. The linear model in (a4) is commonly encountered in estimation and in a number of cases it even accurately approximates non-linear mappings; e.g., via a first-order Taylor expansion in target tracking applications. Although confining ourselves to Gaussian vectors $\mathbf{x}_i(t)$ is of interest on its own, following arguments similar to those in [3, p. 134] we can show that the D-R functions obtained in this paper bound from above their counterparts for non-Gaussian sensor data $\mathbf{x}_i(t)$.

Blocks $\mathbf{x}_i^{(n)} := \{\mathbf{x}_i(t)\}_{t=1}^n$, comprising n consecutive time instantiations of the vector $\mathbf{x}_i(t)$, are encoded per sensor to yield each encoder's output $\mathbf{u}_i^{(n)} = \mathbf{f}_i^{(n)}(\mathbf{x}_i^{(n)})$, $i = 1, \ldots, L$. These outputs are communicated through ideal orthogonal channels to the FC. There, $\mathbf{u}_i^{(n)}$'s are decoded to obtain an estimate of $\mathbf{s}^{(n)} := \{\mathbf{s}(t)\}_{t=1}^n$ denoted as $\hat{\mathbf{s}}_R^{(n)}(\mathbf{u}_1^{(n)}, \ldots, \mathbf{u}_L^{(n)}) = \mathbf{g}_R^{(n)}(\mathbf{x}_1^{(n)}, \ldots, \mathbf{x}_L^{(n)})$, since $\mathbf{u}_i^{(n)}$ is a function of $\mathbf{x}_i^{(n)}$. The rate constraint is imposed through a bound on the cardinality of the range of the sensor encoding functions, i.e., the cardinality of the range of $\mathbf{f}_i^{(n)}$ must be no larger than 2^{nR_i}, where R_i is the available rate at the encoder of the ith sensor. The sum rate satisfies the constraint $\sum_{i=1}^L R_i \leq R$, where R is the total available rate shared by the L sensors. Under this rate constraint, we want to determine the minimum possible MSE distortion $(1/n)\sum_{t=1}^n E[\|\mathbf{s}(t) - \hat{\mathbf{s}}_R(t)\|^2]$ for estimating $\mathbf{s}$ in the limit of infinite blocklength n. When $L = 1$, a single-letter information theoretic characterization is known for the latter, but no simplification is known for the distributed multi-sensor scenario.

3.1. Distortion-rate for centralized estimation. We will first determine the D-R function for estimating $\mathbf{s}$(t) in a *single-sensor* setup. The single-letter characterization of the D-R function in this setup allow us to drop the time index. Here, all $\{\mathbf{x}_i\}_{i=1}^L := \mathbf{x}$ are available to a single sensor, and $\mathbf{x} = \mathbf{H}\mathbf{s} + \mathbf{n}$. We let $\rho := \text{rank}(\mathbf{H})$ denote the rank of matrix $\mathbf{H}$. The D-R function in such a scenario provides a lower (non-achievable) bound on the MMSE that can be achieved in a multi-sensor distributed setup, where each $\mathbf{x}_i$ is observed by a different sensor. Existing works treat the case $N = p$ [29, 35], but here we look for the D-R function regardless of N, p, in the linear-Gaussian model framework.

3.1.1. Background on D-R analysis for reconstruction. The D-R function for encoding $\mathbf{x}$, which has probability density function (pdf) $p(\mathbf{x})$, with rate R at an individual sensor, and reconstructing it (in the MMSE sense) as $\hat{\mathbf{x}}$ at the FC, is given by [8, p. 342]:

$$D_x(R) = \min_{\substack{p(\hat{\mathbf{x}}|\mathbf{x}) \\ I(\mathbf{x};\hat{\mathbf{x}}) \leq R}} E_{p(\hat{\mathbf{x}},\mathbf{x})}[\|\mathbf{x} - \hat{\mathbf{x}}\|^2], \tag{3.1}$$

where $\mathbf{x} \in \mathbb{R}^N$ and $\hat{\mathbf{x}} \in \mathbb{R}^N$, and the minimization is w.r.t. the conditional pdf $p(\hat{\mathbf{x}}|\mathbf{x})$. Let $\boldsymbol{\Sigma}_{xx} = \mathbf{Q}_x \boldsymbol{\Lambda}_x \mathbf{Q}_x^T$ denote the eigenvalue decomposition of $\boldsymbol{\Sigma}_{xx}$, where $\boldsymbol{\Lambda}_x = \text{diag}(\lambda_{x,1} \cdots \lambda_{x,N})$ and $\lambda_{x,1} \geq \cdots \geq \lambda_{x,N} > 0$.

For $\mathbf{x}$ Gaussian, $D_x(R)$ can be determined by applying rwf to the pre-whitened vector $\mathbf{x}_w := \mathbf{Q}_x^T \mathbf{x}$ [8, p. 348]. For a prescribed rate R, it turns out that $\exists\ k$ such that the first k entries $\{\mathbf{x}_w(i)\}_{i=1}^k$ of $\mathbf{x}_w$, are encoded and reconstructed independently from each other using rate $\{R_i = 0.5\log_2(\lambda_{x,i}/d(k,R))\}_{i=1}^k$, where $d(k,R) = \left(\prod_{i=1}^k \lambda_{x,i}\right)^{1/k} 2^{-2R/k}$ with $R = \sum_{i=1}^k R_i$; and the last $N-k$ entries of $\mathbf{x}_w$ are assigned no rate; i.e., $\{R_i = 0\}_{i=k+1}^N$. The corresponding MMSE for encoding $\mathbf{x}_w(i)$, the ith entry of $\mathbf{x}_w$, under a rate constraint R_i, is $D_i = E[\|\mathbf{x}_w(i) - \hat{\mathbf{x}}_w(i)\|^2] = d(k,R)$ when $i = 1,\dots,k$ and $D_i = \lambda_{x,i}$ when $i = k+1\dots,N$. The resultant MMSE (D-R function) is:

$$D_x(R) = E[\|\mathbf{x} - \hat{\mathbf{x}}\|^2] = E[\|\mathbf{x}_w - \hat{\mathbf{x}}_w\|^2] = kd(k,R) + \textstyle\sum_{i=k+1}^N \lambda_{x,i}. \quad (3.2)$$

Especially for $d(k,R)$, it follows that $\max(\{\lambda_{x,i}\}_{i=k+1}^N) \leq d(k,R) < \min\{\lambda_{x,1}, \dots, \lambda_{x,k}\}$. Intuitively, $d(k,R)$ is a threshold distortion determining which entries of $\mathbf{x}_w$ are assigned with nonzero rate. The first k entries of $\mathbf{x}_w$ with variance $\lambda_{x,i} > d(k,R)$ are encoded with non-zero rate, but the last $N-k$ ones are discarded in the encoding procedure (are set to zero).

Associated with the rwf principle is the so called test channel; see e.g., [8, p. 345]. The encoder's MSE optimal output is $\mathbf{u} = \mathbf{Q}_{x,k}^T\mathbf{x}+\boldsymbol{\zeta}$, where $\mathbf{Q}_{x,k}$ is formed by the first k columns of $\mathbf{Q}_x$, and $\boldsymbol{\zeta}$ models the distortion noise that results due to the rate-constrained encoding of $\mathbf{x}$. The zero-mean AWGN $\boldsymbol{\zeta}$ is uncorrelated with $\mathbf{x}$ and its diagonal covariance matrix $\boldsymbol{\Sigma}_{\zeta\zeta}$ has entries $[\boldsymbol{\Sigma}_{\zeta\zeta}]_{ii} = \lambda_{x,i}D_i/(\lambda_{x,i} - D_i)$. The part of the test channel that takes as input $\mathbf{u}$ and outputs $\hat{\mathbf{x}}$, models the decoder. The reconstruction $\hat{\mathbf{x}}$ of $\mathbf{x}$ at the decoder output is:

$$\hat{\mathbf{x}} = \mathbf{Q}_{x,k}\boldsymbol{\Theta}_k\mathbf{u} = \mathbf{Q}_{x,k}\boldsymbol{\Theta}_k\mathbf{Q}_{x,k}^T\mathbf{x} + \mathbf{Q}_{x,k}\boldsymbol{\Theta}_k\boldsymbol{\zeta}, \quad (3.3)$$

where $\boldsymbol{\Theta}_k$ is a diagonal matrix with non-zero entries $[\boldsymbol{\Theta}_k]_{ii} = (\lambda_{x,i} - D_i)/\lambda_{x,i}$, $i = 1,\dots,k$.

3.1.2. D-R analysis for estimation. The D-R function for estimating a source $\mathbf{s}$ given observation $\mathbf{x}$ (where the source and observation are probabilistically drawn from the joint pdf $p(\mathbf{x},\mathbf{s})$) with rate R at an individual sensor, and reconstructing it (in the MMSE sense) as $\hat{\mathbf{x}}$ at the FC is given by [3, p. 79]:

$$D_s(R) = \min_{\substack{p(\hat{\mathbf{s}}_R|\mathbf{x}) \\ I(\mathbf{x};\hat{\mathbf{s}}_R)\leq R}} E_{p(\hat{\mathbf{s}}_R,\mathbf{s})}[\|\mathbf{s} - \hat{\mathbf{s}}_R\|^2], \quad (3.4)$$

where $\mathbf{s} \in \mathbb{R}^N$ and $\hat{\mathbf{s}}_R \in \mathbb{R}^N$, and the minimization is w.r.t. the conditional pdf $p(\hat{\mathbf{s}}_R|\mathbf{x})$. In order to achieve the D-R function, one might be tempted to first compress $\mathbf{x}$ by applying rwf at the sensor, without taking into account the data model relating $\mathbf{s}$ with $\mathbf{x}$, and subsequently use the reconstructed $\hat{\mathbf{x}}$ to form the MMSE estimate $\hat{\mathbf{s}}_{ce} = E[\mathbf{s}|\hat{\mathbf{x}}]$ at the FC. An

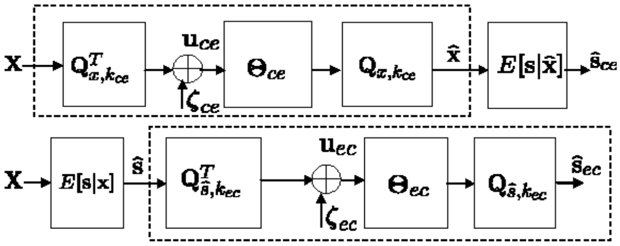
x
$\mathbf{Q}^T_{x,k_{ce}}$
$\mathbf{u}_{ce}$
$\boldsymbol{\zeta}_{ce}$
$\boldsymbol{\Theta}_{ce}$
$\mathbf{Q}_{x,k_{ce}}$
$\hat{\mathbf{x}}$
$E[\mathbf{s}|\hat{\mathbf{x}}]$
$\hat{\mathbf{s}}_{ce}$
x
$E[\mathbf{s}|\mathbf{x}]$
$\hat{\mathbf{s}}$
$\mathbf{Q}^T_{\hat{s},k_{ec}}$
$\mathbf{u}_{ec}$
$\boldsymbol{\zeta}_{ec}$
$\boldsymbol{\Theta}_{ec}$
$\mathbf{Q}_{\hat{s},k_{ec}}$
$\hat{\mathbf{s}}_{ec}$

FIG. 3. *(Top): Test channel for the CE scheme.; (Bottom): Test channel for the EC scheme.*

alternative option would be to first form the MMSE estimate $\hat{\mathbf{s}} = E[\mathbf{s}|\mathbf{x}]$, encode the latter using rwf at the sensor, and after decoding at the FC, obtain the reconstructed estimate $\hat{\mathbf{s}}_{ec}$. Referring to the former option as *Compress-Estimate* (CE), and to the latter as *Estimate-Compress* (EC), we are interested in determining which one yields the smallest MSE under a rate constraint R. Another interesting question is whether any of the CE and EC schemes enjoys MMSE optimality (i.e., achieves (3.4)). With subscripts *ce* and *ec* corresponding to these two options, let us also define the errors $\tilde{\mathbf{s}}_{ce} := \mathbf{s} - \hat{\mathbf{s}}_{ce}$ and $\tilde{\mathbf{s}}_{ec} := \mathbf{s} - \hat{\mathbf{s}}_{ec}$.

For CE, we depict in Fig. 3 (Top) the test channel for encoding $\mathbf{x}$ via rwf, followed by MMSE estimation of $\mathbf{s}$ based on $\hat{\mathbf{x}}$. Suppose that when applying rwf to $\mathbf{x}$ with prescribed rate R, the first k_{ce} components of $\mathbf{x}_w$ are assigned with non-zero rate and the rest are discarded. The MMSE optimal encoder's output for encoding $\mathbf{x}$ is given, as in subsection III-A.1, by $\mathbf{u}_{ce} = \mathbf{Q}^T_{x,k_{ce}}\mathbf{x} + \boldsymbol{\zeta}_{ce}$. The covariance matrix of $\boldsymbol{\zeta}_{ce}$ has diagonal entries $[\boldsymbol{\Sigma}_{\zeta_{ce}\zeta_{ce}}]_{ii} = \lambda_{x,i}D^{ce}_i/(\lambda_{x,i} - D^{ce}_i)$ for $i = 1,\ldots,k_{ce}$, where $D^{ce}_i := E[(\mathbf{x}_w(i) - \hat{\mathbf{x}}_w(i))^2]$. Recalling that $D^{ce}_i = \left(\prod_{i=1}^{k_{ce}} \lambda_{x,i}\right)^{1/k_{ce}} 2^{-2R/k_{ce}}$ when $i = 1,\ldots,k_{ce}$ and $D^{ce}_i = \lambda_{x,i}$, when $i = k_{ce}+1,\ldots,N$, the reconstructed $\hat{\mathbf{x}}$ in CE is [c.f. (3.3)]:

$$\hat{\mathbf{x}} = \mathbf{Q}_{x,k_{ce}}\boldsymbol{\Theta}_{ce}\mathbf{Q}^T_{x,k_{ce}}\mathbf{x} + \mathbf{Q}_{x,k_{ce}}\boldsymbol{\Theta}_{ce}\boldsymbol{\zeta}_{ce}, \tag{3.5}$$

where $[\boldsymbol{\Theta}_{ce}]_{ii} = (\lambda_{x,i} - D^{ce}_i)/\lambda_{x,i}$, for $i = 1,\ldots,k_{ce}$. Letting $\check{\mathbf{x}} := \mathbf{Q}^T_x\hat{\mathbf{x}} = [\check{\mathbf{x}}^T_1 \;\; \mathbf{0}_{1\times(N-k_{ce})}]^T$, with $\check{\mathbf{x}}_1 := \boldsymbol{\Theta}_{ce}\mathbf{Q}^T_{x,k_{ce}}\mathbf{x} + \boldsymbol{\Theta}_{ce}\boldsymbol{\zeta}_{ce}$, we have for the MMSE estimate $\hat{\mathbf{s}}_{ce} = E[\mathbf{s}|\hat{\mathbf{x}}]$:

$$\hat{\mathbf{s}}_{ce} = E[\mathbf{s}|\mathbf{Q}^T_x\hat{\mathbf{x}}] = E[\mathbf{s}|\check{\mathbf{x}}_1] = \boldsymbol{\Sigma}_{s\check{x}_1}\boldsymbol{\Sigma}^{-1}_{\check{x}_1\check{x}_1}\check{\mathbf{x}}_1, \tag{3.6}$$

since $\mathbf{Q}^T_x$ is unitary and the last $N - k_{ce}$ entries of $\check{\mathbf{x}}$ are useless for estimating $\mathbf{s}$. We have shown in [30] that the covariance matrix $\boldsymbol{\Sigma}_{\tilde{s}_{ce}\tilde{s}_{ce}} := E[(\mathbf{s} - \hat{\mathbf{s}}_{ce})(\mathbf{s} - \hat{\mathbf{s}}_{ce})^T] = \boldsymbol{\Sigma}_{ss} - \boldsymbol{\Sigma}_{s\check{x}_1}\boldsymbol{\Sigma}^{-1}_{\check{x}_1\check{x}_1}\boldsymbol{\Sigma}_{\check{x}_1 s}$ of $\tilde{\mathbf{s}}_{ce}$ is:

$$\boldsymbol{\Sigma}_{\tilde{s}_{ce}\tilde{s}_{ce}} = \boldsymbol{\Sigma}_{ss} - \boldsymbol{\Sigma}_{sx}\boldsymbol{\Sigma}^{-1}_{xx}\boldsymbol{\Sigma}_{xs} + \boldsymbol{\Sigma}_{sx}\mathbf{Q}_x\boldsymbol{\Delta}_{ce}\mathbf{Q}^T_x\boldsymbol{\Sigma}_{xs}, \tag{3.7}$$

where $\mathbf{\Delta}_{ce} := \text{diag}\left(D_1^{ce}\lambda_{x,1}^{-2}\cdots D_N^{ce}\lambda_{x,N}^{-2}\right)$.

In Fig. 3 (Bottom) we depict the test channel for the EC scheme. The MMSE estimate $\hat{\mathbf{s}} = E[\mathbf{s}|\mathbf{x}]$ is followed by the test channel that results when applying rwf to a pre-whitened version of $\hat{\mathbf{s}}$, with rate R. Let $\mathbf{\Sigma}_{\hat{s}\hat{s}} = \mathbf{Q}_{\hat{s}}\mathbf{\Lambda}_{\hat{s}}\mathbf{Q}_{\hat{s}}^T$ be the eigenvalue decomposition for the covariance matrix of $\hat{\mathbf{s}}$, where $\mathbf{\Lambda}_{\hat{s}} = \text{diag}(\lambda_{\hat{s},1}\cdots\lambda_{\hat{s},p})$ and $\lambda_{\hat{s},1} \geq \cdots \geq \lambda_{\hat{s},p}$. Suppose now that the first k_{ec} entries of $\hat{\mathbf{s}}_w = \mathbf{Q}_{\hat{s}}^T\hat{\mathbf{s}}$ are assigned with non-zero rate and the rest are discarded. The MSE optimal encoder's output is given by $\mathbf{u}_{ec} = \mathbf{Q}_{\hat{s},k_{ec}}^T\hat{\mathbf{s}} + \boldsymbol{\zeta}_{ec}$, and the estimate $\hat{\mathbf{s}}_{ec}$ is:

$$\hat{\mathbf{s}}_{ec} = \mathbf{Q}_{\hat{s},k_{ec}}\mathbf{\Theta}_{ec}\mathbf{Q}_{\hat{s},k_{ec}}^T\hat{\mathbf{s}} + \mathbf{Q}_{\hat{s},k_{ec}}\mathbf{\Theta}_{ec}\boldsymbol{\zeta}_{ec}, \tag{3.8}$$

where $\mathbf{Q}_{\hat{s},k_{ec}}$ is formed by the first k_{ec} columns of $\mathbf{Q}_{\hat{s}}$. For the $k_{ec} \times k_{ec}$ diagonal matrices $\mathbf{\Theta}_{ec}$ and $\mathbf{\Sigma}_{\zeta_{ec}\zeta_{ec}}$ we have $[\mathbf{\Theta}_{ec}]_{ii} = (\lambda_{\hat{s},i} - D_i^{ec})/\lambda_{\hat{s},i}$ and $[\mathbf{\Sigma}_{\zeta_{ec}\zeta_{ec}}]_{ii} = \lambda_{\hat{s},i}D_i^{ec}/(\lambda_{\hat{s},i} - D_i^{ec})$, where $D_i^{ec} := E[(\hat{\mathbf{s}}_w(i) - \hat{\mathbf{s}}_{ec,w}(i))^2]$, and $\hat{\mathbf{s}}_{ec,w} := \mathbf{Q}_{\hat{s}}^T\hat{\mathbf{s}}_{ec}$. Recall also that $D_i^{ec} = \left(\prod_{i=1}^{k_{ec}}\lambda_{\hat{s},i}\right)^{1/k_{ec}} 2^{-2R/k_{ec}}$ when $i = 1, \ldots, k_{ec}$ and $D_i^{ec} = \lambda_{\hat{s},i}$, for $i = k_{ec}+1, \ldots, p$. Upon defining $\mathbf{\Delta}_{ec} := \text{diag}\left(D_1^{ec}\cdots D_p^{ec}\right)$, the covariance matrix of $\tilde{\mathbf{s}}_{ec}$ is given by [30]:

$$\mathbf{\Sigma}_{\tilde{s}_{ec}\tilde{s}_{ec}} = \mathbf{\Sigma}_{ss} - \mathbf{\Sigma}_{sx}\mathbf{\Sigma}_{xx}^{-1}\mathbf{\Sigma}_{xs} + \mathbf{Q}_{\hat{s}}\mathbf{\Delta}_{ec}\mathbf{Q}_{\hat{s}}^T. \tag{3.9}$$

The MMSE associated with CE and EC is given, respectively, by [c.f. (3.7) and (3.9)]:

$$\begin{aligned} D_{ce}(R) &:= \text{trace}(\mathbf{\Sigma}_{\tilde{s}_{ce}\tilde{s}_{ce}}) = J_o + \epsilon_{ce}(R), \\ D_{ec}(R) &:= \text{trace}(\mathbf{\Sigma}_{\tilde{s}_{ec}\tilde{s}_{ec}}) = J_o + \epsilon_{ec}(R), \end{aligned} \tag{3.10}$$

where $\epsilon_{ce}(R) := \text{trace}(\mathbf{\Sigma}_{sx}\mathbf{Q}_x\mathbf{\Delta}_{ce}\mathbf{Q}_x^T\mathbf{\Sigma}_{xs})$, $\epsilon_{ec}(R) := \text{trace}(\mathbf{Q}_{\hat{s}}\mathbf{\Delta}_{ec}\mathbf{Q}_{\hat{s}}^T)$, and $J_o := \text{trace}(\mathbf{\Sigma}_{ss} - \mathbf{\Sigma}_{sx}\mathbf{\Sigma}_{xx}^{-1}\mathbf{\Sigma}_{xs})$ is the MMSE achieved when estimating $\mathbf{s}$ based on $\mathbf{x}$, without source encoding ($R \to \infty$). Since J_o is common to both EC and CE it is important to compare $\epsilon_{ce}(R)$ with $\epsilon_{ec}(R)$ in order to determine which estimation scheme achieves the smallest MSE. The following proposition provides such an asymptotic comparison:

Proposition 3.1. *If $R > R_{th} := 0.5\max\left\{\log_2\left(\left(\prod_{i=1}^{\rho}\lambda_{x,i}\right)/\sigma^{2\rho}\right), \log_2\left(\left(\prod_{i=1}^{\rho}\lambda_{\hat{s},i}\right)/(\lambda_{\hat{s},\rho})^{\rho}\right)\right\}$, then it holds that $\epsilon_{ce}(R) = \gamma_1 2^{-2R/N}$ and $\epsilon_{ec}(R) = \gamma_2 2^{-2R/\rho}$, where γ_1 and γ_2 are constants.*

An immediate consequence of Proposition 3 is that the MSE for EC converges as $R \to \infty$ to J_o with rate $O(2^{-2R/\rho})$. The MSE of CE converges likewise, but with rate $O(2^{-2R/N})$. For the typical case $N > \rho$, EC approaches the lower bound J_o faster than CE, implying correspondingly a more efficient usage of the available rate R. This is intuitively reasonable since CE compresses $\mathbf{x}$, which contains the noise $\mathbf{n}$. Since the last $N - \rho$ eigenvalues of $\mathbf{\Sigma}_{xx}$ equal the noise variance σ^2, part of the available rate is consumed to compress the noise. On the contrary, the MMSE estimator $\hat{\mathbf{s}}$ in EC suppresses significant part of the noise.

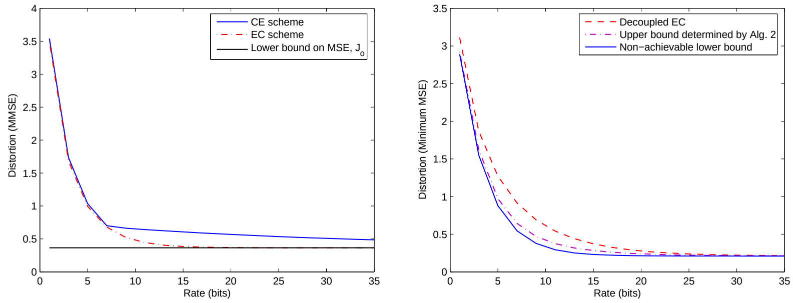

FIG. 4. *(Left): D-R region for EC and CE at SNR $= 2$; (Right): Distortion-rate bounds for estimating* $\mathbf{s}$ *in a two-sensor setup.*

Let us examine now some special cases to gain more insight about Proposition 3.

Scalar model ($p = 1$, $N = 1$): Let $x = hs + n$, where h is fixed, while s, n are uncorrelated with $s \sim \mathcal{N}\ (0, \sigma_s^2)$, $n \sim \mathcal{N}\ (0, \sigma_n^2)$, and $\sigma_x^2 = h^2\sigma_s^2 + \sigma_n^2$. With $\sigma^2_{\tilde{s}_{ce}}$ and $\sigma^2_{\tilde{s}_{ec}}$ denoting the variances of $\tilde{s}_{ce}$ and $\tilde{s}_{ec}$, respectively, we have shown in [30] that:

PROPOSITION 3.2. *For $N = p = 1$, it holds that $\sigma^2_{\tilde{s}_{ce}} = \sigma^2_{\tilde{s}_{ec}}$ and hence the D-R functions for EC and CE are identical; i.e., $D_{ec}(R) = D_{ce}(R)$.*

Vector model ($p = 1$, $N > 1$): With $\mathbf{x} = \mathbf{h}s + \mathbf{n}$ and after setting $R_{th} := 0.5 \log_2 (1+\ \sigma_s^2 \|\mathbf{h}\|^2/\sigma^2)$, we have established that [30]:

PROPOSITION 3.3. *For $R \leq R_{th}$ it holds that $\epsilon_{ce}(R) = \epsilon_{ec}(R)$ and thus $D_{ec}(R) = D_{ce}(R)$. For $R > R_{th}$, we have $\epsilon_{ce}(R) > \epsilon_{ec}(R)$ and thus EC uses more efficiently the available rate.*

We define the signal-to-noise ratio (SNR) as SNR $= \text{trace}(\mathbf{H}\boldsymbol{\Sigma}_{ss}\mathbf{H}^T)/N\sigma^2$, and compare in Fig. 4 (Left) the MMSE when estimating $\mathbf{s}$ using the CE and EC schemes. With $\boldsymbol{\Sigma}_{ss} = \sigma_s^2\mathbf{I}_p$, $p = 4$ and $N = 40$, we observe that beyond a threshold rate, the distortion of EC converges to J_o faster than that of CE, which corroborates Proposition 3.

Our analysis so far raises the question whether EC is MSE optimal. We have shown that this is the case when estimating $\mathbf{s}$ with a given rate R and without forcing any assumption about N and p. A related claim has been reported in [29, 35] for $N = p$, but the extension to $N \neq p$ is not obvious. We have established that [30]:

PROPOSITION 3.4. *The D-R function when estimating* $\mathbf{s}$ *based on* $\mathbf{x}$ *can be expressed as*

$$D_s(R) = \min_{\substack{p(\hat{\mathbf{s}}_R|\mathbf{x}) \\ I(\mathbf{x};\hat{\mathbf{s}}_R) \leq R}} E[\|\mathbf{s} - \hat{\mathbf{s}}_R\|^2] = E[\|\tilde{\mathbf{s}}\|^2] + \min_{\substack{p(\hat{\mathbf{s}}_R|\hat{\mathbf{s}}) \\ I(\hat{\mathbf{s}};\hat{\mathbf{s}}_R) \leq R}} E[\|\hat{\mathbf{s}} - \hat{\mathbf{s}}_R\|^2], \tag{3.11}$$

where $\hat{\mathbf{s}} = \boldsymbol{\Sigma}_{sx}\boldsymbol{\Sigma}_{xx}^{-1}\mathbf{x}$ is the MMSE estimator, and $\tilde{\mathbf{s}}$ is the corresponding MMSE.

Proposition 6 reveals that the optimal means of estimating $\mathbf{s}$ is to first form the optimal MMSE estimate $\hat{\mathbf{s}}$ and then apply optimal rate-distortion encoding to this estimate. The lower bound on this distortion when $R \to \infty$, is $J_o = E[\|\tilde{\mathbf{s}}\|^2]$, which is intuitively appealing. The D-R function in (3.11) is achievable, because the rightmost term in (3.11) corresponds to the D-R function for reconstructing the MMSE estimate $\hat{\mathbf{s}}$ which is known to be achievable using random coding; see e.g., [3, p. 66].

3.2. Distortion-rate for distributed estimation. Let us now consider the D-R function for estimating $\mathbf{s}$ in a multi-sensor setup, under a total available rate R which has to be shared among all sensors. Because analytical specification of the D-R function in this case remains intractable, we will develop an alternating algorithm that numerically determines an achievable upper bound for it. Combining this upper bound with the non-achievable lower bound corresponding to an equivalent single-sensor setup, and applying the MMSE optimal EC scheme, will provide a (hopefully tight) region where the D-R function lies in. For simplicity in exposition, we confine ourselves to a two-sensor setup, but our results apply to any finite $L > 2$.

To this end, we consider the following single-letter characterization of the upper bound on the D-R function:

$$\bar{D}(R) = \min_{\substack{p(\mathbf{u}_1|\mathbf{x}_1), p(\mathbf{u}_2|\mathbf{x}_2), \hat{\mathbf{s}}_R \\ I(\mathbf{x};\mathbf{u}_1,\mathbf{u}_2) \leq R}} E_{p(\mathbf{s},\mathbf{u}_1,\mathbf{u}_2)}[\|\mathbf{s} - \hat{\mathbf{s}}_R(\mathbf{u}_1,\mathbf{u}_2)\|^2], \tag{3.12}$$

where the minimization is w.r.t. $\{p(\mathbf{u}_i|\mathbf{x}_i)\}_{i=1}^2$ and $\hat{\mathbf{s}}_R := \hat{\mathbf{s}}_R(\mathbf{u}_1,\mathbf{u}_2)$. Achievability of $\bar{D}(R)$ can be established by readily extending to the vector case the scalar results in [7]. To carry out the minimization in (3.12), we develop an alternating scheme whereby $\mathbf{u}_2$ is treated as side information that is available at the decoder when optimizing (3.12) w.r.t. $p(\mathbf{u}_1|\mathbf{x}_1)$ and $\hat{\mathbf{s}}_R(\mathbf{u}_1,\mathbf{u}_2)$. The side information $\mathbf{u}_2$ is considered as the output of an optimal rate-distortion encoder applied to $\mathbf{x}_2$ for estimating $\mathbf{s}$, without taking into account $\mathbf{x}_1$. Since $\mathbf{x}_2$ is Gaussian, the side information will have the form (c.f. subsection III-A.2) $\mathbf{u}_2 = \mathbf{Q}_2\mathbf{x}_2 + \boldsymbol{\zeta}_2$, where $\mathbf{Q}_2 \in \mathbb{R}^{k_2 \times N_2}$ and $k_2 \leq N_2$, due to the rate constrained encoding of $\mathbf{x}_2$. Recall that the $k_2 \times 1$ vector $\boldsymbol{\zeta}_2$ is uncorrelated with $\mathbf{x}_2$ and Gaussian; i.e., $\boldsymbol{\zeta}_2 \sim \mathcal{N}\ (\mathbf{0}, \boldsymbol{\Sigma}_{\zeta_2\zeta_2})$.

Based on $\boldsymbol{\psi} := [\mathbf{x}_1^T \ \ \mathbf{u}_2^T]^T$, the optimal estimator for $\mathbf{s}$ is the MMSE one: $\hat{\mathbf{s}} = E[\mathbf{s}|\mathbf{x}_1,\mathbf{u}_2] = \boldsymbol{\Sigma}_{s\psi}\boldsymbol{\Sigma}_{\psi\psi}^{-1}\boldsymbol{\psi} = \mathbf{L}_1\mathbf{x}_1 + \mathbf{L}_2\mathbf{u}_2$, where $\mathbf{L}_1$, $\mathbf{L}_2$ are $p \times N_1$ and $p \times k_2$ matrices such that $\boldsymbol{\Sigma}_{s\psi}\boldsymbol{\Sigma}_{\psi\psi}^{-1} = [\mathbf{L}_1 \ \mathbf{L}_2]$. If $\tilde{\mathbf{s}}$ is the corresponding MSE, then $\mathbf{s} = \hat{\mathbf{s}} + \tilde{\mathbf{s}}$, where $\tilde{\mathbf{s}}$ is uncorrelated with $\boldsymbol{\psi}$ due to the orthogonality principle. Noticing also that $\hat{\mathbf{s}}_R(\mathbf{u}_1,\mathbf{u}_2)$ is uncorrelated with $\tilde{\mathbf{s}}$ because it is a function of $\mathbf{x}_1$ and $\mathbf{u}_2$, we have $E[\|\mathbf{s} - \hat{\mathbf{s}}_R(\mathbf{u}_1,\mathbf{u}_2)\|^2] = E[\|\hat{\mathbf{s}} - \hat{\mathbf{s}}_R(\mathbf{u}_1,\mathbf{u}_2)\|^2] + E[\|\tilde{\mathbf{s}}\|^2]$, or,

$$E[\|\mathbf{s} - \hat{\mathbf{s}}_R(\mathbf{u}_1,\mathbf{u}_2)\|^2] = E[\|\mathbf{L}_1\mathbf{x}_1 - (\hat{\mathbf{s}}_R(\mathbf{u}_1,\mathbf{u}_2) - \mathbf{L}_2\mathbf{u}_2)\|^2] + E[\|\tilde{\mathbf{s}}\|^2]. \tag{3.13}$$

Clearly, it holds that $I(\mathbf{x};\mathbf{u}_1,\mathbf{u}_2) = R_2 + I(\mathbf{x}_1;\mathbf{u}_1) - I(\mathbf{u}_2;\mathbf{u}_1)$, where $R_2 := I(\mathbf{x};\mathbf{u}_2)$ is the rate consumed to form the side information $\mathbf{u}_2$ and the rate constraint in (3.12) becomes $I(\mathbf{x};\mathbf{u}_1,\mathbf{u}_2) \leq R \Leftrightarrow I(\mathbf{x}_1;\mathbf{u}_1) - I(\mathbf{u}_2;\mathbf{u}_1) \leq R - R_2 := R_1$. The new signal of interest in (3.13) is $\mathbf{L}_1\mathbf{x}_1$; thus, $\mathbf{u}_1$ has to be a function of $\mathbf{L}_1\mathbf{x}_1$. Using the fact that $\mathbf{x}_1 \to \mathbf{L}_1\mathbf{x}_1 \to \mathbf{u}_1$, constitutes a Markov chain, we show in [30] that $I(\mathbf{x}_1;\mathbf{u}_1) = I(\mathbf{L}_1\mathbf{x}_1;\mathbf{u}_1)$. Using the latter, we obtain:

$$I(\mathbf{x}_1;\mathbf{u}_1) - I(\mathbf{u}_2;\mathbf{u}_1) = I(\mathbf{L}_1\mathbf{x}_1;\mathbf{u}_1) - I(\mathbf{u}_2;\mathbf{u}_1). \tag{3.14}$$

From the RHS of (3.14), we deduce the equivalent constraint $I(\mathbf{L}_1\mathbf{x}_1;\mathbf{u}_1) - I(\mathbf{u}_2;\mathbf{u}_1) \leq R_1$. Combining the latter with (3.13) and (3.12), we arrive at the D-R upper bound:

$$\bar{\bar{D}}(R_1) = E[\|\tilde{\mathbf{s}}\|^2] + \min_{\substack{p(\mathbf{u}_1|\mathbf{L}_1\mathbf{x}_1),\hat{\mathbf{s}}_R \\ I(\mathbf{L}_1\mathbf{x}_1;\mathbf{u}_1)-I(\mathbf{u}_1;\mathbf{u}_2)\leq R_1}} E[\|\mathbf{L}_1\mathbf{x}_1 - \tilde{\mathbf{s}}_{R,12}(\mathbf{u}_1,\mathbf{u}_2)\|^2], \tag{3.15}$$

where $\tilde{\mathbf{s}}_{R,12}(\mathbf{u}_1,\mathbf{u}_2) := \hat{\mathbf{s}}_R(\mathbf{u}_1,\mathbf{u}_2) - \mathbf{L}_2\mathbf{u}_2$. Through (3.15) we can determine an achievable D-R region, having available rate R_1 at the encoder and side information $\mathbf{u}_2$ at the decoder. Since $\mathbf{x}_1$ and $\mathbf{u}_2$ are jointly Gaussian, we can apply the Wyner-Ziv result [36], which allows us to consider that $\mathbf{u}_2$ is available both at the decoder and the encoder. This, in turn, permits re-writing the first term in (3.15) as:

$$\min_{\substack{p(\hat{\mathbf{s}}_R|\mathbf{L}_1\mathbf{x}_1,\mathbf{u}_2) \\ I(\mathbf{L}_1\mathbf{x}_1;\hat{\mathbf{s}}_R|\mathbf{u}_2)\leq R_1}} E[\|\mathbf{L}_1\mathbf{x}_1 - [\hat{\mathbf{s}}_R(\mathbf{u}_1,\mathbf{u}_2) - \mathbf{L}_2\mathbf{u}_2]\|^2]. \tag{3.16}$$

If $\hat{\mathbf{s}}_1 := E[\mathbf{L}_1\mathbf{x}_1|\mathbf{u}_2] = \mathbf{L}_1\boldsymbol{\Sigma}_{x_1u_2}\boldsymbol{\Sigma}_{u_2u_2}^{-1}\mathbf{u}_2$ and $\tilde{\mathbf{s}}_1$ is the corresponding MSE, then we can write $\mathbf{L}_1\mathbf{x}_1 = \hat{\mathbf{s}}_1 + \tilde{\mathbf{s}}_1$. For the rate constraint in (3.16), we have:

$$\begin{aligned} I(\mathbf{L}_1\mathbf{x}_1;\hat{\mathbf{s}}_R|\mathbf{u}_2) &= I(\mathbf{L}_1\mathbf{x}_1 - \hat{\mathbf{s}}_1;\hat{\mathbf{s}}_R - \mathbf{L}_2\mathbf{u}_2 - \hat{\mathbf{s}}_1|\mathbf{u}_2) \\ &= I(\tilde{\mathbf{s}}_1;\hat{\mathbf{s}}_R - \mathbf{L}_2\mathbf{u}_2 - \hat{\mathbf{s}}_1), \end{aligned} \tag{3.17}$$

where the first equality is true because $\mathbf{u}_2$ is known; while the second one holds since $\mathbf{u}_2$ is uncorrelated with $\tilde{\mathbf{s}}_1$, due to the orthogonality principle, and likewise $\mathbf{u}_2$ is uncorrelated with $\hat{\mathbf{s}}_{R,12}(\mathbf{u}_1,\mathbf{u}_2) := \hat{\mathbf{s}}_R(\mathbf{u}_1,\mathbf{u}_2) - \mathbf{L}_2\mathbf{u}_2 - \hat{\mathbf{s}}_1$. Utilizing (3.16) and (3.17), we arrive at:

$$\bar{\bar{D}}(R_1) = \min_{\substack{p(\hat{\mathbf{s}}_{R,12}|\tilde{\mathbf{s}}_1) \\ I(\tilde{\mathbf{s}}_1;\hat{\mathbf{s}}_{R,12})\leq R_1}} E[\|\tilde{\mathbf{s}}_1 - \hat{\mathbf{s}}_{R,12}(\mathbf{u}_1,\mathbf{u}_2)\|^2] + E[\|\tilde{\mathbf{s}}\|^2]. \tag{3.18}$$

Notice that (3.18) is the D-R function for reconstructing the MSE $\tilde{\mathbf{s}}_1$ with rate R_1. Since $\tilde{\mathbf{s}}_1$ is Gaussian, we can readily apply rwf to the pre-whitened $\mathbf{Q}_{\tilde{s}_1}^T\tilde{\mathbf{s}}_1$ for determining $\bar{\bar{D}}(R_1)$ and the corresponding test channel that achieves $\bar{\bar{D}}(R_1)$. Through the latter, and considering the next eigenvalue decomposition $\boldsymbol{\Sigma}_{\tilde{s}_1\tilde{s}_1} = \mathbf{Q}_{\tilde{s}_1}\ \mathrm{diag}(\lambda_{\tilde{s}_1,1}\cdots\lambda_{\tilde{s}_1,p})\mathbf{Q}^T_{\tilde{s}_1}$, we find that the first encoder's output that minimizes (3.12) has the form:

$$\mathbf{u}_1 = \mathbf{Q}^T_{\tilde{s}_1,k_1}\mathbf{L}_1\mathbf{x}_1 + \boldsymbol{\zeta}_1 = \mathbf{Q}_1\mathbf{x}_1 + \boldsymbol{\zeta}_1, \tag{3.19}$$

where $\mathbf{Q}_{\tilde{s}_1,k_1}$ denotes the first k_1 columns of $\mathbf{Q}_{\tilde{s}_1}$, k_1 is the number of $\mathbf{Q}^T_{\tilde{s}_1}\tilde{\mathbf{s}}_1$ entries that are assigned with non-zero rate, and $\mathbf{Q}_1 := \mathbf{Q}^T_{\tilde{s}_1,k_1}\mathbf{L}_1$. The $k_1\times 1$ AWGN $\boldsymbol{\zeta}_1 \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Sigma}_{\zeta_1\zeta_1}\right)$ is uncorrelated with $\mathbf{x}_1$. Additionally, we have $[\boldsymbol{\Sigma}_{\zeta_1\zeta_1}]_{ii} = \lambda_{\tilde{s}_1,i}D^1_i/(\lambda_{\tilde{s}_1,i} - D^1_i)$, where $D^1_i = \left(\prod_{i=1}^{k_1}\lambda_{\tilde{s}_1,i}\right)^{1/k_1} 2^{-2R_1/k_1}$, for $i = 1,\ldots,k_1$, and $D^1_i = \lambda_{\tilde{s}_1,i}$ when $i = k_1+1,\ldots,p$. This way, we are able to determine also $p(\mathbf{u}_1|\mathbf{x}_1)$. The reconstruction function has the form:

$$\begin{aligned}\hat{\mathbf{s}}_R(\mathbf{u}_1,\mathbf{u}_2) =& \mathbf{Q}_{\tilde{s}_1,k_1}\boldsymbol{\Theta}_1\mathbf{u}_1 + \mathbf{L}_1\boldsymbol{\Sigma}_{x_1u_2}\boldsymbol{\Sigma}^{-1}_{u_2u_2}\mathbf{u}_2 + \mathbf{L}_2\mathbf{u}_2\\ &- \mathbf{Q}_{\tilde{s}_1,k_1}\boldsymbol{\Theta}_1\mathbf{Q}^T_{\tilde{s}_1,k_1}\mathbf{L}_1\boldsymbol{\Sigma}_{x_1u_2}\boldsymbol{\Sigma}^{-1}_{u_2u_2}\mathbf{u}_2,\end{aligned} \tag{3.20}$$

where $[\boldsymbol{\Theta}_1]_{ii} = \lambda_{\tilde{s}_1,i}D^1_i/(\lambda_{\tilde{s}_1,i} - D^1_i)$, and the MMSE is $\bar{\bar{D}}(R_1) = \sum_{j=1}^p D^1_j + E[\|\tilde{\mathbf{s}}\|^2]$.

The approach in this subsection can be applied in an alternating fashion from sensor to sensor in order to determine appropriate $p(\mathbf{u}_i|\mathbf{x}_i)$, for $i = 1,2$, and $\hat{\mathbf{s}}_R(\mathbf{u}_1,\mathbf{u}_2)$ that at best globally minimize (3.15). The conditional pdfs can be determined by finding the appropriate covariances $\boldsymbol{\Sigma}_{\zeta_i\zeta_i}$. Furthermore, by specifying the optimal $\mathbf{Q}_1$ and $\mathbf{Q}_2$, we have a complete characterization of the encoders' structure. The resultant algorithm is summarized next:

Algorithm 2 :

```
Initialize $\mathbf{Q}_1^{(0)}, \mathbf{Q}_2^{(0)}, \boldsymbol{\Sigma}^{(0)}_{\zeta_1\zeta_1}, \boldsymbol{\Sigma}^{(0)}_{\zeta_2\zeta_2}$ by applying optimal D-R encoding to each sensor's test channel independently. For a total rate $R$, generate $M$ random increments $\{r(m)\}_{m=0}^M$, such that $0 \le r(m) \le R$ and $\sum_{m=0}^M r(m) = R$. Set $R_1(0) = R_2(0) = 0$.
for $j = 1,M$ do
  Set $R(j) = \sum_{l=0}^j r(l)$
  for $i = 1,2$ do
    $\bar{i} = \mod(i,2) + 1$ %The complementary index
    $R_0(j) = I(\mathbf{x};\mathbf{u}_{\bar{i}}^{(j)})$
    We use $\mathbf{Q}_{\bar{i}}^{(j-1)}, \boldsymbol{\Sigma}^{(j-1)}_{\zeta_{\bar{i}}\zeta_{\bar{i}}}, R(j), R_0(j)$ to determine $\mathbf{Q}_i^{(j)}$, $\boldsymbol{\Sigma}^{(j)}_{\zeta_i\zeta_i}$ and distortion $\bar{\bar{D}}(R_i(j))$
  end for
  Update matrices $\mathbf{Q}_l^{(j)}, \boldsymbol{\Sigma}^{(j)}_{\zeta_l\zeta_l}$ that result the smallest distortion $\bar{\bar{D}}(R_l(j))$, with $l \in [1,2]$
  Set $R_l(j) = R(j) - I(\mathbf{x};\mathbf{u}_l^{(j)})$ and $R_{\bar{l}}(j) = I(\mathbf{x};\mathbf{u}_l^{(j)})$.
end for
```

In Fig. 4 (Right), we plot the non-achievable lower bound which corresponds to one sensor having available the entire $\mathbf{x}$ and using the optimal EC scheme. Moreover, we plot an achievable D-R upper bound determined by letting the i-th sensor form its local estimate $\hat{\mathbf{s}}_i = E[\mathbf{s}|\mathbf{x}_i]$, and then apply optimal rate-distortion encoding to $\hat{\mathbf{s}}_i$. If $\hat{\mathbf{s}}_{R,1}$ and $\hat{\mathbf{s}}_{R,2}$ are the reconstructed versions of $\hat{\mathbf{s}}_1$ and $\hat{\mathbf{s}}_2$, respectively, then the decoder at the FC forms the final estimate $\hat{\mathbf{s}}_R = E[\mathbf{s}|\hat{\mathbf{s}}_{R,1}, \hat{\mathbf{s}}_{R,2}]$. We also plot the achievable D-R region determined numerically by the alternating algorithm. For each rate, we keep the smallest distortion returned after 500 executions of the algorithm simulated with $\mathbf{\Sigma}_{ss} = \mathbf{I}_p$, $p = 4$, and $N_1 = N_2 = 20$, at SNR = 2. We observe that the algorithm provides a tight upper bound for the achievable D-R region. Using also the non-achievable lower bound (solid line), we have effectively reduced the 'uncertainty region' where the D-R function lies.

4. Distributed quantization-estimation. Consider a WSN consisting of N sensors deployed to estimate a deterministic $p \times 1$ vector parameter $\boldsymbol{\theta}$. The n^{th} sensor observes an $M \times 1$ vector of noisy observations

$$\mathbf{x}(n) = \mathbf{f}_n(\boldsymbol{\theta}) + \mathbf{w}(n), \qquad n = 0, 1, \ldots, N-1\ , \tag{4.1}$$

where $\mathbf{f}_n : \mathbf{R}^p \to \mathbf{R}^M$ is a known (generally nonlinear) function and $\mathbf{w}(n)$ denotes zero-mean noise with pdf $p_{\mathbf{w}}(\mathbf{w})$, that is known possibly up to a finite number of unknown parameters. We further assume that $\mathbf{w}(n_1)$ is independent of $\mathbf{w}(n_2)$ for $n_1 \neq n_2$; i.e., noise variables are independent across sensors. We will use $\mathbf{J}_n$ to denote the Jacobian of the differentiable function $\mathbf{f}_n$ whose $(i,j)^{th}$ entry is given by $[\mathbf{J}_n]_{ij} = \partial[\mathbf{f}_n]_i / \partial[\boldsymbol{\theta}]_j$.

Due to bandwidth limitations, the observations $\mathbf{x}(n)$ have to be quantized and estimation of $\boldsymbol{\theta}$ can only be based on these quantized values. We will henceforth think of quantization as the construction of a set of indicator variables

$$b_k(n) = \mathbf{1}\{\mathbf{x}(n) \in B_k(n)\}, \qquad k = 1, \ldots, K\ , \tag{4.2}$$

taking the value 1 when $\mathbf{x}(n)$ belongs to the region $B_k(n) \subset \mathbf{R}^M$, and 0 otherwise. Estimation of $\boldsymbol{\theta}$ will rely on this set of *binary* variables $\{b_k(n), k = 1, \ldots, K\}_{n=0}^{N-1}$. The latter are Bernoulli distributed with parameters $q_k(n)$ satisfying

$$q_k(n) := \Pr\{b_k(n) = 1\} = \Pr\{\mathbf{x}(n) \in B_k(n)\}. \tag{4.3}$$

In the ensuing sections, we will derive the Cramér-Rao Lower Bound (CRLB) to benchmark the variance of all unbiased estimators $\hat{\boldsymbol{\theta}}$ constructed using the binary observations $\{b_k(n), k = 1, \ldots, K\}_{n=0}^{N-1}$. We will further show that it is possible to find Maximum Likelihood Estimators (MLEs) that (at least asymptotically) are known to achieve the CRLB. Finally, we will reveal that the CRLB based on $\{b_k(n), k = 1, \ldots, K\}_{n=0}^{N-1}$ can come surprisingly close to the clairvoyant CRLB based on $\{x(n)\}_{n=0}^{N-1}$ in certain applications of practical interest.

4.1. Scalar parameter estimation – Parametric approach. Consider the case where $\boldsymbol{\theta} \leftrightarrow \theta$ is a scalar ($p = 1$), $x(n) = \theta + w(n)$, and $p_w(w) \leftrightarrow p_w(w, \sigma)$ is known, with σ denoting the noise standard deviation. Seeking first estimators $\hat{\theta}$ when the possibly non-Gaussian noise pdf is known, we move on to the case where σ is unknown, and prove that in both cases the variance of $\hat{\theta}$ based on a single bit per sensor can come close to the variance of the sample mean estimator, $\bar{x} := N^{-1} \sum_{n=0}^{N-1} x(n)$.

4.1.1. Known noise pdf. When the noise pdf is known, we will rely on a single region $B_1(n)$ in (4.2) to generate a single bit $b_1(n)$ per sensor, using a threshold τ_c common to all N sensors: $B_1(n) := B_c = (\tau_c, \infty)$, $\forall n$. Based on these binary observations, $b_1(n) := \mathbf{1}\{\mathbf{x}(n) \in (\tau_c, \infty)\}$ received from all N sensors, the fusion center seeks estimates of θ.

Let $F_w(u) := \int_u^\infty p_w(w)\, dw$ denote the Complementary Cumulative Distribution Function (CCDF) of the noise. Using (4.3), we can express the Bernoulli parameter as, $q_1 = \int_{\tau_c - \theta}^\infty p_w(w) dw = F_w(\tau_c - \theta)$; and its MLE as $\hat{q}_1 = N^{-1} \sum_{n=0}^{N-1} b_1(n)$. Invoking now the invariance property of MLE, it follows readily that the MLE of θ is given by [27][1]:

$$\hat{\theta} = \tau_c - F_w^{-1} \left(\frac{1}{N} \sum_{n=0}^{N-1} b_1(n) \right) . \tag{4.4}$$

Furthermore, it can be shown that the CRLB, that bounds the variance of any unbiased estimator $\hat{\theta}$ based on $b_1(n)_{n=0}^{N-1}$ is [27]

$$\mathrm{var}(\hat{\theta}) \geq \frac{1}{N} \frac{F_w(\tau_c - \theta)[1 - F_w(\tau_c - \theta)]}{p_w^2(\tau_c - \theta)} \;:=\; B(\theta) \ . \tag{4.5}$$

If the noise is Gaussian, and we define the σ*-distance* between the threshold τ_c and the (unknown) parameter θ as $\Delta_c := (\tau_c - \theta)/\sigma$, then (4.5) reduces to

$$B(\theta) \;=\; \frac{\sigma^2}{N} \frac{2\pi Q(\Delta_c)[1 - Q(\Delta_c]}{e^{-\Delta_c}} \;:=\; \frac{\sigma^2}{N} D(\Delta_c), \tag{4.6}$$

with $Q(u) := (1/\sqrt{2\pi}) \int_u^\infty e^{-w^2/2}\, dw$ denoting the Gaussian tail probability function.

The bound $B(\theta)$ is the variance of $\bar{x}$, scaled by the factor $D(\Delta_c)$; recall that $\mathrm{var}(\bar{x}) = \sigma^2/N$ [13, p.31]. Optimizing $B(\theta)$ with respect to Δ_c, yields the optimum at $\Delta_c = 0$ and

$$B_{\min} = \frac{\pi}{2} \frac{\sigma^2}{N}, \tag{4.7}$$

[1] Although related results are derived in [27, Prop.1] for Gaussian noise, it is straightforward to generalize the referred proof to cover also non-Gaussian noise pdfs.

the minimum CRLB. Eq. (4.7) reveals something unexpected: relying on a single bit per $x(n)$, the estimator in (4.4) incurs a minimal (just a $\pi/2$ factor) increase in its variance relative to the clairvoyant $\bar{x}$ which relies on the unquantized data $x(n)$. But this minimal loss in performance corresponds to the ideal choice $\Delta_c = 0$, which implies $\tau_c = \theta$ and requires perfect knowledge of the unknown θ for selecting the quantization threshold τ_c.

A closer look at $B(\theta)$ in (4.5) will confirm that the loss can be huge if $\tau_c - \theta \gg 0$. Indeed, as $\tau_c - \theta \to \infty$ the denominator in (4.5) goes to zero faster than its numerator, since F_w is the integral of the non-negative pdf p_w; and thus, $B(\theta) \to \infty$ as $\tau_c - \theta \to \infty$. The implication of the latter is twofold: i) since it shows up in the CRLB, the potentially high variance of estimators based on quantized observations is inherent to the possibly severe bandwidth limitations of the problem itself and is not unique to a particular estimator; ii) for any choice of τ_c, the fundamental performance limits in (4.5) are dictated by the end points $\tau_c - \Theta_1$ and $\tau_c - \Theta_2$ when θ is confined to the interval $[\Theta_1, \Theta_2]$. On the other hand, how successful the τ_c selection is depends on the dynamic range $|\Theta_1 - \Theta_2|$ which makes sense because the latter affects the error incurred when quantizing $x(n)$ to $b_1(n)$. Notice that in such joint quantization-estimation problems one faces two sources of error: quantization and noise. To account for both, the proper figure of merit for estimators based on binary observations is what we will term quantization signal-to-noise ratio (Q-SNR):

$$\gamma := \frac{|\Theta_1 - \Theta_2|^2}{\sigma^2}; \tag{4.8}$$

Notice that contrary to common wisdom, the smaller Q-SNR is, the easier it becomes to select τ_c judiciously. Furthermore, the variance increase in (4.5) relative to the variance of the clairvoyant $\bar{x}$ is smaller, for a given σ. This is because as the Q-SNR increases the problem becomes more difficult in general, but the rate at which the variance increases is smaller for the CRLB in (4.5) than for $\mathrm{var}(\bar{x}) = \sigma^2/N$.

4.1.2. Known noise pdf with unknown variance. No matter how small the variance in (4.5) can be made by properly selecting τ_c, the estimator $\hat{\theta}$ in (4.4) requires perfect knowledge of the noise pdf which may not be always justifiable. A more realistic approach is to assume that the noise pdf is known (e.g., Gaussian) but some of its parameters are unknown. A case frequently encountered in practice is when the noise pdf is known except for its variance $\mathrm{E}[w^2(n)] = \sigma^2$. Introducing the standardized variable $v(n) := w(n)/\sigma$ we write the signal model as

$$x(n) = \theta + \sigma v(n). \tag{4.9}$$

Let $p_v(v)$ and $F_v(v) := \int_v^\infty p_v(u)du$ denote the known pdf and CCDF of $v(n)$. Note that according to its definition, $v(n)$ has zero mean, $\mathrm{E}[v^2(n)] =$

1, and the pdfs of v and w are related by $p_w(w) = (1/\sigma)p_v(w/\sigma)$. Note also that all two parameter pdfs can be standardized likewise.

To estimate θ when σ is also unknown while keeping the bandwidth constraint to 1 bit per sensor, we divide the sensors in two groups each using a different region (i.e., threshold) to define the binary observations:

$$B_1(n) := \begin{cases} (\tau_1, \infty) := B_1, & \text{for } n = 0, \ldots, (N/2) - 1 \\ (\tau_2, \infty) := B_2, & \text{for } n = (N/2), \ldots, N. \end{cases} \tag{4.10}$$

That is, the first $N/2$ sensors quantize their observations using the threshold τ_1, while the remaining $N/2$ sensors rely on the threshold τ_2. Without loss of generality, we assume $\tau_2 > \tau_1$.

The Bernoulli parameters of the resultant binary observations can be expressed in terms of the CCDF of $v(n)$ as:

$$q_1(n) := \begin{cases} F_v\left[\dfrac{\tau_1 - \theta}{\sigma}\right] := q_1 & \text{for } n = 0, \ldots, (N/2) - 1, \\ F_v\left[\dfrac{\tau_2 - \theta}{\sigma}\right] := q_2 & \text{for } n = (N/2), \ldots, N. \end{cases} \tag{4.11}$$

Given the noise independence across sensors, the MLEs of q_1, q_2 can be found, respectively, as

$$\hat{q}_1 = \frac{2}{N}\sum_{n=0}^{N/2-1} b_1(n), \qquad \hat{q}_2 = \frac{2}{N}\sum_{n=N/2}^{N-1} b_1(n). \tag{4.12}$$

Mimicking (4.4), we can invert F_v in (4.11) and invoke the invariance property of MLEs, to obtain the MLE $\hat{\theta}$ in terms of $\hat{q}_1$ and $\hat{q}_2$. This result is stated in the following proposition that also derives the CRLB for this estimation problem[2].

PROPOSITION 4.1. *Consider estimating θ in* (4.9)*, based on binary observations constructed from the regions defined in* (4.10).
(a) *The MLE of θ is*

$$\hat{\theta} = \frac{F_v^{-1}(\hat{q}_2)\tau_1 - F_v^{-1}(\hat{q}_1)\tau_2}{F_v^{-1}(\hat{q}_2) - F_v^{-1}(\hat{q}_1)}, \tag{4.13}$$

with F_v^{-1} denoting the inverse function of F_v, and $\hat{q}_1$, $\hat{q}_2$ given by (4.12).
(b) *The variance of any unbiased estimator of θ,* $\text{var}(\hat{\theta})$*, based on $\{b_1(n)\}_{n=0}^{N-1}$ is bounded by*

$$B(\theta) := \frac{2\sigma^2}{N}\left(\frac{\Delta_1\Delta_2}{\Delta_2 - \Delta_1}\right)^2 \left[\frac{q_1(1-q_1)}{p_v^2(\Delta_1)\Delta_1^2} + \frac{q_2(1-q_2)}{p_v^2(\Delta_2)\Delta_2^2}\right] \tag{4.14}$$

[2] Omitted due to space considerations, proofs pertaining to claims in this section can be found in [28].

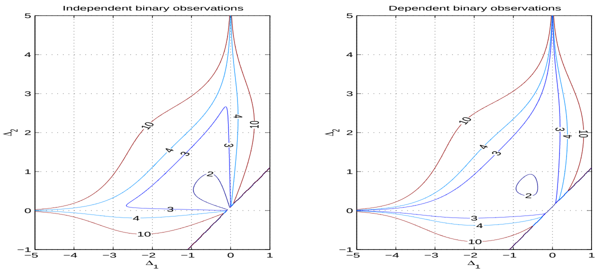

FIG. 5. *Per bit CRLB when the binary observations are independent (Section 4.1.2) and dependent (Section 4.1.3), respectively. In both cases, the variance increase with respect to the sample mean estimator is small when the σ-distances are close to* 1*, being slightly better for the case of dependent binary observations (Gaussian noise).*

where q_k is given by (4.11)*, and*

$$\Delta_k := \frac{\tau_k - \theta}{\sigma}, \qquad k = 1, 2, \tag{4.15}$$

is the σ-distance between θ and the threshold τ_k.

Eq. (4.14) is reminiscent of (4.5), suggesting that the variances of the estimators they bound are related. This implies that even when the known noise pdf contains unknown parameters the variance of $\hat{\theta}$ can come close to the variance of the clairvoyant estimator $\bar{x}$, provided that the thresholds τ_1, τ_2 are chosen close to θ relative to the noise standard deviation (so that Δ_1, Δ_2, and $\Delta_2 - \Delta_1$ in (4.15) are ≈ 1). For the Gaussian pdf, Fig. 5 shows the contour plot of $B(\theta)$ in (4.14) normalized by $\sigma^2/N := \mathrm{var}(\bar{x})$. Notice that in the low Q-SNR regime $\Delta_1, \Delta_2 \approx 1$, and the relative variance increase $B(\theta)/\mathrm{var}(\bar{x})$ is less than 3.

4.1.3. Dependent binary observations.

In the previous subsection, we restricted the sensors to transmit only 1 bit per $x(n)$ datum, and divided the sensors in two classes each quantizing $x(n)$ using a different threshold. A related approach is to let each sensor use two thresholds:

$$\begin{aligned} B_1(n) &:= B_1 = (\tau_1, \infty), \qquad n = 0, 1, \ldots, N-1, \\ B_2(n) &:= B_2 = (\tau_2, \infty), \qquad n = 0, 1, \ldots, N-1 \end{aligned} \tag{4.16}$$

where $\tau_2 > \tau_1$. We define the per sensor vector of binary observations $\mathbf{b}(n) := [b_1(n), b_2(n)]^T$, and the vector Bernoulli parameter $\mathbf{q} := [q_1(n), q_2(n)]^T$, whose components are as in (4.11).

Note the subtle differences between (4.10) and (4.16). While each of the N sensors generates 1 binary observation according to (4.10), each sensor creates 2 binary observations as per (4.16). The total number of bits from all sensors in the former case is N, but in the latter $N \log_2 3$, since our constraint $\tau_2 > \tau_1$ implies that the realization $\mathbf{b} = (0, 1)$ is impossible. In addition, all bits in the former case are independent, whereas correlation is present in the latter since $b_1(n)$ and $b_2(n)$ come from the same $x(n)$. Even though one would expect this correlation to complicate matters, a property of the binary observations defined as per (4.16), summarized in the next lemma, renders estimation of θ based on them feasible.

LEMMA 4.1. *The MLE of* $\mathbf{q} := (q_1(n), q_2(n))^T$ *based on the binary observations* $\{\mathbf{b}(n)\}_{n=0}^{N-1}$ *constructed according to* (4.16) *is given by*

$$\hat{\mathbf{q}} = \frac{1}{N}\sum_{n=0}^{N-1}\mathbf{b}(n). \tag{4.17}$$

Interestingly, (4.17) coincides with (4.12), proving that the corresponding estimators of θ are identical; i.e., (4.13) yields also the MLE $\hat{\theta}$ even in the correlated case. However, as the following proposition asserts, correlation affects the estimator's variance and the corresponding CRLB.

PROPOSITION 4.2. *Consider estimating* θ *in* (4.9)*, when* σ *is unknown, based on binary observations constructed from the regions defined in* (4.16)*. The variance of any unbiased estimator of* θ*,* $\mathrm{var}(\hat{\theta})$*, based on* $\{b_1(n), b_2(n)\}_{n=0}^{N-1}$ *is bounded by*

$$\begin{aligned} B_D(\theta) := \frac{\sigma^2}{N}&\left(\frac{\Delta_1\Delta_2}{\Delta_2-\Delta_1}\right)^2 \\ &\left[\frac{q_1\,(1-q_1)}{p_v^2(\Delta_1)\Delta_1^2}+\frac{q_2\,(1-q_2)}{p_v^2(\Delta_2)\Delta_2^2}-\frac{q_2\,(1-q_1)}{p_v(\Delta_1)p(\Delta_2)\Delta_1\Delta_2}\right], \end{aligned} \tag{4.18}$$

where the subscript D *in* $B_D(\theta)$ *is used as a mnemonic for the dependent binary observations this estimator relies on [c.f.* (4.14)*].*

Unexpectedly, (4.18) is similar to (4.14). Actually, a fair comparison between the two requires compensating for the difference in the total number of bits used in each case. This can be accomplished by introducing the per-bit CRLBs for the independent and correlated cases respectively,

$$C(\theta) = NB(\theta), \qquad C_D(\theta) = N\log_2(3)B_D(\theta)\ , \tag{4.19}$$

which lower bound the corresponding variances achievable by the transmission of 1 bit.

Evaluation of $C(\theta)/\sigma^2$ and $C_D(\theta)/\sigma^2$ follows from (4.14), (4.18) and (4.19) and is depicted in Fig. 5 for Gaussian noise and σ-distances Δ_1, Δ_2 having amplitude as large as 5. Somewhat surprisingly, both approaches yield very similar bounds with the one relying on dependent binary observations being slightly better in the achievable variance; or correspondingly, in requiring a smaller number of sensors to achieve the same CRLB.

4.2. Unknown noise pdf. In certain applications it may not be reasonable to assume knowledge about the noise pdf $p_w(w)$. These cases require *non - parametric* approaches as the one pursued in this section.

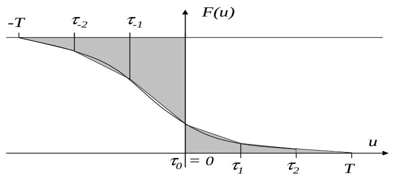

FIG. 6. *When the noise pdf is unknown numerically integrating the CCDF using the trapezoidal rule yields an approximation of the mean.*

We assume that $p_w(w)$ has zero mean so that θ in (4.1) is identifiable. Let $p_x(x)$ and $F_x(x)$ denote the pdf and CCDF of the observations $x(n)$. As θ is the mean of $x(n)$, we can write

$$\theta := \int_{-\infty}^{+\infty} x p_x(x)\ dx = -\int_{-\infty}^{+\infty} x \frac{\partial F_x(x)}{\partial x}\ dx = \int_0^1 F_x^{-1}(v)\ dv\ , \quad (4.20)$$

where in establishing the second equality we used the fact that the pdf is the negative derivative of the CCDF, and in the last equality we introduced the change of variables $v = F_x(x)$. But note that the integral of the inverse CCDF can be written in terms of the integral of the CCDF as (see also Fig. 6)

$$\theta = -\int_{-\infty}^{0} [1 - F_x(u)]\ du + \int_0^{+\infty} F_x(u)\ du, \quad (4.21)$$

allowing one to express the mean θ of $x(n)$ in terms of its CCDF. To avoid carrying out integrals with infinite range, let us assume that $x(n) \in (-T, T)$ which is always practically satisfied for T sufficiently large, so that we can rewrite (4.21) as

$$\theta = \int_{-T}^{T} F_x(u)\ du\ -\ T. \quad (4.22)$$

Numerical evaluation of the integral in (4.22) can be performed using a number of known techniques. Let us consider an ordered set of interior points $\{\tau_k\}_{k=1}^{K}$ along with end-points $\tau_0 = -T$ and $\tau_{K+1} = T$. Relying on the fact that $F_x(\tau_0) = F_x(-T) = 1$ and $F_x(\tau_{K+1}) = F_x(T) = 0$, application of the trapezoidal rule for numerical integration yields (see also Fig. 6),

$$\theta = \frac{1}{2}\sum_{k=1}^{K}(\tau_{k+1} - \tau_{k-1})F_x(\tau_k)\ -\ T + e_a, \quad (4.23)$$

with e_a denoting the approximation error. Certainly, other methods like Simpson's rule, or the broader class of Newton-Cotes formulas, can be used to further reduce e_a.

Whichever the choice, the key is that binary observations constructed from the region $B_k := (\tau_k, \infty)$ have Bernoulli parameters

$$q_k := \Pr\{x(n) > \tau_k\} = F_x(\tau_k). \tag{4.24}$$

Inserting the non-parametric estimators $\hat{F}_x(\tau_k) = \hat{q}_k$ in (4.23), our parameter estimator when the noise pdf is unknown takes the form:

$$\hat{\theta} = \frac{1}{2}\sum_{k=1}^{K} \hat{q}_k(\tau_{k+1} - \tau_{k-1}) \; - \; T. \tag{4.25}$$

Since $\hat{q}_k$'s are unbiased, (4.23) and (4.25) imply that $\mathrm{E}(\hat{\theta}) = \theta + e_a$. Being biased, the proper performance indicator for $\hat{\theta}$ in (4.25) is the Mean Squared Error (MSE), not the variance.

Maintaining the bandwidth constraint of 1 bit per sensor (i.e. $K = 1$), let us divide the N sensors in K subgroups containing N/K sensors each, and define the regions

$$B_1(n) := B_k = (\tau_k, \infty), \quad n = (k-1)(N/K), \ldots, k(N/K) - 1; \tag{4.26}$$

the region $B_1(n)$ will be used by sensor n to construct and transmit the binary observation $b_1(n)$. Herein, the unbiased estimators of the Bernoulli parameters q_k are

$$\hat{q}_k = \frac{1}{(N/K)} \sum_{n=(k-1)(N/K)}^{k(N/K)-1} b_1(n), \qquad k = 1, \ldots, K, \tag{4.27}$$

and are used in (4.25) to estimate θ. It is easy to verify that $\mathrm{var}(\hat{q}_k) = q_k(1 - q_k)/(N/K)$, and that $\hat{q}_{k_1}$ and $\hat{q}_{k_2}$ are independent for $k_1 \neq k_2$.

The resultant MSE, $\mathrm{E}[(\theta - \hat{\theta})^2]$, will be bounded as follows[3].

Proposition 4.3. *Consider the estimator $\hat{\theta}$ given in (4.25), with $\hat{q}_k$ as in (4.27). Assume that for T sufficiently large and known $p_x(x) = 0$, for $|x| \geq T$; the noise pdf has bounded derivative $\dot{p}_w(u) := \partial p_w(w)/\partial w$; and define $\tau_{\max} := \max_k\{\tau_{k+1} - \tau_k\}$ and $\dot{p}_{\max} := \max_{u \in (-T,T)}\{\dot{p}_w(u)\}$. The MSE is given by,*

$$\mathrm{E}[(\theta - \hat{\theta})^2] = |e_a|^2 + \mathrm{var}(\hat{\theta}), \tag{4.28}$$

with the approximation error e_a and $\mathrm{var}(\hat{\theta})$, satisfying

$$|e_a| \leq \frac{T\dot{p}_{\max}}{6}\tau_{\max}^2, \tag{4.29}$$

$$\mathrm{var}(\hat{\theta}) = \sum_{k=1}^{K} \frac{(\tau_{k+1} - \tau_{k-1})^2}{4} \frac{q_k(1 - q_k)}{N/K}, \tag{4.30}$$

[3] Omitted due to space considerations, proofs pertaining to claims in this work can be found in [28].

with $\{\tau_k\}_{k=1}^K$ *a grid of thresholds in* $(-T,T)$ *and* $\{q_k\}_{k=1}^K$ *as in* (4.24).

Note from (4.30) that the larger contributions to $\mathrm{var}(\hat{\theta})$ occur when $q_k \approx 1/2$, since this value maximizes the coefficients $q_k(1-q_k)$; equivalently, this happens when the thresholds satisfy $\tau_k \approx \theta$ [c.f. (4.24)]. Thus, as with the case where the noise pdf is known, when θ belongs to an a priori known interval $[\Theta_1, \Theta_2]$, this knowledge must be exploited in selecting thresholds around the likeliest values of θ.

On the other hand, note that the $\mathrm{var}(\hat{\theta})$ term in (4.28) will dominate $|e_a|^2$, because $|e_a|^2 \propto \tau_{\max}^4$ as per (4.29). To clarify this point, consider an equispaced grid of thresholds with $\tau_{k+1} - \tau_k = \tau = \tau_{\max}$, $\forall k$, such that $\tau_{\max} = 2T/(K+1) < 2T/K$. Using the (loose) bound $q_k(1-q_k) \leq 1/4$, the MSE is bounded by [c.f. (4.28) - (4.30)]

$$\mathrm{E}[(\theta - \hat{\theta})^2] < \frac{4T^6 \dot{p}_{\max}^2}{9K^4} + \frac{T^2}{N}. \tag{4.31}$$

The bound in (4.31) is minimized by selecting $K = N$, which amounts to having *each sensor use a different region* to construct its binary observation. In this case, $|e_a|^2 \propto N^{-4}$ and its effect becomes practically negligible. Moreover, most pdfs have relatively small derivatives; e.g., for the Gaussian pdf we have $\dot{p}_{\max} = (2\pi e \sigma^4)^{-1/2}$. The integration error can be further reduced by resorting to a more powerful numerical integration method, although its difference with respect to the trapezoidal rule will not have any impact in practice.

Since $K = N$, the selection $\tau_{k+1} - \tau_k = \tau$, $\forall k$, yields

$$\hat{\theta} \;=\; \tau \sum_{n=0}^{N-1} b_1(n) - T \;=\; T\left[\frac{2}{N+1}\sum_{n=0}^{N-1} b_1(n) - 1\right], \tag{4.32}$$

that *does not require knowledge of the threshold* used to construct the binary observation at the fusion center of a WSN. This feature allows for each sensor to randomly select its threshold without using values pre-assigned by the fusion center; see also [16] for related random quantization algorithms.

Remark 4.1. While $e_a^2 \propto T^6$ seems to dominate $\mathrm{var}(\hat{\theta}) \propto T^2$ in (4.31), this is not true for the operational low-to-medium Q-SNR range for distributed estimators based on binary observations. This is because the support $2T$ over which $F_x(x)$ in (4.22) is non-zero depends on σ and the dynamic range $|\Theta_1 - \Theta_2|$ of the parameter θ. And as the Q-SNR decreases, $T \propto \sigma$. But since $\dot{p}_{\max} \propto \sigma^{-2}$, $e_a^2 \propto \sigma^2/N^4$ which is negligible when compared to the term $\mathrm{var}(\hat{\theta}) \propto \sigma^2/N$.

Remark 4.2. Pdf-unaware bandwidth-constrained distributed estimation was introduced in [16], where it was referred to as universal. At the (relatively minor) restriction of deterministically-assigned thresholds, the estimator in (4.32) achieves a four times smaller variance than the universal estimator in [16] which can afford randomly assigned thresholds –

though it is true that $\hat{\theta}$ in (4.32) can also be implemented with randomly assigned thresholds, its MSE in (4.31) has been derived for deterministically assigned ones. The reason behind this noticeable performance improvement is that the approach here implicitly utilizes the data pdf (through the numerical approximation of the CCDF) in constructing the asymptotic MLE of (4.25). The only extra condition required over [16] is for the pdf to be differentiable, which is typically satisfied in practice. Also, the approach herein is readily generalizable to estimation of vector parameters – a practical scenario where universal estimators like those in [16] are yet to be found.

Apart from providing useful bounds on the finite-sample performance, Eqs. (4.29), (4.30), and (4.31) establish asymptotic optimality of the $\hat{\theta}$ estimators in (4.25) and (4.32) as summarized in the following:

Corollary 4.1. *Under the assumptions of Propositions 4.3 and the conditions: i) $\tau_{\max} \propto K^{-1}$; and ii) $T^2/N, T^6/K^4 \to 0$ as $T, K, N \to \infty$, the estimators $\hat{\theta}$ in* (4.25) *and* (4.32) *are asymptotically (as $K, N \to \infty$) unbiased and consistent in the mean-square sense.*

The estimators in (4.25) and (4.32) are consistent even if the support of the data *pdf* is infinite, as long as we guarantee a proper rate of convergence relative to the number of sensors and thresholds.

Remark 4.3. To compare the estimators in (4.4) and (4.32), consider that $\theta \in [\Theta_1, \Theta_2] = [-\sigma, \sigma]$, and that the noise is Gaussian with variance σ^2, yielding a Q-SNR $\gamma = 4$. No estimator can have variance smaller than $\text{var}(\bar{x}) = \sigma^2/N$; however, for the (medium) $\gamma = 4$ Q-SNR value they can come close. For the known pdf estimator in (4.4), the variance is $\text{var}(\hat{\theta}) \approx 2\sigma^2/N$. The unknown pdf estimator in (4.32) requires an assumption about the essentially non-zero support of the Gaussian pdf. If we suppose that the noise pdf is non-zero over $[-2\sigma, 2\sigma]$, the corresponding variance becomes $\text{var}(\hat{\theta}) \approx 9\sigma^2/N$. The penalties due to the transmission of a single bit per sensor with respect to $\bar{x}$ are approximately 2 and 9. While the increasing penalty is expected as the uncertainty about the noise pdf increases, the relatively small loss is rather unexpected.

4.3. Vector parameter generalization. Let us now return to the general problem we started with in Section 2. We begin by defining the per sensor vector of binary observations $\mathbf{b}(n) := (b_1(n), \ldots, b_K(n))^T$, and note that since its entries are binary, realizations $\boldsymbol{\beta}$ of $\mathbf{b}(n)$ belong to the set

$$\mathcal{B} := \{\boldsymbol{\beta} \in \mathbf{R}^K \mid [\boldsymbol{\beta}]_k \in \{0, 1\},\ k = 1, \ldots, K\}, \tag{4.33}$$

where $[\boldsymbol{\beta}]_k$ denotes the k^{th} component of $\boldsymbol{\beta}$. With each $\boldsymbol{\beta} \in \mathcal{B}$ and each sensor we now associate the region

$$\mathbf{B}_{\beta}(n) := \bigcap_{[\boldsymbol{\beta}]_k = 1} B_k(n) \bigcap_{[\boldsymbol{\beta}]_k = 0} \bar{B}_k(n), \tag{4.34}$$

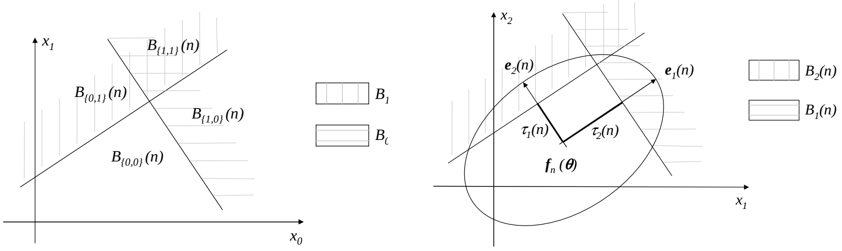

FIG. 7. *(Left): The vector of binary observations* $\mathbf{b}$ *takes on the value* $\{y_1, y_2\}$ *if and only if* $x(n)$ *belongs to the region* $B_{\{y_1,y_2\}}$*; (Right): Selecting the regions* $B_k(n)$ *perpendicular to the covariance matrix eigenvectors results in independent binary observations.*

where $\bar{B}_k(n)$ denotes the set-complement of $B_k(n)$ in $\mathbf{R}^M$. Note that the definition in (4.34) implies that $x(n) \in \mathbf{B}_\beta(n)$ if and only if $\mathbf{b}(n) = \boldsymbol{\beta}$; see also Fig. 7 (Left) for an illustration in $\mathbf{R}^2$ ($M = 2$). The corresponding probabilities are:

$$q_\beta(n) := \Pr\{\mathbf{b}(n) = \boldsymbol{\beta}\} = \int_{\mathbf{B}_\beta(n)} p_{\mathbf{w}}[\mathbf{u} - \mathbf{f}_n(\boldsymbol{\theta}); \boldsymbol{\psi}]\ d\mathbf{u}, \tag{4.35}$$

with $\mathbf{f}_n$ as in (4.1), and $\boldsymbol{\psi}$ containing the unknown parameters of the known noise pdf. Using definitions (4.35) and (4.33), we can write the pertinent log-likelihood function as

$$L(\boldsymbol{\theta}, \boldsymbol{\psi}) = \sum_{n=0}^{N-1} \sum_{y \in \mathcal{B}} \delta(\mathbf{b}(n) - \boldsymbol{\beta}) \ln q_\beta(n), \tag{4.36}$$

and the MLE of $\boldsymbol{\theta}$ as

$$\hat{\boldsymbol{\theta}} = \arg\max_{(\boldsymbol{\theta}, \boldsymbol{\psi})} L(\boldsymbol{\theta}, \boldsymbol{\psi})\ . \tag{4.37}$$

The nonlinear search needed to obtain $\hat{\boldsymbol{\theta}}$ could be challenging. Fortunately, as the following proposition asserts, under certain conditions that are usually met in practice, $L(\boldsymbol{\theta}, \boldsymbol{\psi})$ is concave which implies that computationally efficient search algorithms can be invoked to find its global maximum.

PROPOSITION 4.4. *If the MLE problem in* (4.37) *satisfies the conditions:*

- **[c1]** *The noise pdf* $p_{\mathbf{w}}(\mathbf{w}; \boldsymbol{\psi}) \leftrightarrow p_{\mathbf{w}}(\mathbf{w})$ *is log-concave [6, p.104], and* $\boldsymbol{\psi}$ *is known.*
- **[c2]** *The functions* $\mathbf{f}_n(\boldsymbol{\theta})$ *are linear; i.e.,* $\mathbf{f}_n(\boldsymbol{\theta}) = \mathbf{H}_n\boldsymbol{\theta}$*, with* $\mathbf{H}_n \in \mathbf{R}^{(M \times p)}$.
- **[c3]** *The regions* $B_k(n)$ *are chosen as half-spaces.*

then $L(\boldsymbol{\theta})$ *in* (4.36) *is a concave function of* $\boldsymbol{\theta}$.

Note that [c1] is satisfied by common noise pdfs, including the multivariate Gaussian [6, p.104]; and also that [c2] is typical in parameter estimation. Moreover, even when [c2] is not satisfied, linearizing $\mathbf{f}_n(\boldsymbol{\theta})$ using Taylor's expansion is a common first step, typical in e.g., parameter tracking applications. On the other hand, [c3] places a constraint in the regions defining the binary observations, which is simply up to the designer's choice.

4.3.1. Colored Gaussian noise. Analyzing the performance of the MLE in (4.37) is only possible asymptotically (as N or SNR go to infinity). Notwithstanding, when the noise is Gaussian, simplifications render variance analysis tractable and lead to interesting guidelines for constructing the estimator $\hat{\boldsymbol{\theta}}$.

Restrict $p_{\mathbf{w}}(\mathbf{w}; \boldsymbol{\psi}) \leftrightarrow p_{\mathbf{w}}(\mathbf{w})$ to the class of multivariate Gaussian pdfs, and let $\mathbf{C}(n)$ denote the noise covariance matrix at sensor n. Assume that $\{\mathbf{C}(n)\}_{n=0}^{N-1}$ are known and let $\{(\mathbf{e}_m(n),\ \sigma_m^2(n))\}_{m=1}^{M}$ be the set of eigenvectors and associated eigenvalues:

$$\mathbf{C}(n) = \sum_{m=1}^{M} \sigma_m^2(n)\mathbf{e}_m(n)\mathbf{e}_m^T(n). \tag{4.38}$$

For each sensor, we define a set of $K = M$ regions $B_k(n)$ as half-spaces whose borders are hyper-planes perpendicular to the covariance matrix eigenvectors; i.e.,

$$B_k(n) = \{\mathbf{x} \in \mathbf{R}^M \mid \mathbf{e}_k^T(n)\mathbf{x} \geq \tau_k(n)\}, \quad k = 1, \ldots, K = M, \tag{4.39}$$

Fig. 7 (Right) depicts the regions $B_k(n)$ in (4.39) for $M = 2$. Note that since each entry of $\mathbf{x}(n)$ offers a distinct scalar observation, the selection $K = M$ amounts to a bandwidth constraint of 1 *bit per sensor per dimension.*

The rationale behind this selection of regions is that the resultant binary observations $b_k(n)$ are independent, meaning that $\Pr\{b_{k_1}(n)b_{k_2}(n)\} = \Pr\{b_{k_1}(n)\}\ \Pr\{b_{k_2}(n)\}$ for $k_1 \neq k_2$. As a result, we have a total of MN independent binary observations to estimate $\boldsymbol{\theta}$.

Herein, the Bernoulli parameters $q_k(n)$ take on a particularly simple form in terms of the Gaussian tail function,

$$q_k(n) = \int_{\mathbf{e}_k^T(n)\mathbf{u} \geq \tau_k(n)} p_{\mathbf{w}}(\mathbf{u} - \mathbf{f}_n(\boldsymbol{\theta}))\, d\mathbf{u} = Q\left(\frac{\tau_k(n) - \mathbf{e}_k^T(n)\mathbf{f}_n(\boldsymbol{\theta})}{\sigma_k(n)}\right), \tag{4.40}$$

where we introduced the *σ-distance* between $\mathbf{f}_n(\boldsymbol{\theta})$ and the corresponding threshold $\Delta_k(n) := [\tau_k(n) - \mathbf{e}_k^T(n)\mathbf{f}_n(\boldsymbol{\theta})]/\sigma_k(n)$. Moreover, for simplicity we denote the Q function in (4.40) as $Q(\Delta_k(n))$.

Due to the independence among binary observations we have $p(\mathbf{b}(n)) = \prod_{k=1}^{K} [q_k(n)]^{b_k(n)}[1-q_k(n)]^{1-b_k(n)}$, leading to

$$L(\boldsymbol{\theta}) = \sum_{n=0}^{N-1}\sum_{k=1}^{K} b_k(n)\ln q_k(n) + [1-b_k(n)]\ln[1-q_k(n)], \tag{4.41}$$

whose NK *independent* summands replace the $N2^K$ *dependent* terms in (4.36).

Since the regions $B_k(n)$ are half-spaces, Proposition 4.4 applies to the maximization of (4.41) and guarantees that the numerical search for the $\hat{\boldsymbol{\theta}}$ estimator in (4.41) is well-conditioned and will converge to the global maximum, at least when the functions $\mathbf{f}_n$ are linear. More important, it will turn out that these regions render finite sample performance analysis of the MLE in (4.37), tractable. In particular, it is possible to derive a closed-form expression for the Fisher Information Matrix (FIM) [13, p.44], as we establish next.

Proposition 4.5. *The FIM,* $\mathbf{I}$, *for estimating* $\boldsymbol{\theta}$ *based on the binary observations obtained from the regions defined in* (4.39), *is given by*

$$\mathbf{I} = \sum_{n=0}^{N-1}\mathbf{J}_n^T\left[\sum_{k=1}^{K}\frac{e^{-\Delta_k^2(n)}\mathbf{e}_k(n)\mathbf{e}_k^T(n)}{2\pi\sigma_k^2(n)Q(\Delta_k(n))[1-Q(\Delta_k(n))]}\right]\mathbf{J}_n, \tag{4.42}$$

where $\mathbf{J}_n$ *denotes the Jacobian of* $\mathbf{f}_n(\boldsymbol{\theta})$.

Inspection of (4.42) shows that the variance of the MLE in (4.37) depends on the signal function containing the parameter of interest (via the Jacobians), the noise structure and power (via the eigenvalues and eigenvectors), and the selection of the regions $B_k(n)$ (via the σ-distances). Among these three factors only the last one is inherent to the bandwidth constraint, the other two being common to the estimator that is based on the original $\mathbf{x}(n)$ observations.

The last point is clarified if we consider the FIM $\mathbf{I}_x$ for estimating $\boldsymbol{\theta}$ given the unquantized vector observations $\mathbf{x}(n)$. This matrix can be shown to be (see [28, Apx. D]),

$$\mathbf{I}_x = \sum_{n=0}^{N-1}\mathbf{J}_n^T\left[\sum_{m=1}^{M}\frac{\mathbf{e}_m(n)\mathbf{e}_m^T(n)}{\sigma_m^2(n)}\right]\mathbf{J}_n^T. \tag{4.43}$$

If we define the equivalent noise powers as

$$\rho_k^2(n) := \frac{2\pi Q(\Delta_k(n))[1-Q(\Delta_k(n))]}{e^{-\Delta_k^2(n)}}\sigma_k^2(n), \tag{4.44}$$

we can rewrite (4.42) in the form

$$\mathbf{I} = \sum_{n=0}^{N-1}\mathbf{J}_n^T\left[\sum_{k=1}^{K}\frac{\mathbf{e}_k(n)\mathbf{e}_k^T(n)}{\rho_k^2(n)}\right]\mathbf{J}_n^T, \tag{4.45}$$

which except for the noise powers has form identical to (4.43). Thus, comparison of (4.45) with (4.43) reveals that from a performance perspective, *the use of binary observations is equivalent to an increase in the noise variance* from $\sigma_k^2(n)$ to $\rho_k^2(n)$, while the rest of the problem structure remains unchanged.

Since we certainly want the equivalent noise increase to be as small as possible, minimizing (4.44) over $\Delta_k(n)$ calls for this distance to be set to zero, or equivalently, to select thresholds $\tau_k(n) = \mathbf{e}_k^T(n)\mathbf{f}_n(\boldsymbol{\theta})$. In this case, the equivalent noise power is

$$\rho_k^2(n) = \frac{\pi}{2}\sigma_k^2(n). \tag{4.46}$$

Surprisingly, even in the vector case a judicious selection of the regions $B_k(n)$ results in a very small penalty $(\pi/2)$ in terms of the equivalent noise increase. Similar to Sections 4.1.1 and 4.1.2, we can thus claim that while requiring the transmission of 1 bit per sensor per dimension, the variance of the MLE in (4.37), based on $\{\mathbf{b}(n)\}_{n=0}^{N-1}$, yields a variance close to the clairvoyant estimator's variance –based on $\{\mathbf{x}(n)\}_{n=0}^{N-1}$– for low-to-medium Q-SNR problems.

5. Simulations.

In this section we provide numerical results for the distributed estimation schemes developed in Sections I and III.

5.1. Distributed dimensionality reduction.

We first test the MMSE performance versus k for the EC scheme and the estimator returned by Algorithm 1. To assess the difference in handling noise effects, we also compare EC and Algorithm 1 with the schemes in [38] and [37], which we abbreviate as C$'$E and C$''$E because they perform compression (C) followed by estimation (E). Although C$'$E and C$''$E have been derived under ideal link conditions, we modify them here to account for $\mathbf{D}_i$. Our comparisons will further include an option we term CE, which compresses first the data and reconstructs them at the FC using $\mathbf{C}^o$ and $\mathbf{B}^o$ found by (2.10) after setting $\mathbf{s} = \mathbf{x}$, and then estimates $\mathbf{s}$ based on the reconstructed data vector $\hat{\mathbf{x}}$. For benchmarking purposes, we also plot J_o, achieved when estimating $\mathbf{s}$ based on uncompressed data transmitted over ideal links.

Test Case 1 (EC with uncorrelated sensor data): We consider first the decoupled case of Section 3, where MMSE performance is characterized by the single sensor ($L = 1$) setup. Fig. 8 (Left) depicts the MMSE versus k for J_o, EC, CE, C$'$E and C$''$E for a linear model $\mathbf{x} = \mathbf{H}\mathbf{s} + \mathbf{n}$, where $N = 50$ and $p = 10$. The matrices $\mathbf{H}, \boldsymbol{\Sigma}_{ss}$ and $\boldsymbol{\Sigma}_{nn}$, are selected randomly such that $\text{tr}(\mathbf{H}\boldsymbol{\Sigma}_{ss}\mathbf{H}^T)/\text{tr}(\boldsymbol{\Sigma}_{nn}) = 2$, while $\mathbf{s}$ and $\mathbf{n}$ are uncorrelated. We set $\boldsymbol{\Sigma}_{zz} = \sigma_z^2\mathbf{I}_k$, and select P such that $10\log_{10}(P/\sigma_z^2) = 7$dB. As expected J_o benchmarks all curves, while the worst performance is exhibited by C$'$E. Albeit suboptimal, CE comes close to the optimal EC. The monotonic decrease of MMSE with k for EC corroborates Corollary 2. Contrasting it with the increase C$''$E exhibits in MMSE beyond a certain k, we can

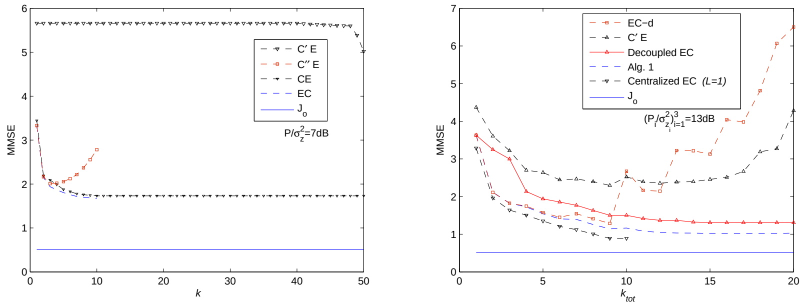

FIG. 8. *MMSE comparisons versus k for a centralized, $L = 1$ (Left), and a distributed 3-sensor setup (Right).*

appreciate the importance of coping with noise effects. This increase is justifiable since each entry of the compressed data in C$''$E is allocated a smaller portion of the given power as k grows. In EC however, the quality of channel links and the available power determine the number of the compressed components (which might lie in a vector space of dimensionality $\kappa \leq k$), and allocate power optimally among them.

Test Case 2 (Algorithm 1 with correlated sensor data): Here we consider a 3-sensor setup using the same linear model as in Test Case 1, while setting $N_1 = N_2 = 17$ and $N_3 = 16$. FC noise $\mathbf{z}_i$ is white with variance $\sigma_{z_i}^2$. The power P_i and variance $\sigma_{z_i}^2$ are chosen such that $10\log_{10}(P/\sigma_{z_i}^2) = 13$dB, for $i = 1, 2, 3$, and the tolerance quantity for the Algorithm 1 is set to $\epsilon = 10^{-3}$. Fig. 8 (Right) depicts the MMSE as a function of the total number $k_{tot} = \sum_{i=1}^{3} k_i$ of compressed entries across sensors for: i) a centralized EC setup for which a single (virtual) sensor ($L = 1$) has available the data vectors of all three sensors; ii) the estimator returned by Algorithm 1; iii) the decoupled EC estimator which ignores sensor correlations; iv) the C$'$E and v) an iterative estimator developed in [31], denoted here as EC-d, which similar to C$'$E accounts for fading but ignores noise. Interestingly, our decentralized Algorithm 1 comes very close to the hypothetical single-sensor bound of the centralized EC estimator, while outperforming the decoupled EC one. Also worth noting is that EC-d performs close to Algorithm 1 for small values of k_{tot}, but as k_{tot} increases it behaves as bad as C$'$E.

5.2. Scalar parameter estimation – parametric approach. We begin by simulating the estimator in (4.13) for scalar parameter estimation in the presence of AWGN with unknown variance. Results are shown in Fig. 9 for two different sets of σ-distances, Δ_1, Δ_2, corroborating the values predicted by (4.14) and the fact that the performance loss with respect to the clairvoyant sample mean estimator, $\bar{x}$, is indeed small.

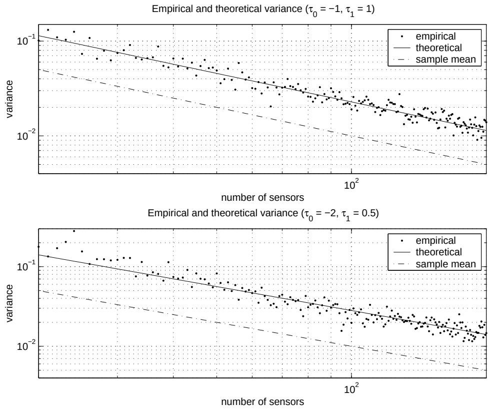

FIG. 9. *Noise of unknown power estimator. The simulation corroborates the close to clairvoyant variance prediction of* (4.14) *($\sigma = 1$, $\theta = 0$, Gaussian noise).*

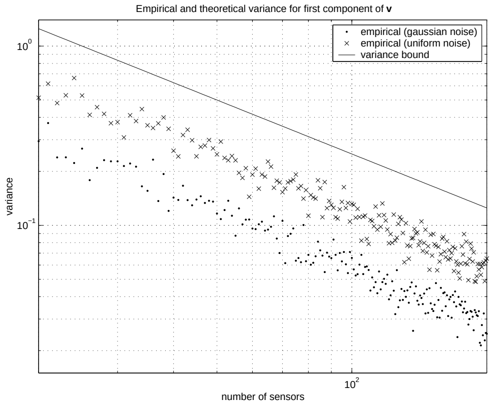

FIG. 10. *The variance of the estimators in* (4.4) *and* (4.32) *are close to the sample mean estimator variance ($\sigma^2 := \mathrm{E}[w^2(n)] = 1$, $T = 3$, $\theta \in [-1, 1]$).*

5.3. Scalar parameter estimation - unknown noise pdf. Fig. 10 depicts theoretical bounds and simulated variances for the estimators (4.4) and (4.32) for an example Q-SNR $\gamma = 4$. The sample mean estimator variance, $\mathrm{var}(\bar{x}) = \sigma^2/N$, is also depicted for comparison purposes. The simulations corroborate the implications of Remark 3, reinforcing the idea that for low to medium Q-SNR problems quantization to a single bit per observation leads to minimal losses in variance performance. Note that for this particular example the unknown pdf variance bound, (4.31), overestimates the variance by a factor of roughly 1.2 for the uniform case and roughly 2.6 for the Gaussian case.

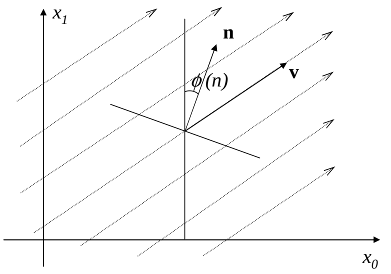

FIG. 11. *The vector flow* $\mathbf{v}$ *incises over a certain sensor capable of measuring the normal component of* $\mathbf{v}$.

5.4. Vector parameter estimation – A motivating application. In this section, we illustrate how a problem involving vector parameters can be solved using the estimators of Section 4.3.1. Suppose we wish to estimate a vector flow using incidence observations. With reference to Fig. 11, consider the flow vector $\mathbf{v} := (v_0, v_1)^T$, and a sensor positioned at an angle $\phi(n)$ with respect to a known reference direction. We will rely on a set of so called incidence observations $\{x(n)\}_{n=0}^{N-1}$ measuring the component of the flow normal to the corresponding sensor

$$x(n) := \langle \mathbf{v}, \mathbf{n} \rangle + w(n) = v_0 \sin[\phi(n)] + v_1 \cos[\phi(n)] + w(n), \tag{5.1}$$

where $\langle , \rangle$ denotes inner product, $w(n)$ is zero-mean AWGN, and $n = 0, 1, \ldots, N-1$ is the sensor index. The model (5.1) applies to the measurement of hydraulic fields, pressure variations induced by wind and radiation from a distant source [20].

Estimating $\mathbf{v}$ fits the framework of Section 4.3.1 requiring the transmission of a single binary observation per sensor, $b_1(n) = \mathbf{1}\{x(n) \geq \tau_1(n)\}$. The FIM in (4.45) is easily found to be

$$\mathbf{I} = \sum_{n=0}^{N-1} \frac{1}{\rho_1^2(n)} \begin{pmatrix} \sin^2[\phi(n)] & \sin[\phi(n)]\cos[\phi(n)] \\ \sin[\phi(n)]\cos[\phi(n)] & \cos^2[\phi(n)] \end{pmatrix}. \tag{5.2}$$

Furthermore, since $x(n)$ in (5.1) is linear in $\mathbf{v}$ and the noise pdf is log-concave (Gaussian) the log-likelihood function is concave as asserted by Proposition 4.4.

Suppose that we are able to place the thresholds optimally as implied by $\tau_1(n) = v_0 \sin[\phi(n)] + v_1 \cos[\phi(n)]$, so that $\rho_1^2(n) = (\pi/2)\sigma^2$. If we also make the reasonable assumption that the angles are random and uniformly distributed, $\phi(n) \sim U[-\pi, \pi]$, then the average FIM turns out to be:

$$\bar{\mathbf{I}} = \frac{2}{\pi\sigma^2} \begin{pmatrix} N/2 & 0 \\ 0 & N/2 \end{pmatrix}. \tag{5.3}$$

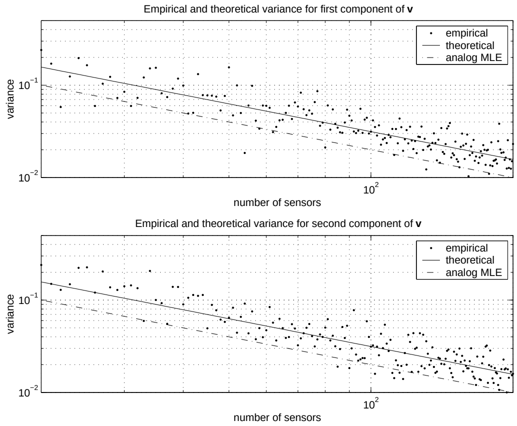

FIG. 12. *Average variance for the components of* $\mathbf{v}$. *The empirical as well as the bound* (5.4) *are compared with the analog observations based MLE* ($\mathbf{v} = (1, 1)$, $\sigma = 1$).

But according to the law of large numbers $\mathbf{I} \approx \bar{\mathbf{I}}$, and the estimation variance will be approximately given by

$$\operatorname{var}(v_0) = \operatorname{var}(v_1) = \frac{\pi\sigma^2}{N}. \tag{5.4}$$

Fig. 12 depicts the bound (5.4), as well as the simulated variances $\operatorname{var}(\hat{v}_0)$ and $\operatorname{var}(\hat{v}_1)$ in comparison with the clairvoyant MLE based on $\{x(n)\}_{n=0}^{N-1}$, corroborating our analytical expressions.

6. Conclusions. We considered the problem of estimation in wireless sensor networks showing that the seemingly unrelated problems of dimensionality reduction, compression, quantization and estimation are actually intertwined due to the distributed nature of the WSN.

We started by deriving algorithms for estimating stationary random signals based on reduced-dimensionality observations collected by power-limited wireless sensors linked with a fusion center. We dealt with non-ideal channel links that are characterized by multiplicative fading and additive noise. When data across sensors are uncorrelated, we established global mean-square error optimal schemes in closed-form and proved that they implement estimation followed by compression per sensor. Besides distributed estimation with reduced dimensionality decoupled observations, such closed-form solutions are valuable for all applications principal components and canonical correlation analysis are sought in the presence of multiplicative and additive noise. For correlated sensor observations, we developed an algorithm that relies on block coordinate descent iterations which are guaranteed to converge at least to a local stationary point of the associate mean-square error cost. The optimal estimators allocate properly the prescribed power following a waterfilling-like principle to balance judiciously channel effects and additive noise at the fusion center with the degree of dimensionality reduction that can be afforded.

Continuing, with digital-amplitude data transmission we determined the distortion −rate (D-R) function for estimating a random vector in a single-sensor setup and established the optimality of the estimate-first compress-afterwards (EC) approach along with the suboptimality of a compress-first estimate afterwards (CE) alternative. When it comes to estimation using multiple sensors, the corresponding D-R function can be bounded from below using the single-sensor D-R function achieved using the EC scheme. An alternating algorithm was also derived for determining numerically an achievable D-R upper bound in the distributed multi-sensor setup. Using this upper bound in combination with the non-achievable lower bound we obtained a tight region, where the D-R function for distributed estimation lies in.

We finally developed parameter estimators for realistic signal models and derived their fundamental variance limits under severe bandwidth constraints. The latter were adhered to by quantizing each sensor's observation to one or a few bits. By jointly accounting for the unique quantization-estimation tradeoffs present, these bit(s) per sensor were first used to derive distributed maximum likelihood estimators (MLEs) for scalar mean-location parameters in the presence of generally non-Gaussian noise when the noise pdf is completely known; subsequently, when the pdf is known except for a number of unknown parameters; and finally, when the noise pdf is unknown. The unknown pdf case was tackled through a non-parametric estimator of the unknown complementary cumulative distribution function based on quantized (binary) observations. In all three cases, the resulting estimators turned out to exhibit comparable variances that can come surprisingly close to the variance of the clairvoyant estimator which relies on unquantized observations. This happens when the SNR capturing both quantization and noise effects assumes low-to-moderate values. Analogous claims were established for practical generalizations that were pursued in the multivariate and colored noise cases for distributed estimation of vector parameters under bandwidth constraints. Therein, MLEs were formed via numerical search but the log-likelihoods were proved to be concave thus ensuring fast convergence to the unique global maximum.

REFERENCES

[1] M. Abdallah and H. Papadopoulos, *Sequential signal encoding and estimation for distributed sensor networks,* in Proc. of the International Conference on Acoustics, Speech, and Signal Processing, 4: 2577–2580, Salt Lake City, Utah, May 2001.

[2] B. Beferull-Lozano, R.L. Konsbruck, and M. Vetterli, *Rate-Distortion problem for physics based distributed sensing,* in Proc. of the International Conference on Acoustics, Speech, and Signal Processing, 3: 913–916, Montreal, Canada, May 2004.

[3] T. Berger, *Rate Distortion Theory: A Mathematical Basis for Data Compression.* Prentice Hall, 1971.

[4] T. Berger, *Multiterminal Source Coding,* in Lectures Presented at CISM Summer School on the Info. Theory Approach to Comm., July 1977.

[5] D. Blatt and A. Hero, *Distributed maximum likelihood estimation for sensor networks,* in Proc. of the International Conference on Acoustics, Speech, and Signal Processing, 3: 929–932, Montreal, Canada, May 2004.

[6] S. Boyd and L. Vandenberghe, *Convex Optimization.* Cambridge University Press, 2004.

[7] J. Chen, X. Zhang, T. Berger, and S.B. Wicker, *An Upper Bound on the Sum-Rate Distortion Function and Its Corresponding Rate Allocation Schemes for the CEO Problem,* IEEE Journal on Selected Areas in Communications, pp. 406–411, August 2004.

[8] T. Cover and J. Thomas, *Elements of Information Theory.* John Wiley and Sons, 2nd edition ed., 1991.

[9] E. Ertin, R. Moses, and L. Potter, *Network parameter estimation with detection failures,* in Proc. of the Intnl. Conference on Acoustics, Speech, and Signal Processing, 2: 273–276, Montreal, Canada, May 2004.

[10] M. Gastpar, P.L. Draggoti, and M. Vetterli, *The distributed Karhunen-Loève transform,* IEEE Transactions on Information Theory, submitted Nov. 2004 (available at http://www.eecs.berkeley.edu/∼gastpar/).

[11] J. Gubner, *Distributed Estimation and Quantization,* IEEE Transactions on Information Theory, 39: 1456–1459, 1993.

[12] P. Ishwar, R. Puri, K. Ramchadran, and S. Pradhan, *On Rate-Constrained Distributed Estimation in Unreliable Sensor Networks,* IEEE Journal on Selected Areas in Communications, pp. 765–775, April 2005.

[13] S.M. Kay, *Fundamentals of Statistical Signal Processing - Estimation Theory.* Prentice Hall, 1993.

[14] S. Kumar, F. Zao, and D. Shepherd, eds., *Special issue on collaborative information processing,* Vol. 19 of IEEE Signal Proc. Magazine, March 2002.

[15] W. Lam and A. Reibman, *Quantizer design for decentralized systems with communication constraints,* IEEE Transactions on Communications, 41: 1602–1605, Aug. 1993.

[16] Z.-Q. Luo, *An isotropic universal decentralized estimation scheme for a bandwidth constrained ad hoc sensor network,* IEEE Journal on Selected Areas in Communications, 23: 735–744, April 2005.

[17] Z.-Q. Luo, *Universal Decentralized Estimation in a Bandwidth Constrained Sensor Network,* IEEE Transactions on Information Theory, 51: 2210–2219, June 2005.

[18] Z.-Q. Luo, G.B. Giannakis, and S. Zhang, *Optimal linear decentralized estimation in a bandwidth constrained sensor network,* in Proc. of the Intl. Symp. on Info. Theory, pp. 1441–1445, Adelaide, Australia, Sept. 4–9 2005.

[19] Z.-Q. Luo and J.-J. Xiao, *Decentralized estimation in an inhomogeneous sensing environment,* IEEE Transactions on Information Theory, 51: 3564 –3575, October 2005.

[20] A. Mainwaring, D. Culler, J. Polastre, R. Szewczyk, and J. Anderson, *Wireless sensor networks for habitat monitoring,* in Proc. of the 1st ACM International Workshop on Wireless Sensor Networks and Applications, 3: 88–97, Atlanta, Georgia, 2002.

[21] R.D. Nowak, *Distributed EM algorithms for density estimation and clustering in sensor networks,* IEEE Transactions on Signal Processing, 51: 2245–2253, August 2002.

[22] Y. Oohama, *The Rate-Distortion Function for the Quadratic Gaussian CEO Problem,* IEEE Transactions On Information Theory, pp. 1057–1070, May 1998.

[23] A. Pandya, A. Kansal, G. Pottie, and M. Srivastava, *Fidelity and Resource Sensitive Data Gathering,* in Proc. of the 42nd Allerton Conference, Allerton, IL, September 2004.

[24] H. Papadopoulos, G. Wornell, and A. Oppenheim, *Sequential signal encoding from noisy measurements using quantizers with dynamic bias control,* IEEE Transactions on Information Theory, 47: 978–1002, 2001.

[25] S.S. Pradhan, J. Kusuma, and K. Ramchandran, *Distributed compression in a dense microsensor network,* IEEE Signal Processing Magazine, 19: 51–60, March 2002.

[26] M.G. Rabbat and R.D. Nowak, *Decentralized source localization and tracking,* in Proc. of the 2004 IEEE Intnl. Conference on Acoustics, Speech, and Signal Processing, 3: 921–924, Montreal, Canada, May 2004.

[27] A. Ribeiro and G.B. Giannakis, *Bandwidth-Constrained Distributed Estimation for Wireless Sensor Networks, Part I: Gaussian Case,* IEEE Transactions on Signal Processing, 54: 1131–1143, March 2006.

[28] A. Ribeiro and G.B. Giannakis, *Bandwidth-Constrained Distributed Estimation for Wireless Sensor Networks, Part II: Unknown pdf,* IEEE Transactions on Signal Processing, 2006, to appear.

[29] D.J. Sakrison, *Source encoding in the presence of random disturbance,* IEEE Transactions on Information Theory, pp. 165–167, January 1968.

[30] I.D. Schizas, G.B. Giannakis, and N. Jindal, *Distortion-Rate Analysis for Distributed Estimation with Wireless Sensor Networks,* IEEE Transactions On Information Theory, submitted December 2005 (available at http://spincom.ece.umn.edu/).

[31] I.D. Schizas, G.B. Giannakis, and Z.-Q. Luo, *Distributed estimation using reduced dimensionality sensor observations,* IEEE Transactions on Signal Processing, submitted November 2005 (available at http://spincom.ece.umn.edu/).

[32] Y. Sung, L. Tong, and A. Swami, *Asymptotic locally optimal detector for large-scale sensor networks under the Poisson regime,* in Proc. of the International Conference on Acoustics, Speech, and Signal Processing, 2: 1077–1080, Montreal, Canada, May 2004.

[33] P.K. Varshney, *Distributed Detection and Data Fusion.* Springer-Verlag, 1997.

[34] H. Viswanathan and T. Berger, *The Quadratic Gaussian CEO Problem, IEEE Transactions on Information Theory*, pp. 1549–1559, September 1997.

[35] J. Wolf and J. Ziv, *Transmission of noisy information to a noisy receiver with minimum distortion,* IEEE Transactions on Information Theory, pp. 406–411, July 1970.

[36] A. Wyner and J. Ziv, *The Rate-Distortion Function for Source Coding with Side Information at the Decoder,* IEEE Trans. on Info. Theory, pp. 1–10, January 1976.

[37] K. Zhang, X.R. Li, P. Zhang, and H. Li, *Optimal linear estimation fusion–Part VI: Sensor data compression,* in Proc. of the Intl. Conf. on Info. Fusion, pp. 221–228, Queensland, Australia 2003.

[38] Y. Zhu, E. Song, J. Zhou, and Z. You, *Optimal dimensionality reduction of sensor data in multisensor estimation fusion,* IEEE Transactions on Signal Processing, 53: 1631–1639, May 2005.