

Distributed Quantization Networks

John Z. Sun, *Student Member, IEEE*, and Vivek K Goyal, *Senior Member, IEEE*

Abstract

Several key results in distributed source coding offer the intuition that little improvement in compression can be gained from intersensor communication when the information is coded in long blocks. However, when sensors are restricted to code their observations in small blocks (e.g., 1), intelligent collaboration between sensors can greatly reduce distortion. For networks where sensors are allowed to “chat” using a side channel that is unobservable at the fusion center, we provide asymptotically-exact characterization of distortion performance and optimal quantizer design in the high-resolution (low-distortion) regime using a framework called distributed functional scalar quantization (DFSQ). The key result is that chatting can dramatically improve performance even when intersensor communication is at very low rate, especially if the fusion center desires fidelity of a nonlinear computation applied to source realizations rather than fidelity in representing the sources themselves. We also solve the rate allocation problem when communication links have heterogeneous costs and provide a detailed example to demonstrate the theoretical and practical gains from chatting. This example for maximum computation gives insight on the gap between chatting and distributed networks, and how to optimize the intersensor communication.

Index Terms

distributed source coding, high-resolution quantization, sensor networks, side information

I. INTRODUCTION

A longstanding consideration in distributed compression systems is whether sensors wishing to convey information to a fusion center should communicate with each other to improve efficiency. Architectures that only allow communication between individual sensors and the fusion center simplify the network’s communication protocol and decrease sensor responsibilities. Moreover, information theoretic results such as the Slepian–Wolf theorem show that distributed compression can perform as well as joint compression for lossless communication of correlated information sources [1]. Although this surprising and beautiful result does not extend fully, comparable results for lossy coding show that the rate loss from separate encoding can be small using Berger–Tung coding (see, e.g., [2]), again suggesting that communication between sensors has little or no utility.

Although it is tempting to use results from information theory to justify simple communication topologies, it is important to note the Slepian–Wolf result is dependent on large blocklength; in the finite-blocklength regime, the optimality of distributed encoding does not hold [3]. This paper examines the use of communication among sensors when the compression blocklength is 1, a regime where collaboration, called *chatting* in this work, can greatly decrease the aggregate communication from sensors to the fusion center to meet a distortion criterion as compared to a distributed network. We analyze chatting networks using the distributed functional scalar quantization (DFSQ) framework, which constrains sensors to using scalar quantizers to compress their observations and generalizes the fusion center’s objective to desire fidelity in computing a function of the sources rather than determining the sources themselves [4], [5]. Our problem model is shown in Fig. 1, where N correlated but memoryless continuous-valued, discrete-time stochastic processes produce scalar realizations $X_1^N(t) = (X_1(t), \dots, X_N(t))$ for $t \in \mathbb{Z}$. For each t , realizations of these sources are scalar quantized by sensors and transmitted to a fusion center at rates R_1^N . To aid this communication, sensors can collaborate with each other via a side channel that is unobservable to the fusion center. Since the quantization is scalar and the sources are memoryless, we remove the time index and model the sources as being drawn from a joint distribution $f_{X_1^N}$ at each t .

The side channel facilitating intersensor communication has practical implications. In typical communication systems, the transmission power needed for reliable communication increases superlinearly with distance and bandwidth [6]. Hence, it is much cheaper to design short and low-rate links between sensors than reliable and high-rate links to a fusion center. Moreover, milder transmission requirements provide more flexibility in determining the transmission media or communication modalities employed, which can allow intersensor communication to be orthogonal to the main network. One such example is cognitive radio, a paradigm where the wireless spectrum can have secondary users that communicate only when the primary users are silent [7]. This means secondary users have less priority and hence lower reliability and rate, which is adequate for intersensor communication.

The main contributions of the paper are to precisely characterize the distortion performance of a distributed network when chatting is allowed and to identify the optimal quantizer design for each sensor. We show that collaboration can have significant impact on performance; in some cases, it can dramatically reduce distortion even when the chatting has extremely low rate. We also give necessary conditions on the chatting topology and protocol for successful decodability in the DFSQ framework,

J. Z. Sun is with the Department of Electrical Engineering and Computer Science and the Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, MA 02139 USA (e-mail: johnsun@mit.edu).

V. K. Goyal is with the Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, MA 02139 USA (e-mail: v.goyal@ieee.org).

This material is based upon work supported by the National Science Foundation under Grant No. 1115159.

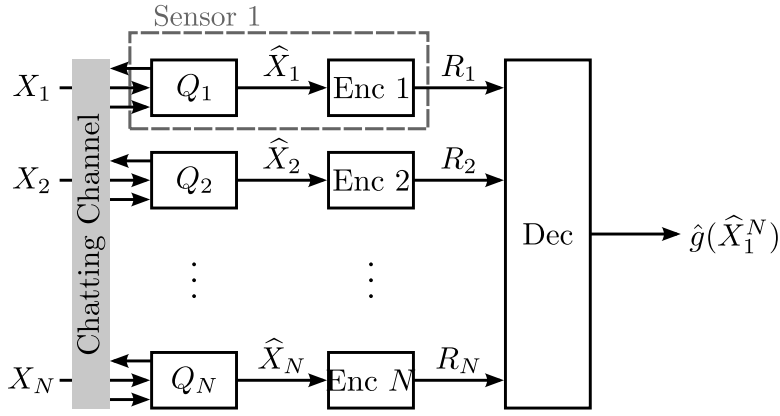


Fig. 1. A distributed computation network, where N sensors (comprising quantizer and encoder) observe realizations of correlated sources. Each observation X_n is encoded and communicated over rate-limited links to a fusion center. Simultaneously, each sensor can interact with a subset of other sensors using a noiseless but rate-limited chatting channel to improve compressibility. The decoder at the fusion center computes an estimate of the function $g(X_1^n) = g(X_1, X_2, \dots, X_n)$ from the received data using a reconstruction function $\hat{g}(\hat{X}_1^n)$ but cannot observe messages communicated on the chatting channel.

thus providing insight into the architecture design for chatting networks. Finally, we recognize that intersensor communication can occur on low-cost channels and solve the rate allocation problem in networks with heterogeneous links and different costs of transmission. The basic concepts of this work were introduced in [8]; this paper provides more complete and definitive coverage, including more results on rate allocation, a discussion on generalizing chatting messages, and details on the impact of various optimizations.

We begin by introducing related work, notation and prerequisite results in Section II. In Section III, we analyze the performance of chatting networks and discuss how to optimize the communication that occurs. We then determine the proper rate allocation for chatting networks in Section IV. Finally, we develop intuition for the behavior of chatting by considering a maximum computation network in Section V; this specific example demonstrates the incremental gains achieved by incorporating the different optimizations discussed in the paper.

II. PRELIMINARIES

A. Previous Work

There is a large body of literature studying asymptotic performance of the distributed network in Fig. 1 without the chatting channel; a comprehensive review of these works and their connections to DFSQ appears in [4]. Similarly, connections to coding for computing (e.g. [9], [10]) are discussed there as well. Recent work on the finite-blocklength regime [11] has led to extensions in source coding [3], [12], [13]. In general, this analysis technique is meaningful for blocklengths as low as 100, but is unsuitable for regimes traditionally considered in high-resolution theory.

We review results that relate to the chatting channel, focusing on Shannon-theoretic results. Kaspi and Berger provided inner bounds for the rate region of a two-encoder problem where one encoder can send information for the other using compress-and-forward techniques [14]. Recently, this bound has been generalized in [15], but the exact rate region is still unknown except in special cases. Chatting is related to source coding problems such as interaction [16], [17], omniscience [18] and data exchange [19]. However, these settings are more naturally suited for discrete-alphabet sources and existing results rely on large-blocklength analysis.

There are also strong connections between this work and distortion side information [20] and vector quantization with alternative distortion measures [21].

B. Quantization

The focus of this work is on compression of continuous-valued, finite-support sources using small blocks of data. Here, performance results from Shannon theory are overly optimistic since tools such as joint-typicality encoding and decoding are not reliable without operating far from the distortion–rate bound. Instead, we consider the complementary asymptotic of high resolution, where the blocklength is small and the compression rate R is large [22]–[24]. Before introducing the high-resolution asymptotic, we summarize the quantization model for the case of blocklength 1 and set up the notation used for the rest of the paper.

A scalar quantizer Q_K is a mapping from the real line to a set of K points $\mathcal{C} = \{c_k\}_{k=1}^K \subset \mathbb{R}$ called the codebook, where $Q_K(x) = c_k$ if $x \in P_k$ and the cells $\{P_k\}_{k=1}^K$ form a partition of \mathbb{R} . The quantizer is called *regular* if the partition cells are intervals containing the corresponding codewords. For simplicity, the codebook and partition are indexed from smallest to

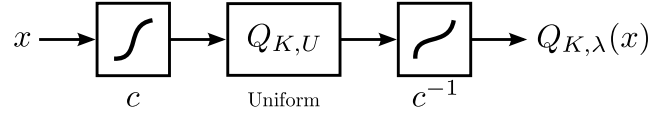


Fig. 2. The companding model is a method to construct nonuniform quantizers using a monotonic nonlinearity c satisfying $\lim_{x \rightarrow -\infty} c(x) = 0$ and $\lim_{x \rightarrow \infty} c(x) = 1$. The notation $Q_{K,U}$ is used to describe the canonical uniform quantizer with K codewords in the granular region $[0, 1]$.

largest, implying $p_0 < c_1 \leq p_1 < c_2 \leq \dots < c_K \leq p_K$ if $P_k = (p_{k-1}, p_k]$, with $p_0 = -\infty$ and $p_K = \infty$. Define the *granular* region as (c_1, c_K) and its complement $(-\infty, c_1] \cup [c_K, \infty)$ as the *overload* region.

Uniform quantization, where partition cells in the granular region have equal length, is most common in practice, but nonuniform quantization can be better for compression if the source can be modeled properly. One way of constructing a nonuniform quantizer is using the compander model, where the scalar source is transformed using a nondecreasing and smooth *compressor* function $c : \mathbb{R} \rightarrow [0, 1]$, then quantized using a uniform quantizer comprising K levels on the granular region $[0, 1]$, and finally passed through the *expander* function c^{-1} (Fig. 2). Compressor functions are defined such that $\lim_{x \rightarrow -\infty} c(x) = 0$ and $\lim_{x \rightarrow \infty} c(x) = 1$. It is convenient to define a *point density function* as $\lambda(x) = c'(x)$. Because of the boundary conditions on c , there is a one-to-one correspondence between λ and c ; hence, a companding quantizer can be uniquely specified using a point density function and codebook size, and is denoted $Q_{K,\lambda}$ in this work. The conversion of point density functions to finite-codeword quantizers is described in more detail in [5, Section II-B].

C. High-resolution Theory

It is generally difficult to determine the distortion of a scalar quantizer for any codebook size K . However, the performance of $Q_{K,\lambda}$ can be precisely analyzed as the number of codewords K becomes large, which is the basis of high-resolution theory. Assume a source X is a continuous random variable, and define the mean squared error (MSE) distortion as

$$D_{\text{mse}}(K, \lambda) = \mathbb{E}[|X - Q_{K,\lambda}(X)|^2], \quad (1)$$

where the expectation is with respect to the source density f_X . Under the additional assumption that the tails of f_X decay sufficiently fast,

$$D_{\text{mse}}(K, \lambda) \simeq \frac{1}{12K^2} \mathbb{E}[\lambda^{-2}(X)], \quad (2)$$

where \simeq indicates that the ratio of the two expressions approaches 1 as K increases [25], [26]. Hence, the MSE performance of a scalar quantizer can be approximated by a simple relationship between the source distribution, point density, and codebook size, and this relation becomes more precise with increasing K . In fact, companding quantizers are *asymptotically optimal*, meaning that the quantizer optimized over λ has distortion that approaches the performance of the best Q_K found by any means [27]–[29]. Experimentally, the high-resolution approximation is accurate even for moderate K [23], [30].

When the quantized values are to be communicated or stored, it is natural to map each codeword to a string of bits and consider the trade-off between performance and communication rate R , defined to be the expected number of bits per sample. In the simplest case, the codewords are indexed with equal-length labels and the communication rate is $R = \log_2(K)$; this is called *fixed-rate* or *codebook-constrained* quantization. Since the distortion's dependence on the shape of the quantizer λ is explicit in the asymptote, calculus techniques can be used to optimize companders. For fixed rate, Hölder's inequality can show the optimal point density satisfies

$$\lambda_{\text{mse,fr}}(x) \propto f_X^{1/3}(x), \quad (3)$$

and the resulting distortion is

$$D_{\text{mse,fr}}^*(R) \simeq \frac{1}{12} \|f_X\|_{1/3} 2^{-2R}, \quad (4)$$

with the notation $\|f\|_p = (\int_{-\infty}^{\infty} f^p(x) dx)^{1/p}$ [31]. The limit conditions on $c(x)$ imply the integral of $\lambda(x)$ is unity. Thus, (3) specifies the point density uniquely; for clarity, we omit the normalization when presenting point density results.

In general, the codeword indices can be coded to produce bit strings of different lengths based on probabilities of occurrence; this is referred to as *variable-rate* quantization. If the decoding latency is allowed to be large, one can employ block entropy coding and the communication rate approaches $H(Q_{K,\lambda}(X))$. This particular scenario, called *entropy-constrained* quantization, can be analyzed using Jensen's inequality to show the optimal point density $\lambda_{\text{mse,ec}}^*$ is constant on the support of the input distribution [31]. The optimal quantizer is thus uniform, and the resulting distortion is

$$D_{\text{mse,ec}}^*(R) \simeq \frac{1}{12} 2^{-2(R-h(X))}. \quad (5)$$

Note that block entropy coding suggests that the sources are transmitted in blocks even though the quantization is scalar. As such, (5) is an asymptotic result and serves as a lower bound on practical entropy coders with finite blocklengths that match the latency restrictions of a network.

D. Distributed Functional Scalar Quantization

When the goal of acquisition is to approximate some computation applied to the sources, optimizing the compression to the source distribution can be suboptimal and potentially worse than uniform quantization. This is most evident in distributed networks since each sensor cannot determine the overall computation at the encoder. The distributed functional scalar quantization (DFSQ) framework accounts for the computational task at the fusion center, and the resulting quantizers can be substantially better than naive designs [4], [5]. In this setting, the distortion criterion is functional MSE (fMSE):

$$D_{\text{fmse}}(K_1^N, \lambda_1^N) = \mathbb{E} \left[|g(X_1^N) - \hat{g}(Q_{K_1^N, \lambda_1^N}(X_1^N))|^2 \right], \quad (6)$$

where g is a scalar function of interest, \hat{g} is the decoding function and $Q_{K_1^N, \lambda_1^N}$ is scalar quantization performed on a vector such that

$$Q_{K_1^N, \lambda_1^N}(x_1^N) = (Q_{K_1, \lambda_1}(x_1), \dots, Q_{K_N, \lambda_N}(x_N)).$$

Before understanding how a quantizer changes fMSE, it is convenient to define how a computation locally affects distortion.

Definition 1 ([4]). The n th *functional sensitivity profile* of a multivariate function g is defined as

$$\gamma_n(x) = \left(\mathbb{E} [|g_n(X_1^N)|^2 \mid X_n = x] \right)^{1/2}, \quad (7)$$

where $g_n(x)$ is the partial derivative of g with respect to its n th argument evaluated at the point x .

Given the sensitivity profile, the main result of DFSQ [4] says the distortion of a set of N companding quantizers has the asymptotic form

$$D_{\text{fmse}}(K_1^N, \lambda_1^N) \simeq \sum_{n=1}^N \frac{1}{12K_n^2} \mathbb{E} \left[\left(\frac{\gamma_n(X_n)}{\lambda_n(X_n)} \right)^2 \right], \quad (8)$$

with conditional expectation decoder

$$\hat{g}(x_1^N) = \mathbb{E} \left[g(X_1^N) \mid Q_{K_1^N, \lambda_1^N}(X_1^N) = Q_{K_1^N, \lambda_1^N}(x_1^N) \right], \quad (9)$$

provided the following conditions are satisfied:

MF1. The function g is Lipschitz continuous and twice differentiable in every argument except possibly on a set of Jordan measure 0.

MF2. The source pdf $f_{X_1^N}$ is continuous, bounded, and supported on $[0, 1]^N$.

MF3. The function g and set of point densities λ_1^N allow $\mathbb{E}[(\gamma_n(X_n)/\lambda_n(X_n))^2]$ to be defined and finite for all n . Similar conditions are given in [5] for infinite-support distributions and a simpler decoder.

Following the same recipes to optimize over λ_1^N as in the MSE setting, the relationship between distortion and communication rate is found. In both cases, the sensitivity acts to shift quantization points to where they can reduce the distortion in the computation. For fixed-rate quantization, the asymptotic minimum distortion is

$$D_{\text{fmse,fr}}^*(R_1^N) \simeq \sum_{n=1}^N \frac{1}{12} \|\gamma_n f_{X_n}\|_{1/3} 2^{-2R_n}, \quad (10)$$

where f_{X_n} is the marginal distribution of X_n and each optimal point density satisfies

$$\lambda_{n,\text{fmse,fr}}^*(x) \propto (\gamma_n(x) f_{X_n}(x))^{1/3}. \quad (11)$$

Meanwhile, for entropy-constrained quantization, the asymptotic minimum distortion is

$$D_{\text{fmse,ec}}^*(R_1^N) \simeq \sum_{n=1}^N \frac{1}{12} 2^{2h(X_n) + 2\mathbb{E}[\log_2 \gamma(X_n)]} 2^{-2R_n}, \quad (12)$$

which results from point densities satisfying

$$\lambda_{n,\text{fmse,ec}}^*(x) \propto \gamma_n(x). \quad (13)$$

E. Don't-care intervals

When the computation induces the sensitivity to be 0 on some subintervals of the support, the high-resolution assumptions are violated and the asymptotic distortion performance may not be described by (8). This issue is addressed by carefully coding when the source is in such a “don't-care” interval [4, Section VII] and then applying traditional high-resolution theory to the remaining support. This consideration is particularly relevant because chatting among sensors can often induce the conditional sensitivity to be 0, and proper coding can lead to greatly improved performance.

Consider L_n don't-care intervals in γ_n and let A_n be the event that the source realization is *not* in the unions of them. In the fixed-rate setting, one codeword is allocated to each don't-care interval, and the remaining $K_n - L_n$ codewords are used to form reconstruction points in the nonzero intervals. There is a small degradation in performance from the loss corresponding to L_n , but this quickly becomes negligible as K_n increases. In the entropy-constrained case, the additional flexibility in coding allows for the encoder to split its message and reduce cost. The first part is an indicator variable revealing whether the source is in a don't-care interval and can be coded at rate $I_A \equiv H_B(P(A_n))$, where H_B is the binary entropy function. The actual reconstruction message is only sent if event A_n occurs, and its rate is amplified to $(R_n - I_A)/P(A_n)$ to meet the average rate constraint. The multiplicative factor $1/P(A_n)$ is called the *rate amplification*.

F. Chatting

In [4, Section VIII], chatting is introduced in the setting where one sensor sends exactly one bit to another sensor. Under fixed-rate quantization, this collaboration can at most decrease the distortion by a factor of 4 using a property of $\mathcal{L}_{1/3}$ quasi-norms. Because utilizing that bit to send additional information to the fusion center would decrease distortion by exactly a factor of 4, this is considered a negative result. Here, there is an implicit assumption that links have equal cost per bit and the network wishes to optimize a total cost budget. In the entropy-constrained setting, chatting may be useful even when links have equal costs. One example was given to demonstrate a single bit of chatting can decrease the distortion by an unbounded amount; more generally, the benefit of chatting varies depending on the source joint distribution and decoder computation.

In previous work, there is no systematic theory on performance and quantizer design of chatting. Moreover, collaboration in larger networks was still an open problem. In this paper, we extend previous results and provide a more complete discussion on how a chatting channel affects a distributed quantization network. A sample result is that chatting can be beneficial in the fixed-rate setting if the cost of communicating a bit to another sensor is lower than the cost of communicating a bit to the fusion center.

III. PERFORMANCE AND DESIGN OF CHATTING NETWORKS

We model the chatting channel in Fig. 1 as a directed graph $\mathcal{G}^c = (\mathcal{V}, \mathcal{E})$, where the set of nodes \mathcal{V} is the set of all sensors and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ is the set of noiseless, directed chatting links. If $(i, n) \in \mathcal{E}$, then for each source realization, Sensor i sends to Sensor n a chatting message $M_{i \rightarrow n}$ with codebook size $K_{i \rightarrow n}$. The parent and children sets of a sensor $n \in \mathcal{V}$ are denoted $\mathcal{N}_p(n)$ and $\mathcal{N}_c(n)$ respectively; when $(i, n) \in \mathcal{E}$, i is a parent of n and n is a child of i . The set of all chatting messages is $M^c = \{M_{i \rightarrow n}\}_{(i,n) \in \mathcal{E}}$ and the set of corresponding codebook sizes is $K^c = \{K_{i \rightarrow n}\}_{(i,n) \in \mathcal{E}}$. The chatting messages are communicated according to a schedule that the sensors and the fusion center know in advance; the set of chatting messages M^c can therefore also be thought of as a sequence. We assume chatting occurs quickly in that all communication is completed before the next discrete time instant (at which point new realizations of X_1^N are measured). After chatting is complete, Sensor n compresses its observation X_n into a message M_n using a codebook dependent on the information gathered from chatting messages, which is noiselessly communicated to the fusion center with a message $M_n(M^c)$ with codebook size $K_n(M^c)$.

We now present fMSE performance of $Q_{K_1^N, \lambda_1^N}$ in the fixed-rate and entropy-constrained settings, and we show how to optimize λ_1^N given K_1^N and K^c . We first analyze the network assuming the fusion center can successfully infer the codebook used by each sensor and hence recover the quantized values from messages M_1^N . Later in Section III-D, we provide conditions on the chatting graph \mathcal{G}^c and set of chatting messages M^c such that the fusion center is successful with zero error, having benefited from already understanding the quantizer design.

Before studying fMSE, we need to extend the definition of functional sensitivity.

Definition 2. Let $\mathcal{N}_p(n) \subseteq \mathcal{V}$ be the set of parents of Sensor n in the graph \mathcal{G}^c induced by chatting. The n th conditional sensitivity profile of computation g given all chatting messages M^c is

$$\gamma_{n|M^c}(x|m) = \left(\mathbb{E} [|g_n(X_1^N)|^2 \mid X_n = x, M_{i \rightarrow n} = m_{i \rightarrow n} \text{ for all } i \in \mathcal{N}_p(n)] \right)^{1/2}. \quad (14)$$

Notice only messages from parent sensors are relevant to $\gamma_{n|M^c}$. Intuitively, chatting messages reveal information about the parent sensors' quantized values and reshape the sensitivity appropriately. Depending on the encoding of chatting messages, this may induce don't-care intervals in the conditional sensitivity (where $\gamma_{n|M^c} = 0$).

The distortion dependence on the number of codeword points and the conditional sensitivity profiles is given in the following theorem:

Theorem 1. Given the source distribution $f_{X_1^N}$, computation g , and point densities $\lambda_1^N(M^c)$ satisfying conditions MF1–3 for every possible realization of M^c , the asymptotic distortion of the conditional expectation decoder (9) given codeword allocation K_1^N and K^c is

$$D_{\text{fmse}}(K_1^N, K^c, \lambda_1^N) \simeq \mathbb{E}_{M^c} \left[\sum_{n=1}^N \mathbb{E}_{X_n|M^c} \left[\frac{1}{12K_n^2(m)} \frac{\gamma_{n|M^c}^2(X_n|m)}{\lambda_{n|M^c}^2(X_n|m)} \mid M^c = m \right] \right]. \quad (15)$$

Proof: Extend the proof of [4, Theorem 17] using the Law of Total Expectation. \blacksquare

Compared to the DFSQ result, the performance of a chatting network can be substantially more difficult to compute since the conditional sensitivity may be different with each realization of M^c and affects the choice of the point density and codebook size. However, Sensor n 's dependence on M^c is through a subset of messages from its parent nodes. In Section V, we will see how structured architectures lead to tractable computations of fMSE. Following the techniques in [5], the theorem can be expanded to account for infinite-support distributions and a simpler decoder. Some effort is necessary to justify the use of normalized point densities in the infinite-support case, especially in the entropy-constrained setting, but high-resolution theory applies in this case as well.

A. Don't-Care Intervals

We have already alluded to the fact that chatting can induce don't-care intervals in the conditional sensitivity profiles of certain sensors. In this case, we must properly code for these intervals to ensure the high-resolution assumptions hold, as discussed in Section II-D.

For fixed-rate coding where $R_n = \log_2(K_n)$, this means shifting one codeword to the interior of each don't-care interval and applying standard high-resolution analysis over the union of all intervals where $\gamma_n(x) > 0$. The resulting distortion of a chatting network is then given as:

Corollary 1. Assume the source distribution $f_{X_1^N}$, computation g , and point densities $\lambda_1^N(M^c)$ satisfying conditions MF1–3 for every possible realization of M^c , with the additional requirement that $\lambda_n(x|m) = 0$ whenever $\gamma_{n|M^c}(x|m) = 0$. Let $L_n(m)$ be the number of don't-care intervals in the conditional sensitivity of Sensor n when $M^c = m$. The asymptotic distortion of such a chatting network where communication links utilize fixed-rate coding is

$$D_{\text{fmse}}(R_1^N, K^c, \lambda_1^N) \simeq \mathbb{E}_{M^c} \left[\sum_{n=1}^N \mathbb{E}_{X_n|M^c} \left[\frac{1}{12(2^{R_n} - L_n(m))} \frac{\gamma_{n|M^c}^2(X_n|m)}{\lambda_{n|M^c}^2(X_n|m)} \mid M^c = m \right] \right]. \quad (16)$$

In the entropy-constrained setting where $R_n = H(\hat{X}_n)$, we must code first the event $A_n(m)$ that the source is not in a don't-care interval given the chatting messages, and then coding the source realization only if A_n occurs. The resulting distortion of a chatting network is:

Corollary 2. Assume the source distribution $f_{X_1^N}$, computation g , and point densities $\lambda_1^N(M^c)$ satisfying conditions MF1–3 for every possible realization of M^c , with the additional requirement that $\lambda_n(x|m) = 0$ whenever $\gamma_{n|M^c}(x|m) = 0$. Let $A_n(m)$ be the event that X_n is not in a don't-care interval given $M^c = m$. The asymptotic distortion of such a chatting network where communication links utilize entropy coding is

$$D_{\text{fmse}}(R_1^N, K^c, \lambda_1^N) \simeq \mathbb{E}_{M^c} \left[\sum_{n=1}^N \mathbb{E}_{X_n|M^c} \left[\frac{P(A_n(m))}{12} 2^{2h(X_n|A_n(m)) + 2\mathbb{E}[\log_2 \lambda_n(X_n)|A_n(m)]} \cdot \frac{\gamma_{n|M^c}^2(X_n|m)}{\lambda_{n|M^c}^2(X_n|m)} 2^{-2(R_n(m) - H_B(A_n(m)))/P(A_n(m))} \mid M^c = m \right] \right].$$

We will use both corollaries in optimizing the design of $\lambda_1^N(M^c)$ in the remainder of the paper.

B. Fixed-rate Quantization Design

We mirror the method used to determine (11) in the DFSQ setup but now allow the sensor to choose from a set of codebooks depending on the incoming messages from parent sensors. The mapping between chatting messages and codebooks is known to the decoder of the fusion center, and each codebook corresponds to the optimal quantizer for a given conditional sensitivity induced by the incoming message. Let $Z_n(M^c)$ be the union of the don't-care intervals of a particular conditional sensitivity. Then using Corollary 1, the optimal point density for fixed-rate quantization satisfies

$$\lambda_{n,\text{fmse},\text{fr},\text{chat}}^*(x|m) \propto \begin{cases} (\gamma_{n|M^c}(x|m) f_{X_n|M^c}(x|m))^{1/3}, & x \notin Z_n(m) \text{ and } f_{X_n|M^c}(x|m) > 0; \\ 0, & \text{otherwise.} \end{cases} \quad (17)$$

Recall that the point density is the derivative of the compressor function $c(x)$ in the compander model. Hence, codewords are placed at the solutions to $c(x) = (k-1)/(K-L)$ for $k = 1, \dots, (K-L)$. In addition, one codeword must be placed in each of the L don't-care interval.

C. Entropy-constrained Quantization Design

Using Corollary 2, the optimal point density when entropy coding is combined with scalar quantization has the form

$$\lambda_{n,\text{fmse,ec,chat}}^*(x|m) \propto \begin{cases} \gamma_{n|M^c}(x|m), & x \notin Z_n(m) \text{ and } f_{X_n|M^c}(x|m) > 0; \\ 0, & \text{otherwise.} \end{cases} \quad (18)$$

Note that rate amplification can arise through chatting, and this can allow distortion terms to decay at rates faster than 2^{-2R_n} . However, there is also a penalty from proper coding of don't-care intervals, corresponding to $H_B(P(A_n))$. This loss is negligible in the high-resolution regime but may become important for moderate rates.

D. Conditions on Chatting Graph

We have observed that chatting can influence optimal design of scalar quantizers through the conditional sensitivity, and that sensors will vary their quantization codebooks depending on the incoming messages from parent sensors. Under the assumption that the fusion center does not have access to M^c , success of compression is contingent on the fusion center identifying the codebook employed by every sensor from the messages M_1^N .

Definition 3. A chatting network is *codebook identifiable* if the fusion center can determine the codebooks of $Q_{K_1^N, \lambda_1^N}$ using the messages it receives from each sensor. That is, it can determine $\mathcal{C}_n(M^c)$ from M_1^N for each time instant.

We have argued that a chatting network can successfully communicate its compressed observations if it is codebook identifiable. The following are sufficient conditions on the chatting graph \mathcal{G}^c and messages M^c such that the network is codebook identifiable:

- C1. The chatting graph \mathcal{G}^c is a directed acyclic graph.
- C2. The causality in the chatting schedule matches \mathcal{G}^c , meaning for every n , Sensor n sends its chatting message after it receives messages from all parent sensors.
- C3. The quantizer at Sensor n is a function of the source joint distribution and all incoming messages from parent sensors in $\mathcal{N}_p(n)$.
- C4. At any discrete time, the message transmitted by Sensor n is a function of M_n and incoming messages from parent sensors in $\mathcal{N}_p(n)$.

When each sensor's quantizer is regular and encoder only operates on the quantized values \hat{X}_n , matching the DFSQ setup, the chatting message can only influence the choice of codebook. In this setting, the above conditions become necessary as well. Alternatively, if sensors can locally fuse messages from parents with their own observation, there may exist other conditions for a network to be codebook identifiable.

IV. RATE ALLOCATION IN CHATTING NETWORKS

A consequence of chatting is that certain sensors can exploit their neighbors' acquisitions to refine their own. Moreover, a sensor can potentially utilize this side information to adjust its communication rate in addition to changing its quantization if the network is codebook identifiable. These features of chatting networks suggest intelligent rate allocation across sensors can yield significant performance gains. In addition, a strong motivation for intersensor interaction is that sensors may be geographically closer to each other than a fusion center and hence require less transmit power, or can utilize low-bandwidth orthogonal channels that do not interfere with the main communication network. As a result, the cost of communicating a bit may vary in a network.

This section explores proper rate allocation to minimize the total cost of transmission in a chatting network, allowing asymmetry of the information content at each sensor and heterogeneity of the communication links. Consider the distributed network in Fig. 1. The cost per bit of the communication link and the resource allocation between Sensor n and the fusion center are denoted by α_n and b_n respectively, leading to a communication rate of $R_n = b_n/\alpha_n$ from Sensor n to the fusion center. Similarly, for a chatting link between Sensors i and n , the cost per bit and resource allocation are denoted by $\alpha_{i \rightarrow n}$ and $b_{i \rightarrow n}$ respectively, corresponding to a chatting rate of $R_{i \rightarrow n} = b_{i \rightarrow n}/\alpha_{i \rightarrow n}$. Consistent with previous notation, we denote the set of costs per chatting bit, resource allocations on chatting links, and chatting rates by $\alpha^c = \{\alpha_{i \rightarrow n}\}_{(i,n) \in \mathcal{E}}$, $b^c = \{b_{i \rightarrow n}\}_{(i,n) \in \mathcal{E}}$, and $R^c = \{R_{i \rightarrow n}\}_{(i,n) \in \mathcal{E}}$.

Given a total resource budget C , how should the rates be allocated among these links? For simplicity, assume all chatting links employ fixed-rate quantization; this implies that $K_n = 2^{R_n}$ for all $n \in \{1, 2, \dots, N\}$ and $K_{i \rightarrow n} = 2^{R_{i \rightarrow n}}$ for all $(i, n) \in \mathcal{E}$. The distortion-cost trade-off is then expressed as

$$D(C) = \inf_{\substack{b_1^N, b^c, \lambda_1^N: \\ \sum_{n=1}^N b_n + \sum_{(i,n) \in \mathcal{E}} b_{i \rightarrow n} = C}} D_{\text{fmse}}(K_1^N, K^c, \lambda_1^N). \quad (19)$$

In general, this optimization is extremely difficult to describe analytically since the distortion contribution of each sensor is dependent in a nontrivial way on the conditional sensitivity, which in turn is dependent on the design of the chatting messages.

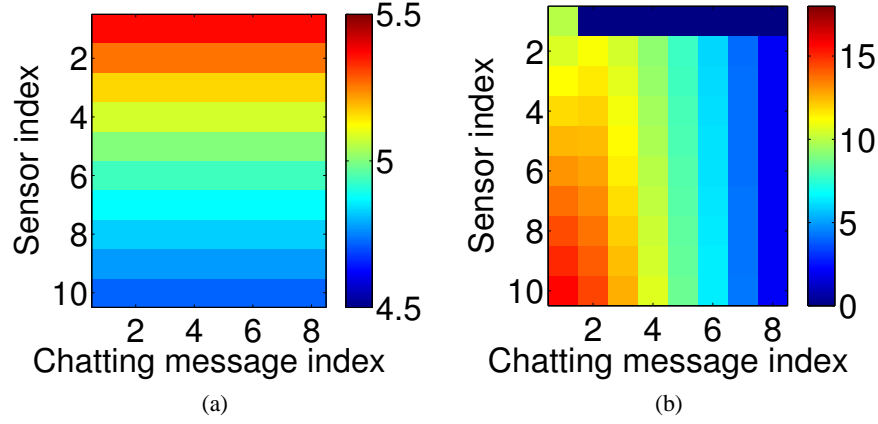


Fig. 3. Cost allocation for a maximum computation network, as described in Section V. In this case, $N = 10$, $C = 5N$, $R_c = 3$, $\alpha_c = 0$, and $\alpha_n = 1$. In the fixed-rate setting (a), the sensors are allowed to have different communication rates but cannot adjust the rate with the received chatting message. In the entropy-constrained setting (b), each sensor except sensor 1 receives chatting messages and can adjust its communication rate appropriately.

However, the relationship between b_1^N and the overall system distortion is much simpler, as described in Theorem 1. Hence, once the chatting allocations b^c is fixed, the optimal b_1^N is easily determined using extensions of traditional rate allocation techniques described in Appendix A. In particular, the optimal b_1^N can be found by applying Lemmas 3 and 4 with a total cost constraint

$$C' = C - \sum_{(i,n) \in \mathcal{E}} b_{i \rightarrow n}. \quad (20)$$

A brute-force search over b^c then provides the best allocation, but this procedure is computationally expensive. More realistically, network constraints may limit the maximum chatting rate, which greatly reduces the search space.

In Fig. 3, we show optimal communication rates for the network described in Section V. We delay description of the specific network properties and aim only to illustrate how the cost allocations $b_n(m)$ may change depending with sensors or chatting messages. Under fixed-rate coding, b_n varies depending on the chatting graph. In the entropy-constrained setting, the allocation can also vary with the chatting messages, except for Sensor 1. This increased flexibility allows for a wider range of rates, as well as improved performance in many situations.

V. MAXIMUM COMPUTATION

The results in the previous sections hold generally, and we now build some intuition about chatting using a specific distributed network performing a maximum computation. The choice of this computation is not arbitrary; we will show that it allows for a particular chatting architecture that makes it convenient to study large networks. Moreover, this network reveals some surprising insights into the behavior of chatting. While this paper restricts its attention solely to the maximum computation, more examples are discussed in [8].

A. Problem Model

We consider a network where the fusion center aims to reproduce the maximum of N sources, where each X_n is independent and uniformly distributed on $[0, 1]$. The sensors measuring these sources are allowed to chat in a serial chain, meaning each sensor has at most one parent and one child (see Fig. 4). Initially, we will consider the simplest such network with the following assumptions:

- 1) The chatting is serial, meaning the sequence of chatting messages is $\{M_{(n-1) \rightarrow n}\}_{n=2}^N$.
- 2) Each chatting link is identical and has rate R_c , codebook size $K_c = 2^{R_c}$ and cost α_c .
- 3) The communication links between sensors and the fusion center are allowed to have different rates. For simplicity, we assume them to be homogeneous and normalize the cost to be $\alpha_n = 1$.
- 4) The outgoing chatting message at Sensor 1 is the index of a uniformly quantized version of its observation with K_c levels.
- 5) For $n > 1$, the chatting message from Sensor n is the maximum of the index of Sensor n 's own uniformly quantized observation and the chatting message from its parent.

Under this architecture, the chatting messages effectively correspond to a uniformly quantized observation of the maximum of all ancestor nodes:

$$M_{(n-1) \rightarrow n} = \mathcal{I}(Q_{K_c, U}(\max(X_1^{n-1}))), \quad (21)$$

where \mathcal{I} is the index of the quantization codeword and can takes values $\{1, \dots, K_c\}$. The simplicity of the chatting message here arises from the permutation-invariance of the maximum function. We will exploit this structure to provide precise characterizations of system performance.

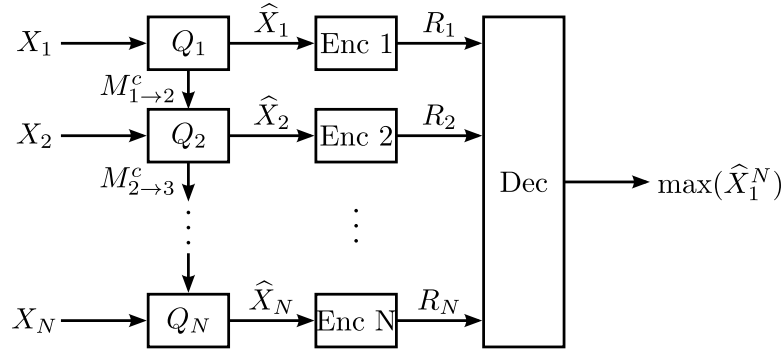


Fig. 4. A fusion center wishes to determine the maximum of N iid uniform sources and receives messages M_n from each sensor n at rate R_n . The sensors are allowed to chat serially down the network using messages $M_{(n-1) \rightarrow n}$ at rate R_c .

B. Quantizer Design

Using (7), we find the max function has sensitivity $\gamma_n^2(x) = x^{N-1}$ for all n . Without chatting, each sensor's quantizer would be the same with a point density that is a function of the source distribution and sensitivity. Moreover, since the cost per bit of transmitting to the fusion center is the same, the solution of the resource allocation problem assigns equal weight to each link. Hence, minimizing (10) yields the optimal fixed-rate distortion-cost trade-off:

$$D_{\max, \text{fr}}(C) \simeq \frac{N}{12} \left(\frac{3}{N+2} \right)^3 2^{C/N}. \quad (22)$$

Similarly, the minimum of (12) leads to the optimal entropy-constrained distortion-cost trade-off

$$D_{\max, \text{ec}}(C) \simeq \frac{N}{12} e^{-N+1} 2^{C/N}. \quad (23)$$

These high-resolution expressions provide scaling laws on how the distortion relates to the number of sensors. They require the total cost C increase linearly with N to hold.

With chatting, we first need to determine the conditional sensitivity, which is given below for uniform sources:

Lemma 1. Given $K_c = 2^{R_c}$, the sensitivity profile corresponding to a received chatting message $M_{(n-1) \rightarrow n} = k$ is

$$\gamma_{n|M_{(n-1) \rightarrow n}}^2(x|k) = \begin{cases} 0, & x < \frac{k-1}{K_c}; \\ \frac{(K_c x)^{n-1} - (k-1)^{n-1}}{k^{n-1} - (k-1)^{n-1}} x^{N-n}, & \frac{k-1}{K_c} \leq x < \frac{k}{K_c}; \\ x^{N-n}, & x \geq \frac{k}{K_c}. \end{cases} \quad (24)$$

Proof: See Appendix B. ■

We have already noted the incident chatting message of Sensor n is a uniformly quantized observation of $Y_n = \max(X_1^{n-1})$, where $f_Y(y) = (n-1)y^{n-2}$. Hence,

$$\mathbb{P}(M_{(n-1) \rightarrow n} = k) = \left(\frac{k}{K_c} \right)^{n-1} - \left(\frac{k-1}{K_c} \right)^{n-1}. \quad (25)$$

Below, we give distortion asymptotics for the serial chatting network under both fixed-rate and entropy-constrained quantization.

1) *Fixed-rate case:* From Theorem 1, the asymptotic total fMSE distortion is

$$\sum_{n=1}^N \beta_n 2^{-2R_n}, \quad (26)$$

where $\beta_n = \frac{1}{12} \|\gamma_{n|M_c}^2\|_{1/3}$. Because Sensor 1 has no incoming chatting messages, its sensitivity is $\gamma_1^2(x) = x^{N-1}$ and the resulting distortion constant is

$$\beta_1 = \frac{1}{12} \left(\frac{3}{N+2} \right)^3.$$

For other sensors, the distortion contribution is

$$\beta_n = \frac{1}{12} \sum_{k=1}^{K_c} \mathbb{P}(M_{(n-1) \rightarrow n} = k) \|\gamma_{n|M_{(n-1) \rightarrow n}=k}^2\|_{1/3}.$$

For Sensor n with $n > 1$, all incoming messages besides $k = 1$ induce a don't-care interval, so one of the 2^{R_n} codewords is placed exactly at $(k-1)/K$.

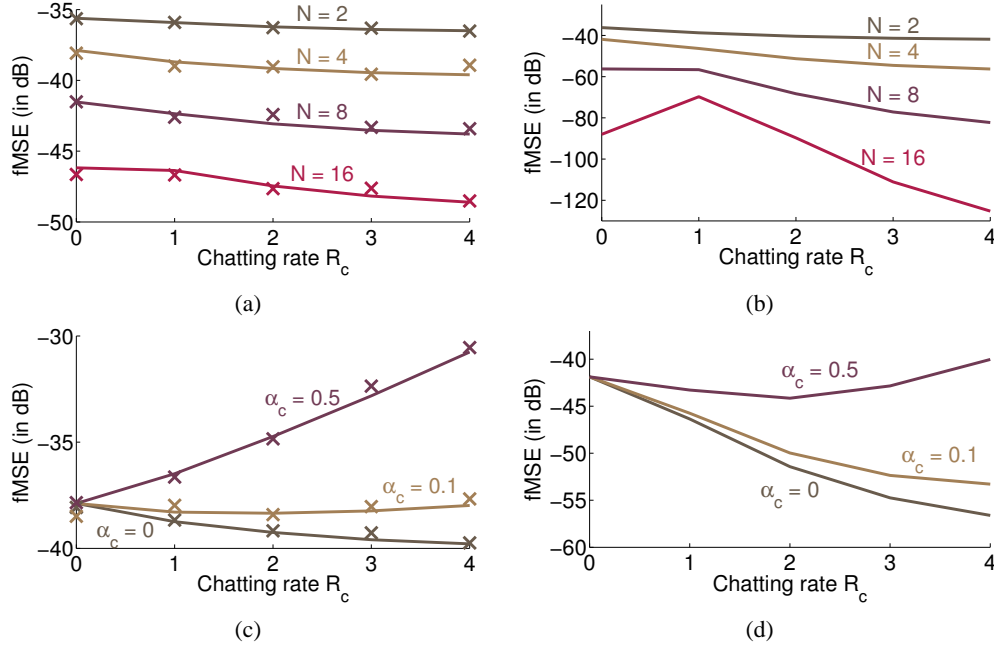


Fig. 5. Performance of the maximum computation network in both the fixed-rate (left plots) and entropy-constrained (right plots) settings. Plots (a) and (b) illustrate the trade-off between fMSE and chatting rate for choices of N assuming total cost $C = 4N$ and $\alpha_c = 0.01$. Plots (c) and (d) illustrate the trade-off between fMSE and chatting rate for choices of α_c assuming $N = 4$ sensors and total cost $C = 4N$. In all cases, the cost of communication is $\alpha_n = 1$. For the fixed-rate setting, we validate the distortion through simulated runs on real quantizers designed using (17). We observe that high-resolution theory predicts actual performance at rates as low as 4 bits/sample, as shown by crosses in the fixed-rate plots.

We study the trade-off between chatting rate R_c and fMSE for several choices of N and α_c using optimal cost allocation as determined by Lemma 3. In Fig. 5a, we observe that increasing the chatting rate yields improvements in fMSE. As the number of sensors increases, this improvement becomes more pronounced. However, this is contingent on the chatting cost α_c being low. As discussed in Section II-D, chatting can lead to worse system performance if the cost of chatting is on the same order as the cost of communication given a total resource budget, as demonstrated by Fig. 5c. Although the main results of this work are asymptotic, we have asserted the distortion equations are reasonable at finite rates. To demonstrate this, we design real quantizers under the same cost constraint and demonstrate that the resulting performance is comparable to high-resolution approximations of Theorem 1. This is observed in Figs. 5a and c, which shows the asymptotic prediction of the distortion-rate trade-off is accurate even at 4 bits/sample.

2) *Entropy-constrained case:* Generally, the total distortion in the entropy-constrained case is

$$\sum_{n=1}^N \mathbb{E} [\beta_{n,k} 2^{-2R_{n,k}} \mid M_{(n-1) \rightarrow n} = k], \quad (27)$$

noting each sensor is allowed to vary its communication rate with the chatting messages it receives. Like in the fixed-rate setting, an incoming message k will induce a don't-care interval of $[0, (k-1)/K]$ in the conditional sensitivity. If $A_{n,k}$ is the event that X_n is not in a don't-care interval when receiving message k , then

$$\beta_{n,k} = \frac{1}{12} \mathbb{P}(M_{(n-1) \rightarrow n} = k) 2^{2h(X_n | A_{n,k}) + 2\mathbb{E}[\log_2 \gamma_n \mid M_{(n-1) \rightarrow n}(X_n | k)]} \quad (28)$$

and $R_{n,k} = (R_n - H_B(\mathbb{P}(A_{n,k}))) / \mathbb{P}(A_{n,k})$.

Like in the fixed-rate setting, we study the relationship between the chatting rate R_c and fMSE, this time using the probabilistic allocation optimization of Lemma 4 in Appendix A. Due to the extra flexibility of allowing a sensor to vary its communication to the fusion center with the chatting messages it receives, we observe that increasing the chatting rate can improve performance more dramatically than in the fixed-rate case (see Fig. 5b). Surprisingly, chatting can also lead to inferior performance for some combinations of R_c and N , even when α_c is small. This phenomenon will be discussed in greater detail below. In Fig. 5d, we compare different choices of α_c to see how performance changes with the chatting rate. Unlike for fixed rate, in the entropy-constrained setting, chatting can be useful even when its cost is close to the cost of communication to the fusion center.

C. Generalizing the Chatting Messages

We have considered the case where a chatting message is the uniform quantization of the maximum of all ancestor nodes, as shown in (21). Although simple, this coding of chatting messages is not optimal. Here, we generalize chatting messages to

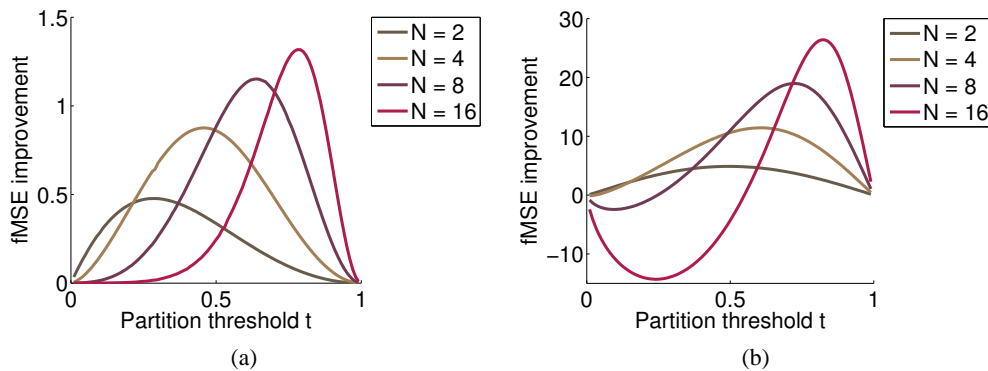


Fig. 6. Distortion improvement compared to no chatting in the maximum computation network for the fixed-rate (left plot) and entropy-constrained (right plot) settings when varying the partition boundary p_1 . We assume chatting is free, i.e., $\alpha_c = 0$, but the chatting rate is limited to one bit.

understand how the performance can change with this design choice.

We begin by considering the same network under the restriction that the chatting rate is $R_c = 1$, but allow the single partition boundary p_1 to vary rather than setting it to $1/2$. Currently, we keep the coding consistent for every sensor such that a chatting message $k = 1$ implies $\max(X_1^{n-1}) \in [0, p_1]$ and $k = 2$ means $\max(X_1^{n-1}) \in (p_1, 1]$. Distortions for a range of N and p_1 are shown in Fig. 6.

From these performance results, we see that the choice of p_1 should increase with the size of the network, but precise characterization of the best p_1 is difficult because of the complicated effect the conditional sensitivity has on both the distortion constants and rate allocation. We can recover some of the results of Fig. 5 by considering $p_1 = 1/2$. It is now evident that this choice of p_1 can be very suboptimal, especially as N becomes large. In fact, we observe that for certain choices of the partition with entropy coding, the distortion with chatting can be larger than from a traditional distributed network even though the chatting cost is 0. This unintuitive fact arises because the system's reliance on the conditional sensitivity is fixed, and the benefits of a don't-care interval are mitigated by creating a more unfavorable conditional sensitivity. We emphasize that this phenomenon disappears as the rate becomes very large.

Since the flexibility in the choice of the chatting encoder's partitions can lead to improved performance when $R_c = 1$, we can expect even more gains when the chatting rate is increased. However, the only method for optimizing the choice of partition boundaries developed currently involve brute-force search using the conditional sensitivity derived in Appendix B. Another extension that leads to improved performance is to allow chatting encoders to employ different partitions. This more general framework yields strictly improved results, but some of the special structure of the serial chatting network is lost as the chatting message is no longer necessarily the maximum of all ancestor sensors. The added complexity of either of these extensions make their performances difficult to quantify.

D. Optimizing a Chatting Network

In this paper, we have formulated a framework allowing low-rate collaboration between sensors in a distributed network. We have introduced several methods to optimize such a network, including nonuniform quantization, rate allocation, and design of chatting messages. Here, we combine these ingredients and see how each one impacts fMSE.

We will continue working with the maximum computation network from Fig. 4 assuming $R_c = 1$, $\alpha_c = 0$, $N = 5$ and $C = 5N$. We further assume the coding of chatting messages is the same for every sensor on the serial chain. We will then consider the following scenarios:

- 1) A chatting network with $R_n = 5$ for all n and chatting designed by (21).
- 2) A chatting network with rate allocation and chatting designed by (21).
- 3) A chatting network with rate allocation and optimization over chatting messages.

We analyze the fMSE of each scenario compared to a distributed network without chatting ($R_c = 0$). From Fig. 7, we can see that incorporating rate allocation and chatting optimization yields substantial gains in the entropy-constrained setting. For fixed rate, the most meaningful improvement comes from allowing chatting, while additional optimization provides little additional benefit. Up to this point, we have limited chatting to have fixed codebook size and did not allow entropy coding. Lifting these restrictions increase system complexity and can provide even greater compression gain.

VI. CONCLUSIONS

In this work, we explored how intersensor communication—termed *chatting*—can improve approximation of a function of sensed data in a distributed network constrained to scalar quantization. We have motivated chatting from two directions: providing an analysis technique for distortion performance when low-blocklength limitations make Shannon theory too optimistic,

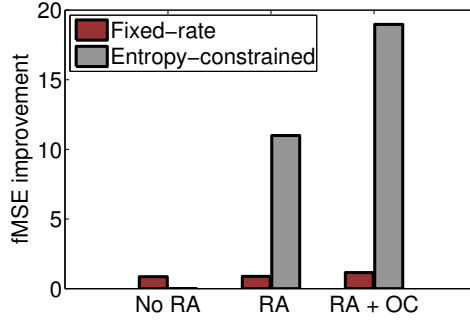


Fig. 7. Distortion improvement for Scenarios 1–3 over a distributed network without chatting. Both rate allocation (RA) and optimized chatting (OC) are considered.

and illustrating the potential gains over simplistic practical designs. There are many opportunities to leverage heterogeneous network design to aid information acquisition using the tools of high-resolution theory, and we provide precise characterizations of distortion performance, quantizer design, and cost allocation to optimize distributed networks. Many challenges remain in analyzing chatting networks. Some future directions that are meaningful include a more systematic understanding of how to design chatting messages and applications where chatting may be feasible and beneficial.

One can consider “sensors” being distributed in time rather than space, with the decoder computing a function of samples from a random process. Connections of this formulation to structured vector quantizers are of independent interest.

APPENDIX A RATE ALLOCATION FOR DISTRIBUTED NETWORKS

Consider the distributed network in Fig. 1 without the chatting channel. The cost per bit of the communication link and the cost allocation between Sensor n and the fusion center is denoted by α_n and b_n respectively, leading to a communication rate of $R_n = b_n/\alpha_n$. Below, we solve the cost allocation problem under the assumption that companding quantizers are used and noninteger rates are allowed.

Lemma 2. *The optimal solution to*

$$D(C) = \min_{\sum b_n = C, b_n \geq 0} \sum_{n=1}^N \beta_n 2^{-2b_n/\alpha_n} \quad (29)$$

has cost allocation

$$b_n^* = \max \left(0, \frac{1}{2} \log_2 \frac{\beta_n/\alpha_n}{\tilde{\beta}} \right), \quad (30)$$

where $\tilde{\beta}$ is chosen such that $\sum b_n^* = C$.

Proof: This lemma extends the result from [32] or can be derived directly from the KKT conditions. ■

Each β_n is calculated using only the functional sensitivity γ_n and marginal source pdf f_{X_n} . Although Lemma 2 is always true, we emphasize that its effectiveness in predicting the proper cost allocation in a distributed network is only rigorously shown for high cost (i.e. high rate) due to its dependence on (8). However, it can be experimentally verified that costs corresponding to moderate communication rates still yield near-optimal allocations.

When the solution of Lemma 2 is strictly positive, a closed-form expression exists:

Lemma 3. *Assuming each b_n^* in (30) is strictly positive, it can be expressed as*

$$b_n^* = \frac{\alpha_n}{\tilde{\alpha}} C + \frac{\alpha_n}{2} \log_2 \frac{\beta_n/\alpha_n}{\left(\prod_j (\beta_j/\alpha_j)^{\alpha_j} \right)^{1/\sum \alpha_i}}. \quad (31)$$

Proof: The proof uses Lagrangian optimization. ■

If Sensor n is allowed to vary the communication rate depending on the side information $M_{\text{si},n}$ it receives, further gains can be enjoyed. This situation is natural in chatting networks, where the side information is the low-rate messages passed by neighboring sensors. Here, we introduce *probabilistic cost allocation*, yielding a distortion–cost trade-off

$$D(C) = \min_{\substack{\sum \mathbb{E}[b_n(M_{\text{si},n})] = C \\ b_n(m) \geq 0}} \sum_{n=1}^N \mathbb{E} \left[\beta_n(M_{\text{si},n}) 2^{-2b_n(M_{\text{si},n})/\alpha_n} \right], \quad (32)$$

where the expectation is taken with respect to $M_{\text{si},n}$. Each link will have a cost allocation $b_n(m)$ for every possible message m while satisfying an average cost constraint. An analogous result to Lemma 2 can be derived; for the situation where the optimal allocation is strictly positive, it can again be expressed in closed form:

Lemma 4. Assume the side information $M_{\text{si},n}$ received at Sensor n is $m \in \mathcal{M}_n$ and the cost per bit of the communication link may vary with m . Assuming each allocation $b_n^*(m)$ in the solution to (32) is strictly positive, it can be expressed as

$$b_n^*(m) = \frac{\alpha_n(m)}{\tilde{\alpha}} C + \frac{\alpha_n(m)}{2} \log_2 \frac{\beta_n(m)/\alpha_n(m)}{\prod_j \prod_l ((\beta_j(l)/\alpha_j(l))^{\alpha_j(l)/\tilde{\alpha}})}, \quad (33)$$

where $\tilde{\alpha} = \sum_n \sum_m f_{M_{\text{si},n}}(m) \alpha_n(m)$.

Here, we extended previous known rate allocation results [22], [32] to account for heterogeneity in distributed networks. Although these results do not account for chatting, we see in Section IV that they become important tools in optimizing performance in such networks.

APPENDIX B SENSITIVITY OF MAXIMUM COMPUTATION NETWORK

Assuming iid uniform sources on the support $[0, 1]$, the sensitivity of each sensor in the maximum computation network in Fig. 4 without chatting is

$$\begin{aligned} \gamma_n^2(x) &= \mathbb{E}[|g_n(X_1^N)|^2 | X_n = x] \\ &= \mathbb{P}(\min(X_1^N) = X_n | X_n = x) \\ &= \mathbb{P}(X_1 < x) \cdots \mathbb{P}(X_{n-1} < x) \mathbb{P}(X_{n+1} < x) \cdots \mathbb{P}(X_N < x) \\ &= x^{N-1}. \end{aligned}$$

When the chatting graph is a serial chain, Sensor n has some lossy version of the information collected by its ancestor sensors. For the max function, chatting reduces the support of the estimate of $\max(X_1^{n-1})$ by Sensor n . Hence, the message $M_{(n-1) \rightarrow n}$ reveals the max of the ancestor sensors is in the range $[s_l, s_u]$. This side information forms three distinct intervals in the conditional sensitivity. First, in the interval $x < s_l$, X_n is assuredly less than $\max(X_1^{n-1})$ and hence sensitivity is 0 since the information at Sensor n is irrelevant at the fusion center. Second, if $x > s_u$, X_n is greater than $\max(X_1^{n-1})$ and the sensitivity should only depend on the number of descendant sensors, leading to a sensitivity of x^{N-n} . Finally, when $s_l \leq x < s_u$, Sensor n must take into consideration both ancestors and descendants, yielding sensitivity

$$\begin{aligned} &\mathbb{P}(\min(X_1^N) = X_n | X_n = x, \max(X_1^{n-1}) \in [s_l, s_u]) \\ &= \mathbb{P}(\max(X_1^{n-1}) < x | \max(X_1^{n-1}) \in [s_l, s_u]) \mathbb{P}(\max(X_{n+1}^N) < x) \\ &= \frac{x^{n-1} - s_l^{n-1}}{s_u^{n-1} - s_l^{n-1}} x^{N-n}. \end{aligned}$$

More specific to the case when messages correspond to uniform quantization, we define $K_c = 2^{R_c}$ and denote each received message $M_{(n-1) \rightarrow n}$ as k_n . Setting $s_l = (k_n - 1)/K_c$ and $s_u = k_n/K_c$ gives Lemma 1.

REFERENCES

- [1] D. Slepian and J. K. Wolf, "Noiseless coding of correlated information sources," *IEEE Trans. Inform. Theory*, vol. IT-19, pp. 471–480, July 1973.
- [2] R. Zamir, "The rate loss in the Wyner–Ziv problem," *IEEE Trans. Inform. Theory*, vol. 42, pp. 2073–2084, Nov. 1996.
- [3] V. Y. F. Tan and O. Kosut, "On the dispersions of three network information theory problems." arXiv:1201.3901v2 [cs.IT], Feb. 2012.
- [4] V. Misra, V. K. Goyal, and L. R. Varshney, "Distributed scalar quantization for computing: High-resolution analysis and extensions," *IEEE Trans. Inform. Theory*, vol. 57, pp. 5298–5325, Aug. 2011.
- [5] J. Z. Sun, V. Misra, and V. K. Goyal, "Distributed functional scalar quantization simplified." arXiv:1206.1299v1 [cs.IT], June 2012.
- [6] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. Cambridge, UK: Cambridge University Press, 2005.
- [7] T. Yucek and H. Arslan, "A survey of spectrum sensing algorithms for cognitive radio applications," *IEEE Comm. Surveys Tutorials*, vol. 11, no. 1, pp. 116–130, 2009.
- [8] J. Z. Sun and V. K. Goyal, "Chatting in distributed quantization networks," in *Proc. 50th Ann. Allerton Conf. on Commun., Control and Comp.*, (Monticello, IL), Oct. 2012.
- [9] A. Orlitsky and J. R. Roche, "Coding for computing," *IEEE Trans. Inform. Theory*, vol. 47, pp. 903–917, Mar. 2001.
- [10] H. Feng, M. Effros, and S. A. Savari, "Functional source coding for networks with receiver side information," in *Proc. 42nd Annu. Allerton Conf. Commun. Control Comput.*, pp. 1419–1427, Sept. 2004.
- [11] Y. Polyanskiy, H. V. Poor, and S. Verdú, "Channel coding rate in the finite blocklength regime," *IEEE Trans. Inform. Theory*, vol. 56, pp. 2307–2359, May 2010.
- [12] A. Ingber and Y. Kochman, "The dispersion of lossy source coding," in *Proc. IEEE Data Compression Conf.*, (Snowbird, Utah), pp. 53–62, Mar. 2011.
- [13] V. Kostina and S. Verdú, "Fixed-length lossy compression in the finite blocklength regime," *IEEE Trans. Inform. Theory*, vol. 58, pp. 3309–3338, June 2012.
- [14] A. H. Kaspi and T. Berger, "Rate–distortion for correlated sources with partially separated encoders," *IEEE Trans. Inform. Theory*, vol. IT-28, pp. 828–840, Nov. 1982.

- [15] M. Sefidgaran and A. Tchamkerten, "On cooperation in multi-terminal computation and rate distortion," in *Proc. IEEE Int. Symp. Inform. Theory*, (Cambridge, MA), pp. 771–775, July 2012.
- [16] N. Ma and P. Ishwar, "Some results on distributed source coding for interactive function computation," *IEEE Trans. Inform. Theory*, vol. 57, pp. 6180–6195, Sept. 2011.
- [17] N. Ma, P. Ishwar, and P. Gupta, "Interactive source coding for function computation in collocated networks," *IEEE Trans. Inform. Theory*, vol. 58, pp. 4289–4305, July 2012.
- [18] S. Nitinawarat and P. Narayan, "Perfect omniscience, perfect secrecy, and Steiner tree packing," *IEEE Trans. Inform. Theory*, vol. 56, pp. 6490–6500, Dec. 2010.
- [19] T. Courtade and R. Wesel, "Efficient universal recovery in broadcast networks," in *Proc. 48th Ann. Allerton Conf. on Commun., Control and Comp.*, (Monticello, IL), pp. 1542–1549, Oct. 2010.
- [20] E. Martinian, G. W. Wornell, and R. Zamir, "Source coding with distortion side information," *IEEE Trans. Inform. Theory*, vol. 54, pp. 4638–4665, Oct. 2008.
- [21] T. Linder, R. Zamir, and K. Zeger, "High-resolution source coding for non-difference distortion measures: Multidimensional companding," *IEEE Trans. Inform. Theory*, vol. 45, pp. 548–561, Mar. 1999.
- [22] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*. Boston, MA: Kluwer Acad. Pub., 1992.
- [23] D. L. Neuhoff, "The other asymptotic theory of lossy source coding," in *Coding and Quantization* (R. Calderbank, G. D. Forney, Jr., and N. Moayeri, eds.), vol. 14 of *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, pp. 55–65, American Mathematical Society, 1993.
- [24] R. M. Gray and D. L. Neuhoff, "Quantization," *IEEE Trans. Inform. Theory*, vol. 44, pp. 2325–2383, Oct. 1998.
- [25] W. R. Bennett, "Spectra of quantized signals," *Bell Syst. Tech. J.*, vol. 27, pp. 446–472, July 1948.
- [26] P. F. Panter and W. Dite, "Quantizing distortion in pulse-count modulation with nonuniform spacing of levels," *Proc. IRE*, vol. 39, pp. 44–48, Jan. 1951.
- [27] J. A. Bucklew and G. L. Wise, "Multidimensional asymptotic quantization theory with r th power distortion measures," *IEEE Trans. Inform. Theory*, vol. IT-28, pp. 239–247, Mar. 1982.
- [28] S. Cambanis and N. L. Gerr, "A simple class of asymptotically optimal quantizers," *IEEE Trans. Inform. Theory*, vol. IT-29, pp. 664–676, Sept. 1983.
- [29] T. Linder, "On asymptotically optimal companding quantization," *Prob. Contr. Inform. Theory*, vol. 20, no. 6, pp. 475–484, 1991.
- [30] V. K. Goyal, "High-rate transform coding: How high is high, and does it matter?," in *Proc. IEEE Int. Symp. Inform. Theory*, (Sorrento, Italy), p. 207, June 2000.
- [31] R. M. Gray and A. H. Gray, Jr., "Asymptotically optimal quantizers," *IEEE Trans. Inform. Theory*, vol. IT-23, pp. 143–144, Feb. 1977.
- [32] A. Segall, "Bit allocation and encoding for vector sources," *IEEE Trans. Inform. Theory*, vol. IT-22, pp. 162–169, Mar. 1976.