A Low-Complexity Universal Scheme for Rate-Constrained Distributed Regression Using a Wireless Sensor Network

Avon L. Fernandes, *Student Member, IEEE*, Maxim Raginsky, *Member, IEEE*, and Todd P. Coleman, *Member, IEEE*

***Abstract*—We propose a scheme for rate-constrained distributed nonparametric regression using a wireless sensor network. The scheme is universal across a wide range of sensor noise models, including unbounded and nonadditive noise; it has low complexity, requiring simple operations such as uniform scalar quantization with dither and message passing between neighboring nodes in the network, and attains minimax optimality for regression functions in common smoothness classes. We present theoretical results on the tradeoff between the compression rate, communication complexity of encoding, and the MSE and demonstrate empirical performance of the scheme using simulations.**

***Index Terms*—Conditional rate-distortion theory, distributed estimation, distributed sequential entropy coding, dithered scalar quantization, low-complexity schemes, minimax-optimal estimators, nonparametric regression, sensor networks, universal orthogonal series estimators.**

I. Introduction

In many practical applications of wireless sensor networks, very little prior knowledge is available about the phenomenon being sensed. Consider, for example, the following environmental monitoring scenario. A large number of sensors are deployed over a geographical region of interest, where they measure the concentration of a pollutant in the air. They then transmit their measurements via a wireless channel to a fusion center (FC) whose task is to produce an accurate image of air pollution in the region. The relationship between the position and the measurement of each sensor is probabilistic due to ambient noise, inaccuracies in measurement acquisition and signal transduction, etc. For instance, it might be of the standard "signal-plus-noise" form, where the signal (i.e., the concentration) is an unknown deterministic function of position. Thus, the basic inferential task of the FC is to *learn* the relationship between the sensor positions and their measurements, a problem well known in statistics under the heading of *regression*.

We seek the minimum-mean-square-error (MMSE) estimator of the measurement of a sensor from its location, which is given by the *regression function*, i.e., the conditional mean of the measurement given location. The regression function depends on the joint probability distribution of locations and measurements, which is either unknown or known imprecisely. The only information the FC has about it is contained in the data it has received from the network, and we expect that, as the number of sensors increases, the function estimated by the FC converges to the regression function in the mean-square sense. Moreover, for greater flexibility in modeling complex phenomena it is preferable to take the *nonparametric* approach [1], [2]—instead of assuming that the regression function is described by a vector of parameters (which is the case for linear functions or for polynomials of fixed degree), we suppose that it lies in some sufficiently broad infinite-dimensional space (e.g., monotone, r-times differentiable, Lipschitz, etc.).

Wireless sensor networks typically operate under tight energy constraints. The majority of all energy-intensive operations are involved in transmitting the data from the network to the FC and in any preliminary exchange of information among the sensors. In fact, energy expenditures for communicating a single bit over a wireless medium are orders of magnitude above the expenditures for carrying out an elementary computational step [3]. One way to conserve communication resources is to reduce the amount of communication from the network to the FC, while ensuring that the latter can still do a good job at inference. For instance, each sensor can *quantize* its measurement, so what the FC receives is a compressed "summary" of the network state. This *decentralized*, or *distributed*, setup is quite different from the traditional, *centralized*, scenario, where the FC has full observation of the sensor measurements without any compression. Additional energy savings can be gained by reducing the communication complexity of the network-side preprocessing. These energy considerations ultimately translate into a tradeoff between the compression rate and the achievable MSE.

In this paper, we propose, analyze and evaluate a scheme for rate-constrained distributed nonparametric regression using a wireless sensor network with randomly deployed sensors. The

Manuscript received January 23, 2008; accepted November 26, 2008. First published January 23, 2009; current version published April 15, 2009. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Deniz Erdogmus. This work was supported by the Beckman Foundation Fellowship and by the DARPA ITMANET program via US Army RDECOM contract W911NF-07-1-0029. Preliminary version of this work was presented at the IEEE International Conference on Acoustics, Speech and Signal Processing, Las Vegas, NV, March–April 2008.

A. L. Fernandes was with the Department of Electrical and Computer Engineering, University of Illinois, Urbana, IL 61801 USA. He is now with Microsoft Corporation, Redmond, WA 98052 USA (e-mail: alfernan@uiuc.edu).

M. Raginsky was with the Beckman Institute for Advanced Science and Technology, University of Illinois, Urbana, IL 61801 USA. He is now with the Department of Electrical and Computer Engineering, Duke University, Durham, NC 27708 USA (e-mail: m.raginsky@duke.edu).

T. P. Coleman is with the Department of Electrical and Computer Engineering, University of Illinois, Urbana, IL 61801 USA (e-mail: colemant@uiuc.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSP.2009.2013897

scheme consists of three successive stages, namely 1) quantization and encoding of sensor observations at the network side, 2) decoding of sensor observations by the FC, and 3) estimation of the regression function, and has the following attractive characteristics:

- **universality**: only very generic assumptions are made about the underlying joint distribution of the physical location of a sensor and its measurement;
- **low complexity**: the compression involves standard operations, such as uniform scalar quantization commonly used in A/D converters, as well as simple message passing between neighboring sensors;
- **minimax optimality**: the estimation procedure achieves minimax rates of convergence for certain broad classes of regression functions.

We give information-theoretic bounds on the average number of bits transmitted to the FC and bound the communication complexity of the network-side preprocessing and the rate at which the estimate of the regression function converges to the true one as the number of sensors increases.

The paper is organized as follows. Related work is surveyed in the remainder of this section. The problem is formulated in Section II. Section III offers a rough outline of the proposed scheme. Next, we describe and analyze the compression part of the scheme in Section IV and the estimation part in Section V. Section VI discusses minimax optimality. Experimental results are presented in Section VII. Section VIII summarizes the contributions of the paper. Appendix A contains a short summary of conditional rate-distortion theory. Proofs of all lemmas and theorems are relegated to Appendix B.

A. Related Work

The idea to cast distributed inference tasks for sensor networks (such as regression, set estimation, object tracking, or self-localization) in the framework of learning from random samples was proposed by Simić [4]. The approach of [4] centered around a distributed implementation of a popular algorithm for nonparametric regression based on reproducing-kernel Hilbert spaces [5], but (apart from considerations of locality) the issue of communication rate constraints was not taken into account. More recently, Predd *et al.* [6] introduced statistically consistent schemes for rate-constrained regression in networks of noncooperating sensors. However, whereas in this paper we are concerned with estimating the entire regression function, the results of [6] are for the *pointwise* MSE criterion. That is, the FC wishes to estimate the regression function *at a given point* and can broadcast the coordinates of that point to all the sensors. Each sensor then uses a local rule mapping its own observation and the point of interest into a short binary message; the FC aggregates the received messages to form the final estimate. (Similar strategies have also been used in the setting of rate-constrained decentralized estimation of a single deterministic parameter, see, e.g., [7] and [8] and references therein.)

Wang and Ishwar [9] proposed a scheme for distributed estimation of spatial fields using networks of randomly deployed, noisy sensors under the constraint that each sensor can send only 1 bit to the FC. Just as in [6], there is no cooperation among the sensors. Our approach is very similar to that of [9], particularly in the use of randomization and series estimation, except that we allow some communication among the sensors and impose no hard constraints on the number of bits transmitted by a sensor to the FC. However, our method is truly universal, requiring minimal knowledge of the joint distribution of sensor measurements and positions, while [9] assumes that a) the sensors are dispersed throughout the observation domain uniformly at random and b) the measurements are of the "signal-plus-noise" form, where both the signal and the noise have bounded dynamic ranges known to the FC. By contrast, our algorithm can handle nonuniformly deployed sensors, as well as unbounded and nonadditive noise.

An alternative approach to rate-constrained estimation in sensor networks uses ideas from distributed source coding [10]–[13]. There, the data collected by the network is interpreted as a noisy realization of a random field (e.g., Gaussian) with known statistics. The tradeoff between the compression rate and the MSE is analyzed using rate-distortion theory. Because the underlying statistical models are assumed to have known parametric form, efficient distributed compression schemes can exploit correlations between the measurements at neighboring sensors. By contrast, we are interested in situations where very little is known *a priori* about the phenomenon being sensed. For this reason, we adopt the nonparametric approach, which precludes the use of distributed source codes tailored to prespecified parametric sources. Alternatively, in keeping with the philosophy that nonparametric estimation is often merely the first step in data analysis [1], our nonparametric scheme can be used as a *training step* followed by model refinement and approximation by a simpler, parametric model that admits efficient distributed source codes.

II. Problem Formulation

A. Sensor Data Model

We make the following assumptions.

1) The network consists of n sensors deployed at random over a compact domain $\mathcal{X} \subset \mathbb{R}^d$ (where d is 1, 2, or 3) according to a fixed probability distribution Pr_X, which is known at the FC. Pr_X is absolutely continuous with a continuous density p_X.
2) Sensor measurements range over some $\mathcal{Y} \subseteq \mathbb{R}$. For each $x \in \mathcal{X}$, the conditional distribution $\mathrm{Pr}_{Y|X}(\cdot|x)$ of the measurement of a sensor given its location is absolutely continuous with a continuous density $p_{Y|X}(\cdot|x)$. $\mathrm{Pr}_{Y|X}$ is presumed unknown.
3) Conditioned on X, Y is a *sub-Gaussian* random variable: there exist constants $\lambda, \Lambda < \infty$, such that

$$\mathbb{E}\{e^{tY}|X\} \leq \Lambda e^{\lambda^2 t^2/2} \tag{1}$$

holds for all $t \in \mathbb{R}$. The constants Λ, λ are not assumed known to the sensors or to the FC. Examples of sub-Gaussian random variables include Gaussian or uniform.
4) The *regression function* $\eta(x) \triangleq \mathbb{E}\{Y|X = x\}$ is square-integrable w.r.t. Pr_X, i.e., $\eta \in L^2(\mathcal{X}, \mathrm{Pr}_X)$.

5) Let $\Pr_{XY}$ denote the joint distribution of the location and the measurement of a sensor. Let $\boldsymbol{X} = \{X_i\}_{i=1}^n$ and $\boldsymbol{Y} = \{Y_i\}_{i=1}^n$ denote, respectively, the n-tuples of sensor locations and their measurements Then $(X_1, Y_1), \ldots, (X_n, Y_n)$ are n independent and identically distributed (i.i.d.) samples from $\Pr_{XY}$. That is, $\boldsymbol{X}$ and $\boldsymbol{Y}$ have a joint density

$$p_{\boldsymbol{XY}}(\boldsymbol{x}, \boldsymbol{y}) = \prod_{i=1}^{n} p_X(x_i) \prod_{i=1}^{n} p_{Y|X}(y_i|x_i).$$

6) The n-tuple $\boldsymbol{X}$ is known at the FC, and each sensor knows its own location. Self-localization in wireless sensor networks is an active area of research (see, e.g., [14]–[16] and references therein). For example, the sensors can self-localize with the help of so-called *beacon nodes* that are aware of their own position, as well as of their distance from other nodes in the network.
7) The network and the FC have common access to an n-tuple $\boldsymbol{U} = \{U_i\}_{i=1}^n$ of i.i.d. random variables taking values in some set $\mathcal{U}$. Initially, each U_i is known only to the ith sensor and to the FC. The U_i's will be referred to as the *dither signals*. One way to generate dither signals is to use pseudorandom number generators. If each sensor can implement a pseudorandom number generator Ψ that produces samples of U by applying the iterates Ψ^k, $k = 1, 2, \ldots$, to a fixed seed ω, then for $i = 1, 2, \ldots, n$, both the FC and the ith sensor simply compute $U_i = \Psi^i(\omega)$.

As an example, Assumptions 1)–4) hold for the usual "signal-plus-noise" model $Y = f(X) + Z$, where $f \in L^2(\mathcal{X}, \Pr_X)$ is a bounded deterministic function and Z is a zero-mean sub-Gaussian random variable with density p_Z independent of X. Then $p_{Y|X}(y|x) = p_Z(y - f(x))$ and $\eta(x) = f(x)$.

B. Communication Setup

We assume the following about the communication setup.

1) Each sensor is equipped with a transceiver and can send analog messages to other sensors, but only *binary* messages to the FC. This assumption is reasonable in situations when the minimum distance between any sensor and the FC is much larger than the distance between any pair of sensors. In such cases, the effective SNR for intersensor communication is much higher than that for sensor-to-FC communication, so communication within the network is not as energy-intensive as between the network and the FC [17]. Therefore, it is feasible to have energy-efficient short-range analog communication between neighboring sensors, but, because the energy costs for communicating data to the FC over a long distance are much more severe, all data transmitted to the FC will have to be aggressively compressed.
2) After deployment, a routing path is established in the network. The routing path is induced by a linear ordering of the sensors, such that the sensor that occupies the kth position in the path can exchange messages with sensors in positions $k-1$ and $k+1$ (sensors in positions 1 and n can only communicate with sensors in positions 2 and $n-1$ respectively). In large networks with randomly deployed sensors, such paths exist with very high probability and can be discovered using greedy techniques (see, e.g., [18]). Let $X_{(1)}, \ldots, X_{(n)}$ denote the locations of the n sensors arranged according to their position in the routing path, and let $Y_{(1)}, \ldots, Y_{(n)}$ denote the corresponding measurements. This linear ordering need not coincide with the order in which the sensors were deployed. However, for simplicity we shall assume that the choice of the routing path corresponds to a random permutation of the n sensors. Since the joint distribution of $\boldsymbol{X}$ and $\boldsymbol{Y}$ is invariant w.r.t. such a permutation, $\{(X_{(i)}, Y_{(i)})\}_{i=1}^n$ are also i.i.d. according to $\Pr_{XY}$ (conditioned on the permutation). Thus, we will assume that $X_{(i)} = X_i$ for all $1 \leq i \leq n$.

To keep track of the energy costs in operating the network, we will use two different measures, depending on whether we are talking about in-network or out-of-network communication: the former will be measured by the total number of scalar analog messages exchanged by the sensors, while the latter will be measured by the total number of bits transmitted by the sensors to the FC. Both of these metrics are commonly used in the literature (see, e.g., [7], [19], and references therein).

C. Rate-Constrained Distributed Nonparametric Regression

The regression function $\eta(x)$ is the MMSE estimator of the measurement of a sensor placed at $X = x$. The task of the FC is to *estimate* (or to *learn*) η from the locations $\boldsymbol{X}$ and the compressed version of the measurements $\boldsymbol{Y}$. Formally, we define a scheme for rate-constrained distributed nonparametric regression as follows. For each $n = 1, 2, \ldots$, let $\mathcal{F}_n$ be a set of functions from $\mathcal{X}$ into $\mathbb{R}$. Then we have a sequence $\{(e_n, d_n, \widehat{f}_n)\}_{n=1}^{\infty}$, where, for each n, $e_n : \mathcal{X}^n \times \mathcal{Y}^n \times \mathcal{U}^n \to \{0,1\}^*$ is the *encoder*, $d_n : \mathcal{X}^n \times \{0,1\}^* \times \mathcal{U}^n \to \mathcal{Y}^n$ is the *decoder*, and $\widehat{f}_n : \mathcal{X}^n \times \mathcal{Y}^n \to \mathcal{F}_n$ is the *estimator*. $\{0,1\}^*$ is the space of all finite-length binary sequences. The encoding is done on the network side, while the decoding and the estimation are done by the FC. For each n, $\mathcal{F}_n$ consists of all possible candidates for the estimate of η, and the dependence on n allows us to adjust the complexity of the estimate relative to the network size. The three quantities of interest to us are as follows.

1) The average number of bits transmitted to the FC:

$$R_n = \frac{1}{n}\mathbb{E}\left\{\text{length}\left(e_n(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{U})\right)\right\}.$$

2) The average communication complexity of the encoder, i.e., the average number of messages per sensor exchanged among the n sensors during encoding:

$$C_n = \frac{1}{n}\mathbb{E}\left\{\sum_{i=1}^{n} c_i(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{U})\right\}$$

where c_i denotes the number of analog messages transmitted by sensor i to other sensors.
3) The MSE of the estimator:

$$\text{MSE}(\widehat{f}_n, \eta) = \mathbb{E}\left\{\int_{\mathcal{X}} \left(\widehat{f}_n(x) - \eta(x)\right)^2 d\Pr_X(x)\right\}$$

(here we abuse the notation slightly and denote by $\widehat{f}_n$ the function computed by the estimator).

For all three of these, the expectation is with respect to the sensor locations $\boldsymbol{X}$ and measurements $\boldsymbol{Y}$, and the dither $\boldsymbol{U}$.

D. Extension to Multiple Time Steps

In stating these performance metrics, we have ignored the cost of initializing the network (which consists of sensor localization and establishment of the routing path). The reason for doing this is that our setup naturally extends to the situation when the sensors take measurements at discrete times $t = 1, 2, \ldots, T$, where the conditional distribution of a sensor measurement given the location varies with time. In this case the one-time initialization cost becomes negligible as $T \to \infty$. To formulate this extension, let $\Pr_{Y|X}^{(t)}$ be the conditional probability distribution of a sensor measurement at time t given the location, and let $p_{Y|X}^{(t)}(\cdot|x)$ denote the corresponding conditional density given $X = x$. We make the following additional assumptions.

1) The sensor locations $\boldsymbol{X}$ do not change with time.
2) Let $\boldsymbol{Y}^{(t)} = \{Y_i^{(t)}\}_{i=1}^n$ be the n-tuple of sensor measurements at time t. Then, for any two distinct t_1 and t_2, the random variables $\boldsymbol{Y}^{(t_1)}$ and $\boldsymbol{Y}^{(t_2)}$ are conditionally independent given $\boldsymbol{X}$.
3) At each t, a new n-tuple $\boldsymbol{U}^{(t)} = \{U_i^{(t)}\}_{i=1}^n$ of dither signals, independent of $\boldsymbol{X}$, $\boldsymbol{Y}^{(t)}$, and $\boldsymbol{U}^{(1)}, \ldots, \boldsymbol{U}^{(t-1)}$, is available simultaneously at the network and at the FC.

That is, for a given T, the joint distribution of $\boldsymbol{X}$ and $\boldsymbol{Y}^{(1)}, \ldots, \boldsymbol{Y}^{(T)}$ is described by the density

$$p_{\boldsymbol{X},\boldsymbol{Y}^{(1)},\ldots,\boldsymbol{Y}^{(T)}}\left(\boldsymbol{x}, \boldsymbol{y}^{(1)}, \ldots, \boldsymbol{y}^{(T)}\right) = \prod_{i=1}^{n} p_X(x_i) \prod_{t=1}^{T} \prod_{i=1}^{n} p_{Y|X}^{(t)}\left(y_i^{(t)}|x_i\right).$$

Under these assumptions, the MMSE estimator of the measurement sequence $Y^{(1)}, \ldots, Y^{(T)}$ of a sensor placed at $X = x$ is given by the vector-valued function $\boldsymbol{\eta}(x) = [\eta^{(1)}(x), \ldots, \eta^{(T)}(x)]$. We can convert a one-time regression scheme $\{(e_n, d_n, \widehat{f}_n)\}_{n=1}^{\infty}$ into a multi-time scheme $\{(e_n^{(t)}, d_n^{(t)}, \widehat{f}_n^{(t)})\}_{n,t=1}^{\infty}$ by letting $e_n^{(t)} = e_n$, $d_n^{(t)} = d_n$, and $\widehat{f}_n^{(t)} = \widehat{f}_n$. The reason why the same encoder, decoder and estimator are used at all t is that the conditional distribution $\Pr_{Y|X}^{(t)}$ may vary with t in an arbitrary manner. Now let $\eta^{(t)}$ be the regression function at time t, i.e., $\eta^{(t)}(x) = \mathbb{E}\{Y^{(t)}|X = x\}$. The performance of the scheme is measured by the average number of bits transmitted by the network

$$R_{n,T} = \frac{1}{nT}\mathbb{E}\left\{\sum_{t=1}^{T} \text{length}\left(e_n^{(t)}\left(\boldsymbol{X}, \boldsymbol{Y}^{(t)}, \boldsymbol{U}^{(t)}\right)\right)\right\},$$

the average MSE

$$M_{n,T} = \frac{1}{T}\sum_{t=1}^{T} \text{MSE}\left(\widehat{f}_n^{(t)}, \eta^{(t)}\right),$$

and the average communication complexity of the encoder. Owing to the assumed conditional independence, the performance metrics for $T > 1$ can be calculated by adding the corresponding numbers for each t and averaging. However, because all the key ideas are already present when $T = 1$, we focus on this case in the remainder of the paper.

III. Outline of the Proposed Approach and Summary of Results

Our scheme is parametrized by $\varepsilon > 0$, known both to the network and to the FC. Higher values of ε will correspond to lower communication rates. The dither $\boldsymbol{U} = \{U_i\}_{i=1}^n$ is an i.i.d. sequence drawn from the uniform distribution on the interval $\mathcal{U} = [-\sqrt{3\varepsilon}, \sqrt{3\varepsilon}]$ independently of $\boldsymbol{X}$ and $\boldsymbol{Y}$. The main idea is as follows: Encoding consists of randomized uniform scalar quantization of the sensor measurements, where the dither is used for randomization, followed by sequential universal entropy coding of the quantizer indices. Sequential entropy coding has low communication complexity, measured by the number of messages exchanged among the sensors. Each sensor then transmits a variable-length binary encoding of its quantizer index. This is, essentially, a distributed implementation of the universal quantization scheme due to Ziv [20], with refinements by Zamir and Feder [21]. Once the FC decodes the quantizer indices, it estimates the regression function using a universal orthogonal series estimator [1], [2]. The overall architecture of the scheme is shown in Fig. 1.

We summarize here the main results to be proved in the following sections:

1) the average number of bits transmitted to the FC is

$$R_n \le R_{Y|X}(\varepsilon) + 0.754 + \frac{K_1(\lambda, \Lambda)\sqrt{\log^5 n}}{\sqrt{\varepsilon} n} + \Delta_n;$$

2) the communication complexity of the encoding is

$$C_n \le \frac{K_2(\lambda, \Lambda)\sqrt{\log^3 n}}{\sqrt{\varepsilon}};$$

3) the MSE of the regression function estimator is

$$\text{MSE}(\widehat{f}_n, \eta) \le \frac{(K_3(\lambda, \Lambda) + \varepsilon) L_n}{n} + \Gamma_n.$$

Here, K_1, K_2, and K_3 are constants that depend only on the sub-Gaussianity parameters λ, Λ of the conditional distribution of Y given X [cf. (1)]; $\{\Delta_n\}$ and $\{\Gamma_n\}$ are decreasing sequences of nonnegative reals that both converge to zero as $n \to \infty$ and depend only on the joint distribution $\Pr_{XY}$; and $\{L_n\}$ is an increasing sequence of positive integers that is chosen by the FC depending on $\Pr_X$ in such a way that $L_n/n \to 0$ as $n \to \infty$.

The quantity $R_{Y|X}(\varepsilon)$ appearing in the bound on R_n is the *conditional rate-distortion function* (CRDF) of Y given X. It gives the minimum amount of information (in bits) that must be provided about Y in order to be able to reconstruct it with an MSE of ε, provided both the encoder and the decoder have access to X as *side information* (see Appendix A for details). For example, when the sensor data model has the form of a signal in Gaussian noise, i.e., $Y = f(X) + Z$, where $Z \sim$

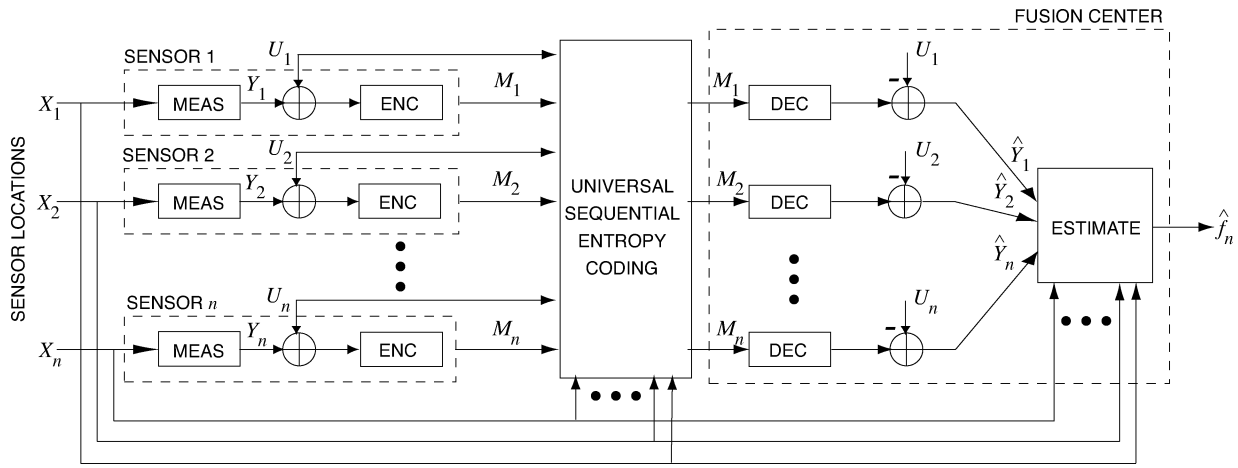

Fig. 1. Overall architecture of our scheme for rate-constrained distributed regression using a wireless sensor network.

$\mathrm{Normal}(0,\sigma^2)$ is independent of X, the CRDF depends only on the noise variance σ^2 and has the form

$$R_{Y|X}(\varepsilon) = \begin{cases} \frac{1}{2}\log\frac{\sigma^2}{\varepsilon}, & \varepsilon \le \sigma^2 \\ 0, & \varepsilon > \sigma^2. \end{cases}$$

In general, this information-theoretic limit can be achieved only asymptotically, provided the encoder can simultaneously encode a large number n of i.i.d. samples $\{(X_i, Y_i)\}_{i=1}^n$, the decoder has access to the side information $\{X_i\}_{i=1}^n$, and the joint distribution of X and Y is known to both the encoder and the decoder. In our case, neither the encoder (i.e., the entire network) nor the decoder (i.e., the FC) know $\Pr_{XY}$. However, the universal strategy of Ziv [20], [21] can be used to achieve the CRDF with an additional penalty of at most 0.754 bits per sample. This penalty is the price that must be paid for using simple scalar quantizers; of course, much more elaborate universal codes based on vector quantizers can be used to eliminate this penalty, but at the cost of vastly increased encoder complexity.

It is worth noting that, in limit $\varepsilon \to 0$, which corresponds to high quantizer resolution, the MSE can be bounded as $(K_3(\lambda,\Lambda)L_n/n) + \Gamma_n$, which is also the MSE achievable by the same orthogonal series estimator in the centralized setting. Hence, quantization of sensor measurements has no effect on the *rate* at which the MSE converges to zero as the number of sensors n tends to infinity. Under certain assumptions on the functional class containing the regression function and with proper choice of $\{L_n\}$, the MSE of the orthogonal series estimator in the centralized setting can be shown to converge to zero at the *minimax rate* as $n \to \infty$. Since the rate of convergence is not affected by quantization of sensor measurements, the estimator is still minimax optimal even when the sensor measurements are quantized at very low resolutions. The issues of minimax optimality are covered in Section VI.

We close this section with a back-of-the-envelope analysis of the tradeoff between communication costs and MSE. The communication costs (per sensor) are dominated by in-network analog communication, and will scale approximately as $\sqrt{\log^3 n}/\sqrt{\varepsilon}$. For regression functions in common smoothness classes (see Section VI), the MSE will scale approximately as $(1+\varepsilon)n^{-\alpha}$, where $\alpha > 0$ is some smoothness constant. Thus, we arrive at the following relation for the tradeoff between MSE and communication costs:

$$\mathrm{MSE} \cdot (\text{Comm. costs})^2 \sim \left(1 + \frac{1}{\varepsilon}\right)\frac{\log^3 n}{n^\alpha}.$$

The term involving ε is the price we pay for quantizing the sensor measurements, while the logarithmic factor $\log^3 n$ is the price we pay for cooperation among sensors.

IV. Encoding and Decoding

Given $\varepsilon > 0$, we define the basic encoder and decoder mappings $E_\varepsilon : \mathbb{R} \to \mathbb{Z}$ and $D_\varepsilon : \mathbb{Z} \to \mathbb{R}$ via $E_\varepsilon(y) \triangleq \lfloor (y + \sqrt{3\varepsilon})/2\sqrt{3\varepsilon} \rfloor$ and $D_\varepsilon(m) \triangleq 2m\sqrt{3\varepsilon}$. The composite mapping $Q_\varepsilon = D_\varepsilon \circ E_\varepsilon$ is a uniform scalar quantizer with the levels $\{2m\sqrt{3\varepsilon}\}_{m\in\mathbb{Z}}$. We further define $\widehat{Y}_i = Q_\varepsilon(Y_i + U_i) - U_i$ for all $1 \le i \le n$. This operation is known in source coding as *uniform quantization with additive dither* (see, e.g., [22, ch. 6]). It was shown by Ziv [20] (see also [21]) that it leads to a universal quantization scheme that can achieve the CRDF within 0.754 bits per sample (see Section IV-A and Appendix B-1) for details). The reconstructions $\widehat{Y}_i$ are computed in two steps.

1) For every $i = 1, \ldots, n$, sensor i computes $M_i = E_\varepsilon(Y_i + U_i)$. Then it transmits a lossless binary encoding $B_i \in \{0,1\}^*$ of M_i to the FC. To get B_i, the ith sensor exchanges messages with the rest of the network.
2) The FC receives the $\boldsymbol{B} = \{B_i\}_{i=1}^n$, decodes $\boldsymbol{M}$, and computes $\widehat{Y}_i = D_\varepsilon(M_i) - U_i$ for every i.

We now describe a distributed algorithm for computing $\boldsymbol{B}$. First, note that $\boldsymbol{M} = \{M_i\}_{i=1}^n$ is an n-tuple of i.i.d. integer-valued random variables. If the joint probability distribution of (M, X, U) were known, then each sensor could independently use a Huffman code [23] to encode its M_i. The resulting per-sensor expected codelength would then be within one bit of $H(M|X,U)$, the conditional entropy of M given X and U, which is the minimum possible expected codelength [23]. In practice, however, both $\boldsymbol{X}$ and $\boldsymbol{U}$ have to be discretized. Thus, for every i, the ith sensor would use a fixed discretization rule to map its (X_i, U_i) to $(\widetilde{X}_i, \widetilde{U}_i)$, where the possible values of $(\widetilde{X}_i, \widetilde{U}_i)$ come from a finite subset of $\mathcal{X} \times \mathcal{U}$, and

then use a Huffman code matched to the joint distribution of $(M, \widetilde{X}, \widetilde{U})$ to describe its M_i to the FC. With judicious choice of the discretization rule (which may, in principle, depend on n), we would have $H(M|\widetilde{X}, \widetilde{U}) \approx H(M|X, U)$. However, the lack of knowledge of Pr_{XY} precludes the use of predesigned codebooks. Instead, one has to resort to a *universal* method which first uses the input sequence to acquire a statistical model for the underlying source, and then encodes that sequence with a code matched to the acquired model. One way of doing this is to use *sequential probability assignment* [24] to build a *predictive model* of the data source. We apply this methodology here.

The key idea behind sequential probability assignment is as follows. Let $\boldsymbol{m}$, $\boldsymbol{x}$ and $\boldsymbol{u}$ denote specific realizations of $\boldsymbol{M}$, $\boldsymbol{X}$ and $\boldsymbol{U}$. Let us use the shorthand $S_i = (\widetilde{X}_i, \widetilde{U}_i)$ for the discretized side information at the ith sensor, and let $s_i = (\widetilde{x}_i, \widetilde{u}_i)$ denote its specific realization. Suppose that sensor i is informed of $m^{i-1} \triangleq (m_1, \ldots, m_{i-1})$ and $s^{i-1} \triangleq (s_1, \ldots, s_{i-1})$, i.e., the quantizer indexes and the discretized locations and dither signals of the "downstream" sensors 1 through $i-1$, and can then run a fixed algorithm to compute a sequence $\{\widehat{P}^i(m|m^{i-1}, s^i)\}_m$ of nonnegative reals satisfying $\sum_m \widehat{P}^{(i)}(m|m^{i-1}, s^i) = 1$. This sequence assigns a *conditional probability model* to M_i given the observed values of M^{i-1} and S^i. The sensor can then use a Huffman code to encode its quantizer index $M_i = m_i$ using approximately $-\log \widehat{P}^{(i)}(m_i|m^{i-1}, s^i)$ bits. The average number of bits transmitted by the network is approximately

$$\frac{1}{n}\mathbb{E}\left\{-\sum_{i=1}^{n} \log \widehat{P}^{(i)}(M_i|M^{i-1}, S^i)\right\}$$

where the expectation is with respect to $\boldsymbol{M}$ and $\boldsymbol{S}$. In other words, the "true" distribution $P(\boldsymbol{m}|\boldsymbol{s}) = \prod_{i=1}^{n} P(m_i|s_i)$ is approximated by the *predictive distribution* $\widehat{P}(\boldsymbol{m}|\boldsymbol{s}) = \prod_{i=1}^{n} \widehat{P}^{(i)}(m_i|m^{i-1}, s^i)$. Decoding of the m_i's is done sequentially: having decoded m^{i-1}, the FC can use that and its knowledge of s^i to compute the predictive model $\widehat{P}^{(i)}(\cdot|m^{i-1}, s^i)$ and generate the corresponding codebook. The idea here is that, as i increases, the predictive model approximates the true distribution more accurately, and, correspondingly, the average number of bits transmitted by each sensor decreases with i.

Now, a naive implementation of this approach would result in $O(n)$ per-sensor communication complexity because the amount of data each sensor would need to receive and/or transmit would grow linearly with the position of the sensor in the routing path. However, using a specific recursive algorithm for computing the predictive models $\widehat{P}^{(i)}(\cdot|m^{i-1}, s^{i-1})$, we can both improve the communication complexity and asymptotically approach the conditional entropy $H(M|X, U)$. In particular, under the assumptions stated in Section II, the expected communication complexity will be on the order of $O(\sqrt{\log^3 n}/\sqrt{\varepsilon})$.

Our encoding algorithm works as follows. First, the sensors exchange messages to determine $\underline{m} = \min m_i$ and $\overline{m} = \max m_i$. If sensor i can send messages to sensors $i-1$ and $i+1$, $\underline{m}$ and $\overline{m}$ can each be determined using simple comparisons and then propagated to all the sensors with no more than $2n$ message passes. A designated sensor (say, the last sensor in the routing path) then uses a fixed encoding of the integers, such as the Elias scheme [25], to communicate the values of $\underline{m}$ and $N \triangleq \overline{m} - \underline{m} + 1$ to the FC. Under the Elias representation, each $r \in \mathbb{N}$ can be encoded using $L(r) = \log r + O(\log\log r)$ bits. Negative numbers can be encoded by prepending a sign bit.[1]

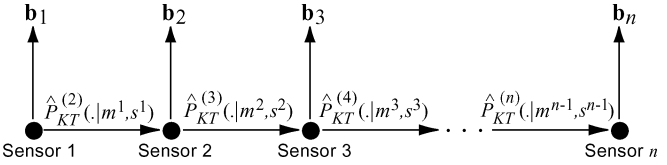

Fig. 2. Forward message passing scheme for entropy coding of the quantizer indices.

Without loss of generality, let us suppose that $\underline{m} = 1$, so that $m_i \in \{1, \ldots, N\}$. To obtain $\boldsymbol{s}$, let $q \triangleq c\lfloor\sqrt{\log n}\rfloor$ (here, c can be any positive integer), then cover $\mathcal{X}$ by q disjoint cubes $\mathcal{C}_1, \ldots, \mathcal{C}_q$ and carve the interval $[-\sqrt{3\varepsilon}, \sqrt{3\varepsilon}]$ into q disjoint subintervals $\mathcal{I}_1, \ldots, \mathcal{I}_q$. For each l, let us pick a unique $x^{(l)} \in \mathcal{C}_l \cap \mathcal{X}$ and let $u^{(l)}$ be the midpoint of $\mathcal{I}_l$. For $i = 1, \ldots, n$, let $\widetilde{x}_i = x^{(l)}$ if $x_i \in \mathcal{C}_l$ and let $\widetilde{u}_i = u^{(k)}$ if $u_i \in \mathcal{I}_k$. Next, for $i = 1, \ldots, n$, define

$$\widehat{P}^{(i)}(m|m^{i-1}, s^i) = \frac{\widehat{P}^{(i)}_{KT}(m, s_i|m^{i-1}, s^{i-1})}{\sum_{m=1}^{N} \widehat{P}^{(i)}_{KT}(m, s_i|m^{i-1}, s^{i-1})} \quad (2)$$

for every $m \in \{1, \ldots, N\}$, where $\widehat{P}^{(i)}_{KT}$ is the so-called *Krichevsky–Trofimov* (KT) estimator [26]

$$\widehat{P}^{(i)}_{KT}(m, s|m^{i-1}, s^{i-1}) \triangleq \frac{\mathsf{n}(m, s|m^{i-1}, s^{i-1}) + \frac{1}{2}}{i - 1 + \frac{A}{2}} \quad (3)$$

where $\mathsf{n}(m, s|m^{i-1}, s^{i-1})$ is the number of occurrences of (m, s) in $\{(m_j, s_j)\}_{j=1}^{i-1}$, and $A = Nq^2$ is the number of values the pair (m, s) can take. Note that $\widehat{P}^{(1)}_{KT}(m, s) = 1/A$ for all (m, s). From (3) it follows that for $1 \le i < n$ the KT estimator admits the recursive representation

$$\widehat{P}^{(i+1)}_{KT}(m, s|m^i, s^i) = (1 - \lambda_i)\widehat{P}^{(i)}_{KT}(m, s|m^{i-1}, s^{i-1}) + \lambda_i 1_{\{(m,s)=(m_i,s_i)\}} \quad (4)$$

with $\lambda_i = (i + A/2)^{-1}$. Thus, the sensors can compute the predictive distributions $\widehat{P}^{(i)}(\cdot|m^{i-1}, s^i)$ using the following forward message passing scheme: for $i = 1, \ldots, n-1$,

1) sensor i computes $\widehat{P}^{(i+1)}_{KT}(m, s|m^i, s^i)$ for all m, s via (4) and passes these to sensor $i+1$;
2) sensor $i+1$ computes the N probabilities $\widehat{P}^{(i+1)}(\cdot|m^i, s^{i+1})$ via (2), designs a Huffman code for M_i given (M^{i-1}, S^i), and encodes m_i using this code.

The total number of messages sent is $A(n-1)$. Fig. 2 shows the resulting information flow. The overall per-sensor communication complexity of computing $\underline{m}$ and $\overline{m}$ and the encoding of $\boldsymbol{m}$ is thus $O(A) = O(N \log n)$.

[1]To save communication resources, the designated sensor can compute the Elias encodings of $\underline{m}$ and $\overline{m}$, transmit a one-bit flag to indicate whether or not $L(\underline{m}) < L(\overline{m})$, and follow this with the Elias encoding of $\underline{m}$ or $\overline{m}$, as appropriate.

Remark 4.1: We note that, for each i, the A values of the KT estimator $\widehat{P}_{KT}^{(i+1)}$ are integer multiples of $2i+A$, and a simple counting argument shows that there are at most $(2i+A)^A$ possible KT estimator tuples. Hence, instead of passing the full tuple to sensor $i+1$, sensor i can send a unique binary encoding of this tuple, which will require at most

$$\begin{aligned} A\log(2i+A) &= O\left(A\log(n+A)\right) \\ &= O\left(N\log n\log(n+N\log n)\right) \end{aligned}$$

bits per sensor.

A. Analysis of Performance

We now analyze the performance of our encoding scheme under the assumptions stated in Section II. We start by stating an information-theoretic bound on the conditional entropy $H(M|X,U)$, which is the minimum per-sensor encoding length for $\boldsymbol{M}$ when the encoding and the decoding are based on the side information $\boldsymbol{X}$ and $\boldsymbol{U}$. The following theorem is a straightforward extension of the results of [21] to the case of extra side information $\boldsymbol{X}$.

Theorem 4.1: Let (X,Y) be jointly distributed according to $\Pr_{XY}$. The conditional entropy of $M \equiv E_\varepsilon(Y+U)$ given X and U satisfies $H(M|X,U) \le R_{Y|X}(\varepsilon) + 0.754$, where $R_{Y|X}$ is the *conditional rate-distortion function* of Y given X.

Proof: See Appendix B. ∎

Next, we show that the expected codelength of our sequential encoding is close to $H(M|\widetilde{X},\widetilde{U})$:

Theorem 4.2: The predictive distribution (2) satisfies

$$\frac{1}{n}\mathbb{E}\left\{-\sum_{i=1}^{n}\log\widehat{P}^{(i)}(M_i|M^{i-1},S^i)\middle|N\right\} \le H(M|\widetilde{X},\widetilde{U}) + \frac{CN\log^2 n}{n} \quad (5)$$

for some constant $C>0$.

Proof: See Appendix B. ∎

Finally, using these two theorems together with the assumptions in Section II, we can characterize the overall performance of the encoder as follows:

Theorem 4.3: Under the assumptions stated in Section II, the following holds:

1) The average number of bits transmitted by the network to the FC is bounded as

$$R_n \le R_{Y|X}(\varepsilon) + 0.754 + \frac{K_1(\lambda,\Lambda)\sqrt{\log^5 n}}{\sqrt{\varepsilon}n} + \Delta_n \quad (6)$$

where the sequence $\{\Delta_n\}$ converges to zero as $n\to\infty$ and depends only on $\Pr_X$ and $\Pr_{Y|X}$.

2) The average communication complexity of the encoder is

$$C_n \le \frac{K_2(\lambda,\Lambda)\sqrt{\log^3 n}}{\sqrt{\varepsilon}}. \quad (7)$$

The constants $K_1(\lambda,\Lambda)$ and $K_2(\lambda,\Lambda)$ depend only on the parameters λ and Λ of $\Pr_{Y|X}$ [cf. (1)].

Proof: See Appendix B. ∎

V. Estimation of the Regression Function

We now describe the strategy the FC will use to estimate the regression function once the data transmitted by the network have been received and decoded. Recall our assumption that $\eta \in L^2(\mathcal{X},\Pr_X)$. Pick a complete orthonormal system $\Phi = \{\varphi_j\}_{j=0}^{\infty}$ in $L^2(\mathcal{X},\Pr_X)$. Then η can be expanded in a Fourier series as $\eta(x) = \sum_{j=0}^{\infty}\theta_j\varphi_j(x)$, where the Fourier coefficients θ_j are given by $\theta_j = \int_{\mathcal{X}}\varphi_j\eta d\Pr_X$. We construct our estimator as follows. For any $J=0,1,2,\ldots$, define $C(J) \triangleq \max_{0\le j\le J}\|\varphi_j\|_\infty^2$, where $\|f\|_\infty = \sup_{x\in\mathcal{X}}|f(x)|$ is the sup norm. Choose an increasing sequence $\{J_n\}_{n=1}^{\infty}$ of nonnegative reals (the *cutoffs*), such that

$$\lim_{n\to\infty}\frac{L_n}{n} = 0 \quad (8)$$

where $L_n = (J_n+1)C(J_n)$. For example, this condition is satisfied in the following cases.

- *Uniformly bounded bases*. If $\|\varphi_j\|_\infty \le C < \infty$ for all j, then we can pick $J_n = \lfloor\sqrt{n}\rfloor$.
- *Wavelet bases*. Let $\mathcal{X} = [0,1]$ and consider, for instance, the *Haar system* [27] $\Phi = \{\psi_{k,l}\}$, defined by $\psi_{0,0}(x) = 1$ and $\psi_{k,l}(x) = 2^{k/2}\psi(2^k x - l)$, $k\in\mathbb{N}$, $0\le l\le 2^k-1$, where $\psi(x) = 1_{[0,1/2]}(x) - 1_{[1/2,1]}(x)$. The index k is the *scale*, while l is the *location*. We can re-index Φ using a single integer by defining $j(k,l) = 2^k - 1 + l$ and setting $\varphi_{j(k,l)} = \psi_{k,l}$. Conversely, by defining $k(j) = \max\{k\ge 0 : 2^k \le j\}$ and $l(j) = j - 2^{k(j)} + 1$, we can write $\varphi_j = \psi_{k(j),l(j)}$. Then $\|\varphi_j\|_\infty = 2^{k(j)/2}$ for every j, so $C(J) = 2^{k(J)}$. Therefore, if we choose $J_n = 2(2^{\lceil(1/4)\log n\rceil} - 1)$, then $J_nC(J_n)/n \le 16/\sqrt{n}\to 0$ as $n\to\infty$. In other words, we retain all wavelet coefficients up to and including scale $\lfloor(1/4)\log n\rfloor$.

For every $0\le j\le J_n$, let the FC estimate θ_j by

$$\widehat{\theta}_j \triangleq \frac{1}{n}\sum_{i=1}^{n}\varphi_j(X_i)\widehat{Y}_i. \quad (9)$$

and then form the *projection estimate*

$$\widehat{f}_n(x) = \sum_{j=0}^{J_n}\widehat{\theta}_j\varphi_j(x). \quad (10)$$

Thus, for each n, the estimates are taken from the (J_n+1)-dimensional subspace $\mathcal{F}_n$ of $L^2(\mathcal{X},\Pr_X)$, spanned by $\{\varphi_j\}_{j=0}^{J_n}$.

A. Analysis of Performance

In this section, we show that the estimate $\widehat{f}_n$ converges to the regression function η in the mean square sense and, moreover, that the quantization of sensor measurements does not affect the rate of convergence. This robustness against quantization errors comes from the fact that, informally speaking, dithering "whitens" the quantization error. Namely, each $\widehat{Y}_i$ can be written as $Y_i + E_i$, where, conditionally on X_i, E_i has mean zero and variance ε, and is independent of Y_i (see Appendix B for details). Therefore, we have $\mathbb{E}\{\widehat{Y}_i|X_i\} = \mathbb{E}\{Y_i|X_i\} \equiv \eta(X_i)$. Moreover, the E_i's are mutually conditionally independent given $\boldsymbol{X}$. Thus, dithering converts the *centralized* regression problem based on the original sensor data $\{(X_i,Y_i)\}_{i=1}^{n}$

into a new regression problem based on noise-corrupted data $\{(X_i, Y_i + E_i)\}_{i=1}^n$, which has the *same* regression function. The effect of this noise is to increase the variance of each $\widehat{\theta}_j$ [which in centralized case would be on the order of $O(1/n)$] by ε/n. In other words, we have the following.

Lemma 5.1: For any j, $\widehat{\theta}_j$ defined in (9) satisfies

$$\mathbb{E}\widehat{\theta}_j = \theta_j \quad \text{and} \quad \mathbb{E}\left\{(\widehat{\theta}_j - \theta_j)^2\right\} \leq \frac{\|\varphi_j\|_\infty^2 (\operatorname{Var} Y + \varepsilon)}{n}.$$

Proof: See Appendix B. ∎

We can now state our result on the MSE of our estimator:

Theorem 5.1: The projection estimator (10) satisfies

$$\operatorname{MSE}(\widehat{f}_n, \eta) \leq \frac{(K_3(\lambda, \Lambda) + \varepsilon)\, L_n}{n} + \Gamma_n \tag{11}$$

where $\Gamma_n \triangleq \sum_{j>J_n}^{\infty} \theta_j^2$, and the constant $K_3(\lambda, \Lambda)$ depends only on the parameters λ, Λ of $\Pr_{Y|X}$. Moreover, $\widehat{f}_n \to \eta$ in mean square as $n \to \infty$.

Proof: See Appendix B. ∎

From (11), we see that the MSE penalty for quantization is on the order of $L_n\varepsilon/n$, and that

$$\lim_{\varepsilon \to 0} \operatorname{MSE}(\widehat{f}_n, \eta) \leq \frac{K_3(\lambda, \Lambda) L_n}{n} + \Gamma_n.$$

That is, quantization has no effect on the rate of convergence of the MSE to zero. Therefore, provided that the number of sensors n is sufficiently large (i.e., the network is sufficiently dense), communication resources can be saved by having the sensors use very coarse quantizers to quantize their measurements. In fact, in our simulations we have found that the degradation of the MSE due to quantization is not very significant (see Section VII), which makes our scheme suitable for low-rate operation. As we have remarked above, this robustness of the scheme to quantization errors is due to deliberate introduction of noise in the form of dither. We comment on this further in Section VIII.

B. Adaptive Choice of the Cutoffs

The above procedure for choosing the cutoffs $\{J_n\}$ is overly conservative, and may result in overfitting. We now suggest an alternative, data-driven procedure for choosing the cutoffs based on the idea of empirical risk minimization [1, sec. 7.4].

If a fixed cutoff J is used, and η is estimated by $\widehat{f}_{n,J} \triangleq \sum_{j=0}^{J} \widehat{\theta}_j \varphi_j$ where $\{\widehat{\theta}_j\}$ are given by (9), then we can use Parseval's identity to write the resulting MSE as

$$\operatorname{MSE}(\widehat{f}_{n,J}, \eta) = \sum_{j=0}^{J} \mathbb{E}\left\{(\widehat{\theta}_j - \theta_j)^2\right\} + \sum_{j>J}^{\infty} \theta_j^2. \tag{12}$$

Defining $d_{n,j} \triangleq n\mathbb{E}\{(\widehat{\theta}_j - \theta_j)^2\}$, we may rewrite (12) as

$$\operatorname{MSE}(\widehat{f}_{n,J}, \eta) = \sum_{j=0}^{J} \left(n^{-1} d_{n,j} - \theta_j^2\right) + \sum_{j=0}^{\infty} \theta_j^2. \tag{13}$$

By Parseval, the second term on the right-hand side of (13) is equal to $\int_{\mathcal{X}} |f|^2 d\Pr_X$, which is independent of $\widehat{f}_{n,J}$. Consider now an *oracle* that knows η and can choose the cutoff J_n^* to minimize $\operatorname{MSE}(\widehat{f}_{n,J}, \eta)$ over all J. This oracle will choose

$$J_n^* = \arg\min_{J} \sum_{j=0}^{J} \left(n^{-1} d_{n,j} - \theta_j^2\right). \tag{14}$$

Note that J_n^* depends on $\{d_{n,j}\}$ and $\{\theta_j^2\}$, which in turn depend on the unknown function η. However, as we now show, the FC can obtain *unbiased* estimates of these quantities from $\boldsymbol{X}$ and $\widehat{\boldsymbol{Y}}$ and hence can *mimic* the oracle.

In particular, from Lemma 5.1 it follows that $d_{n,j} = n\operatorname{Var}\{\widehat{\theta}_j\}$. On the other hand, it is not hard to show that

$$\operatorname{Var}\{\widehat{\theta}_j\} = \operatorname{Var}\left\{\frac{1}{n}\sum_{i=1}^{n} \varphi_j(X_i)\widehat{Y}_i\right\} = \frac{1}{n}\operatorname{Var}\left\{\varphi_j(X)\widehat{Y}\right\}.$$

Hence, $n^{-1}d_{n,j} = n^{-1}\operatorname{Var}\{\varphi_j(X)\widehat{Y}\}$. Now, because

$$\widehat{V}_{n,j} = \frac{1}{n-1}\sum_{i=1}^{n}\left(\varphi_j(X_i)\widehat{Y}_i - \widehat{\theta}_j\right)^2$$

is an unbiased estimator of $\operatorname{Var}\{\varphi_j(X)\widehat{Y}\}$, $n^{-1}\widehat{V}_{n,j}$ is an unbiased estimator of $n^{-1}d_{n,j}$. Moreover, as can be easily shown, $\widehat{\theta}_j^2 - n^{-1}\widehat{V}_{n,j}$ is an unbiased estimator of θ_j^2. Hence, the FC can estimate each summand in (14) by $2n^{-1}\widehat{V}_{n,j} - \widehat{\theta}_j^2$. Choosing an increasing sequence $\{J_n\}$ satisfying (8), the FC can now select the cutoff $\widehat{J}_n^*$ in a data-driven way as

$$\widehat{J}_n^* = \arg\min_{0 \leq J \leq J_n} \sum_{j=0}^{J}\left(2n^{-1}\widehat{V}_{n,j} - \widehat{\theta}_j^2\right). \tag{15}$$

We have found that this adaptive procedure for choosing the cutoff leads to better empirical performance compared to simply using $\{J_n\}$ (see Section VII).

VI. Minimax Optimality

We now demonstrate that, as far as MSE is concerned, our scheme is optimal in a certain sense. We consider the case when $\mathcal{X} = [0,1]^d$ and $\Pr_X$ is the uniform distribution on $\mathcal{X}$, i.e., the sensors are deployed in a cube uniformly at random. We focus on the standard Gaussian noise model $Y = f(X) + Z$, where $Z \sim \operatorname{Normal}(0, \sigma^2)$ is independent of X, and the variance σ^2 is known to the FC. The only available information about f is that it lies in some infinite-dimensional function space $\mathcal{F}$. Here, $\eta(x) = \mathbb{E}\{Y|X = x\} = f(x)$. Hence, we are interested in $\operatorname{MSE}(\widehat{f}_n, f)$.

Consider estimating f based on an i.i.d. sequence $(X_1, Y_1), \ldots, (X_n, Y_n)$ drawn from $\Pr_{XY}$. According to the *minimax paradigm* (see, e.g., [28] or [1] and references therein), we seek sequences of estimators whose MSE approaches asymptotically the MMSE for the hardest-to-estimate function in $\mathcal{F}$. Specifically, if we have a decreasing sequence $\{a_n\}$ of nonnegative integers, such that the *minimax risk*[2]

$$R_n(\mathcal{F}) = \inf_{\widehat{f}_n} \sup_{f \in \mathcal{F}} \operatorname{MSE}(\widehat{f}_n, f) \succeq a_n \tag{16}$$

[2]The notation $a_n \succeq b_n$ means that $a_n \geq b_n$ for n sufficiently large; $a_n \preceq b_n$ is defined similarly.

then any sequence of estimators $\{\widehat{f}_n^*\}$ satisfying $\sup_{f\in\mathcal{F}}\mathrm{MSE}(\widehat{f}_n^*, f) \preceq Ca_n$ with some constant $C > 0$ is said to be *minimax optimal* for $\mathcal{F}$. The sequence $\{a_n\}$ then characterizes the minimax rate of convergence. Note that the minimization domain in (16) includes all estimators based on $(X_1, \widehat{Y}_1), \ldots, (X_n, \widehat{Y}_n)$. In the following, we show that the projection estimator (10) is minimax optimal for two function classes commonly used in the theory of nonparametric estimation, namely analytic and Lipschitz. In each case, the corresponding minimax lower bound is attained by choosing the cutoff to minimize the sum of the variance and the bias terms in the MSE. We also remark that the same reasoning can be used to show that our scheme attains minimax optimality on any function class for which there exists a minimax estimator that uses the unquantized samples to estimate the Fourier coefficients in a suitable basis at the "parametric" $O(1/n)$ rate. This includes, e.g., Sobolev, Hölder, and Besov spaces [1], [27].

A. Analytic Functions

Suppose f lies in the class $A_{\boldsymbol{\gamma},M}$, $M > 0$ and $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_d)$ with each $\gamma_l > 0$, which consists of all functions $h : \mathbb{R}^d \to \mathbb{R}$ that are 1-periodic in each of their arguments and can be analytically continued from $\mathbb{R}^d$ to the strip $S_{\boldsymbol{\gamma}} = \{\boldsymbol{z} \in \mathbb{C}^d : |\mathrm{Im} z_l| < \gamma_l, l = 1, \ldots, d\}$ in such a way that $|h| \le M$ on $S_{\boldsymbol{\gamma}}$. Define the trigonometric basis for $L^2([0,1])$:

$$\left.\begin{array}{l}\varphi_0(x) = 1,\\ \varphi_{2j-1}(x) = \sqrt{2}\sin(2\pi jx)\\ \varphi_{2j}(x) = \sqrt{2}\cos(2\pi jx)\end{array}\right\} \quad j = 1, 2, \ldots \tag{17}$$

and consider the tensor-product basis $\Phi = \{\varphi_{\boldsymbol{j}}\}$ in $L^2([0,1]^d)$, where $\boldsymbol{j} = (j_1, j_2, \ldots, j_d)$ with each $j_l = 0, 1, \ldots$, and $\varphi_{\boldsymbol{j}} = \varphi_{j_1}\varphi_{j_2}\ldots\varphi_{j_d}$. Then the Fourier coefficients $\{\theta_{\boldsymbol{j}}\}$ of f in Φ satisfy $|\theta_{\boldsymbol{j}}| \le Ce^{-\boldsymbol{\gamma}\cdot\boldsymbol{j}}$, where $\boldsymbol{\gamma}\cdot\boldsymbol{j} = \gamma_1 j_1 + \ldots + \gamma_d j_d$, and C is a constant that depends on M and d, but not on f [29]. It can be shown that $R_n(A_{\boldsymbol{\gamma},M}) \succeq C'(\ln n)^d/n$, where C' is a constant that depends on $\boldsymbol{\gamma}$, d and M [29]. Now consider the projection estimator $\widehat{f}_n$ of f from (10). Choosing $J_l^0 = 2\lfloor(1/2)\gamma_l^{-1}\ln n\rfloor$ for $l = 1, 2, \ldots, d$, we see that the cutoff $J_n = J_1^0 \ldots J_d^0$ satisfies condition (8). Thus

$$\mathrm{MSE}(\widehat{f}_n, f) \le C_1\frac{(\ln n)^d}{n} + \sum_{j_1 > J_1^0}\cdots\sum_{j_d > J_d^0}|\theta_{\boldsymbol{j}}|^2 \le C_2\frac{(\ln n)^d}{n}$$

where the first inequality follows from Theorem 5.1, and the second from the bound on $|\theta_{\boldsymbol{j}}|$ for $f \in A_{\boldsymbol{\gamma},M}$, The constants C_1 and C_2 depend on d, $\boldsymbol{\gamma}$, M and σ^2. Hence, $\{\widehat{f}_n\}$ is minimax optimal for $A_{\boldsymbol{\gamma},M}$.

B. Lipschitz Functions

Another commonly used function space is the space of Lipschitz functions. Let $d = 1$, and suppose that f belongs to the class $\mathrm{Lip}_{r,\alpha,M}$ for some $M > 0$, $r \in \{0, 1, 2, \ldots\}$ and $\alpha \in (0, 1]$, which consists of all bounded, 1-periodic functions $h : \mathbb{R} \to \mathbb{R}$ satisfying $|h^{(r)}(x + y) - h^{(r)}(x)| \le M|y|^\alpha$ for all $x, y \in \mathbb{R}$. Here, $h^{(r)}$ is the rth derivative of h. Then the Fourier coefficients $\{\theta_j\}$ of f in the trigonometric basis (17) satisfy $\sum_{j>J}^{\infty}\theta_j^2 \le CJ^{-2\beta}$, where $\beta \triangleq r + \alpha$, and C is a constant that depends on r, α and M, but not on f [30]. When $\beta > 1/2$, it can be shown that $R_n(\mathrm{Lip}_{r,\alpha,M}) \succeq Cn^{-2\beta/(2\beta+1)}$ [1]. Conversely, if we choose $J_n = 2\lfloor n^{1/(2\beta+1)}\rfloor$, then condition (8) is satisfied, and

$$\mathrm{MSE}(\widehat{f}_n, f) \le C_1 n^{-\frac{2\beta}{2\beta+1}} + \sum_{j>J_n}^{\infty}\theta_j^2 \le C_2 n^{-\frac{2\beta}{2\beta+1}}$$

where the first inequality follows from Theorem 5.1, and the second from the bound on $\sum_{j>J}\theta_j^2$ for $f \in \mathrm{Lip}_{r,\alpha,M}$. The constants C_1 and C_2 depend on r, α, M and σ^2. Hence, $\{\widehat{f}_n\}$ is minimax optimal for $\mathrm{Lip}_{r,\alpha,M}$ with $\beta > 1/2$. Similar bounds can be proved, e.g., for functions on $\mathbb{R}^d$ that are Lipschitz in each of their arguments.

VII. Experiments

We have tested the performance of our scheme via simulations. In this section, we report the results for two observation models:

1) additive Gaussian noise: $Y = f(X) + Z$;
2) multiplicative Gaussian noise: $Y = f(X)(1 + Z)$.

The underlying domain is the unit square $[0,1]^2$, and $Z \sim \mathrm{Normal}(0, \sigma^2)$ is independent of X with $\sigma^2 = 0.2$. The function f, shown in Fig. 5 (left), is given by $f(x) = \exp(-x_1^2 - 5x_2) + \sin(10(x_1 - x_2)) + \cos(5x_1 + x_2) + \sin(4\pi x_1) - \cos(6\pi x_2)$, where x_1 and x_2 are the coordinates of a point $x \in [0,1]^2$. This function is a linear combination of a number of sinusoids and a rapidly decaying exponential term, and is Lipschitz with $r = 0$ and $\alpha = 1$, i.e., $\beta = 1$.

We experimented with three network sizes $n = 100$, 500, 1000, where in each case n sensors were placed in $[0,1]^2$ uniformly at random. The projection estimator was based on the tensor-product basis formed from the trigonometric basis (17) for $L^2([0,1])$. Cutoffs were selected using two methods: In the first, we took the minimax-optimal cutoffs $J_n = \lfloor n^{1/(2\beta+2)}\rfloor$, while in the second, we used an adaptive procedure based on (15). For each n, the positions of the sensors and the dither signals were discretized by independently quantizing them into q cells, where $q = 3\lfloor\sqrt{\log n}\rfloor$. Simulation results are shown in Fig. 3. We now highlight the main insights.

For additive noise, Fig. 3(a) shows that, for a given value of ε, the average number of bits per sensor is above the theoretical bound of Theorem 4.1 (note that in this case the conditional rate-distortion function is simply the rate-distortion function of the Gaussian noise; see Appendix A for details). The gap between the empirical performance and the theoretical bound is a combination of an approximation error due to discretization of the sensor locations and the dither signals, and an estimation error of the universal entropy coder. However, this gap closes as we increase the number of sensors or decrease the quantizer resolution. For multiplicative Gaussian noise [Fig. 3(b)], we did not compute the conditional rate-distortion function. However, the empirical results show that the bit rates are already quite low (for $n = 1000$ and $\varepsilon \ge 0.3$, the bit rate is less than 2 bits per sensor).

Fig. 3(c) and (d) shows the effective alphabet size (i.e., $\log N$) versus ε for the chosen values of n both for additive and multiplicative noise. We find, as expected from our theory, that the effective alphabet size increases as we increase the network size

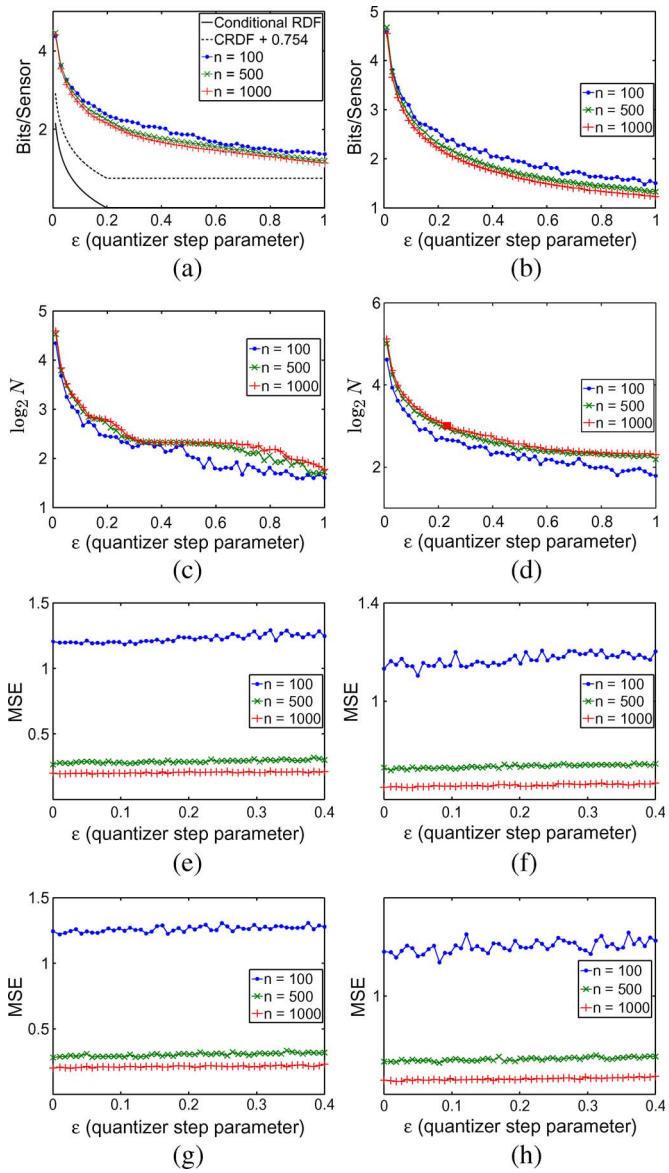

Fig. 3. Simulation results: average bit rate per sensor versus ε (a) with additive Gaussian noise and (b) with multiplicative Gaussian noise; effective alphabet size versus ε (c) with additive Gaussian noise and (d) multiplicative Gaussian noise; MSE versus ε (e) with adaptively selected cutoffs and (f) with minimax-optimal cutoffs in additive Gaussian noise; MSE versus ε (g) with adaptively selected cutoffs and (h) with minimax-optimal cutoffs in multiplicative Gaussian noise.

(for a given quantizer resolution), whereas decreasing the quantizer resolution (for a given network size) decreases the effective alphabet size. In general, for a given quantizer resolution and network size, the effective alphabet size is slightly higher in the presence of multiplicative noise than in the presence of additive noise. Note also that, although N increases with n, the average number of bits transmitted by the network decreases with n and, in particular, is much smaller than $\log N$ for large n and ε.

The encoding algorithm of Section IV is based on a distributed implementation of the KT estimator, where the sensors exchange messages to update their conditional probability models used for lossless coding of quantizer indices. On average, the number of outgoing bits should decrease as messages propagate from lower numbered sensors to higher numbered

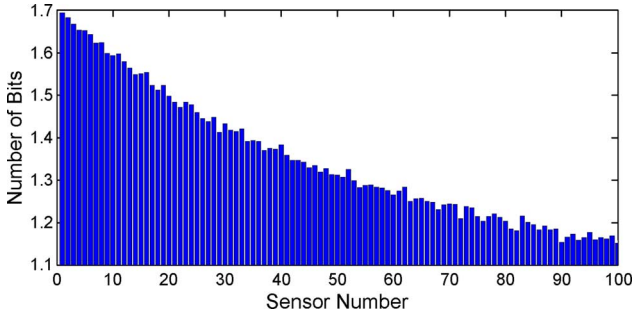

Fig. 4. Average bits per sensor versus sensor number for network size $n = 100$.

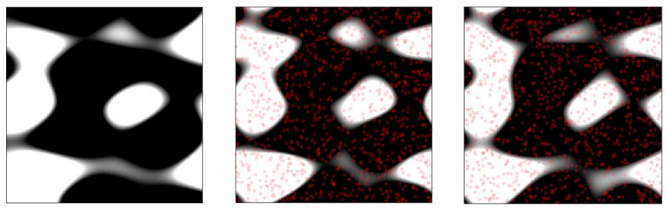

Fig. 5. The original function (left) and its reconstructions using adaptive cutoff selection for the additive (middle) and the multiplicative (right) Gaussian noise models ($\varepsilon = 0.2, n = 1000$). The dots indicate the sensor locations.

ones, as shown in Fig. 4 for $n = 100$ and $\varepsilon = 0.2$. Notice that the average number of bits per sensor is below 2.

Fig. 3(e)–(h) shows that the degradation of the MSE due to quantization is not very appreciable (the MSE versus ε curves are essentially flat). This makes our scheme attractive for situations that call for low communication rates, since we can use low-resolution quantizers in dense networks. We also find that selecting cutoffs adaptively leads to significantly better performance compared to nonadaptive cutoff choice, especially for network sizes $n = 500$ and 1000. For example, in the additive noise case, the adaptive MSE is $\approx$0.3 versus the nonadaptive MSE of $\approx$0.72 for $n = 500$, while for $n = 1000$ the adaptive MSE is $\approx$0.2 versus the nonadaptive MSE of $\approx$0.65. Similar comparisons can be made for the multiplicative case as well, which indicates that choosing cutoffs adaptively based on the data leads to superior performance.

Finally, Fig. 5 shows the reconstructions of the original function (left) with adaptively chosen cutoffs for additive (middle) and multiplicative (right) noise, for $n = 1000$ and $\varepsilon = 0.2$. Note that the reconstruction is not only good in the global sense of having small MSE, but also retains all the major features of the underlying function.

VIII. Summary and Future Directions

We have proposed, analyzed and evaluated a novel scheme for rate-constrained distributed nonparametric regression using a wireless sensor network. The scheme is as follows:

- **low-complexity**, requiring the sensors to carry out standard signal-processing tasks such as uniform scalar quantization, as well as simple in-network message passing;
- **universal**, requiring very minimal assumptions on the joint probability distribution of the locations and the measurements of the sensors;
- **minimax optimal**, attaining minimax MSE convergence rates for commonly used smoothness classes.

TABLE I
A COMPARISON OF SOME PRIOR WORK WITH THE SCHEME OF THIS PAPER

	DSC	Predd et al.	Wang–Ishwar	**this paper**
nonparametric	no	yes	yes	**yes**
rate constraint	fixed	$\log 3$ bps*	1 bps	**variable**
MSE criterion	global	pointwise	global	**global**
nonadditive noise?	no	yes	no	**yes**
unbounded noise?	yes	yes	no	**yes**

* bps = bits per sensor.

A particularly attractive feature of our scheme is that it can support low communication rates yet remain minimax optimal. Hence, in a sufficiently dense network, the sensors can quantize their measurements at very low resolutions, leading to significant savings of communication resources. Our simulations show that the empirical performance of the scheme is close to that predicted by the theory and that the effect of the quantization on the MSE is not very significant when the network is sufficiently dense. This robustness to quantization is essentially due to random dithering. The necessity of using randomized quantizers in distributed estimation has been already recognized, for example, by Xiao and Luo [31] in the parametric setting and, more recently, by Wang and Ishwar [9] in the nonparametric setting. Table I gives a comparative listing of the features of our scheme versus the parametric schemes based on distributed source coding (DSC) [10]–[13] and the nonparametric schemes of Predd *et al.* [6] and Wang and Ishwar [9].

We close by outlining some directions for future work. First, we would like to develop extensions of our scheme that would make it resilient against channel noise and sensor localization errors. Also, our distributed encoding procedure assumes that a routing path has been established in the network beforehand. To eliminate this dependence on routing, we would like to design an efficient iterative procedure in the spirit of belief propagation techniques [19], where every sensor would broadcast its current probability assignment to all sensors in a small neighborhood around it. Finally, a direction worth pursuing is using wireless sensor networks for nonparametric distributed regression of time-varying spatial fields, allowing for dependence across time.

APPENDIX A
CONDITIONAL RATE-DISTORTION FUNCTION

Let Y be a real-valued random variable. Given $d \geq 0$, let $\mathcal{R}_Y(d)$ be the set of all random variables W jointly distributed with Y and satisfying the second-moment constraint $\mathbb{E}\{(Y-W)^2\} \leq d$. The rate-distortion function of Y, defined as [23]

$$R_Y(d) \triangleq \inf_{W \in \mathcal{R}_Y(d)} I(Y;W) \tag{A1}$$

where $I(Y;W)$ is the mutual information between Y and W, is the minimum number of bits needed to describe Y in order to reconstruct it with a MSE of d. For example, if Y is Gaussian with variance σ^2, then [23]

$$R_Y(d) = \begin{cases} \frac{1}{2}\log\frac{\sigma^2}{d}, & d \leq \sigma^2 \\ 0, & d > \sigma^2. \end{cases}$$

Let X be some other random variable jointly distributed with Y, and let $\mathcal{R}_{Y|X}(d)$ denote the set of all random variables W jointly distributed with (X,Y) and satisfying the constraint $\mathbb{E}\{(Y-W)^2\} \leq d$. If both the description and the reconstruction of Y can also depend on X, then the minimum required number of bits is given by the *conditional rate-distortion function* [32], [33]

$$R_{Y|X}(d) \triangleq \inf_{W \in \mathcal{R}_{Y|X}(d)} I(Y;W|X) \tag{A2}$$

where $I(Y;W|X)$ is the conditional mutual information between Y and W given X. The minimization domain in (A2) includes all random variables W that are independent of X, for which $I(Y;W|X) = I(Y;W)$. Moreover, the minimization domain in (A1) can be limited to W independent of X. Hence, $R_{Y|X}(d) \leq R_Y(d)$. A useful fact about conditional rate-distortion functions is that, for any function g of X, $R_{Y'|X}(d) = R_{Y|X}(d)$, where $Y' = Y - g(X)$ [32]. Moreover, if Y' is independent of X, then $R_{Y'|X}(d) = R_{Y'}(d)$. As an example, if $Y = f(X) + U$, where f is a deterministic function of X and U is independent of X, then $R_{Y|X}(d) = R_U(d)$, the rate-distortion function of U.

APPENDIX B
PROOFS

A. *Proof of Theorem 4.1*

We closely follow the arguments of Zamir and Feder [21]. Let $V = Y + U$. We start by proving that

$$H(M|X,U) = I(Y;V|X) = H(V|X) - \log 2\sqrt{3\varepsilon} \tag{B3}$$

where $I(Y;V|X)$ is the conditional mutual information between Y and V given X. For fixed realizations $X = x$ and $U = u$, let $p_m(x,u) \triangleq \Pr(M = m \big| X = x, U = u)$. Then

$$\begin{aligned} p_m(x,u) &= \Pr\left(|Y + u - c_m| \leq \sqrt{3\varepsilon} | X = x\right) \\ &= \int_{c_m - \sqrt{3\varepsilon} - u}^{c_m + \sqrt{3\varepsilon} - u} p_{Y|X}(y|x)dy \end{aligned} \tag{B4}$$

where we have defined $c_m \triangleq 2m\sqrt{3\varepsilon}$. The conditional pdf $p_{V|X}$ of $V = Y + U$ given X is just the convolution

$$p_{V|X}(v|x) = \frac{1}{2\sqrt{3\varepsilon}} \int_{v-\sqrt{3\varepsilon}}^{v+\sqrt{3\varepsilon}} p_{Y|X}(y|x)dy. \tag{B5}$$

Comparing (B4) with (B5), we can write $p_m(x,u) = 2\sqrt{3\varepsilon} q_m(x,u)$ with $q_m(x,u) = p_{V|X}(c_m - u)$. Using these facts and the definition of the conditional entropy, it is easy to show that

$$H(M|X,U) = H(V|X) - \log 2\sqrt{3\varepsilon}.$$

Next, we show that $H(M|X,U)$ can be bounded by

$$H(M|X,U) \leq R_{Y|X}(\varepsilon) + C \tag{B6}$$

where

$$C \stackrel{\Delta}{=} \sup\left\{I(Y;V|X)|p_{YX}\text{s.t.}\mathbb{E}\{Y^2\} \le \varepsilon\right\}. \tag{B7}$$

The supremum is over all joint pdf's of X and Y such that the second moment of Y is bounded by ε. Consider any W jointly distributed with (X,Y) independently of U and satisfying

$$\mathbb{E}\left\{(Y-W)^2\right\} \le \varepsilon. \tag{B8}$$

Now, from (B3) we have $H(M|X,U) = I(Y;V|X)$. Hence, $H(M|X,U) - I(Y;W|X) = I(Y;V|X) - I(Y;W|X)$. Following now exactly the same steps as in [21], but with an additional conditioning on X, we get the bound $H(M|X,U) \le I(Y;W|X) + I(Y-W;V-W|X)$. Now, because W satisfies (B8), $I(Y-W;V-W|X) \le C$ by the definition of C. Hence, $H(M|X,U) \le I(Y;W|X) + C$ holds for every W satisfying (B8), including the W that achieves the conditional rate-distortion function (which can be chosen independently of U). This proves (B6).

It remains to derive an upper bound on C. The supremum in (B7) is over all (X,Y) such that $\mathbb{E}\{Y^2\} \le \varepsilon$. The dither U satisfies $\mathbb{E}\{U^2\} = \varepsilon$, as well, so we see that the channel output $V = Y + U$ must satisfy the constraint

$$\mathbb{E}\{V^2\} = \mathbb{E}\{Y^2\} + \mathbb{E}\{U^2\} \le 2\varepsilon. \tag{B9}$$

Using standard information-theoretic identities, we can write $I(Y;V|X) = H(V|X) - H(U) \le H(V) - H(U)$, where $H(U) = \log 2\sqrt{3\varepsilon}$. Thus, $C \le \sup\{H(V) - H(U)|\mathbb{E}\{V^2\} \le 2\varepsilon\}$. The random variable V^* that maximizes $H(V)$ under the constraint (B9) is Gaussian with $H(V^*) = (1/2)\log(2\pi e \cdot 2\varepsilon)$. Hence, $C \le (1/2)\log(2\pi e \cdot 2\varepsilon) - (1/2)\log 12\varepsilon = (1/2)\log(\pi e/3)$, which is approximately equal to 0.754 bits per sample.

B. Proof of Theorem 4.2

Note first that the KT estimators $\widehat{P}_{KT}^{(i)}$, the predictive distributions $\widehat{P}^{(i)}$, and the "true" joint distribution P_{MS} of M and $S = (\widetilde{X}, \widetilde{U})$ are all conditioned on N. Hence, all expectations w.r.t. $\boldsymbol{M}$ and $\boldsymbol{S}$ are actually conditional expectations given N, but we shall omit this conditioning for brevity. We shall also suppress the dependence of $\widehat{P}_{KT}^{(i)}$ and $\widehat{P}^{(i)}$ on (m^{i-1}, s^{i-1}) and (m^{i-1}, s^i), respectively.

Recall the definition of the *information divergence* (or relative entropy) between two probability distributions P and Q over a finite or countable set [23, ch. 2]: $D(P\|Q) \stackrel{\Delta}{=} \mathbb{E}_P\{\log(P/Q)\}$. We begin by writing down the key property of the KT estimator [26]: there exists a universal constant $C > 0$ such that

$$\frac{1}{n}\sum_{i=1}^{n} D\left(P_{M_iS_i}\|\widehat{P}_{KT}^{(i)}\right) \le \frac{CA\log n}{n} \tag{B10}$$

where $A = Nq^2 \le cN\log n$. This bound holds for every realization of N. Next, we write

$$\frac{P_{M_iS_i}(m_i,s_i)}{\widehat{P}_{KT}^{(i)}(m_i,s_i)} = \frac{P_{M_i|S_i}(m_i|s_i)}{\widehat{P}^{(i)}(m_i)}\frac{P_{S_i}(s_i)}{\widehat{Q}^{(i)}(s_i)} \quad \forall i$$

where we have defined the marginal distributions $\widehat{Q}^{(i)}(s_i) \stackrel{\Delta}{=} \sum_{m=1}^{N}\widehat{P}_{KT}^{(i)}(m,s_i)$. Taking logarithms, summing over i, and taking the expectation with respect to $\boldsymbol{M}$ and $\boldsymbol{S}$, we get

$$\begin{aligned}\sum_{i=1}^{n} & D\left(P_{M_iS_i}\|\widehat{P}_{KT}^{(i)}\right) \\ &= \sum_{i=1}^{n}\mathbb{E}_{S^i}\left\{D\left(P_{M_i|S_i}\|\widehat{P}^{(i)}\right) + D\left(P_{S_i}\|\widehat{Q}^{(i)}\right)\right\} \\ &\ge \sum_{i=1}^{n}\mathbb{E}_{S^i}D\left(P_{M_i|S_i}\|\widehat{P}^{(i)}\right)\end{aligned} \tag{B11}$$

where the inequality follows from the fact that the divergence is nonnegative [23, ch. 2]. On the other hand

$$\begin{aligned}\frac{1}{n}\mathbb{E}&\left\{-\sum_{i=1}^{n}\log\widehat{P}^{(i)}(M_i)\right\} - H(M|\widetilde{X},\widetilde{U}) \\ &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{S^i}D\left(P_{M_i|S_i}\|\widehat{P}^{(i)}\right) \\ &\le \frac{1}{n}\sum_{i=1}^{n}D\left(P_{M_iS_i}\|\widehat{P}_{KT}^{(i)}\right).\end{aligned} \tag{B12}$$

Combining (B10)–(B12), we get (5), which proves the theorem.

C. Proof of Theorem 4.3

The overall binary message sent to the FC consists of two parts: 1) the Elias encoding of $\underline{m}$ and $\overline{m}$ and 2) the concatenation of the individual binary messages sent by the sensors. The lengths of both of these are determined by $N = \overline{m} - \underline{m} + 1$. We have

$$\begin{aligned}\mathbb{E}\{N\} &\le \frac{1}{2\sqrt{3\varepsilon}}\mathbb{E}\left\{\max_{1\le i\le n}|Y_i|\right\} \\ &= \frac{1}{2\sqrt{3\varepsilon}}\mathbb{E}\left\{\mathbb{E}\left\{\max_{1\le i\le n}|Y_i|\middle|\boldsymbol{X}\right\}\right\}.\end{aligned}$$

Now, each Y_i, given X_i, is sub-Gaussian with parameters λ and Λ. Using Jensen's inequality and (1), we have

$$\begin{aligned}\exp&\left(t\mathbb{E}\left\{\max_{1\le i\le n}|Y_i|\middle|\boldsymbol{X}\right\}\right) \\ &\le \mathbb{E}\left\{\max_{1\le i\le n}e^{t|Y_i|}\middle|\boldsymbol{X}\right\} \\ &\le \sum_{i=1}^{n}\mathbb{E}\left\{e^{t|Y_i|}\middle|X_i\right\} \\ &\le \sum_{i=1}^{n}\left(\mathbb{E}\left\{e^{-tY_i}|X_i\right\} + \mathbb{E}\left\{e^{tY_i}|X_i\right\}\right) \\ &\le 2n\Lambda e^{\lambda^2t^2/2}\end{aligned}$$

for any $t > 0$. Therefore

$$\mathbb{E}\left\{\max_{1\le i\le n}|Y_i|\middle|\boldsymbol{X}\right\} \le \frac{\ln(2\Lambda n)}{t} + \frac{\lambda^2 t}{2}$$

for every $t > 0$. Choosing t to minimize the right-hand side of this expression, we get

$$\mathbb{E}\left\{\max_{1\le i\le n}|Y_i|\Big|\boldsymbol{X}\right\} \le \lambda\sqrt{2\ln(2\Lambda n)}.$$

Hence, $\mathbb{E}\{N\} \le (\lambda\sqrt{2\ln(2\Lambda n)}/2\sqrt{3\varepsilon})$. The average number of bits needed to transmit the Elias encodings of $\underline{m}$ and $\overline{m}$ to the FC is on the order of $O(\mathbb{E}\{\log N\})$, which can in turn be bounded as $O(\log\mathbb{E}\{N\}) = O(\log\lambda) + O(\log\log n) + O(\log(1/\varepsilon))$. Moreover, by Theorem 4.2, the overall expected length of the individual binary messages from each sensor to the FC is bounded by

$$nH(M|\widetilde{X},\widetilde{U})+C\mathbb{E}\{N\}\log^2 n \le nH(M|\widetilde{X},\widetilde{U})+C'\frac{\sqrt{\log^5 n}}{\sqrt{\varepsilon}},$$

where C' depends on λ and Λ. Because the densities $p_X, p_{Y|X}, p_U$ are continuous functions of their arguments on the interiors of their respective domains, and because the deterministic discretizations $X \mapsto \widetilde{X}$ and $U \mapsto \widetilde{U}$ get increasingly fine as $n \to \infty$, the conditional entropy $H(M|\widetilde{X},\widetilde{U})$ can be written as $H(M|X,U) + \Delta_n$, where $\Delta_n = o(1)$. By Theorem 4.1, we have $H(M|X,U) \le R_{Y|X}(\varepsilon) + 0.754$ bits per sample. Putting together all these facts, we conclude that (6) holds with some $K_1(\lambda,\Lambda) > 0$. Finally, the average communication complexity of the encoder (i.e., the number of messages exchanged by the sensors) is

$$O(n) + O\left(\mathbb{E}\{N\}n\log n\right) \le \frac{K_2(\lambda,\Lambda)n\sqrt{\log^3 n}}{\sqrt{\varepsilon}}$$

for some $K_2(\lambda,\Lambda) > 0$. This yields (7).

D. Proof of Lemma 5.1

We begin by noting that

$$\begin{aligned}\theta_j &= \mathbb{E}\left\{\varphi_j(X)\eta(X)\right\}\\ &= \mathbb{E}\left\{\varphi_j(X)\mathbb{E}\{Y|X\}\right\}\\ &= \mathbb{E}\left\{\varphi_j(X)Y\right\}\\ &= \mathbb{E}\left\{\frac{1}{n}\sum_{i=1}^{n}\varphi_j(X_i)Y_i\right\}\end{aligned}$$

where in the last step we have used the fact that (X_i, Y_i) are i.i.d. according to $\Pr_{XY}$. We shall also need the fact that the output $\widehat{Y} = Q_\varepsilon(Y+U) - U$ of a dithered uniform quantizer can be written as $\widehat{Y} = Y + E$, where the additive noise E satisfies $\mathbb{E}\{E|X\} = 0$, $\mathbb{E}\{YE|X\} = 0$ and $\mathbb{E}\{E^2\} = \varepsilon$ (see, e.g., Lemmas 1 and 2 in [20]). Then

$$\begin{aligned}\mathbb{E}\left\{(\widehat{\theta}_j - \theta_j)\right\} &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left\{\varphi_j(X_i)E_i\right\}\\ &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left\{\varphi_j(X_i)\mathbb{E}\{E_i|X_i\}\right\}\\ &= 0.\end{aligned}$$

Also,

$$\begin{aligned}\mathbb{E}\left\{(\widehat{\theta}_j - \theta_j)^2\right\} &= \mathrm{Var}\{\widehat{\theta}_j\}\\ &= \mathrm{Var}\left\{\frac{1}{n}\sum_{i=1}^{n}\varphi_j(X_i)\widehat{Y}_i\right\}\\ &= \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}\left\{\varphi_j(X_i)(Y_i+E_i)\right\}\\ &\le \frac{\|\varphi_j\|_\infty^2(\mathrm{Var}\,Y+\varepsilon)}{n}.\end{aligned}$$

The lemma is proved.

E. Proof of Theorem 5.1

From Parseval's identity,

$$\int_{\mathcal{X}}\left(\widehat{f}_n(x)-\eta(x)\right)^2 d\Pr_X(x) = \sum_{j=0}^{J_n}(\widehat{\theta}_j-\theta_j)^2 + \Gamma_n.$$

Taking expectations and applying Lemma 5.1, we get

$$\begin{aligned}\mathrm{MSE}(\widehat{f}_n,\eta) &= \sum_{j=0}^{J_n}\mathbb{E}\left\{(\widehat{\theta}_j-\theta_j)^2\right\} + \Gamma_n\\ &\le \frac{L_n(\mathrm{Var}\,Y+\varepsilon)}{n} + \Gamma_n.\end{aligned}$$

Noting that $\mathrm{Var}Y$ can be bounded by some constant $K_3(\lambda,\Lambda)$, we get (11). This, together with the decay condition (8) and the fact that $\Gamma_n \to 0$ as $J_n \to \infty$, in turn implies the mean-square convergence of $\widehat{f}_n$ to η.

ACKNOWLEDGMENT

The authors would like to thank the referees for their constructive criticism and for making numerous suggestions, which helped greatly improve the paper.

REFERENCES

[1] S. Efromovich, *Nonparametric Curve Estimation: Methods, Theory and Applications*. New York: Springer, 1999.
[2] L. Györfi, M. Kohler, A. Krzyżak, and H. Walk, *A Distribution-Free Theory of Nonparametric Regression*. New York: Springer, 2002.
[3] L. Doherty, B. A. Warneke, B. E. Boser, and K. S. J. Pister, "Energy and performance considerations for smart dust," *Int. J. Parallel Distribut. Syst. Netw.*, vol. 4, no. 3, pp. 121–1331, 2001.
[4] S. N. Simić, "A learning-theory approach to sensor networks," *IEEE Pervasive Comput.*, vol. 2, no. 4, pp. 44–49, 2003.
[5] T. Poggio and S. Smale, "The mathematics of learning: Dealing with data," *Notices Amer. Math. Soc.*, vol. 50, pp. 537–544, 2003.
[6] J. B. Predd, S. R. Kulkarni, and H. V. Poor, "Consistency in models for distributed learning under communication constraints," *IEEE Trans. Inf. Theory*, vol. 52, no. 1, pp. 52–63, Jan. 2006.
[7] J.-J. Xiao, A. Ribeiro, Z.-Q. Luo, and G. B. Giannakis, "Distributed compression-estimation using wireless sensor networks," *IEEE Signal Process. Mag.*, vol. 23, no. 4, pp. 27–41, Jul. 2006.
[8] J.-J. Xiao, Z.-Q. Luo, and G. B. Giannakis, "Performance bounds for the rate-constrained universal decentralized estimators," *IEEE Signal Process. Lett.*, vol. 14, no. 1, pp. 47–50, Jan. 2007.
[9] Y. Wang and P. Ishwar, "On non-parametric field estimation using randomly deployed, noisy, binary sensors," presented at the IEEE Int. Symp. Inform. Theory, Nice, France, Jun. 2007.
[10] S. S. Pradhan and K. Ramchandran, "Distributed source coding using syndromes (DISCUS): Design and construction," *IEEE Trans. Inf. Theory*, vol. 49, no. 3, pp. 626–643, Mar. 2003.

[11] D. Marco, E. J. Duarte-Melo, M. Liu, and D. L. Neuhoff, "On the many-to-one transport capacity of a dense wireless sensor network and the compressibility of its data," presented at the 2nd Int. Workshop Inform. Processing Sensor Networks, Palo Alto Research Center (PARC), CA, Apr. 22–23, 2003.
[12] S. C. Draper and G. W. Wornell, "Side information aware coding strategies for sensor networks," *IEEE J. Sel. Areas Commun.*, vol. 22, no. 6, pp. 966–976, Aug. 2004.
[13] S. Servetto, "Lattice quantization with side information: Codes, asymptotics and applications in sensor networks," *IEEE Trans. Inf. Theory*, vol. 53, no. 2, pp. 714–731, Feb. 2007.
[14] R. L. Moses, D. Krishnamurthy, and R. M. Patterson, "Self-localization for wireless networks," *EURASIP J. Appl. Signal Process.*, no. 4, pp. 348–358, 2003.
[15] N. Patwari, A. O. Hero, III, M. Perkins, N. S. Correal, and R. J. O'Dea, "Relative location estimation in wireless sensor networks," *IEEE Trans. Signal Process.*, vol. 51, no. 8, pp. 2137–2148, Aug. 2003.
[16] D. Niculescu, "Positioning in ad hoc sensor networks," *IEEE Network*, vol. 18, no. 4, pp. 24–29, Jul.–August 2004.
[17] F. Zhao, J. Liu, J. Liu, L. Guibas, and J. Reich, "Collaborative signal and information processing: An information-directed approach," *Proc. IEEE*, vol. 19, no. 8, pp. 1199–1209, Aug. 2003.
[18] M. G. Rabbatt and R. D. Nowak, "Quantized incremental algorithms for distributed optimization," *IEEE J. Sel. Areas Commun.*, vol. 23, no. 4, pp. 798–808, Apr. 2005.
[19] M. Çetin, L. Chen, J. W. Fisher, III, A. T. Ihler, R. L. Moses, M. J. Wainwright, and A. S. Willsky, "Distributed fusion in sensor networks," *IEEE Signal Process. Mag.*, vol. 23, no. 4, pp. 42–55, Jul. 2006.
[20] J. Ziv, "On universal quantization," *IEEE Trans. Inf. Theory*, vol. IT-31, no. 3, pp. 344–347, May 1985.
[21] R. Zamir and M. Feder, "On universal quantization by randomized uniform/lattice quantizers," *IEEE Trans. Inf. Theory*, vol. 38, no. 2, pp. 428–436, Mar. 1992.
[22] R. M. Gray, *Source Coding Theory*. Boston, MA: Kluwer, 1990.
[23] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.
[24] N. Merhav and M. Feder, "Universal prediction," *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2124–2147, Oct. 1998.
[25] P. Elias, "Universal codeword sets and representations of the integers," *IEEE Trans. Inf. Theory*, vol. IT-21, no. 2, pp. 194–203, Mar. 1975.
[26] R. E. Krichevsky and V. K. Trofimov, "The performance of universal encoding," *IEEE Trans. Inf. Theory*, vol. IT-27, no. 2, pp. 199–207, Mar. 1981.
[27] Y. Meyer, *Wavelets and Operators*. Cambridge, U.K.: Cambridge Univ. Press, 1992.
[28] D. L. Donoho, I. M. Johnstone, G. Kerkyacharian, and D. Picard, "Wavelet shrinkage: Asymptopia? (With Discussion)," *J. Roy. Stat. Soc. Ser. B*, vol. 57, no. 2, pp. 301–369, 1995.
[29] B. Levit and N. Stepanova, "Efficient estimation of multivariate analytic functions on cube-like domains," *Math. Methods Statist.*, vol. 13, no. 3, pp. 253–281, 2004.
[30] R. A. Devore and G. G. Lorentz, *Constructive Approximation*. New York: Springer-Verlag, 1993.
[31] J.-J. Xiao and Z.-Q. Luo, "Decentralized estimation in an inhomogeneous sensing environment," *IEEE Trans. Inf. Theory*, vol. 51, no. 10, pp. 3564–3575, Oct. 2005.
[32] R. M. Gray, "Conditional rate-distortion theory," Stanford Electronics Laboratories, Stanford, CA, Tech. Rep. 6502-2, 1972.
[33] A. D. Wyner, "The rate-distortion function for source coding with side information at the decoder II: General sources," *Inf. Control*, vol. 38, pp. 60–80, 1978.

Avon L. Fernandes (S'02–M'09) received the B.S. (with highest hons.) and M.S. degrees in electrical engineering from the University of Illinois at Urbana-Champaign in 2005 and 2008, respectively.

He is currently working at Microsoft in the Windows Live division.

Mr. Fernandes has been awarded the FMC Technologies Inc. Graduate Fellowship, I.E.C. William L. Everitt Student Award of Excellence, the Micron Technologies Undergraduate Scholarhsip, the Dad's Association Award for Scholastic Achievement, the Niccols Calvin B.M. Scholarship, and the Fred Eggers Engineering Scholarship.

Maxim Raginsky (S'99–M'00) received the B.S. and M.S. degrees in 2000 and the Ph.D. degree in 2002 from Northwestern University, Chicago, IL, all in electrical engineering.

From 2002 to 2004, he was a Postdoctoral Researcher at the Center for Photonic Communication and Computing at Northwestern University, where he pursued work on quantum cryptography and quantum communication and information theory. From 2004 to 2007, he was a Beckman Foundation Postdoctoral Fellow at the University of Illinois in Urbana-Champaign, where he carried out research on information theory, statistical learning and computational neuroscience. Since September 2007, he has been a Research Scientist at the Department of Electrical and Computer Engineering at Duke University, Durham, NC, as a Research Scientist. His interests include statistical signal processing, information theory, statistical learning and nonparametric estimation.

Todd P. Coleman (S'01–M'05) received the B.S. degrees in electrical engineering (*summa cum laude*), as well as computer engineering (*summa cum laude*) , from the University of Michigan, Ann Arbor, in 2000, and the M.S. and Ph.D. degrees in electrical engineering from the Massachusetts Institute of Technology (MIT), Cambridge, in 2002, and 2005.

During the 2005–2006 academic year, he was a Postdoctoral Scholar at MIT's Department of Brain and Cognitive Sciences and the Massachusetts General Hospital's Neuroscience Statistics Research Laboratory in computational neuroscience. Since fall 2006, he has been on the faculty in the Electrical and Computer Engineering Department and Neuroscience Program at the University of Illinois, Urbana-Champaign. His research interests include information theory, operations research, computational neuroscience, and brain–machine interfaces.

Dr. Coleman, a National Science Foundation Graduate Research Fellowship recipient, was awarded the University of Michigan College of Engineering's Hugh Rumler Senior Class Prize in 1999 and was awarded the MIT EECS Department's Morris J. Levin Award for Best Masterworks Oral Thesis Presentation in 2002. In fall 2008, he was a corecipient of the University of Illinois College of Engineering's Grainger Award in Emerging Technologies for development of a novel, practical timing-based communication technology.