On Rate-constrained Estimation in Unreliable Sensor Networks*

Prakash Ishwar[1], Rohit Puri[1],
S. Sandeep Pradhan[2], and Kannan Ramchandran[1]

[1] Department of Electrical Engineering and Computer Sciences,
University of California, Berkeley, CA 94720, USA.
{ishwar, rpuri, kannanr}@eecs.berkeley.edu
[2] Department of Electrical Engineering and Computer Science,
University of Michigan, Ann Arbor, MI 48103, USA.
pradhanv@eecs.umich.edu

Abstract. We study a network of non-collaborating sensors that make noisy measurements of some physical process X and communicate their readings to a central processing unit. Limited power resources of the sensors severely restrict communication rates. Sensors and their communication links are both subject to failure; however, the central unit is guaranteed to receive data from a minimum fraction of the sensors, say k out of n sensors. The goal of the central unit is to optimally estimate X from the received transmissions under a specified distortion metric. In this work, we derive an information-theoretically achievable rate-distortion region for this network under symmetric sensor measurement statistics.

When all processes are jointly Gaussian and independent, and we have a squared-error distortion metric, the proposed distributed encoding and estimation framework has the following interesting optimality property: *when any k out of n rate-R bits/sec sensor transmissions are received, the central unit's estimation quality matches the best estimation quality that can be achieved from a completely reliable network of k sensors, each transmitting at rate R. Furthermore, when more than k out of the n sensor transmissions are received, the estimation quality strictly improves.*

When the network has clusters of collaborating sensors should clusters compress their raw measurements or should they first try to estimate the source from their measurements and compress the estimates instead. For some interesting cases, we show that there is no loss of performance in the distributed compression of local estimates over the distributed compression of raw data in a rate-distortion sense, i.e., encoding the local sufficient statistics is good enough.

1 Introduction

With the recent proliferation of distributed sensor networks designed to capture physical phenomena of interest to the scientific and military communities, there

* This research was supported by NSF under grant CCR-0219722 and DARPA under grant F30602-00-2-0538.

has arisen the need to develop a fundamental understanding of the performance bounds for different signal processing tasks (such as estimation, detection, and classification) under a distributed regime for sensing, communicating, and processing. More specifically, while the classical signal processing literature has a mature knowledge-base on the problems of estimation and detection [1], the distributed dense sensor network setting imposes additional constraints. These include (i) bandwidth (due to low power requirements), (ii) distributed processing (due to prohibitive inter-sensor communication costs), and (iii) robustness requirements (due to the wireless transmission channel).

This paper addresses the estimation problem under these relevant constraints with the goal of providing fundamental information-theoretic performance bounds for an important class of estimation problems. In our problem formulation, we translate the three sensor network requirements listed above into the specific constraints of (i) bit-rate, (ii) fully distributed processing of the sensors, and (iii) a packet-erasure communication model respectively.

Figure 1 depicts a mathematical abstraction of a sensor network consisting of individual sensors that make synchronous, correlated, noisy measurements of some physical process X (e.g., temperature). Individual sensors are typically lightweight devices that have very limited battery power. Inter-sensor communication has heavy power and protocol overhead that strongly discourages information exchange between sensors. The goal is to communicate sensor measurements to a central query in a manner that allows the central unit to form the best possible estimate of the physical process with respect to a specified (distortion) metric of estimation quality. Note that the central unit is only a logical entity and can also be one of the other sensors which acts as a "cluster-head" with different sensors taking turns.

The wireless communication links from the sensors to the central processing unit are often through multiple hops that are time-varying and characterized by deep fades. Furthermore, the availability of only limited transmission power severely constrains sensor communication rates. Motivated by this, in this work, the effective communication links from individual sensors to the central unit are modeled as independent, discrete, rate-constrained channels that are subject to erasures. The erasure pattern is assumed to remain fixed over some large block length of sensor measurements. We invoke a rate-constrained communication model in keeping with the existing base of digital communication architectures[3].

Prior work on information-theoretic aspects of this sensor network has focused exclusively on the case when all sensors and their effective communication links to the central decoding unit are functional [3]. In [3], Oohama provides a complete characterization of the rate-distortion region when all sensors and their communication links are reliable, and when the source and measurement

[3] Note that in some scenarios where multihop communication is not needed, and when the physical process, the communication channel, and the distortion measure are statistically "matched" to one another [2] then an uncoded "analog" transmission might be desirable if distributed synchronization of these sensor signals can be orchestrated at the physical layer.

processes are independent across sensors and time and are Gaussian distributed. However, in most practical scenarios, the sensor nodes and their communication links are subject to failure.

In Section 2, we present an achievable rate-distortion region for the sensor network when, out of a total of n rate-R sensor transmissions, at least some k are guaranteed to successfully arrive at the decoding unit while the remaining transmissions are "erased". Our analysis assumes symmetric sensor measurement noise statistics, i.e., the measurements of all sensors are statistically indistinguishable. This assumption is motivated by the consideration that in practice, due to lack of information exchange between sensors, it will be difficult to estimate the relative measurement noise power of individual sensors. With this assumption, the reconstruction quality depends only on the number of sensor transmissions received, and not on *which* specific subset of sensors measurements is received. We would like to note, however, that the presented results and methods of analysis can be extended to handle asymmetric sensor noise as well. As discussed above, transmission power and bandwidth limitations impose severe communication rate restrictions on the sensors: it is assumed that each sensor cannot use more than R bits per noisy sample of the physical process being measured. We present an information-theoretic code construction that enables reconstruction of X with a quality that is commensurate with a total network rate of kR bits per source sample when *some* k out of n encoded sensor transmissions are received. The estimation quality strictly improves when more transmissions are received.

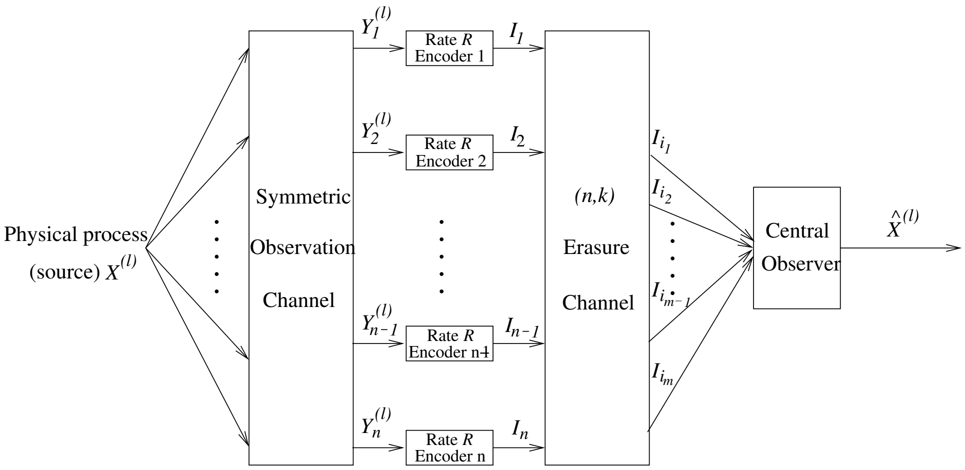

Fig. 1. An unreliable sensor network. A noisy version of source X is observed at each of the n sensors. The noisy observations $Y_1, ..., Y_n$, which are statistically indistinguishable, are encoded and the encoded indices $I_1, ..., I_n$ transmitted. The network delivers *some* $m \geq k$ indices. The decoder obtains the best possible reconstruction of X from the received indices.

There are two important aspects associated with the sensor network of Figure 1. First, the sensor measurements are correlated, and the power and protocol overhead of inter-sensor communication precludes any collaboration between sensors. Since both the transmission power and bandwidth are at a premium, efficient compression is a necessity. This motivates the development of algorithms that provide efficient distributed compression. The problem of distributed compression was first studied within an information-theoretic framework by Slepian and Wolf for lossless distributed compression [4]. The second important aspect associated with the sensor network is that only some k out of the n sensor readings are guaranteed to be received by the central unit. This implies an inherent uncertainty at the sensor encoders as to which measurements will be received by the central unit. The decoder, however, knows which sensor transmissions were successfully received and which got "erased". This can be interpreted as *side information* known only to the decoder but not to the encoders. The encoders, however, do know the network reliability in terms of the number of transmissions that are guaranteed to reach the decoder. Thus the encoders have access to the "statistics" of the side information process but not its particular realization. This observation provides a link to the theory of source coding with side information at the decoder studied by Wyner and Ziv in [5] where the statistics of the side information is available to the source encoder but the actual realization is not. In some interesting cases (e.g., for jointly Gaussian statistics), there is no loss of performance if the side-information is available at only one end (the receiver). Constructive algorithms using error control codes, inspired by information-theoretic random binning (hashing) ideas in [4, 5], have been developed recently in [6, 7].

Central to the problems of distributed compression of correlated sources and source coding with side information at the decoder is the notion of random binning [4, 5]. The concept of random binning involves each encoder sending an index for a *list of messages* (a bin) rather than a unique index for each message. This reduces the rate of transmission. If the bin size is small enough, it is possible to construct message lists for each encoder in a manner that allows the decoder to disambiguate the transmitted messages from the list using the side information available to it. The construction of these bins relies only on the joint statistics of the correlated messages that the encoders need to communicate to the central decoding unit. To further develop intuition for the problem at hand, suppose that the encoder alphabet sizes are all finite, the sensor measurements are symmetrically correlated, and the total rate available to the network as a whole equals the joint entropy [8] of *any* k noisy observations. In this case, a straightforward application of the Slepian-Wolf coding theorem for multiple sources [8] shows that if the total rate is split equally amongst all n sensors, then with the reception of *any* k transmissions, the decoder can reconstruct the corresponding noisy observations perfectly (also see [9]). This shows that it is possible to achieve the same performance in the unreliable case as when the encoders know in advance *which* specific k sensor transmissions will reach the decoder.

In Section 3 we specialize the achievability results of Section 2 to the Gaussian setting. This reveals an interesting optimality property of the proposed robust, distributed encoding and estimation framework: When *any* k out of n *unreliable* sensor transmissions are received, the central observer's estimation quality can be as good as the best reconstruction quality that can be achieved by deploying only k *reliable* sensors and the central decoding unit is able to receive the encoded noisy observations from all k sensors. Furthermore, when more than k out of the n sensor transmissions are received, the estimation quality strictly improves.

Section 4 discusses an important question associated with distributed, rate-constrained estimation: if sensors could collaborate in small local clusters, should they compress their raw measurements or should they first try to estimate the source from their measurements and compress the estimates instead. We show, under suitable conditions, that distributed compression of local estimates is optimal.

2 Problem set-up and main result

Random quantities are denoted by capital letters while their specific realizations are denoted by small letters. Let $\{X(t)\}_{t=1}^{\infty}$ be an i.i.d. sequence of source symbols and $Y_1(t), ..., Y_n(t)$ the noisy measurements of the n sensors at time t. The noisy sensor measurements take values in a common finite alphabet $\mathcal{Y}$ and are assumed to be symmetrically correlated conditioned on the source values[4]. Let $\{\mathbf{Y}(t)\}_{t=1}^{\infty}$ be an i.i.d. sequence of noisy sensor measurements where $\mathbf{Y}(t) := (Y_1(t), ..., Y_n(t))$. Let

$$F_i^{(l)} : \mathcal{Y}^l \longrightarrow \mathcal{I}_{2^{lR}},$$

be the rate-R encoding function for the i^{th} sensor ($i = 1, \ldots, n$) for block-length l, where $\mathcal{I}_m := \{1, ..., m\}$ and $\mathcal{Y}^l$ is the l-fold Cartesian product of the observation alphabet $\mathcal{Y}$. In other words, sensor observations are encoded in blocks of length l at a time by each sensor. Every block of observations is mapped to an index in the set $\{1, \ldots, 2^{lR}\}$. The encoding rate is therefore R bits per sample. Let $\widehat{\mathcal{X}}$ be the reconstruction alphabet and let

$$G_m^{(l)} : \mathcal{I}_{2^{lR}}^m \longrightarrow \widehat{\mathcal{X}}^l$$

be the decoding function when measurements from $m \geq k$ sensors are received by the central observer. Here, $\widehat{X}^{(l)} := \widehat{X}(1), ..., \widehat{X}(l)$. In other words, the decoder at the central unit maps a set of $m \geq k$ indices it receives to a reconstruction vector of length l. Let $d : \mathcal{X} \times \widehat{\mathcal{X}} \longrightarrow \mathbb{R}^+$ be a bounded distortion metric (i.e., $d(x, \widehat{x}) \leq d_{\max} < \infty$ for all $(x, \widehat{x}) \in \mathcal{X} \times \widehat{\mathcal{X}}$). The distortion metric on $\mathcal{X}^l \times \widehat{\mathcal{X}}^l$ is defined as $d^{(l)}(x^{(l)}, \widehat{x}^{(l)}) := \frac{1}{l} \sum_{t=1}^{l} d(x(t), \widehat{x}(t))$.

[4] All subsets of random variables of the same cardinality have the same joint conditional mass function.

Definition 1. *A rate-distortion tuple* $(R, D_k, ..., D_n)$ *is said to be achievable if for every* $\epsilon > 0$ *there exists a sufficiently long block length* $l(\epsilon)$ *and rate-*R *encoders* $\{F_i^{(l)}\}_{i=1}^n$ *and decoders* $\{G_m^{(l)}\}_{m=k}^n$ *such that*

$$E\left[d^{(l)}\left(X^{(l)}, G_m^{(l)}\left(F_{i_1}^{(l)}(Y_{i_1}^{(l)}), ..., F_{i_m}^{(l)}(Y_{i_m}^{(l)})\right)\right)\right] \leq D_m + \epsilon,$$

for all $k \leq m \leq n$.

Here, E denotes the expectation operator. The expected distortion depends only on m, and not on the specific values of $i_1, ..., i_m$ due to symmetric observation statistics. In plain words, a rate-distortion tuple is achievable if for arbitrarily stringent tolerance specifications ϵ there exist rate-R encoders and decoders for which the expected estimation quality at the reception of m sensor readings is within ϵ of D_m, for all $m = k, \ldots, n$.

Conditioned on $\mathbf{Y}$, let $U_1, ..., U_n$ be independent random variables taking values in a common finite alphabet $\mathcal{U}$ with

$$p_{\mathbf{U}|\mathbf{Y}}(u_1, ..., u_n | y_1, ..., y_n) := \prod_{i=1}^{n} p_{U|Y}(u_i | y_i), \tag{1}$$

where $\mathbf{U} := (U_1, ..., U_n)$. The random variables U_i may be thought of as arising from a rate-distortion quantization of Y_i. Define

$$R_k^*(p_{\mathbf{U}|\mathbf{Y}}) := \frac{1}{k}\left[H(U_1, ..., U_k) - \sum_{i=1}^{k} H(U_i | Y_i)\right], \tag{2}$$

$g_m : \mathcal{U}^m \longrightarrow \widehat{\mathcal{X}}$, $k \leq m \leq n$, and

$$D_m^*(p_{\mathbf{U}|\mathbf{Y}}, g_m) := E\left[d(X, g_m(U_{i_1}, ..., U_{i_m}))\right]. \tag{3}$$

Due to the symmetry of the underlying probability distributions, $H(U_i|Y_i)$ is the same for all i and the expected distortion in (3) only depends on m, g_m, and $p_{\mathbf{U}|\mathbf{Y}}$ and not on the specific values of $i_1, ..., i_m$. The main result of this paper is contained in the following theorem whose proof is outlined in Section 4.

Theorem 1. *If* $R > R_k^*(p_{\mathbf{U}|\mathbf{Y}})$ *and* $D_m \geq D_m^*(p_{\mathbf{U}|\mathbf{Y}}, g_m)$ *for all* $m \geq k$ *then the rate distortion tuple* $(R, D_k, ..., D_n)$ *is achievable.*

The above theorem describes the estimation quality that can be achieved at the central processing unit for a given bit budget. In the above description, the dependence of estimation quality on the bit-budget is expressed in terms of the quantization strategy (governed by $p_{\mathbf{U}|\mathbf{Y}}$) and the estimation algorithm governed by $\{g_m\}_{m=k}^n$. By varying the quantization and estimation strategies, one can trade off the required rate R and estimation quality $\{D_m\}_{m=k}^n$.

Figure 2 shows the detailed structure of sensor encoders and the central decoder that can asymptotically achieve the rate-distortion region of Theorem 1 for long block lengths. The encoders have two parts. The first stage is a rate R', block-length l, rate-distortion quantizer matched to the statistics of the sensor measurements. The second stage involves a random binning that exploits the correlation between the quantized observations across different sensors to get a rebate of $R' - R$ bits per sample. The decoder first reconstructs the quantized sensor observations from the bin indices of different sensors and then forms the best estimate of the physical process of interest.

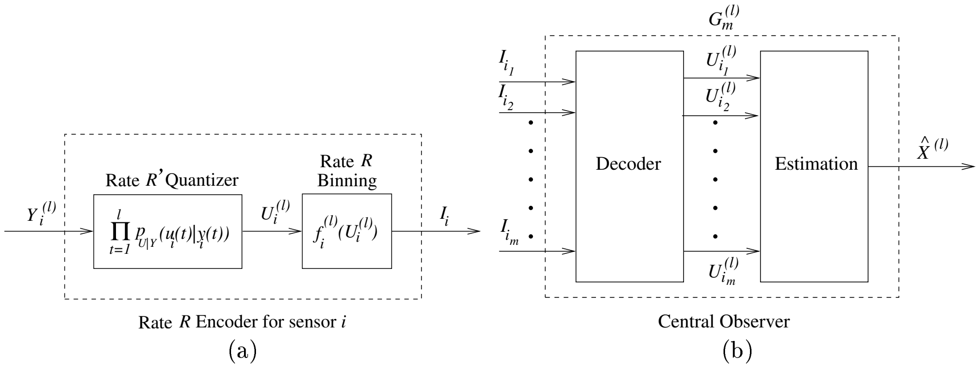

Fig. 2. Structure of sensor encoders: (a) and the central observer decoder: (b) that can achieve the rate-distortion performance of Theorem 1. The encoder of each sensor consists of a rate R' quantizer, for block length l, followed by a rate R random binning function. The central observer first decodes the received indices to intermediate representations, of length l, and then forms the best estimate of the source. Here, $m \geq k$.

3 The Gaussian sensor network

The result of the previous section can be extended to continuous alphabet sources – with differential entropy [8] replacing discrete entropy in all expressions – using techniques similar to those used in [10]. In this section, we specialize our results to the setting described in [3]: $\{X(t)\}_{t=1}^{\infty} \sim \mathcal{N}(0, \sigma_X^2)$ is an i.i.d. Gaussian source and $Y_i(t) = X(t) + N_i(t), \forall i \in \mathcal{I}_n$. The observation noise processes $N_i(t)$ are independent of each other and the source process. Across time, the noise processes are i.i.d., zero-mean, and Gaussian distributed with variance σ_N^2. With this setup note that the noisy sensor observations are conditionally independent and symmetrically correlated given the source samples. Estimation quality is measured by mean squared error, i.e., $d(x, \widehat{x}) = (x - \widehat{x})^2$. Our problem differs from the one considered by Oohama in [3] in the following important aspect: In [3], the transmissions from *all* sensors are assumed to be available to the central observer; in our problem only *some* k out of n transmissions are guaranteed to be available.

Let $U_i(t) := Y_i(t) + q_i(t)$ for each sensor i and time t where $\{q_i(t)\}_{t=1}^{\infty}$ is a zero-mean, i.i.d. Gaussian quantization noise process for sensor i with common variance σ_q^2. Hence (we refer the reader to [8] for expressions of joint and conditional differential entropies of Gaussian random variables),

$$\sum_{i=1}^{k} h(U_i|Y_i) = \frac{k}{2}\log_2(2\pi e\ \sigma_N^2\ \mathrm{DNR}), \tag{4}$$

where DNR is the Distortion to Noise Ratio given by σ_q^2/σ_N^2. The eigenvalues of the correlation matrix of $U_1, \ldots, U_k$ can be shown to be $\sigma_N^2(1+\mathrm{DNR}+k\cdot\mathrm{SNR})$ with multiplicity 1 and $\sigma_N^2(1+\mathrm{DNR})$ with multiplicity $(k-1)$, where σ_X^2/σ_N^2 is the Signal to Noise Ratio (SNR). Hence,

$$h(U_1, \ldots, U_k) = \frac{k}{2}\log_2\left[2\pi e\ \sigma_N^2\left(1+\frac{k\cdot\mathrm{SNR}}{1+\mathrm{DNR}}\right)^{\frac{1}{k}}\left(1+\frac{1}{\mathrm{DNR}}\right)\right].$$

Using these expressions in (2), yields

$$R_k^* = \frac{1}{2}\log_2\left[\left(1+\frac{k\cdot\mathrm{SNR}}{1+\mathrm{DNR}}\right)^{\frac{1}{k}}\left(1+\frac{1}{\mathrm{DNR}}\right)\right]. \tag{5}$$

On the reception of any $m \geq k$ transmissions, the central observer first decodes the appropriate set of quantized sensor observations $u_{i_1}^{(l)} \ldots u_{i_m}^{(l)}$ and then computes the minimum mean squared error (MMSE) estimate of $X^{(l)}$. The MMSE estimate is the conditional mean of the posterior distribution of $X^{(l)}$ given the quantized observations [1]. Since the source, observation noise, and quantization noise processes are independent and Gaussian, the MMSE estimate is a linear function of the quantized observations [1]:

$$\widehat{x}_m^{(l)} = \frac{\mathrm{SNR}}{1+\mathrm{DNR}+m\cdot\mathrm{SNR}}\sum_{j=1}^{m} u_{i_j}^{(l)}.$$

The optimal mean square error for this estimate is given by

$$D_m^* = \frac{\sigma_X^2(1+\mathrm{DNR})}{1+\mathrm{DNR}+m\cdot\mathrm{SNR}}. \tag{6}$$

Suppose that $k = n\alpha$ and $m = n\beta$ where $0 < \alpha \leq \beta \leq 1$. Here, α and β respectively represent the fraction of guaranteed and the fraction of received sensor measurements. For fixed α and β as $n \longrightarrow \infty$, the asymptotic rate reduces to $\frac{1}{2}\log_2\left(1+\frac{1}{\mathrm{DNR}}\right)$. The distortion with the reception of a fraction β of the sensor observations decays as $\frac{\sigma_X^2(1+\mathrm{DNR})}{n\beta}$ which is similar, up to proportionality constants, to the decay of distortion when the central decoder has direct access to *all* the uncoded sensor measurements.

In [3] Oohama proves the following result: if there are there exactly k sensors, all their transmissions reach the central unit, and the total rate available to the

k-sensor network is kR_k^* (given by (5)), then the best estimation quality that can be achieved is D_k^* (given by (6)) under the Gaussian statistics assumptions of this section. However, R_k^* and D_k^* are respectively the rate and distortion achievable when $n \geq k$ sensors are used and it is not a priori known exactly which of the k sensor readings will be available to the central unit. Hence, regardless of which k sensor observations reach the decoding unit, the performance is as good as that of k completely reliable transmissions. Thus we get robustness to erasures without any loss in performance. Furthermore, the estimation quality strictly improves with the reception of more sensor observations: D_m^* is a strictly decreasing function of m; see (6). The source of this optimality property can be traced to the structure of the encoders in [3]. The encoder and decoder structures are further elaborated in Section 6.1.

4 Compressing raw data versus compressing local estimates

In some scenarios, small clusters of sensors have the communication resources to locally collaborate. In other situations, a sensor node may consist of a sensor-array or sensors with multiple modalities (acoustic, seismic, infra red, etc.). Both these scenarios can be captured by allowing vector-valued sensor data $\{\mathbf{Y}_i\}_{i=1}^n$ in the model of Figure 1 (also see Figure 3) . Individual components of the vector-valued observations would then represent the scalar measurements of collaborating sensors in each cluster or the component measurements in sensor-arrays or the measurements corresponding to different modalities in multi-modal sensors. A natural question that arises in this context is: how should these vector measurements be processed and encoded by the sensor-clusters/multi-modal sensors to be rate-distortion optimal. The approach in the scalar case would suggest doing a rate-distortion quantization of the vector measurements followed by random binning that exploits the correlation between measurements across sensor-clusters. When all associated processes are jointly Gaussian, rate-distortion quantization amounts to doing a reverse water-filling on the locally decorrelated components of the measurements, i.e., a Karhunen-Loève transform followed by reverse water-pouring [8].

A different approach would be to first form local estimates $T_i(\mathbf{Y}_i)$ of X at each cluster and then compress the estimates in a distributed manner. Under what conditions will there be no loss in (rate-distortion) performance? There are two possible sources of loss. First, compression of $\{T_i\}_{i=1}^n$ might be sub-optimal even when infinite rate is available, i.e., the best estimate of X based on $\{T_i(\mathbf{Y}_i)\}_{i=1}^n$ is not the same as the best estimate based on $\{\mathbf{Y}_i\}_{i=1}^n$. The availability of finite rate is another potential source of loss. However,

Proposition 1. *For MSE as the distortion metric, if there is no performance loss in encoding local statistics at high (infinite) rates, i.e., if*

$$Z := E\left[X|\mathbf{Y}_1, \ldots, \mathbf{Y}_n\right] = h\left(T_1(\mathbf{Y}_1), \ldots, T_n(\mathbf{Y}_n)\right), \tag{7}$$

for some function $h()$ of local statistics, then there will be no performance loss for any finite rate R.

Proposition 1, proved in the Appendix, is a generalization of a result due to Sakrison in [11], where it was shown that for a point-to-point communication scenario, i.e., when there is only one sensor, estimation (forming $T_1(\mathbf{Y}_1) := E[X|\mathbf{Y}_1]$) followed by quantization is optimal in a rate-distortion sense. This is a consequence of the orthogonality principle[5] and the structure of optimal, fixed-rate, vector quantizers. Note that (7) and the orthogonality principle together imply that $h(T_1, \ldots, T_n) = E\left[X|T_1(\mathbf{Y}_1), \ldots, T_n(\mathbf{Y}_n)\right]$ (cf. Appendix for details).

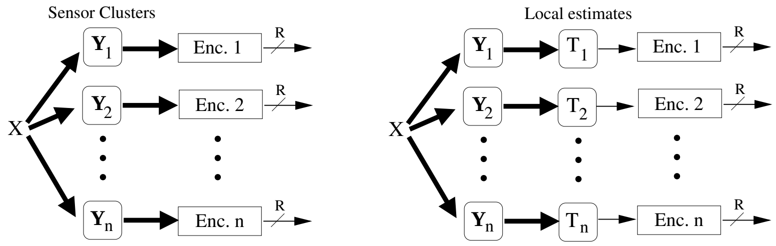

Fig. 3. Distributed compression of raw data (left) versus distributed compression of local estimates (right).

Thus, under the assumption of no performance loss at high rates given by (7) the problem of compressing correlated, vector-valued observations reduces to a problem of compressing local, scalar-valued statistics T_i. These ideas can be illustrated by considering a simple vector-version of the Gaussian CEO problem studied by Oohama in [3]. Here, the source X is a discrete, zero-mean, Gaussian random process. The noise processes are assumed to be i.i.d., zero-mean, and Gaussian across all sensor-clusters and across all components in each cluster. The local and global observation statistics in this example have the same structure. Hence, intuitively, it is clear that the form of the local and global MMSE estimators will also be the same. In fact, all (local as well as global) MMSE estimators in this example will compute the simple average of the constituent observations followed by scaling. From this it follows that the global MMSE estimate can be expressed as a function of local MMSE estimates so that condition (7) holds. Hence, there will be no loss in rate-distortion performance in first forming local averages at each sensor-cluster. The vector-version of the CEO problem is thereby reduced to the regular scalar Gaussian CEO problem for which the optimal distributed compression encoder structure is known (see Section 6.1 and [3]). In general, we have the following result which is proved in the Appendix.

5 The MMSE estimation error $(X - E[X|\text{data}])$ is orthogonal to any measurable statistic of the data [1].

Proposition 2. *If* $(X, \mathbf{Y}_1, \ldots, \mathbf{Y}_n)$ *are jointly Gaussian and* $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ *are conditionally independent given* X *then (7) holds with* $T_i(\mathbf{Y}_i) := E[X|\mathbf{Y}_i]$, $\forall i$.

Thus, there will be no loss in compressing local estimates even when clusters have arbitrary local noise correlation structure as long as the noise processes are independent across clusters and everything is jointly Gaussian.

How do things change if condition (7) does not hold? In other words, what if the clusters went ahead and formed local estimates even when doing so is not optimal? Clearly, there would be additional distortion due to the fact that the best MMSE estimate of X based on all data is not the same as the best estimate based on local statistics. The process of forming local estimates looses some information that is useful for estimating X. However, will there be additional rate-dependent performance loss due to (finite-rate) compression? Intuitively, although the process of compression looses information with respect to the estimation of source X, the local estimates should be easier to encode for the same rate provided that the statistics are "well-behaved". The local estimates are in some sense "less random" than the raw data and should be easier to compress. Hence, we strongly believe that finite-rate compression would not introduce any more performance loss than that at high (infinite) rates, for well-behaved local estimates, i.e., there will be no rate-dependent excess loss. *For jointly Gaussian statistics, and linear local estimators, this result can be shown to be true.* The proof is similar to that of Proposition 1 with the additional observation that both the global MMSE estimate based on raw data and the global MMSE estimate based on local linear estimates are Gaussian and the latter is easier to compress because it has smaller variance. Whenever there is no rate-dependent excess loss, the task of designing distributed estimation algorithms for sensor networks with finite rate constraints can be decoupled into two tasks: (i) designing good distributed estimation algorithms assuming infinite rates and (ii) designing good distributed compression algorithms. The decoupling is in the sense that the second task will not introduce any rate-dependent excess distortion.

5 Concluding remarks

This work represents a step towards developing a fundamental understanding of the performance limits for estimation in a distributed, rate-constrained, unreliable environment. We have provided an information-theoretically achievable rate-distortion region for the sensor network setup of Figure 1. For a Gaussian sensor network, we have shown how robustness can be incorporated without sacrificing rate-distortion performance. Our ongoing work includes evaluating fundamental estimation performance limits for increasingly complex source and sensor statistics and channel models. More realistic models for sensor networks should incorporate correlation between sensor noise processes and also have the ability to deal with a vector-valued source process having general spatio-temporal dependencies. We are also looking at practical code constructions to be used in sensor networks for robust, distributed estimation. This involves constructing quantizers for capturing independent signal components to mimic the performance of independent random codes. This can be done using quantizers that are shifted versions of one another (see [7] for details). A structured way of realizing independent binning of multiple codebooks presented in this paper is by using algebraic trellis and lattice codes along the lines of [7, 12]. The key idea is to construct "linearly independent" partitions of different codebooks under consideration. The algebraic structure of trellis and lattice codes is amenable to such partitions. Ongoing work involves finding computationally efficient constructions for such partitions, and fast algorithms for doing index assignment.

6 Appendix

6.1 Proof of Theorem 1

We now present the key ideas in the proof of the theorem. For clarity, we shall only present a sketch of the proof and omit technical details. Consider a symmetric distribution $p_{\mathbf{U}|\mathbf{Y}}$ of the form given by (1) and decoders $\{g_m\}_{m=k}^n$. Let $R' \geq R$.

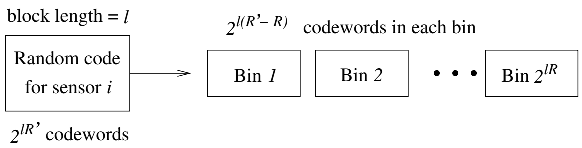

Fig. 4. Random code construction: n independent random codebooks of block length l are constructed each containing $2^{lR'}$ codewords. Each codebook is randomly partitioned into 2^{lR} bins each with approximately $2^{l(R'-R)}$ codewords.

Random Coding: For each sensor $i \in \mathcal{I}_n$, generate a codebook $\mathcal{C}_i$ containing $2^{lR'}$ codewords of blocklength l. The codewords of $\mathcal{C}_i$ are generated by drawing i.i.d. symbols from the marginal distribution p_{U_i} The rate R' has to be large enough to ensure that there exists at least one codeword in $\mathcal{C}_i$ that is jointly (strongly) typical [8, 13] with the sensor observation $y_i^{(l)}$ Using properties of strongly typical sequences, the probability that there exists a codeword which is jointly typical with $y_i^{(l)}$, to the first order with respect to l in the exponent, is $2^{-lI(Y_i;U_i)} = \frac{2^{lH(Y_i,U_i)}}{2^{lH(Y_i)}2^{lH(U_i)}}$. Intuitively, the desired probability will be close to one if the expected number of jointly typical codewords: $2^{lR'}2^{-lI(Y_i;U_i)}$ is exponentially large. This can be ensured if

$$R' > I(Y_i; U_i). \tag{8}$$

Note: since distributions are symmetric,

$$R' > I(Y_i; U_i) = \frac{1}{k}\sum_{j=1}^{k} I(Y_j; U_j) = \frac{1}{k}\sum_{j=1}^{k}\left[H(U_j) - H(U_j|Y_j)\right], \tag{9}$$

for all $i = 1, \ldots, n$.
Random Binning: For each sensor, construct 2^{lR} bins, each having approximately $2^{l(R'-R)}$ codewords (see Figure 4). The codewords are drawn independently, uniformly and with replacement from the corresponding codebook. Random binning is a technique that enables each encoder to transmit data at the *reduced* rate R by leveraging the fact that the quantized noisy observations of all sensors are correlated and the decoder will have access to $k-1$ additional correlated sensor observations.
Encoding: Given a noisy sensor observation $y_i^{(l)}$ find a codeword $u_i^{(l)}$ from $\mathcal{C}_i$ that is jointly typical with it. If (8) is satisfied, there exists at least one such codeword. The index I_i of the bin containing the codeword $u_i^{(l)}$ is transmitted.
Decoding: Let the decoder receive some k bin indices $I_{i_1}, ..., I_{i_k}$. The decoder searches through the corresponding bins to identify a k tuple of codewords that is jointly typical. If (8) holds, there is at least one such k tuple, namely the one that was transmitted. The size of each bin should be small enough so that the probability of finding another jointly typical k tuple is very small. There are $2^{kl(R'-R)}$ combinations of k-tuples of codewords from k bins. The probability that a random k tuple of codewords is jointly typical is, to the first order in the exponent w.r.t. l, given by $p \approx \frac{2^{lH(U_1,\ldots,U_k)}}{2^{lH(U_1)}2^{lH(U_2)}\ldots 2^{lH(U_k)}}$. Intuitively, the desired probability goes to zero if the expected number of jointly typical k-tuples: $p \cdot 2^{kl(R'-R)}$ is exponentially small. This can be ensured if the exponent satisfies

$$k(R'-R) < \sum_{i=1}^{k} H(U_i) - H(U_1, \ldots, U_k). \tag{10}$$

Combining (9) and (10) we get the desired result: $R > R_k^*$ (cf. (2)). With these conditions, the decoder successfully recovers the quantized noisy observations $u_{i_1}^{(l)} ... u_{i_k}^{(l)}$ with high probability and can optimally estimate the source from them.

If $m > k$ indices are received, the above decoding procedure can be repeatedly used on subsets of k indices at a time to recover all m quantized noisy observations with high probability.

6.2 Proof of Proposition 1

First note that the both $X - Z$ and $Z - E[Z|T_1, \ldots, T_n]$ are orthogonal to $\{T_i\}_{i=1}^n$. Hence, $X - E[Z|T_1, \ldots, T_n]$ is also orthogonal to the local statistics and $E[X|T_1, \ldots, T_n] = E[Z|T_1, \ldots, T_n] = h(T_1, \ldots, T_n)$, where the last equality follows from (7).

Let X take values in $\mathbb{R}$, $\mathbf{Y}_i$ take values in $\mathcal{Y}_i$ and consider l i.i.d. realizations of the tuple $(X, \mathbf{Y}_1, \ldots, \mathbf{Y}_n)$ denoted by $X^{(l)}$, and $\mathbf{Y}_i^{(l)}$ for $i = 1, \ldots, n$.

Let $T_i : \mathcal{Y}_i :\rightarrow \mathbb{R}, i = 1, \ldots, n$ be some measurable maps for which (7) holds. Let $T_i^{(l)}(\mathbf{Y}_i^{(l)})$ denote $(T_i(\mathbf{Y}_i(1)), \ldots, T_i(\mathbf{Y}_i(l)))^T$. Since the realizations are independent and (7) holds, we have

$$Z^{(l)} := E\left[X^{(l)} | \mathbf{Y}_1^{(l)}, \ldots, \mathbf{Y}_n^{(l)}\right] = h^{(l)}\left(X^{(l)} | T_1^{(l)}, \ldots, T_n^{(l)}\right),$$

where in fact,

$$Z^{(l)}(j) = E[X(j)|\mathbf{Y}_1(j), \ldots, \mathbf{Y}_n(j)] = h(T_1(\mathbf{Y}_1(j)), \ldots, T_n(\mathbf{Y}_n(j))),$$

for all $j = 1, \ldots, l$. Now consider a set of n rate-R encoders and decoders (distributed vector-quantizers) of block-length l:

$$\begin{aligned} f_i &: \mathcal{Y}_i^l \longrightarrow \mathcal{I}_{2^{lR}},\ i = 1, \ldots, n, \\ g &: (\mathcal{I}_{2^{lR}})^n \longrightarrow \mathbb{R}^l. \end{aligned}$$

Let the set of reconstruction vectors in $\mathbb{R}^l$ be denoted by $\{\mathbf{v}_1, \ldots, \mathbf{v}_{2^{nlR}}\}$. We shall show that if the $\{f_i\}$'s and g are optimized for yielding the smallest MSE, then it is necessary that for each sensor i, observations $\mathbf{Y}_i^{(l)}$ which map to the same value of the statistic $T_i^{(l)}$, should have the same encoding index. By the orthogonality principle,

$$E\left|\left|X^{(l)} - g(F_1, \ldots, F_n)\right|\right|^2 = E\left|\left|X^{(l)} - Z^{(l)}\right|\right|^2 + E\left|\left|Z^{(l)} - g(F_1, \ldots, F_n)\right|\right|^2.$$

The first term on the right-hand side of the last equation does not depend on the encoders and decoder used. The second term is of the form

$$E\left|\left|h^{(l)}\left(T_1^{(l)}, \ldots, T_n^{(l)}\right) - g\left(F_1, \ldots, F_n\right)\right|\right|^2.$$

If g is optimal, the reconstruction vectors in $\{\mathbf{v}_j\}_{j=1}^{2^{nlR}}$ must be all distinct otherwise one could use additional reconstruction vectors to reduce the MSE without increasing the rate. Now,

$$\begin{aligned} &E\left|\left|h^{(l)}\left(T_1^{(l)}, \ldots, T_n^{(l)}\right) - g\left(F_1, \ldots, F_n\right)\right|\right|^2 = \\ &\quad E\left[E\left[\left|\left|h^{(l)}\left(T_1^{(l)}, \ldots, T_n^{(l)}\right) - g\left(F_1, \ldots, F_n\right)\right|\right|^2 \middle| T_1^{(l)} = \mathbf{t}_1, \ldots, T_n^{(l)} = \mathbf{t}_n\right]\right]. \end{aligned}$$

Let

$$\mathbf{v}(\mathbf{t}_1, \ldots, \mathbf{t}_n) := \arg \min_{\mathbf{v} \in \{\mathbf{v}_1, \ldots, \mathbf{v}_{2^{nlR}}\}} \left|\left| h^{(l)}(\mathbf{t}_1, \ldots, \mathbf{t}_n) - \mathbf{v}\right|\right|.$$

Then it is clear that over the region of space $\bigotimes_{i=1}^n \mathcal{Y}_i^{(l)}$ where $T_i^{(l)} = \mathbf{t}_i, i = 1 \ldots, n$, we must have $g(F_1, \ldots, F_n) = \mathbf{v}(\mathbf{t}_1 \ldots, \mathbf{t}_n)$ otherwise the MSE will be larger. Since different indices map to different reconstruction vectors, it follows that all points of $\bigotimes_{i=1}^n \mathcal{Y}_i^{(l)}$ for which $T_i^{(l)}(\mathbf{Y}_i^{(l)}) = \mathbf{t}_i, i = 1 \ldots, n$, have the same the n-tuple of encoding indices as was claimed.

6.3 Proof of Proposition 2

First note that $E[\mathbf{Y}_i|X] = \frac{E[X\mathbf{Y}_i]}{E[X^2]}$. From the conditional independence assumption we have

$$E[\mathbf{Y}_i\mathbf{Y}_j^T] = E[E[\mathbf{Y}_i\mathbf{Y}_j^T]|X] = E[E[\mathbf{Y}_i|X]E[\mathbf{Y}_j^T|X]] = \frac{E[X\mathbf{Y}_i]E[X\mathbf{Y}_j^T]}{E[X^2]}.$$

The MMSE estimator of X based on all observations has the form $\sum_{i=1}^n \mathbf{a}_i^T\mathbf{Y}_i$. The local MMSE estimator of X based on $\mathbf{Y}_i$ is given by $E[X|\mathbf{Y_i}] = E[X\mathbf{Y}_i^T]K_i^{-1}\mathbf{Y}_i$, where $K_i := E[\mathbf{Y}_i\mathbf{Y}_i^T]$. From the orthogonality principle and the conditional independence assumption we obtain for $j = 1, \ldots, n$,

$$\begin{aligned} E[X\mathbf{Y}_j^T] &= \mathbf{a}_j^T K_j + \sum_{i \neq j} \mathbf{a}_i^T E[\mathbf{Y}_i\mathbf{Y}_j^T], \\ &= \mathbf{a}_j^T K_j + \beta_j E[X\mathbf{Y}_j^T], \end{aligned}$$

where $\beta_j := \sum_{i\neq j} \mathbf{a}_i^T E[X\mathbf{Y}_i]/E[X^2]$. This shows that

$$\mathbf{a}_j^T\mathbf{Y}_j = (1-\beta_j)E[X\mathbf{Y}_j^T]K_j^{-1}\mathbf{Y}_j = (1-\beta_j)E[X|\mathbf{Y}_j]$$

for all j. Hence, the global MMSE estimate has the form $\sum_i (1-\beta_i)E[X|\mathbf{Y}_i]$ which conforms to (7) as claimed.

References

1. H. V. Poor, *An Introduction to Signal Detection and Estimation.* New York, NY: Springer–Verlag, 1994.
2. M. Gastpar, B. Rimoldi, and M. Vetterli, "To code, or not to code: Lossy source-channel communication revisited," *IEEE Transactions on Information Theory.* To appear.
3. Y. Oohama, "The Rate-Distortion Function for the Quadratic Gaussian CEO Problem," *IEEE Transactions on Information Theory*, vol. 44, pp. 55–67, May 1998.
4. D. Slepian and J. K. Wolf, "Noiseless Coding of Correlated Information Sources," *IEEE Transactions on Information Theory*, vol. 19, pp. 471–480, July 1973.
5. A. Wyner and J. Ziv, "The Rate-Distortion Function for Source Coding with Side Information at the Decoder," *IEEE Transactions on Information Theory*, vol. 22, pp. 1–10, January 1976.
6. S. S. Pradhan and K. Ramchandran, "Distributed Source Coding Using Syndromes (DISCUS): Design and Construction," *Proceedings of the Data Compression Conference (DCC)*, March, Snowbird, UT, 1999.
7. S. S. Pradhan and K. Ramchandran, "Distributed Source Coding: Symmetric Rates and Applications to Sensor Networks," *Proceedings of the Data Compression Conference (DCC)*, March, Snowbird, UT, 2000.
8. T. M. Cover and J. A. Thomas, *Elements of Information Theory.* New York: John Wiley and Sons, 1991.
9. S. S. Pradhan, R. Puri, and K. Ramchandran, "n-Channel Symmetric Multiple Descriptions - Part I: (n,k) Source-Channel Erasure Codes," *IEEE Transactions on Information Theory.* To appear.

10. R. G. Gallager, *Information Theory and Reliable Communication.* New York: John Wiley and Sons, 1968.
11. D. J. Sakrison, "Source Encoding in the Presence of Random Disturbance," *IEEE Trans. on Information Theory*, vol. IT-14, Jan. 1968.
12. S. S. Pradhan and K. Ramchandran, "Generalized Coset Codes for Symmetric Distributed Coding," *IEEE Transactions on Information Theory*, 2003. Submitted.
13. T. Berger, "Multiterminal Source Coding," *Information Theory Approach to Communication, (CISM Courses and Lecture Notes No. 229)*, G. Longo, Ed., Wien and New York: Springer-Verlag, 1977.