Side Information Aware Coding Strategies for Sensor Networks

Stark C. Draper, *Member, IEEE,* and Gregory W. Wornell, *Fellow, IEEE*

***Abstract*— We develop coding strategies for estimation under communication constraints in tree-structured sensor networks. The strategies have a modular and decentralized architecture. This promotes the flexibility, robustness, and scalability that wireless sensor networks need to operate in uncertain, changing, and resource-constrained environments. The strategies are based on a generalization of Wyner-Ziv source coding with decoder side information. We develop solutions for general trees, and illustrate our results in serial (pipeline) and parallel (hub-and-spoke) networks. Additionally, the strategies can be applied to other network information theory problems. They have a successive coding structure that gives an inherently less complex way to attain a number of prior results, as well as some novel results, for the CEO problem, multiterminal source coding, and certain classes of relay channels.**

***Index Terms* — sensor networks, distributed estimation, data fusion, side information, Wyner-Ziv coding, rate distortion theory, CEO problems, multiterminal source coding, distributed detection, relay channels.**

I. Introduction

Starting from a set of architectural principles appropriate for wireless sensor networks, we develop and analyze efficient coding techniques for estimation under communication constraints. We base our approach on information-theoretic ideas of source coding with decoder side information.

The central characteristic differentiating estimation in sensor networks from more traditional contexts is that data is not co-located. Limits on communication between sensor nodes typically prevent us from conveying all data losslessly to a central location for processing. Hence, many standard estimation techniques cannot be directly applied. Instead, we must determine what is the most important information for nodes to share, and design quantizers to encode that information. This leads to a required coupling of the estimation and communication subtasks for efficient implementation.

Manuscript received July 15, 2003; revised February 4, 2004. This work was supported in part by the National Science Foundation under Grant No. CCR-0073520, Microsoft Research, Hewlett-Packard through the MIT/HP Alliance, and Texas Instruments through the Leadership Universities Program. This work was presented in part at the Allerton Conference on Communication, Control, and Computing, Urbana, IL, October 2001, and at the International Symposium on Information Theory, Lausanne, Switzerland, June 2002.

S. C. Draper is with the Department of Electrical Engineering and Computer Science, University of California, Berkeley, Berkeley, CA 94720 USA (e-mail: sdraper@eecs.berkeley.edu).

G. W. Wornell is with the Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139 USA (e-mail: gww@allegro.mit.edu).

While one might think that coupling estimation and communication would require global coordination, we show how to effect such coupling in a distributed manner. We present a modular and decentralized strategy that jointly addresses the quantization, communication, and estimation aspects of the problem. Our design significantly outperforms strategies where data communication and signal estimation are decoupled.

As the simplest model from which we can obtain useful insight, we consider the source $\mathbf{x} = x^n$ to be estimated to be a length-n vector of independent identically distributed (i.i.d.) random variables, where n is sufficiently large. Many of the insights derived carry over naturally to more elaborate source models. The sensor network consists of L sensor nodes where node l measures $\mathbf{y}_l$. The source and observations are jointly distributed, but memoryless.

We illustrate our ideas in the context of tree-structured networks. In a "sensor tree", the tree implies a data routing from "leaf" nodes to a "root" node. Figure 1 depicts such a network. Each node receives messages from neighboring nodes above it in the tree, and sends a message to the next node down the tree. In the figure, Node 3 measures $\mathbf{y}_3$, receives messages from Nodes 1 and 2, and sends message m_3 to Node 6. Depending on the application, our objective may be to estimate $\mathbf{x}$ at all nodes, or perhaps only at the root node.

We focus on a digital model for inter-node communications, consisting of fixed-rate links. Thus, in Fig. 1, if message m_1 is limited to rate R_1 bits per observation sample, then $m_1 \in \{1, \ldots, 2^{nR_1}\}$. This model decouples the application-layer estimation problem from the physical-layer communication problem, and allows us to focus directly on the effect of communication constraints on estimate quality. The algorithms we develop can be implemented on top of any physical layer. Naturally, more advanced physical layer implementations will lead to higher data rates, and better estimation performance.

In this paper we concentrate on scenarios where the sensor tree is given. Routing and rate allocations would have to be managed by a network layer protocol. While we do not focus on network layer issues, we do make some brief observations about them in the context of our estimation strategies.

A. Architectural Issues

To ensure system flexibility, robustness, and scalability, we design modular systems that have decentralized knowledge requirements. Modular systems consist of functionally interchangeable sensor nodes. The common functionality of the nodes means that the network can be more easily reconfigured

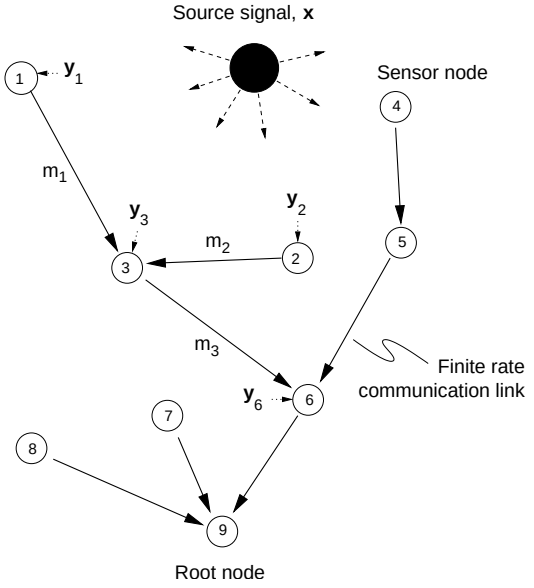

Fig. 1. A sensor network is shown consisting of nine sensor nodes and tree-structured data routing. Sensor node l observes $\mathbf{y}_l$ (shown explicitly for nodes 1,2,3,6), receives messages from neighboring nodes higher up the tree, and sends a rate-constrained message m_l to the next node down the tree.

for unanticipated or changing tasks, e.g., when individual nodes are unreliable, or when the location of phenomena under observation changes.

We design our algorithms to have decentralized knowledge requirements so that the network can operate successfully even when each node's knowledge is restricted to local network conditions. In contrast, if each node were required to have knowledge of global network conditions, then the cost of collecting and disseminating this information in an up-to-date manner could be prohibitive. At a minimum, neighboring nodes must coordinate their communication rates and must have some statistical knowledge of the relationship between their data sets (perhaps, e.g., in the form of signal-to-noise or distortion-to-noise ratios). Without the former they could not agree on a communication protocol, and without the latter they would have no basis on which to combine data to estimate the source. We focus on algorithms that rely on this bare minimum of knowledge.

B. Related Work

The coding strategies of this paper build, in part, upon Wyner and Ziv's [38] approach to coding with decoder side information. In this context, this paper ties in with a growing body of work focusing on side-information coding fundamentals, constructions, and dualities (see, e.g., [1], [8], [25], [26], [42]). A related set of work considers the CEO problem [4]. In the CEO problem a number of sensor nodes make noisy observations of an underlying source signal. Each then sends a message to a central hub node (the "Chief Executive/Estimation Officer") that estimates the source. The CEO problem is studied further in [34], [21], [41]. The perspective taken in this paper can be thought of as viewing the CEO problem as a generalization of Wyner-Ziv to multiple indirect (noisy) observations, and as extending the hub-and-spoke CEO model to sensor trees. In Sec. V-A we show that the coding strategy we introduce gives a novel and inherently less complex way to achieve the rate-distortion optimal CEO results of [21]. We also note that in certain situations where sources, channels, and distortion measures are well matched, very efficient joint source-channel coding approaches [13] present an attractive alternative to rate-constrained schemes.

Multiterminal source coding is another research area related to the problems and approaches considered herein. In this case, the goal is to jointly estimate all observations $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_L$, rather than some underlying source signal. This vein of research was initiated by Slepian and Wolf [31] for the lossless encoding of a distributed pair of correlated source signals. Their elegant solution motivated many extensions, both lossless [36], [16], [15], [2] and lossy [33], [3], [20], [41]. The full solution to the latter remains unsolved. In comparison with multiterminal source coding, in our context we have no specific interest in the observations $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_L$, other than in how they can be used to estimate $\mathbf{x}$. In Sec. V-B, however, we show how to attain the rate-distortion region of [33], [3] using the coding strategies developed in this paper.

Finally, a related thread of work focuses on problems of detection with distributed sensors (e.g., [32], [35]). In this case, the objective is to make a decision about the source, rather than an estimate of it. We briefly connect to distributed detection problems in Sec. V-C where we apply our results to certain classes of relay channels [9].

C. Paper Insights and Contributions

The insights and contributions of this paper are both architectural and technical. First, we show that techniques of coding with decoder side information (and the "binning" ideas that underlie them) have an important role to play in the design of statistical inference algorithms for communication-constrained sensor networks. While a similar point is made, e.g., in the context of the CEO problem, this paper shows how to apply these ideas to the more general topologies of sensor trees. We employ well-understood random coding techniques but, as discussed in Sec. II-B, the extension to trees relies on a less well-known generalization of Berger's Markov Lemma [3].

Our second point is that modularity and decentralization are important principles underlying the design of flexible, robust, and scalable sensor networks. While in some cases (though, as we show, not all) performance may be lowered in comparison with non-modular and centralized designs, gains in these other system criteria will often outweigh such losses.

Thirdly, the estimation strategies we present are examples of "soft" coupling across traditional network layers. The strategies require an "awareness" both of what is going on at the physical layer of the network, in terms of communication rate, and at the application layer of the network, in terms of side information quality. We show that such coupling can lead to significant performance gains, but can be implemented without violating the traditional layering paradigm of networks.

Finally, as mentioned in Sec. I-B, we show that the coding strategies introduced can be profitably applied to a number of

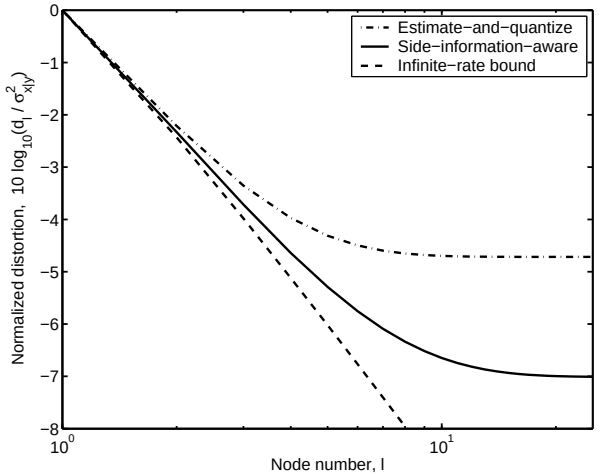

Fig. 2. Mean-squared estimation error versus node number in a data-pipeline example with constant fixed-rate links.

related problems. The resulting solutions have novel structures that displays attractive complexity and scalability properties.

D. Illustrative Results

To give a sense of the performance improvement that can be effected by making a sensor network side information "aware", we now give illustrative results for both side-information-aware and "unaware" strategies. We consider a data-pipeline or "serial" network, a class of networks discussed in some depth later in the paper and illustrated by node grouping $(4, 5, 6)$ in Fig. 1 (and shown schematically later in Fig. 3). The source $\mathbf{x}$ to be estimated is an i.i.d. Gaussian sequence of variance $\sigma_x^2 = 4$, observed at each node in independent equal-variance additive white Gaussian noise of variance $= 4/3$. The inter-node communication rates are set equal to $R = 2.5$ bits per observation sample for all node pairs.

"Estimate-and-quantize" is a strategy that uses decoder side information at the application layer during estimation, but not during communication. Communication occurs in a multi-step fashion where each node forms a source estimate based on its observation and the message it received from the node just upstream. It quantizes that estimate at a rate equal to its communication rate, and sends the corresponding quantization index to the next node downstream. In Fig. 2 we plot the mean-squared estimation error of estimate-and-quantize for the data-pipeline example with the dash-dotted curve.

The side-information-aware strategies we present in this paper are more efficient than estimate-and-quantize because the whole process — quantization, communication, and estimation — is designed to make use of the encoder's statistical knowledge of the decoder's data as decoder side information in the sense of Wyner and Ziv. In Fig. 2, the performance achieved by the side-information-aware strategy is plotted with the solid curve. For a target distortion, the number of nodes required by estimate-and-quantize can far exceed the number required by the side-information-aware strategy. In the example, the estimation performance of both strategies saturates because the pipeline is set to have equal-rate links. As a final point of comparison, the dashed line plots the infinite-rate bound, $\sigma^2_{x|y_1,\ldots y_l}$, which is only achievable by relaxing all communication constraints.

E. Paper Outline and Notation

The paper is organized as follows. Sec. II describes the main results needed to develop the side-information-aware strategies. In Sec. III we apply these results to sensor trees and develop simple cut-set bounds. Sec. IV discusses Gaussian sources with quadratic (mean-squared) distortion measures, and presents illustrative examples. Sec. V discusses connections to other network information theory problems: the CEO problem, multiterminal source coding, and relay channels.

We use $I(\cdot;\cdot)$ to denote mutual information, and $\mathcal{T}_x$ to denote the set of all ϵ-strongly-typical sequences of length-n with respect to $p_x(x)$ (using standard definitions as presented, e.g., in [10]). The superscript ^{c} applied to an event denotes its complement, $|\cdot|$ applied to a set denotes its cardinality, $\emptyset$ denotes the null set, $\leftrightarrow$ is used to denote Markov chain relationships, and $E[\cdot]$ denotes expectation.

II. Side-Information-Aware Coding

In this section we present the results that underlie our coding strategies developed in Sec. III. In Sec. II-A we present an achievable distortion-rate trade-off for our canonical one-step coupled communication and estimation problem. In Sec. II-B we describe how this result relies on a generalization of the Markov Lemma, which we term the "Serial" Markov Lemma to distinguish between the two. Finally, in Sec. II-C, we discuss how these results relate to earlier work.

A. One-Step Problem

The simplest communication-constrained sensor network consists of a single encoder and a single decoder. The source $\mathbf{x}$, to be estimated is observed as $\mathbf{y}_\mathrm{E}$ at the encoder, and as $\mathbf{y}_\mathrm{D}$ at the decoder. Based on its observation $\mathbf{y}_\mathrm{E}$ the encoder transmits a message m over a fixed-rate bit pipe to the decoder. The decoder produces source estimate $\hat{\mathbf{x}}$ as a function of m and its observation or "side information" $\mathbf{y}_\mathrm{D}$. In order to more easily apply the results we develop to larger networks, we include a third source observation $\mathbf{y}_\mathrm{N}$, not available at either encoder or decoder. Eventually $\mathbf{y}_\mathrm{N}$ will correspond to source observations elsewhere in the network. For this setup we have the following result:

Theorem 1: Let a set of random source and observation vectors $(\mathbf{x}, \mathbf{y}_\mathrm{E}, \mathbf{y}_\mathrm{D}, \mathbf{y}_\mathrm{N})$, and a distortion measure $D(\cdot,\cdot)$, be given such that:

(a) $(\mathbf{x}, \mathbf{y}_\mathrm{E}, \mathbf{y}_\mathrm{D}, \mathbf{y}_\mathrm{N}) \in \mathcal{T}_{x,y_\mathrm{E},y_\mathrm{D},y_\mathrm{N}}$ for some $p_{x,y_\mathrm{E},y_\mathrm{D},y_\mathrm{N}}$
(b) $D(\mathbf{x}, \hat{\mathbf{x}}) = \frac{1}{n}\sum_{i=1}^{n} D(x_i, \hat{x}_i)$.

A sequence of length-n block encoder-decoder pairs can be designed such that if $\mathbf{y}_\mathrm{E}$ is encoded at rate R then, with arbitrarily high probability as n grows to infinity, $\mathbf{x}$ can be recovered to within any average distortion d satisfying

$$d \geq \min_{f,u \in \mathbf{\Pi}} E\left[D(x, f(y_\mathrm{D}, u))\right]. \qquad (1)$$

The minimization is over all functions $f : \mathcal{Y}_D \times \mathcal{U} \to \hat{\mathcal{X}}$, and the set $\mathbf{\Pi}$ consists of all random variables u such that

(i) $u \leftrightarrow y_E \leftrightarrow x, y_D, y_N$,
(ii) $R > I(y_E; u) - I(y_D; u)$.

Theorem 1 is an achievability result. Subsequently we show examples of certain networks where Thm. 1 leads to rate-distortion optimal performance.

We now describe how to achieve (1) to highlight the particular role played by a variant of the Markov Lemma. The argument is a relatively straightforward generalization of earlier Wyner-Ziv type source coding with decoding side information approaches to accommodate the lack of direct source observations. First, construct a code $\mathcal{C}$ consisting of $2^{n(I(y_E;u)+\epsilon)}$ codewords $\mathbf{u}(s)$, $s \in \{1, 2, \ldots, 2^{n(I(y_E;u)+\epsilon)}\}$, each selected uniformly from the set $\mathcal{T}_u$. The codewords are randomly and uniformly partitioned into 2^{nR} cosets or "bins". There are approximately $2^{n(I(y_E;u)-R+\epsilon)}$ codewords per coset. The observation $\mathbf{y}_E$ is block encoded (according to $p_{y_E,u}$) to a jointly typical $\mathbf{u}(s)$, an element of some coset. This encoding is successful since $|\mathcal{C}| > 2^{nI(y_E;u)}$. The index m of the coset containing $\mathbf{u}(s)$ is sent to the decoder. At the decoder, the codeword in coset m that is jointly typical with the side information $\mathbf{y}_D$ is selected as the transmitted $\mathbf{u}(s)$. As we discuss next in Sec. II-B, because $u \leftrightarrow y_E \leftrightarrow (x, y_D, y_N)$ the Serial Markov Lemma ensures that $(\mathbf{u}(s), \mathbf{y}_E, \mathbf{x}, \mathbf{y}_D, \mathbf{y}_N)$ are jointly typical, whence $\mathbf{y}_D$ and the transmitted $\mathbf{u}(s)$ are jointly typical. Because all other codewords in bin m are chosen independently of $\mathbf{y}_D$, by choosing $R > I(y_E; u) - I(y_D; u) + 3\epsilon$ we ensure that none of these non-transmitted codewords is jointly typical with $\mathbf{y}_D$. Because $(\mathbf{u}(s), \mathbf{x}, \mathbf{y}_D) \in \mathcal{T}_{u,x,y_D}$, the empirical joint distribution is close to the chosen distribution p_{u,x,y_D}. Therefore, a source estimate formed element-wise as $\hat{x}_i = f(y_{D,i}, u_i(s))$ has an expected distortion close to d.

B. Serial Markov Lemma

The Serial Markov Lemma is required for the proof of Thm. 1 because the set of vectors $(\mathbf{x}, \mathbf{y}_E, \mathbf{y}_D, \mathbf{y}_N)$ is assumed to be jointly strongly typical, but not memoryless. If the vectors were memoryless, i.e., if $p_{\mathbf{x},\mathbf{y}_E,\mathbf{y}_D,\mathbf{y}_N}(\mathbf{x}, \mathbf{y}_E, \mathbf{y}_D, \mathbf{y}_N) = \prod_{i=1}^{n} p_{x,y_E,y_D,y_N}(x_i, y_{E,i}, y_{D,i}, y_{N,i})$, then Berger's Markov Lemma [3] suffices to assert the joint typicality of $(\mathbf{y}_D, \mathbf{u}(s))$. Instead we need the following natural extension of the Markov Lemma developed by Chang and by Kaspi [5], [17], [18].

Lemma 1: Let $p_{x,y,z}(x, y, z) = p_x(x)p_{y|x}(y|x)p_{z|y}(z|y)$ define a Markov chain over finite alphabets $\mathcal{X}$, $\mathcal{Y}$, $\mathcal{Z}$. Let $\mathcal{C}_z$ be a set of sequences chosen equally likely from $\mathcal{T}_z$ with replacement. Then the deterministic mapping $\mathbf{z}(\mathbf{y} = \mathbf{y}) = \mathbf{z}_i$, where $\mathbf{z}_i$ is the first $\mathbf{z} \in \mathcal{C}_z$ (assume some arbitrary ordering) that is ϵ-jointly-strongly-typical with $\mathbf{y}$ satisfies

$$\Pr[\mathbf{x}, \mathbf{y}, \mathbf{z}(\mathbf{y}) \in \mathcal{T}_{x,y,z} | (\mathbf{x}, \mathbf{y}) \in \mathcal{T}_{x,y} \text{ and } (\mathbf{y}, \mathbf{z}(\mathbf{y})) \in \mathcal{T}_{y,z}] \to 1$$

as n grows to infinity. The probability is taken over the source distribution and the random selection of $\mathcal{C}_z$.

Lemma 1 is used in Thm. 1 by setting $x = (x, y_D, y_N)$, $y = y_E$, and $z = u$. In [7] a dither-encoding rule is introduced that avoids randomization over the selection of $\mathcal{C}_z$ by randomizing the $\mathbf{z}(\mathbf{y})$ mapping over codewords jointly typical with $\mathbf{y}$.

C. Relation to Earlier Results

The one-step problem discussed in Sec. II-A can be thought of as an "indirect" (i.e., noisy encoder observations) Wyner-Ziv problem. A memoryless version is posed by replacing the strong typicality condition of Thm. 1 with the memoryless condition discussed in Sec. II-B, and by setting $\mathbf{y}_N = \emptyset$. This memoryless version is discussed by Yamamoto and Itoh in [40]. They present the single-letter rate-distortion frontier, and discuss the binary-Hamming and quadratic-Gaussian cases. Because of the lack of availability of [39], which [40] cites for the development of its results, we give our full derivation of the rate-distortion frontier for this memoryless case in [11]. Flynn and Gray [12] also consider this system, focusing on achievability results.

III. Strategies for Sensor Trees

In this section we describe how to apply Thm. 1 in an iterative manner to develop strategies for sensor trees. Then, in Sec. III-B, we develop a cut-set bound on estimation error.

A. Achievability

We describe a strategy for sensor trees based on the observation that tree networks can be factored into a succession of canonical one-step estimation and communication problems of the form described by Thm. 1. As discussed in the introduction, we assume that the source and observations are jointly distributed and memoryless so that $p_{\mathbf{x},\mathbf{y}_1,\ldots,\mathbf{y}_L}(\mathbf{x}, \mathbf{y}_1, \ldots, \mathbf{y}_L) = \prod_{i=1}^{n} p_{x,y_1,\ldots,y_L}(x_i, y_{1,i}, \ldots, y_{L,i})$. This ensures that with high probability $(\mathbf{x}, \mathbf{y}_1, \ldots, \mathbf{y}_L) \in \mathcal{T}_{x,y_1,\ldots,y_L}$.

We first sketch three steps of the successive coding process for the network depicted in Fig. 1. Say that in Step A, Node 1 transmits and Node 3 receives. In the notation of Thm. 1, set $x = x$, $y_E = y_1$, $y_D = y_3$, $y_N = (y_2, y_4, \ldots, y_9)$, and $u = u_1$. The two conditions of the theorem define a restricted set $\mathbf{\Pi}_a$ of random variables u_1 that satisfy the Markov condition $u_1 \leftrightarrow y_1 \leftrightarrow x, y_2, \ldots, y_9$ and the rate constraint $R_1 \geq I(y_1; u_1) - I(y_3; u_1)$. Any distortion d_a satisfying $d_a \geq \min_{f_a, u_1 \in \mathbf{\Pi}_a} E[D(x, f_a(y_3, u_1))]$ is achievable. Furthermore, the Serial Markov Lemma guarantees that at the end of the step $(\mathbf{x}, \mathbf{y}_1, \ldots, \mathbf{y}_L, \mathbf{u}_1)$ are strongly jointly typical which sets up Step B.

Say that in Step B, Node 2 transmits and Node 3 again receives. Then we apply Thm. 1 a second time, but with different variables playing the role of encoder and decoder observations. This time $y_E = y_2$, $y_D = (y_3, u_1)$, $y_N = (y_1, y_4, \ldots, y_9)$, and $u = u_2$. The random variable u_2 is restricted to the set $\mathbf{\Pi}_b$ of random variables that satisfy the Markov condition $u_2 \leftrightarrow y_2 \leftrightarrow x, y_1, y_3, \ldots, y_9, u_1$, and $R_2 \geq I(y_2; u_2) - I(y_3, u_1; u_2)$. Any distortion d_b satisfying $d_b \geq \min_{f_b, u_2 \in \mathbf{\Pi}_b} E[D(x, f_b(y_3, u_1, u_2))]$ is achievable.

Finally, in Step C let Node 3 transmit and Node 6 receive. This time set $y_E = (y_3, u_1, u_2)$, $y_D = y_6$, $y_N = (y_1, y_2, y_4, \ldots, y_9)$ and $u = u_3$. The set $\mathbf{\Pi}_c$ consists of random variables u_3 that satisfy the Markov condition $u_3 \leftrightarrow y_3, u_1, u_2 \leftrightarrow x, y_1, y_2, y_4, \ldots, y_9$, and the rate constraint $R_3 \geq$

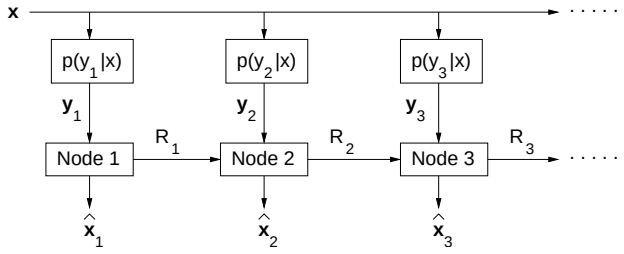

Fig. 3. Example of a serial network. At stage l, the rate R_{l-1} message m_{l-1} and the observation $\mathbf{y}_l$, form the encoder's indirect source knowledge, while $\mathbf{y}_{l+1}$ is the decoder side information.

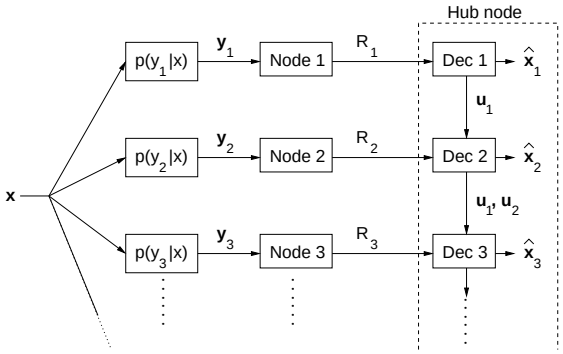

Fig. 4. Example of a parallel network. At stage l, $\mathbf{y}_l$ is the encoder observation, while messages previously decoded by the hub node serve as decoder side information.

$I(y_3, u_1, u_2; u_3) - I(y_6; u_3)$. Any distortion d_c satisfying $d_c \geq \min_{f_c, u_3 \in \mathbf{\Pi}_c} E\left[D(x, f_c(y_6, u_3))\right]$ is achievable.

Generally then, when a sensor node encodes a message, it considers two things. First, its encoding is based both on its observations and all the messages it has received. Second, it takes into account its statistical knowledge of the decoder's observations, and the messages that the decoder has already received, as decoder side information. The strategy is modular because each stage takes the form of an application of Thm. 1, and decentralized because each stage requires knowledge sharing only between encoder and decoder.

It is straightforward to extend this process to any sensor tree. Communication is delayed until the n observations are made, and then begins at leaf nodes. Each non-leaf non-root node in the tree waits until it has received messages from all incoming branches. It then sends a message toward the root. Once the root node has received all incoming messages, it makes its final estimate. Whenever multiple branches feed into a single common node there is a degree of freedom in message ordering (e.g., in the example, the ordering of Node 1 and 2's messages could have been reversed).

To further illustrate this strategy, we present the results of using it in serial and parallel networks.

1) Serial Networks: A serial network has a data-pipeline, or chain structure, as is illustrated in Fig. 3. Encoding starts with the first node in the chain. Generally, at the lth step node l is the encoder and node $l+1$ is the decoder. The strategy leads to the following result:

Proposition 1: Let a set of source and observation vectors $(\mathbf{x}, \mathbf{y}_1, \ldots, \mathbf{y}_L)$, and a distortion measure $D(\cdot, \cdot)$ be given such that (a) $p_{\mathbf{x},\mathbf{y}_1,\ldots,\mathbf{y}_L}(\mathbf{x}, \mathbf{y}_1, \ldots, \mathbf{y}_L) = \prod_{i=1}^{n} p_{x,y_1,\ldots,y_L}(x_i, y_{1,i}, \ldots, y_{L,i})$, and (b) $D(\mathbf{x}, \hat{\mathbf{x}}) = \frac{1}{n}\sum_{i=1}^{n} D(x_i, \hat{x}_i)$. For a serial network consisting of L nodes, a sequence of $L-1$ length-n block encoder-decoder pairs can be designed such that if at node l the pair $(\mathbf{y}_l, \mathbf{u}_{l-1}) \in \mathcal{T}_{y_l, u_{l-1}}$ is encoded at rate R_l then, with arbitrarily high probability as n grows to infinity, $\mathbf{x}$ can be recovered at node $l+1$ to within any average distortion d_{l+1} that satisfies

$$d_{l+1} \geq \min_{f_{l+1}, u_l \in \mathbf{\Pi}_l} E\left[D(x, f_{l+1}(y_{l+1}, u_l))\right].$$

The minimization is over all functions $f_{l+1} : \mathcal{Y}_{l+1} \times \mathcal{U}_l \to \hat{\mathcal{X}}$ and $u_l \in \mathbf{\Pi}_l$ for all l where the set $\mathbf{\Pi}_l$ consists of all random variables u_l such that

(i) $u_l \leftrightarrow y_l, u_{l-1} \leftrightarrow x, y_1^{l-1}, y_{l+1}^{L}, u_1^{l-2}$,

(ii) $R_l \geq I(y_l, u_{l-1}; u_l) - I(y_{l+1}; u_l)$.

2) Parallel Networks: A parallel network has a hub-and-spoke structure, as is illustrated in Fig. 4. Encoding and decoding are done successively. In general, by step l the hub has decoded messages from nodes 1 through $l-1$, getting $\mathbf{u}_1, \ldots, \mathbf{u}_{l-1}$. In each step the decoded message is used to improve the source estimate.

Proposition 2: Let a set of source and observation vectors $(\mathbf{x}, \mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_L)$, and a distortion measure $D(\cdot, \cdot)$ be given such that (a) $p_{\mathbf{x},\mathbf{y}_1,\ldots,\mathbf{y}_L}(\mathbf{x}, \mathbf{y}_1, \ldots, \mathbf{y}_L) = \prod_{i=1}^{n} p_{x,y_1,\ldots,y_L}(x_i, y_{1,i}, \ldots, y_{L,i})$, and (b) $D(\mathbf{x}, \hat{\mathbf{x}}) = \frac{1}{n}\sum_{i=1}^{n} D(x_i, \hat{x}_i)$. For a parallel network consisting of L sensor and 1 hub node, a sequence of L length-n block encoder-decoder pairs can be designed such that if $\mathbf{y}_l$ is encoded at rate R_l, and messages are decoded in order $1, 2, \ldots, L$, then with arbitrarily high probability as n grows to infinity, after the first l messages have been decoded by the hub node, $\mathbf{x}$ can be recovered to within any average distortion d_l that satisfies

$$d_l \geq \min_{f_l, u_l \in \mathbf{\Pi}_l} E\left[D(x, f_l(u_1^l))\right].$$

The minimization is over all functions $f_l : \mathcal{U}_1 \times \ldots \times \mathcal{U}_l \to \hat{\mathcal{X}}$, and $u_l \in \mathbf{\Pi}_l$ for all l, where the set $\mathbf{\Pi}_l$ consists of all random variables u_l such that

(i) $u_l \leftrightarrow y_l \leftrightarrow x, y_1^{l-1}, y_{l+1}^{L}, u_1^{l-1}$,

(ii) $R_l \geq I(y_l; u_l) - I(u_1^{l-1}; u_l)$.

B. Cut-Set Bound

In this section we use the fact, discussed in Sec. II-C, that Thm. 1 is tight for the memoryless scenario to derive a cut-set bound on estimation performance at the root of a sensor tree. Partition the nodes into two groups: $\mathcal{A}$ and its complement $\mathcal{A}^c$, where $\mathcal{A}^c$ contains the root node. Each group is allowed to convene and share observations losslessly. Group $\mathcal{A}$ then transmits a message to group $\mathcal{A}^c$ at a rate equal to the sum of the rates of all links that connect a node in $\mathcal{A}$ to one in $\mathcal{A}^c$. The observations $\mathbf{y}^{\mathcal{A}}$ of the convened nodes in $\mathcal{A}$ form a vector of encoder source observations, while the observations $\mathbf{y}^{\mathcal{A}^c}$ of the convened nodes in $\mathcal{A}^c$ form a vector of decoder side information. Since a number of inter-node rate constraints must be relaxed to allow the convening of nodes

in any partition, each partition provides a lower-bound on the achievable distortion. Any achievable distortion must satisfy all possible partitions. This gives the following theorem, which follows directly from Thm. 1, and so is stated without proof.

Theorem 2: Let a set of L nodes make up a tree-structured sensor network. Let R_{ij} be the link rate from node i to node j. Furthermore, let $(\mathbf{x}, \mathbf{y}_1, \ldots, \mathbf{y}_L)$ and $D(\cdot,\cdot)$ be given such that (a) $p_{\mathbf{x},\mathbf{y}_1,\ldots,\mathbf{y}_L}(\mathbf{x}, \mathbf{y}_1, \ldots, \mathbf{y}_L) = \prod_{i=1}^{n} p_{x,y_1,\ldots,y_L}(x_i, y_{1,i}, \ldots, y_{L,i})$, and (b) $D(\mathbf{x}, \hat{\mathbf{x}}) = \frac{1}{n}\sum_{i=1}^{n} D(x_i, \hat{x}_i)$. Then, if a sequence of $L-1$ length-n block encoding and decoding rules can be designed to recover $\mathbf{x}$ to within average distortion d at the root node, with arbitrarily high probability as n grows to infinity, d must satisfy the following inequality

$$d \geq \max_{\mathcal{A}} \min_{f, u \in \mathbf{\Pi}} E\left[D(x, f(y^{\mathcal{A}^c}, u))\right]. \tag{2}$$

The maximization is over all partitions of the nodes into the two sets $\mathcal{A}$ and $\mathcal{A}^c$, such that $\mathcal{A}^c$ forms a subtree of the network that contains the root node.[1] The minimization is over all functions $f : \mathcal{Y}^{\mathcal{A}^c} \times \mathcal{U} \to \hat{\mathcal{X}}$, and the set $\mathbf{\Pi}$ consists of all random variables u such that

(i) $u \leftrightarrow y^{\mathcal{A}} \leftrightarrow x, y^{\mathcal{A}^c}$,

(ii) $\sum_{i\in\mathcal{A}, j\in\mathcal{A}^c} R_{ij} > I(y^{\mathcal{A}}; u) - I(y^{\mathcal{A}^c}; u)$.

While fairly loose in general, this bound can, in certain settings, identify when a scheme is good. The bound is tightest when there is a particular cut-set that serves as a choke point. For example, consider a serial network with constant-rate links, and let the first node in the network observe the source directly, $\mathbf{y}_1 = \mathbf{x}$. A distortion-minimizing solution is to apply regular Wyner-Ziv coding to this problem, where intermediate nodes simply forward the first node's message to the last node. The cut-set bound confirms this by grouping together all nodes except the last. However, note that this performance is attained at the cost of violating our architectural principles since all intermediate nodes must coordinate to forward the initial message unchanged.

IV. Quadratic-Gaussian Cases

In this section we discuss quadratic-Gaussian problems, which give useful insight into practical scenarios. In Sec. IV-A we specify an achievable rate-distortion trade-off for the one-step problem of Thm. 1. In Sections IV-B and IV-C we discuss the multi-step serial and parallel networks, respectively.

A. One-step problem

The rate-distortion trade-off for the one-step problem has a particularly simple form when the source and observations are jointly Gaussian and the distortion measure is quadratic (mean-squared error). In App. A we specify a test channel for this problem that results in the following rate-distortion trade-off:

$$R(d) = \frac{1}{2}\log\left[\frac{\sigma^2_{x|y_D} - \sigma^2_{x|y_E,y_D}}{d - \sigma^2_{x|y_E,y_D}}\right], \tag{3}$$

where $\sigma^2_{x|y_E,y_D} \leq d \leq \sigma^2_{x|y_D}$. The conditional variance $\sigma^2_{x|y_D}$ is the minimum mean-squared estimation error given the decoder observation y_D, while $\sigma^2_{x|y_E,y_D}$ is similarly defined given both observations. The distortion-rate form is

$$d(R) = \sigma^2_{x|y_E,y_D} + (\sigma^2_{x|y_D} - \sigma^2_{x|y_E,y_D})2^{-2R}. \tag{4}$$

When the source and observations are memoryless, (3) specifies the rate-distortion frontier. Full derivations of this result, as well as for the binary-Hamming case, can be found in [40], [11].

B. Serial Networks

The distortion-rate performance of the successive coding strategy in serial networks has a simple iterative form in the quadratic-Gaussian case. In this example, the memoryless Gaussian source $\mathbf{x}$ is observed at each node in independent additive white Gaussian noise. Specifically, node l observes $\mathbf{y}_l = \mathbf{x} + \mathbf{v}_l$ where $\mathbf{x} \sim \mathcal{N}(0, \sigma_x^2\mathbf{I})$ and $\mathbf{v}_l \sim \mathcal{N}(0, N_l\mathbf{I})$ are independent. It sends a message at rate R_l to node $l+1$. In App. B we show that the following distortion-rate trade-off can be achieved by our approach:

$$d_l = \frac{N_l\, d_{l-1}}{N_l + d_{l-1}} + \sigma^2_{x|y_l}\frac{\left(1 - \frac{d_{l-1}}{\sigma_x^2}\right)}{\left(1 + \frac{d_{l-1}}{N_l}\right)}2^{-2R_{l-1}}, \tag{5}$$

where $d_1 = \sigma^2_{x|y_1}$. As all link rates become arbitrarily large, the second term of (5) converges to zero, and the first term generates the infinite-rate bound $\sigma^2_{x|y_1,\ldots,y_l}$. Generally, the finiteness of R_{l-1} slows the decrease of d_l with l.

While in this paper we concentrate on developing estimation strategies for situations where the network is prescribed, if one had the flexibility to design the structure of the network, and allocate link rates, then we can use the results of this section to make observations on what are better and worse choices. For example, we can ask how the link rates must grow if one is to obtain the full benefit of the observations, and avoid the type of saturation effects illustrated in Fig. 2. In particular, we determine the rate allocation needed to stay within a constant multiple $(1+\Delta) \geq 1$ of the lower-bound $\sigma^2_{x|y_1,\ldots,y_l}$ in the case of constant $\mathrm{SNR} = \sigma_x^2/N_l$ for all l. The rate allocation can be found by setting $d_l = (1+\Delta)\sigma^2_{x|y_1,\ldots,y_l}$, and using (5) to solve for the rate R_l such that $d_{l+1} = (1+\Delta)\sigma^2_{x|y_1,\ldots,y_{l+1}}$:

$$\begin{aligned} R_l &= \frac{1}{2}\log\left[\frac{[(l+1)\,\mathrm{SNR}+1][l\,\mathrm{SNR}-\Delta]}{\mathrm{SNR}(1+\mathrm{SNR})\Delta(1+\Delta)}\right] \\ &\leq \frac{1}{2}\log\left[\frac{\mathrm{SNR}}{1+\mathrm{SNR}}\right] + \log\left[\frac{l}{\Delta}\right] + o_l(1). \end{aligned}$$

This reveals that rate must increase logarithmically.

To help quantify the rate savings of side-information-awareness, we calculate the extra rate required by estimate-and-quantize to achieve the same target distortion levels. If $R_{\mathrm{EQ},l}$ is the rate required by estimate-and-quantize[2] to achieve

[1] The subtree condition ensures there is no communication cycle between node groups — a more complex topology than Wyner-Ziv coding allows for.

[2] Due to lack of space, we do not include derivation of estimate-and-quantize performance. The derivation which, on the whole, is similar to the derivation of App. B can be found in [11].

$d_l = (1+\Delta)\sigma^2_{x|y_1,\ldots,y_l}$, then the rate-savings is

$$R_{\mathrm{EQ},l} - R_l = \frac{1}{2}\log\left[\left(1 - \frac{(1+\Delta)\,\mathrm{SNR}}{(l+1)\,\mathrm{SNR}+1}\right)(1+\mathrm{SNR})\right].$$

This difference decreases with increasing Δ, since a larger Δ means that the target distortion is more easily met. As l increases, however, the difference converges to a constant

$$\lim_{l\to\infty}[R_{\mathrm{EQ},l} - R_l] = \frac{1}{2}\log[1+\mathrm{SNR}] = \frac{1}{2}\log\left[\frac{\sigma_x^2}{\sigma^2_{x|y}}\right].$$

This limit equals the rate needed by a standard quantizer to achieve the quality $\sigma^2_{x|y}$ of the decoder side information when that side information is ignored.

C. Parallel Networks

The performance of the successive coding strategy when applied to parallel networks also has a particularly tractable form in the quadratic-Gaussian case. Let the source $\mathbf{x}$ and observations $\mathbf{y}_l$ be defined as in Sec. IV-B. Denote by d_l the distortion in the hub node's estimate after it has received the first l messages, where $d_0 = \sigma_x^2$. If the hub node has its own observation, account for it through an additional observation node without a rate constraint. Then, in App. C, we show that the following trade-off is attained:

$$d_l = \frac{N_l\, d_{l-1}}{N_l + d_{l-1}} + \frac{d_{l-1}^2}{N_l + d_{l-1}} 2^{-2R_l}. \tag{6}$$

Just as in serial networks, we can determine the rate allocation needed to stay within the constant multiple $(1+\Delta)$ of $\sigma^2_{x|y_1,\ldots,y_l}$. In contrast to the serial network, the needed R_l now decreases with l. This is because as the hub node accumulates messages, its side information improves, and later nodes can communicate more efficiently.

One way to choose a message ordering is to use (6) to sort the nodes via a sequence of pair-wise decisions. Express (6) compactly as $d_l = g(d_{l-1}, N_l, R_l)$. Then, given two nodes with noise levels N_a and N_b, and communication rates R_a and R_b, the estimation error each ordering achieves starting from distortion d are

$$d_{ab} = g[g(d, N_a, R_a), N_b, R_b]$$
$$d_{ba} = g[g(d, N_b, R_b), N_a, R_a]$$

If, e.g., $d_{ab} < d_{ba}$ then it is best for node a to encode its message assuming no side information, and for node b to encode its message treating a's message as decoder side information. A sequence of pair-wise orderings extends this sorting to more than two nodes.

To further illustrate successive coding performance, we show that this strategy achieves the previously unknown rate-distortion frontier for a two node, sum-rate constrained problem. Let the two nodes have equal-variance independent additive white Gaussian noise observations, and assume that the hub does not have a source observation. Applying (6) twice, with $R_1 = \lambda R_{\mathrm{sum}}$ and $R_2 = (1-\lambda)R_{\mathrm{sum}}$ where $0 \le \lambda \le 1$ gives d_2 as a function of $\mathrm{SNR} = \sigma_x^2/N$, R_{sum}, and λ. Minimizing with respect to λ gives

$$\lambda = \frac{1}{2R_{\mathrm{sum}}}\log\left[\frac{-\mathrm{SNR}^2 + \gamma(1+\mathrm{SNR})}{(1+2\,\mathrm{SNR})}\right], \tag{7}$$

where $\gamma = \sqrt{\mathrm{SNR}^2 + (1+2\,\mathrm{SNR})2^{2R_{\mathrm{sum}}}}$. Note that $\lambda \ge 0.5$ and that $\lambda \to 0.5$ as R_{sum} gets either very small or very large. Using the fractional rate allocation (7) gives

$$d_2 = \frac{\sigma_x^2\, 2^{2R_{\mathrm{sum}}}}{(\gamma - \mathrm{SNR})^2}. \tag{8}$$

We can show that (8) is the distortion-rate frontier for the two node problem by using a bound that Oohama develops for the CEO problem in [21]. As discussed further in Sec. V-A, the CEO problem is a parallel network where the number of nodes increases to infinity. However, Oohama's bound is also applicable to systems with finite numbers of nodes. Using it with two nodes gives the distortion achieved in (8).

The optimization of this section can be generalized to nodes with differing SNRs. The resulting expressions are more complex and one node may receive the full sum-rate and the other zero rate. We conjecture that, given an appropriate rate allocation, our coding strategy can achieve the rate-distortion frontier for larger networks. In correspondence with Oohama [24] we have learned that he is also further investigating estimation problems for parallel networks with a finite number of nodes, and different SNRs. He claims to have found the rate-distortion frontier by using an inherently different (joint) decoding structure [22], [23]. Our results confirm one another for the two node case.

V. Applications to other Network Information Theory Problems

In this section we show that the successive coding strategies of Sec. III lead to novel solutions for a number of previously explored problems. In Sec. V-A we show how to achieve the rate-distortion frontier for the quadratic-Gaussian CEO problem. In Sec. V-B we show how to achieve the best known rate-distortion region for multiterminal source coding. Finally, in Sec. V-C we show how to apply the strategies to relay channel communications. While Sections V-A and V-B do not produce new results, they demonstrate alternate, simpler, and therefore potentially more useful approaches to the same results.

A. The CEO Problem

In this section we specify a rate-allocation for the successive coding strategy that achieves the rate-distortion bound of the quadratic-Gaussian CEO problem [34], [21]. The CEO problem has the same hub-and-spoke topology as the parallel network where the CEO acts as the hub node. There is a sum-rate constraint R_{sum} on all links, and the objective is to find the rate-distortion frontier as L, the number of nodes, grows to infinity. Hence, the average per-node rate R_{sum}/L goes to zero. As noted in Sec. IV-C, in the two-node problem, as R_{sum} gets very small, an equal per-node rate allocation is optimal.

We show that using our coding strategy with an equal per-node rate allocation $R_l = R_{\rm sum}/L$ achieves the rate-distortion frontier for the quadratic-Gaussian CEO problem with equal-SNR observations, a frontier first achieved by Oohama in [21].

Using $R_l = R_{\rm sum}/L$, defining the distortion-to-noise ratio at node l as $x_l = d_l/N$, and working in nats for convienience, we can rewrite (6) as $\frac{x_l - x_{l-1}}{1-e^{-2R_{\rm sum}/L}} = -\frac{x_{l-1}^2}{1+x_{l-1}}$. For L large, $1-\exp(-2R_{\rm sum}/L) \simeq 2R_{\rm sum}/L$ which we use to get $\frac{x_l - x_{l-1}}{R_{\rm sum}/L} \simeq \frac{-2x_{l-1}^2}{1+x_{l-1}}$. For large L, this can be approximated to arbitrary precision [11] by the differential equation $\frac{dx}{dR} = \frac{-2x^2}{1+x}$, where $dR = R_{\rm sum}/L$ is the per-node rate increase and $dx = x_l - x_{l-1}$ is the per-node decrease in distortion-to-noise ratio. Solving this differential equation gives

$$R_{\rm sum} = \int_0^{R_{\rm sum}} dR = \int_{\frac{\sigma_x^2}{N}}^{\frac{d}{N}} \left(-\frac{1}{2x^2} - \frac{1}{2x}\right) dx = \frac{N}{2\sigma_x^2}\left[\frac{\sigma_x^2}{d} - 1\right] + \frac{1}{2}\log\frac{\sigma_x^2}{d}. \quad (9)$$

Equation (9) is the rate-distortion frontier for the problem [21].

The successive coding framework suggested here may well better fit the architectural constraints of sensor networks than approaches based on joint decoding [21]. First, in joint decoding, decoding cannot begin before all messages are received. The successive coding technique we propose allows incremental increases in estimate quality as each message is decoded. Secondly, joint decoding requires multiple messages be decoded simultaneously, which is an exponentially more complex task than a sequence of single-message decoding problems. Finally, while joint decoding requires coordination between all encoders to ensure that the hub can decode, successive coding require coordination only between each encoder and the hub at each step.

B. Multiterminal Source Coding

We now show how to use successive coding to reproduce the best known achievable rate region for the multiterminal source coding problem [3], [33], [20]. In multiterminal source coding L sources $\mathbf{y}_1, \ldots, \mathbf{y}_L$ are observed at L separate encoders where $p_{\mathbf{y}_1,\ldots,\mathbf{y}_L}(\mathbf{y}_1,\ldots,\mathbf{y}_L) = \prod_{i=1}^n p_{y_1,\ldots,y_L}(y_{1,i},\ldots,y_{L,i})$. Encoder l sends a message at rate R_l bits per source sample to a central decoding hub. The hub decodes all messages and makes estimates of all sources, $\hat{\mathbf{y}}_1, \ldots, \hat{\mathbf{y}}_L$. This is a parallel network, but the problem objective has changed — we now estimate the observations, rather than a single underlying signal.

Theorem 3: Let a L-tuple of source vectors $(\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_L)$, and a set of distortion measures $\{D_1(\cdot,\cdot), \ldots, D_L(\cdot,\cdot)\}$ be given such that (a) $p_{\mathbf{y}_1,\ldots,\mathbf{y}_L}(\mathbf{y}_1,\ldots,\mathbf{y}_L) = \prod_{i=1}^n p_{y_1,\ldots,y_L}(y_{1,i},\ldots,y_{L,i})$, and (b) $D_l(\mathbf{y}_l, \hat{\mathbf{y}}_l) = \frac{1}{n}\sum_{i=1}^n D_l(y_{l,i}, \hat{y}_{l,i})$. Then a sequence of L length-n block encoder-decoder pairs can be designed such that if $\mathbf{y}_l$ is encoded at rate R_l, and the messages are decoded in order $1, 2, \ldots, L$, then with arbitrarily high probability as n grows to infinity, $\mathbf{y}_l$ can be recovered to within any average distortion d_l that satisfies

$$d_l \geq \min_{f_l, u_l \in \mathbf{\Pi}_l} E\left[D_l(y_l, f_l(u_1^L))\right]. \quad (10)$$

The minimization is over all functions $f_l : \mathcal{U}_1 \times \ldots \times \mathcal{U}_L \to \hat{\mathcal{Y}}_l$, and $u_l \in \mathbf{\Pi}_l$ for all l where the set $\mathbf{\Pi}_l$ consists of all random variables u_l such that

(i) $u_l \leftrightarrow y_l \leftrightarrow y_1^{l-1}, y_{l+1}^L, u_1^{l-1}$.
(ii) $R_l \geq I(y_l; u_l) - I(u_1^{l-1}; u_l)$.

The achieved rate region is found by taking the convex hull over all rate points resulting from different choices of the $p_{u_l|y_l}(u_l|y_l)$ satisfying (i) and (ii), and transmission orderings.[3]

The proof of this theorem is basically identical to that of Prop. 2 and so is omitted. The sum-rate used by encoders $1, \ldots, k$ has a particularly simple form, $\sum_{l=1}^k R_l \geq I(y_1^k; u_1^k)$. To see this, note that it is true for $k = 1$ by Thm. 3. We show the result for $k > 1$ by induction as follows,

$$\begin{aligned}\sum_{l=1}^k R_l &\geq \sum_{l=1}^k [I(y_l; u_l) - I(u_1^{l-1}; u_l)] = \sum_{l=1}^k I(y_l; u_l | u_1^{l-1}) \\ &= I(y_k; u_k | u_1^{k-1}) + I(y_1^{k-1}; u_1^{k-1}) \\ &= I(y_1^k; u_k | u_1^{k-1}) + I(y_1^k; u_1^{k-1}) = I(y_1^k; u_1^k). \end{aligned} \quad (11)$$

where the first line follows from Markov chain, the second from the induction assumption, and the third because conditioned on y_k, u_k is independent of all other variables, and because $u_1^{k-1} \leftrightarrow y_1^{k-1} \leftrightarrow y_k$.

Consider the two-terminal case, $L = 2$, first investigated by Berger and Tung [3], [33]. There are two possible orderings: (a) the message from terminal 1 is designed to be decoded first, and (b) the message from terminal 2 is designed to be decoded first. Ordering (a) and Thm. 3 gives:

$$R_1 \geq I(y_1; u_1), \quad R_2 \geq I(y_2; u_2 | u_1), \quad (12)$$

$$R_1 + R_2 \geq I(y_1, y_2; u_1, u_2). \quad (13)$$

Ordering (b) gives the same rates as (12)–(13) with the subscripts interchanged. From (11) we know the rate pair $(R_1, R_2) = (I(y_1; u_1), I(y_2; u_2|u_1))$ lies on the sum-rate bound given by (13). Allowing time sharing between orderings (a) and (b), and each choice of valid joint distribution $p(y_1, y_2, u_1, u_2)$ achieves a rate-distortion region identical to that given in [3], [33].

C. "Estimate-and-Detect" for Relay Channels

Finally, in a rather different direction from the rest of the paper, in this section we consider distributed detection problems. We design a two part "estimate-and-detect" strategy for the relay channel whereby we first estimate the codeword using the distributed estimation techniques developed herein, and then detect the message based on the estimate. In the case of a single relay and additional direct path, as discussed in [6], the scheme reduces to one presented in [9].

We focus on a parallel Gaussian two-relay network [29]. The transmitted codeword is constrained to power P, measured by each relay in additive white Gaussian noise of variance N. The relays send rate-constrained messages to a central decoder

[3]In the theorem we define the d_l to be measured after all messages are received. This makes it easier to compare our results to earlier results. However, it is also possible to make estimates after each decoding step giving, e.g., $d_{l,k}$, the distortion in the estimate of $\mathbf{y}_l$ made after the first k codewords have been decoded.

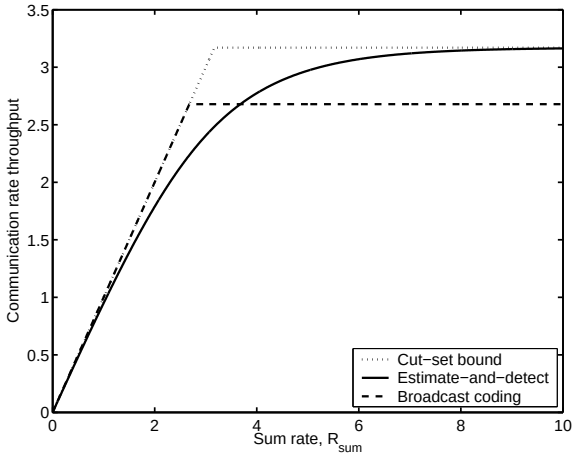

Fig. 5. Communication rate throughput achieved by estimate-and-detect with two additive white Gaussian noise observations at the relays, both with $\mathrm{SNR} = 40$.

under a sum-rate constraint R_{sum}. A baseline approach is to use broadcast codes to communicate losslessly to the two relays. Since the noise powers are identical, the maximal reliable sum transmission rate to the relays is $R_{\mathrm{BC}} = 0.5 \log[1 + \mathrm{SNR}]$ where $\mathrm{SNR} = P/N$. The communication throughput of this strategy is $\min\{R_{\mathrm{BC}}, R_{\mathrm{sum}}\}$. On the other hand, estimate-and-detect builds on the results of Sec. IV-C. Codewords are generated in an i.i.d. Gaussian manner. Based on their observations, the relays send bit streams to the decoder which makes an estimate of the codeword using the techniques of this paper. Since the resulting estimate and codeword are jointly typical, we can detect the message using standard typicality decoding. The resulting throughput is

$$\begin{aligned} R_{\mathrm{EstDet}} &= I(x; u_1, u_2) = 0.5 \log[2\pi e \sigma_x^2] - 0.5 \log[2\pi e \sigma^2_{x|u_1,u_2}] \\ &= 0.5 \log [P/d_2] = 0.5 \log \left[(\gamma - \mathrm{SNR})^2 \, 2^{-2R_{\mathrm{sum}}} \right], \end{aligned}$$

where u_1 and u_2 are the auxiliary random variables of the two relays, $\gamma = \sqrt{\mathrm{SNR}^2 + (1 + 2\,\mathrm{SNR}) 2^{2R_{\mathrm{sum}}}}$, and d_2 is the distortion achieved in (8) with $\sigma_x^2 = P$.

Figure 5 plots the communication throughput of the two schemes versus R_{sum}. Broadcast coding does better for small R_{sum} since estimate-and-detect introduces extra quantization noise. On the other hand, for large enough R_{sum} estimate-and-detect outperforms broadcast coding since it is able to exploit the diversity of the relay observations. For comparison we plot the minimal cut-set bound: the minimum of the information flow to the relays, and R_{sum}.

In [30] Schein discusses strategies for this situation where the relays communicate to the decoder over rate-constrained channels. He derives qualitative results similar to those of Fig. 5, but the explicit rate evaluation presented herein is new. Other recent work in this area includes, e.g., [27], [14], [19].

VI. Future Directions

Many aspects of communication-constrained estimation algorithms remain to be explored. First, we would like to derive tighter converses, e.g., for the serial problem. The multi-step structure of the problem differentiates it from other problems where tight converses are known. Second, we would like to develop practical encoders and decoders. Recent progress on building side information coding systems (see, e.g., [26], [42]) should prove useful. Finally, there is a host of interesting network-layer issues in deciding how to choose the sensor tree, allocate rates, and how to manage the network to be robust to the failure of individual nodes. Some recent work in this direction has appeared, e.g., in [28].

Appendix

A. One-step problem

We do not formally extend the finite-alphabet results of Theorem 1 to continuous alphabets This extension can be made using tools developed, e.g., in [37], [20]. Given this extension, we specify a test channel that gives the rate-distortion trade-off of (1).

We consider the case where x, y_{E}, y_{D}, and y_{N} are jointly Gaussian random variables. Define the auxiliary random variable $u = \alpha y_{\mathrm{E}} + e$ where $e \sim N(0, \alpha d^*)$ is independent of x, y_{E}, y_{D}, and y_{N}. For this choice of u, $I(y_{\mathrm{E}}; u) - I(y_{\mathrm{D}}; u) = \frac{1}{2} \log \left[1 + \frac{\alpha}{d^*} \sigma^2_{y_{\mathrm{E}}|y_{\mathrm{D}}} \right]$. The minimum mean-squared estimation error for x given y_{D} and u is

$$\sigma^2_{x|y_{\mathrm{D}},u} = \frac{\frac{\alpha}{d^*} \sigma^2_{x|y_{\mathrm{E}},y_{\mathrm{D}}} + \frac{\sigma^2_{x|y_{\mathrm{D}}}}{\sigma^2_{y_{\mathrm{E}}|y_{\mathrm{D}}}}}{\frac{\alpha}{d^*} + \frac{1}{\sigma^2_{y_{\mathrm{E}}|y_{\mathrm{D}}}}}. \tag{14}$$

Setting (14) equal to the target distortion d, and solving for α/d^* gives $\frac{\alpha}{d^*} = \frac{1}{\sigma^2_{y_{\mathrm{E}}|y_{\mathrm{D}}}} \left(\frac{\sigma^2_{x|y_{\mathrm{D}}} - d}{d - \sigma^2_{x|y_{\mathrm{E}},y_{\mathrm{D}}}} \right)$. Substituting this into the expression for $I(y_{\mathrm{E}}; u) - I(y_{\mathrm{D}}; u)$ gives (3). In [11] we specify the data-fusion function $f(\cdot, \cdot)$ and, for the case of a memoryless source and observations, give a converse that shows (3) is the rate-distortion frontier.

B. Serial Networks

Assuming that encoding and decoding are accomplished without error up to node $l-1$, then $\hat{x}_{l-1}$ and x are jointly typical. We use an innovations form to rewrite the relationship between $\hat{x}_{l-1}$ and x as $\hat{x}_{l-1} = \alpha x + \tilde{v}_{l-1}$, where $\alpha = \left(1 - \frac{d_{l-1}}{\sigma_x^2}\right)$ and $\tilde{v}_{l-1} \sim \mathcal{N}(0, \alpha d_{l-1})$. For the purpose of encoding, define node $l-1$'s source observation to be $z_{l-1} = \frac{\hat{x}_{l-1}}{\alpha} = x + \frac{\tilde{v}_{l-1}}{\alpha}$.

The encoding node's observation z_{l-1} can be treated as the source in additive white Gaussian noise, $\frac{1}{\alpha}\tilde{v}_{l-1}$, of variance $\frac{\sigma_x^2 d_{l-1}}{\sigma_x^2 - d_{l-1}}$. The decoding node's observation y_l serves as decoder side information. We can therefore use the distortion-rate trade-off (4) with $y_{\mathrm{E}} = z_{l-1}$, $y_{\mathrm{D}} = y_l$, $R = R_{l-1}$, and $d = d_l$. This results in an achieved distortion $d_l = \sigma^2_{x|y_l,z_{l-1}} + (\sigma^2_{x|y_l} - \sigma^2_{x|y_l,z_{l-1}}) 2^{-2R_{l-1}}$. Finally, using the relation $\sigma^2_{x|y_l,z_{l-1}} = \frac{1}{\frac{\sigma_x^2 - d_{l-1}}{\sigma_x^2 d_{l-1}} + \frac{1}{N_l} + \frac{1}{\sigma_x^2}} = \frac{N_l d_{l-1}}{N_l + d_{l-1}}$, we get (5).

C. Parallel Networks

As in App. B, we start by defining z_{l-1} in the same way. In the parallel network, however, this side information is known at the hub node, the decoder. Node l is the encoder and measures $x + v_l$ where $v_l \sim \mathcal{N}(0, N_l)$. Again, we use the distortion-rate form (4), but with $y_{\mathrm{E}} = y_l$, $y_{\mathrm{D}} = z_{l-1}$, $R = R_l$, and $d = d_{l-1}$. Simplification results in (6).

ACKNOWLEDGMENTS

The authors wish to thank Prof. Ram Zamir for helpful interactions, and Dr. Paul Algoet for pointing out the earlier work on remote Wyner-Ziv coding by Yamamoto and Itoh [40]. They also want to thank the reviewers for their careful reading of the manuscript and suggestions for improvement.

REFERENCES

[1] R. J. Barron, B. Chen, and G. W. Wornell. The duality between information embedding and source coding with side information and some applications. *IEEE Trans. Inform. Theory*, 49, 2003.

[2] J. Barros and S. D. Servetto. Reachback capacity with non-interfering nodes. In *Proc. Int. Symp. Inform. Theory*, page 366, Yokohama, Japan, June 2003.

[3] T. Berger. Multiterminal source coding. In G. Longo, editor, *The Information Theory Approach to Communications*, chapter 4. Springer-Verlag, 1977.

[4] T. Berger, Z. Zhang, and H. Viswanathan. The CEO problem. *IEEE Trans. Inform. Theory*, 42:887–902, May 1996.

[5] M. U. Chang. *Rate-Distortion with a Fully Informed Decoder and a Parially Informed Encoder*. PhD thesis, Cornell University, 1978.

[6] B. Chen, S. C. Draper, and G. W. Wornell. Information embedding and related problems: Recent results and applications. In *Proc. 39th Allerton Conf. on Communication, Control and Computing*, Allerton House, Monticello, IL, October 2001.

[7] A. Cohen, S. C. Draper, E. Martinian, and G. W. Wornell. Stealing bits from a quantized source. *Submitted to IEEE Trans. Inform. Theory*, 2003.

[8] T. M. Cover and M. Chiang. Duality between channel capacity and rate distortion with two-sided side information. *IEEE Trans. Inform. Theory*, 48:1629–1638, June 2002.

[9] T. M. Cover and A. A. El Gamal. Capacity theorems for the relay channel. *IEEE Trans. Inform. Theory*, 25:572–584, September 1979.

[10] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley and Sons, 1991.

[11] S. C. Draper and G. W. Wornell. Side information aware coding strategies for estimation under communication constraints. Technical Report 704, MIT Research Laboratory of Electronics, 2004.

[12] T. J. Flynn and R. M. Gray. Encoding of correlated observations. *IEEE Trans. Inform. Theory*, 33:773–787, November 1987.

[13] M. Gastpar and M. Vetterli. Source-channel communication in sensor networks. In *2nd Int. Workshop on Inform. Processing in Sensor Networks*, pages 162–177, Palo Alto, CA, April 2003.

[14] P. Gupta and P. R. Kumar. Toward an information theory of large networks: An achievable rate region. *IEEE Trans. Inform. Theory*, 49:1877–1894, August 2003.

[15] T. S. Han. Slepian-Wolf-Cover theorem for networks of channels. *Inform. and Contr.*, 47:67–83, 1980.

[16] T. S. Han and K. Kobayashi. A unified achievable rate region for a general class of multiterminal source coding systems. *IEEE Trans. Inform. Theory*, 26:277–288, May 1980.

[17] A. Kaspi. *Rate Distortion for Correlated Sources with Partially Separated Encoders*. PhD thesis, Cornell University, 1979.

[18] A. H. Kaspi and T. Berger. Rate-distortion for correlated sources with partially separated encoders. *IEEE Trans. Inform. Theory*, 28:828–840, November 1982.

[19] G. Kramer, M. Gastpar, and P. Gupta. Capacity theorems for wireless relay channels. In *Proc. 41st Allerton Conf. on Communication, Control and Computing*, pages 1074–1083, Allerton House, Monticello, IL, October 2003.

[20] Y. Oohama. Gaussian multiterminal source coding. *IEEE Trans. Inform. Theory*, 43:1912–1923, November 1997.

[21] Y. Oohama. The rate-distortion fuction for the quadratic Gaussian CEO problem. *IEEE Trans. Inform. Theory*, 44:1057–1070, May 1998.

[22] Y. Oohama. Multiterminal source coding for correlated memoryless Gaussian sources with several side informations at the decoder. In *IEEE Information Theory and Communications Workshop, Kruger National Park, South Africa*, page 100, June 1999.

[23] Y. Oohama. Rate-distortion theory for Gaussian multiterminal source coding systems with several side informations at the decoder. *Submitted to IEEE Trans. Info. Theory*, 2001.

[24] Y. Oohama. Personal communication. 2002.

[25] S. S. Pradhan, J. Chou, and K. Ramchandran. Duality between source coding and channel coding and its extension to the side information case. *IEEE Trans. Inform. Theory*, 49:1181–1203, May 2003.

[26] S. S. Pradhan and K. Ramchandran. Distributed source coding using syndromes (DISCUS): Design and construction. *IEEE Trans. Inform. Theory*, 49:626–643, March 2003.

[27] A. Reznik, S. R. Kulkarni, and S. Verdu. Capacity and optimal resource allocation in the degraded Gaussian relay channel with multiple relays. In *Proc. 40th Allerton Conf. on Communication, Control and Computing*, Allerton House, Monticello, IL, October 2002.

[28] A. Scaglione and S. D. Servetto. On the interdependence of routing and data compression in multi-hop sensor networks. In *Proc. 8th ACM Int. Conf. Mobile Computing and Networking*, Atlanta, GA, September 2002.

[29] B. Schein and R. Gallager. The Gaussian parallel relay channel. In *Proc. Int. Symp. Inform. Theory*, page 22, Sorrento, Italy, June 2000.

[30] B. E. Schein. *Distributed Coordination in Network Information Theory*. PhD thesis, Mass. Instit. of Tech., 2001.

[31] D. Slepian and J. K. Wolf. Noiseless coding of correlated information sources. *IEEE Trans. Inform. Theory*, 19:471–480, July 1973.

[32] J. N. Tsitsiklis. Decentralized detection. In H. V. Poor and J. B. Thomas, editors, *Advances in Statistical Signal Processing, vol. 2*, pages 297–344. JAI Press, 1993.

[33] S. Tung. *Multiterminal Source Coding*. PhD thesis, Cornell University, 1978.

[34] H. Viswanathan and T. Berger. The quadratic Gaussian CEO problem. *IEEE Trans. Inform. Theory*, 43:1549–1559, September 1997.

[35] R. Viswanathan and P. K. Varshney. Distributed detection with multiple sensors: Part I – Fundamentals. *Proc. IEEE*, 85:54–63, January 1997.

[36] A. D. Wyner. On source coding with side information at the decoder. *IEEE Trans. Inform. Theory*, 21:294–300, May 1975.

[37] A. D. Wyner. The rate-distortion function for source coding with side information at the decoder–II: General sources. *Inform. and Contr.*, 38:60–80, 1978.

[38] A. D. Wyner and J. Ziv. The rate-distortion function for source coding with side information at the decoder. *IEEE Trans. Inform. Theory*, 22:1–10, January 1976.

[39] H. Yamamoto. *Source Coding Theory for multiterminal communication systems*. PhD thesis, Univ. of Tokyo, 1980.

[40] H. Yamamoto and K. Itoh. Source coding theory for multiterminal communication systems with a remote source. *Trans. of the IECE of Japan*, 1980.

[41] R. Zamir and T. Berger. Multiterminal source coding with high resolution. *IEEE Trans. Inform. Theory*, 45:106–117, January 1999.

[42] R. Zamir, S. Shamai, and U. Erez. Nested linear/lattice codes for structured multiterminal binning. *IEEE Trans. Inform. Theory*, 48:1250–1276, June 2002.

Stark Draper (S'99–M'03) received the B.A. and B.S. degrees in history and electrical engineering, respectively, from Stanford University, Stanford, CA, and the M.S. and Ph.D. degrees from the Massachusetts Institute of Technology (MIT), Cambridge, MA.

From fall 2002 to spring 2004 he held the Information Processing Laboratory Postdoctoral Fellowship at the University of Toronto, Toronto, ON, Canada. Since spring 2004 he has been a postdoctoral associate in the Department of Electrical Engineering and Computer Science, University of California, Berkeley, Berkeley, CA. He has held industrial positions at a variety of places, including Arraycomm, San Jose, CA, and Draper Laboratory, Cambridge, MA.

Among several awards, Dr. Draper has received the MIT Carlton E. Tucker Teaching Award, an Intel Graduate Fellowship, Stanford's Frederick E. Terman Engineering Scholastic Award, and a Fulbright Fellowship. His research interests and activities span several aspects of signal processing, communications, estimation, information theory, queuing, and networking.

Gregory Wornell (S'83–M'91–SM'00–F'04) received the B.A.Sc. degree from the University of British Columbia, Vancouver, BC, Canada, and the M.S. and Ph.D. degrees from the Massachusetts Institute of Technology (MIT), Cambridge, all in electrical engineering and computer science, in 1985, 1987 and 1991, respectively.

Since 1991 he has been on the MIT faculty, where he is Professor of Electrical Engineering and Computer Science, Co-director of the Center for Wireless Networking, and Chair of the department's Graduate Area I (Systems, Communication, Control, and Signal Processing) within the department's doctoral program. He has held visiting appointments at the former AT&T Bell Laboratories, Murray Hill, NJ, the University of California, Berkeley, and Hewlett-Packard Laboratories, Palo Alto, CA. His research interests and publications span the areas of signal processing, digital communication, and information theory, and include algorithms and architecture for wireless networks, broad-band systems, and multimedia environments.

Dr. Wornell has been involved in the Signal Processing and Information Theory societies of the IEEE in a variety of capacities, and maintains a number of close industrial relationships and activities. He has won a number of awards for both his research and teaching.