A LOW-COMPLEXITY UNIVERSAL SCHEME FOR RATE-CONSTRAINED DISTRIBUTED REGRESSION USING A WIRELESS SENSOR NETWORK

Avon Loy Fernandes[1]*, Maxim Raginsky*[2]*, Todd Coleman*[1]

[1]Department of Electrical and Computer Engineering, University of Illinois, Urbana, IL 61801
[2]Department of Electrical and Computer Engineering, Duke University, Durham, NC 27708

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 trade-off between the compression rate and the MSE and demonstrate empirical performance of the scheme using simulations.

Index Terms— Sensor networks, nonparametric estimation, rate-distortion theory, message-passing algorithms

1. INTRODUCTION

Consider the problem of distributed estimation using a wireless sensor network. To save power, we should limit the amount of communication between the sensors and the fusion center by having the sensors quantize their measurements. The fusion center will use quantized data to learn the model of the phenomenon being sensed. In this paper, we adopt the minimum-mean-squared-error (MMSE) framework, where the MMSE estimator of the sensor's measurement from its location is given by the *regression function*, i.e., the conditional mean. It is often useful to model complex phenomena *nonparametrically* [1]: instead of assuming that the regression function is described by a fixed number of parameters, we suppose that it lies in some infinite-dimensional class of functions. Clearly, there is a trade-off between the compression rate and the achievable MSE. This paper proposes a scheme for rate-constrained distributed nonparametric regression using a wireless sensor network with randomly deployed sensors, which has the following attractive characteristics:

• **Universality:** very minimal assumptions are made on the joint distribution of sensor location and (noisy) measurement.

This work was performed while M. Raginsky was with the Beckman Institute for Advanced Science and Technology, University of Illinois, Urbana, IL. Support provided by the Beckman Foundation Fellowship to M.R. and by the DARPA ITMANET program via US Army RDECOM contract W911NF-07-1-0029 to T.C.

• **Low complexity:** the compression involves standard operations such as uniform quantization, 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 by the network to the fusion center, and estimate the rate at which the MSE converges to zero as the network gets denser. Our scheme has some common elements with recent work of Wang and Ishwar [8] on nonparametric distributed estimation using binary noisy sensors. However, they assume that the sensors are dispersed throughout the observation domain uniformly at random, and that the measurements are bounded and corrupted by bounded additive noise. By contrast, we can handle nonuniformly deployed sensors, as well as unbounded and nonadditive noise.

2. PROBLEM STATEMENT

The network consists of n sensors deployed over a compact spatial domain $\mathcal{X}$ according to a fixed and known probability distribution $\mathbb{P}_X$. Their measurements lie in some $\mathcal{Y} \subseteq \mathbb{R}$, and there is uncertainty concerning the conditional distribution $\mathbb{P}_{Y|X}$ of the measurement given location. The signal-plus-noise model $Y = f(X) + Z$, where f is an unknown deterministic function and Z is zero-mean and independent of X, is a special case of this set-up. Let $\mathbb{P}_{XY}$ denote the joint distribution of the location and the measurement of a sensor.

Let $\boldsymbol{X} = \{X_i\}_{i=1}^n, \boldsymbol{Y} = \{Y_i\}_{i=1}^n$ be the sensor locations and their measurements, where $(X_1, Y_1), \ldots, (X_n, Y_n)$ are n i.i.d. samples from $\mathbb{P}_{XY}$. We assume that the network and the fusion center share an n-tuple $\boldsymbol{U} = \{U_i\}_{i=1}^n$ of i.i.d. random variables (the *dither signals*), e.g., by using synchronized pseudorandom number generators, where U_i is held by the ith sensor, and that the fusion center knows $\boldsymbol{X}$. We also assume that each sensor knows its own location X_i (see, e.g., [7] on self-localization in sensor networks) and can send analog (continuous-valued) messages to neighboring sensors and binary messages of arbitrary length to the fusion center.

We assume the regression function $\eta(x) = \mathbb{E}\{Y|X = x\}$ is in $L^2(\mathcal{X}, \mathbb{P}_X)$. We think of $\eta(x)$ as the MMSE estimator of the measurement of a sensor placed at $X = x$. The task

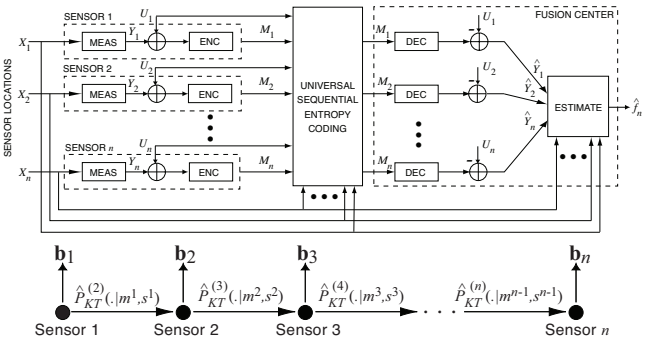

Fig. 1. The overall architecture of the scheme (top) and information flow in sequential universal entropy coding (bottom).

of the fusion center is to *estimate* (or *to learn*) η from $\boldsymbol{X}$ and a compressed version of $\boldsymbol{Y}$. The sensors must collaborate to produce a binary encoding $\boldsymbol{B}$ of $\boldsymbol{Y}$, and they are allowed to use $\boldsymbol{X}$ and $\boldsymbol{U}$ as *side information*: $\boldsymbol{B} = e_n(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{U})$, where e_n is an encoding function. The fusion center receives $\boldsymbol{B}$ and uses its knowledge of $\boldsymbol{X}$ and $\boldsymbol{U}$ to compute the *reconstruction* of $\boldsymbol{Y}$ as $\widehat{\boldsymbol{Y}} = d_n(\boldsymbol{X}, \boldsymbol{B}, \boldsymbol{U})$, where d_n is a decoding function. It then estimates η by $\widehat{f}_n = \widehat{f}_n(\boldsymbol{X}, \widehat{\boldsymbol{Y}}) \in L^2(\mathcal{X}, \mathbb{P}_X)$.

We are interested in the average number of bits transmitted by the network to the fusion center and in the MSE of the estimator, $\mathrm{MSE}(\widehat{f}_n, \eta) = \mathbb{E}\left\{ \int_{\mathcal{X}} \left(\widehat{f}_n(x) - \eta(x)\right)^2 d\mathbb{P}_X(x) \right\}$, where the expectation is with respect to $\boldsymbol{X}$, $\boldsymbol{Y}$, and $\boldsymbol{U}$.

3. DESCRIPTION OF THE SCHEME

Here is the main idea: Let $\varepsilon > 0$ be chosen in advance and revealed both to the network and to the fusion center. The dither $\boldsymbol{U} = \{U_i\}_{i=1}^n$ is an i.i.d. sequence drawn from the uniform distribution on $[-\sqrt{3\varepsilon}, \sqrt{3\varepsilon}]$ independently of $\boldsymbol{X}$ and $\boldsymbol{Y}$. Encoding consists of randomized uniform quantization of sensor measurements using step size $2\sqrt{3\varepsilon}$, followed by sequential universal entropy coding of the quantizer indices. This step is a distributed implementation of the universal quantization scheme of Ziv [11, 10] and has low communication complexity, measured by the number of analog messages exchanged among the sensors. Once the fusion center decodes the indices, it estimates the regression function using a universal orthogonal series estimator [1]. As we shall see, the use of random dither is crucial both for compression and for estimation. The overall architecture is displayed in Fig. 1 (top).

3.1. Encoding and decoding

For each $i = 1, 2, \ldots, n$, define $M_i = E_\varepsilon(Y_i + U_i)$ and $\widehat{Y}_i = D_\varepsilon(M_i) - U_i$, where $E_\varepsilon(y) = \lfloor (y + \sqrt{3\varepsilon})/2\sqrt{3\varepsilon} \rfloor$ and $D_\varepsilon(m) = 2m\sqrt{3\varepsilon}$. Note that $Q_\varepsilon = D_\varepsilon \circ E_\varepsilon$ is a uniform quantizer with step size $2\sqrt{3\varepsilon}$. The mapping $Y_i \mapsto \widehat{Y}_i$ is known as *uniform quantization with additive dither* [10]. For $1 \le i \le n$, the ith sensor computes M_i, then transmits its lossless binary encoding $\boldsymbol{B}_i$ to the fusion center. The latter receives the $\boldsymbol{B}_i$'s, decodes $\boldsymbol{M}$, and computes the $\widehat{Y}_i$'s.

To produce a good encoding of $\boldsymbol{M}$ (i.e., at an average bit rate $\approx H(M|X,U)$) without knowing the joint distribution of (M, X, U), we use a universal scheme based on *sequential probability assignment* [6]. For a distributed implementation, some message passing among the sensors is required, for a total of $O(n \log n)$ messages. Let $\boldsymbol{m} = \{m_i\}_{i=1}^n$, $\boldsymbol{x} = \{x_i\}_{i=1}^n$, and $\boldsymbol{u} = \{u_i\}_{i=1}^n$ denote the realizations of $\boldsymbol{M}$, $\boldsymbol{X}$ and $\boldsymbol{U}$. First, the sensors exchange messages to find $\underline{m} = \min m_i$ and $\overline{m} = \max m_i$. Assuming the ith sensor can send messages to sensors $(i-1) \mod n$ and $(i+1) \mod n$, $\underline{m}$ and $\overline{m}$ can each be found after no more than $8(n + n \log n)$ message passes [3]. Let $N = \overline{m} - \underline{m} + 1$. A designated sensor (say, the nth) then uses a universal encoding of the integers [2] to communicate the values of $\underline{m}$ and N to the fusion center. Assume, without loss of generality, that $\underline{m} = 1$. Then $1 \le m_i \le N$ for all i. Cover $\mathcal{X}$ by N disjoint cubes $\mathcal{C}_1, \ldots, \mathcal{C}_N$ and carve $[-\sqrt{3\varepsilon}, \sqrt{3\varepsilon}]$ into N disjoint subintervals $\mathcal{I}_1, \ldots, \mathcal{I}_N$. For $1 \le i \le n$, let $l_i = l$ if $x_i \in \mathcal{C}_l$ and $k_i = k$ if $u_i \in \mathcal{I}_k$ (we motivate this discretization procedure later). For $1 \le i \le n$, let $s_i = (l_i, k_i)$ and define

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

for all m. $\widehat{P}^{(i)}_{KT}$ is the *Krichesvky–Trofimov* (KT) estimator [4]

$$\widehat{P}^{(i)}_{KT}(m, s|m^{i-1}, s^{i-1}) = \frac{\nu(m, s|m^{i-1}, s^{i-1}) + 1/2}{i - 1 + N^3/2}, \quad (2)$$

where $\nu(m, s|m^{i-1}, s^{i-1})$ is the number of times (m, s) occurs in $(m_1, s_1), \ldots, (m_{i-1}, s_{i-1})$. Note that $\widehat{P}^{(1)}_{KT}(m, s) = 1/N^3, \forall (m, s)$, For $1 \le i \le n$, the ith sensor uses a Huffman code to encode m_i using $-\log \widehat{P}^{(i)}(m_i|m^{i-1}, s^i)$ bits. The decoding is done sequentially: having decoded m^{i-1}, the fusion center uses m^{i-1} and s^i to compute $\widehat{P}^{(i)}(\cdot|m^{i-1}, s^i)$ and generate the right codebook.

In order not to force each sensor to aggregate data from all the downstream sensors, we use the following message passing scheme. From (2) it follows that for $1 \le i < n$ $\widehat{P}^{(i+1)}_{KT}$ can be computed recursively from $\widehat{P}^{(i)}_{KT}$. Therefore, for $1 \le i < n-1$, let sensor i compute $\widehat{P}^{(i+1)}_{KT}(m, s|m^i, s^i)$ for all m, s and pass these N^3 values to sensor $i+1$. Sensor $i+1$ computes the N probabilities $\widehat{P}^{(i+1)}(\cdot|m^i, s^{i+1})$ via (1), designs a Huffman code for M_i given (M^{i-1}, S^i), and encodes m_i. The message passing is shown in Fig. 1 (bottom): horizontal arrows correspond to analog messages exchanged among the sensors, while vertical arrows depict outgoing binary messages. This requires $n-1$ message passes if the transmission of the N^3 values of the KT estimator from one sensor to another is counted as a single (analog) message. The combined communication complexity of computing $\underline{m}$ and $\overline{m}$ and the encoding of $\boldsymbol{m}$ is thus $O(n \log n)$.

3.2. Estimation of the regression function

Let $\Phi = \{\varphi_j\}_{j=0}^\infty$ be an orthonormal basis in $L^2(\mathcal{X}, \mathbb{P}_X)$. Since $\eta \in L^2(\mathcal{X}, \mathbb{P}_X)$, we can expand it in a Fourier series

$\eta(x) = \sum_{j=0}^{\infty} \theta_j \varphi_j(x)$ with $\theta_j = \int_{\mathcal{X}} \varphi_j \eta d\,\mathbb{P}_X$. Now we construct our estimator. Define $C(J) = \max_{0 \le j \le J} \sup_{x \in \mathcal{X}} |\varphi_j(x)|^2$ for all J and choose an increasing sequence $\{J_n\}_{n=1}^{\infty}$ of nonnegative reals (the *cutoffs*) satisfying the condition

$$C(J_n) J_n / n \to 0 \quad \text{as } n \to \infty. \tag{3}$$

It is satisfied, for example, by uniformly bounded bases or by wavelet bases. For every $0 \le j \le J_n$, estimate θ_j by $\widehat{\theta}_j = n^{-1} \sum_{i=1}^{n} \varphi_j(X_i) \widehat{Y}_i$ and then form the *projection estimate*

$$\widehat{f}_n(x) = \sum_{j=0}^{J_n} \widehat{\theta}_j \varphi_j(x). \tag{4}$$

The above choice of $\{J_n\}$ may lead to overfitting. To avoid this, we suggest an alternative, data-driven procedure for cutoff selection via empirical risk minimization [1]. Let $\{J_n\}$ satisfy (3). Then the fusion center can select the cutoff

$$\widehat{J}_n^* = \operatorname*{arg\,min}_{0 \le J \le J_n} \sum_{j=0}^{J} \left(2n^{-1} \widehat{V}_{n,j} - \widehat{\theta}_j^2 \right), \tag{5}$$

where

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

is an unbiased estimator of $\mathrm{Var}\{\varphi_j(X)\widehat{Y}\}$. Derivation is omitted for lack of space. We found that this adaptive rule leads to better empirical performance compared to simply using $\{J_n\}$; see Section 5.

4. PERFORMANCE ANALYSIS

We first analyze the encoding and decoding performance under the assumption that $\underline{m}, \overline{m} \ll n$. This holds, e.g., when the distribution of Y has light tails, so quantizer indices with large absolute values are unlikely, or when low-resolution quantizers are used. Then the $O(\log N)$ overhead due to sending $\underline{m}$ and N to the fusion center is negligible, and we can focus on the average number of bits needed to encode $\boldsymbol{M}$. Using properties of the KT estimator [4], we can prove that

$$n^{-1}\, \mathbb{E}\left\{ \sum_{i=1}^{n} -\log \widehat{P}^{(i)}(M_i | M^{i-1}, S^i) \Big| N \right\} \le H(M|L,K) + O\left(N^3 n^{-1} \log n\right). \tag{6}$$

We now motivate our method for discretizing $\boldsymbol{X}$ and $\boldsymbol{U}$. We would like $H(M|L,K) \approx H(M|X,U)$. Given arbitrary partitions $\{\mathcal{L}_l\}_{l=1}^{L}$ of $\mathcal{X}$ and $\{\mathcal{K}_k\}_{k=1}^{K}$ of $[-\sqrt{3\varepsilon}, \sqrt{3\varepsilon}]$, for each i let $L_i = l$ if $X_i \in \mathcal{L}_l$ and $K_i = k$ if $U_i \in \mathcal{K}_k$. This will replace N^3 in (2) and (6) with NLK. Choosing $L, K \gg N$ will give a good approximation of $H(M|X,U)$ but result in a large excess codelength, while choosing $L, K \ll N$ will keep the excess codelength low but result in a poor approximation of $H(M|X,U)$. This is akin to the trade-off between the estimation and the approximation errors in statistical inference. A good compromise is to let $L = K = N$. For large n, $H(M|L,K) \approx H(M|X,U)$ with high probability. Now, the results of [11, 10] are easily extended to cover the case of additional side information $\boldsymbol{X}$, giving the bound $H(M|X,U) \le R_{Y|X}(\varepsilon) + 0.754$, where $R_{Y|X}$ is the *conditional rate-distortion function* of Y given X [9]. The discretization is heuristic, but, as we show in Section 5, it leads to empirical performance close to our bound on $H(M|X,U)$.

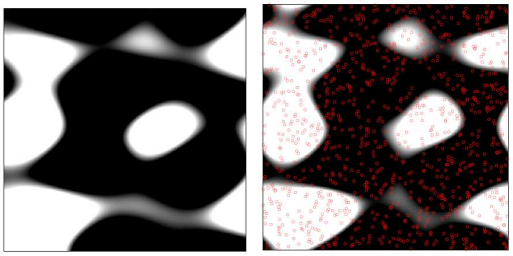

Fig. 2. Original function (left) and reconstruction (right) using adaptive cutoffs ($\varepsilon = 0.2, n = 1000$). The dots show sensor locations.

Next, we show that $\widehat{f}_n$ converges to the regression function η in the mean square sense and that quantization does not affect the rate of convergence. The key here is in the performance of the Fourier coefficient estimator $\widehat{\theta}_j$. Namely, for any j, $\widehat{\theta}_j$ is an unbiased and efficient estimator of θ_j: $\mathbb{E}\{\widehat{\theta}_j\} = \theta_j$ and $\mathbb{E}\{(\widehat{\theta}_j - \theta_j)^2\} \le C(j)(\sigma_Y^2 + \varepsilon) n^{-1}$. This follows from the properties of dithered uniform quantizers. Using this, we can show that the projection estimator (4) satisfies

$$\mathrm{MSE}(\widehat{f}_n, \eta) \le (J_n + 1) C(J_n)(\sigma_Y^2 + \varepsilon) n^{-1} + \Delta(J_n), \tag{7}$$

where $\Delta(J_n) \to 0$ as $J_n \to \infty$, and hence $\widehat{f}_n$ converges to η in the mean square sense. The only effect of quantization is to add ε to the numerator in the right-hand side of (7). That is, quantization does not affect *rate* at which the MSE converges to zero. Therefore, provided that n is sufficiently large (i.e., the network is sufficiently dense), communication resources can be saved by using very coarse quantizers. In fact, in our simulations we found that the degradation of the MSE due to quantization is not very significant (see Section 5), which makes our scheme suitable for low-rate operation.

Finally, we show that the projection estimator $\widehat{f}_n$ is minimax optimal for the additive Gaussian noise model $Y = f(X) + Z$, where $Z \sim \mathrm{Normal}(0, \sigma^2)$ is independent of X, and where sensors are deployed uniformly at random in the unit cube $[0,1]^d$. Here, $\eta = f$. We consider two commonly used function classes, namely analytic and Lipschitz.

Analytic functions: Suppose f belongs to the class $A_{\boldsymbol{\gamma}, M}$, where $M > 0$ 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 $S_{\boldsymbol{\gamma}} = \{\boldsymbol{z} \in \mathbb{C}^d : |\mathrm{Im}\, z_l| < \gamma_l, l = 1, \ldots, d\}$ in such a way that $|h| \le M$ on $S_{\boldsymbol{\gamma}}$. Choosing the tensor-product basis built from the trigonometric basis in $L^2([0,1])$ and $J_n = \lfloor (\gamma_1 \ldots \gamma_d)^{-1} (\ln n)^d \rfloor$, we get $\mathrm{MSE}(\widehat{f}_n, f) \le C (\ln n)^d / n$, where C depends only on $\boldsymbol{\gamma}, M, \sigma^2, \varepsilon$. On the other hand, it

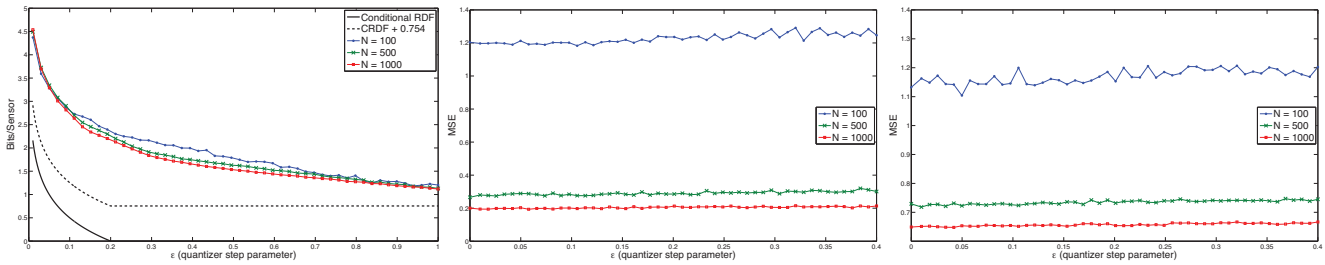

Fig. 3. Simulation results: average bits per sensor vs. ε (left); MSE vs. ε with adaptive cutoff selection (middle); MSE vs. ε with smoothness-based cutoff selection (right). The network sizes are $n = 100, 500, 1000$.

can be shown [5] that

$$\inf_{\tilde{f}_n} \sup_{f \in A_{\gamma,M}} \mathrm{MSE}(\tilde{f}_n, f) \succeq C_1 (\ln n)^d / n.$$

Lipschitz functions: Suppose 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 $|h^{(r)}(x + y) - h^{(r)}(x)| \leq M|y|^\alpha$ for all $x, y \in \mathbb{R}$, where $h^{(r)}$ is the rth derivative of h. Again, choosing the trigonometric tensor-product basis and $J_n = \lfloor n^{1/(2\beta+1)} \rfloor$, we get $\mathrm{MSE}(\widehat{f}_n, f) \leq C n^{-2\beta/(2\beta+1)}$, where C depends only on $r, \alpha, M, \varepsilon$. On the other hand, we have [1]

$$\inf_{\tilde{f}_n} \sup_{f \in \mathrm{Lip}_{r,\alpha,M}} \mathrm{MSE}(\tilde{f}_n, f) \succeq C_2 n^{-2\beta/(2\beta+1)}.$$

In both cases, the infimum is over all estimators of f from n samples. Hence, $\widehat{f}_n$ is minimax.

5. EXPERIMENTS

We have tested the performance of our scheme on the additive Gaussian noise model $Y = f(X) + Z$, where $Z \sim \mathrm{Normal}(0, \sigma^2)$ is independent of X. The underlying domain is the unit square $[0, 1]^2$, $\sigma^2 = 0.2$, and the function f, shown in Fig. 2, is a linear combination of a number of sinusoids and a rapidly decaying exponential term. This function is Lipschitz with $r = 0$ and $\alpha = 1$, i.e., $\beta = 1$.

We have used both a nonadaptive and an adaptive approach to estimate f. For the former, the smoothness constant $\beta = 1$ was used to determine the cutoffs J_n, while for the latter the cutoffs were selected using (5). Simulation results are shown in Fig. 3. As Fig. 3 (left) illustrates, for a given value of ε the average number of bits per sensor is above the conditional rate distortion function of Y given X evaluated at ε plus 0.754 bits (i.e., Ziv's bound with side information), but the gap closes as we increase the number of sensors. Moreover, the degradation of the MSE due to quantization is not very significant (the curves in Fig. 3, middle and right, are essentially flat). This makes the proposed scheme attractive for situations that call for low communication rates, since we can use low-resolution quantizers in dense networks. We also find that the adaptive procedure for determining cutoffs does significantly better than its nonadaptive, smoothness-based counterpart, especially for network sizes $n = 500$ (adaptive MSE ≈ 0.3 vs. nonadaptive MSE ≈ 0.72) and $n = 1000$ (adaptive MSE ≈ 0.2 vs. nonadaptive MSE ≈ 0.65).

6. CONCLUSION

This paper has proposed a scheme for rate-constrained distributed nonparametric regression which is low-complexity, universal, and minimax optimal for commonly used smoothness classes. One particularly attractive feature is that it can support very low communication rates yet still remain minimax optimal. Our simulations show that its empirical performance is close to that predicted by the theory, and confirm the theoretical conclusion that, for sufficiently dense networks, the effect of quantization on the MSE is not very significant. Random dithering is crucial not only for universal quantization, but also for obtaining unbiased and efficient estimators of the Fourier coefficients of the regression function.

7. REFERENCES

[1] S. Efromovich, *Nonparametric Curve Estimation: Methods, Theory and Applications*. New York: Springer, 1999.

[2] P. Elias, "Universal codeword sets and representations of the integers," *IEEE Trans. Inform. Theory*, vol. IT-21, no. 2, pp. 194–203, March 1975.

[3] D. S. Hirschberg and J. B. Sinclair, "Decentralized extrema-finding in circular configurations of processors," *Commun. ACM*, vol. 23, no. 11, pp. 627–628, November 1980.

[4] R. E. Krichevsky and V. K. Trofimov, "The performance of universal encoding," *IEEE Trans. Inform. Theory*, vol. IT-27, no. 2, pp. 199–207, March 1981.

[5] 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.

[6] N. Merhav and M. Feder, "Universal prediction," *IEEE Trans. Inform. Theory*, vol. 44, no. 6, pp. 2124–2147, October 1998.

[7] D. Niculescu, "Positioning in ad hoc sensor networks," *IEEE Network*, vol. 18, no. 4, pp. 24–29, July-August 2004.

[8] Y. Wang and P. Ishwar, "On non-parametric field estimation using randomly deployed, noisy, binary sensors," in *Proc. IEEE Int. Symp. on Inform. Theory*, Nice, France, June 2007.

[9] A. D. Wyner, "The rate-distortion function for source coding with side information at the decoder II: general sources," *Inform. Control*, vol. 38, pp. 60–80, 1978.

[10] R. Zamir and M. Feder, "On universal quantization by randomized uniform/lattice quantizers," *IEEE Trans. Inform. Theory*, vol. 38, no. 2, pp. 428–436, March 1992.

[11] J. Ziv, "On universal quantization," *IEEE Trans. Inform. Theory*, vol. IT-31, no. 3, pp. 344–347, May 1985.