PERFORMANCE BOUNDS FOR SENSOR DATA GATHERING BY CODING IN FINITE FIELDS

Tamara Tošić and Pascal Frossard

Ecole Polytechnique Fédérale de Lausanne (EPFL)
Signal Processing Laboratory (LTS4), Lausanne, 1015-Switzerland
E-mail: {tamara.tosic,pascal.frossard}@epfl.ch.

ABSTRACT

We address the problem of data gathering in adhoc networks. We propose a novel framework where sensor signals are quantized and mapped to a finite field. The network nodes then combine the data from different sensors to form messages that are transmitted towards a receiver. The receiver gathers different messages and reconstructs the original signal. We study the dependence of the signal reconstruction error on the quantization and network parameters. We further compute a bound on the reconstruction error for sparse sensor signals that depends on the number of messages gathered by the receiver. We validate our results with simulations in line array and tree-based sensor networks and show that our new framework leads to effective signal reconstruction with limited transmission costs.

Index Terms— Data Gathering, Network Coding, Finite Fields, Reconstruction, Sensor Network

1. INTRODUCTION

Sensor networks are used in numerous applications, such as environmental or industrial monitoring. Sensors generally have a limited power, which necessitates efficient data gathering and signal processing algorithms. These algorithms should combine low complexity encoding at sensors with ad-hoc transmission of data towards a receiver, which reconstructs the data captured by the sensors. When the data obeys a model that is known by the receiver, sensors should ideally only transmit the innovative or critical information that is necessary for model-based reconstruction. Unfortunately, the sensors generally do not have a priori knowledge required for identification of this critical information.

In this paper, we propose a novel framework for effective gathering of the sensor data using principles similar to network coding [1]. We present a system where the sensor nodes perform finite fields combinations of data measured by different sensors in order to build messages that are forwarded to the receiver. Instead of inefficient strategies where the data from every sensor is sent separately to the receiver, such network coding operations permit to gather information from all sensors in each message. This leads to reduced communication costs as long as the receiver can reconstruct the data from an underdetermined system of coded messages, with help of priors on the sensor data. We analyse our new framework by providing performance bounds on signal recovery. We show the influence of the number of messages, the size of the finite fields and the network size on the data reconstruction error. In particular, we concentrate on sensor data that is sparse and locally correlated and show that the decoding error stays small even if a small number of messages are collected by the receiver.

Numerous works have addressed the problem of effective gathering of data in sensor networks [2], and the data collection algorithms are generally driven by factors such as the sensors' power and the network topology, for example. Rate allocation algorithms for data collection have been proposed in [3], and Distributed Source Coding (DSC) algorithms have been proposed for reducing data redundancy in the network [4], [5]. However, such methods assume that data correlation is known by the sensors and are difficult to implement in fully distributed systems. Network coding ideas have been considered for effective data gathering in [6], where the decoding is performed at the receiver from a full rank matrix of linearly independent messages. Finally, we note that our solution bares some resemblance with compressed sensing ideas [7], where information can be sampled efficiently under the assumption that signal priors can be used in the reconstruction. Differently than the compressed sensing framework though, the combinations in our system are performed in a finite field due to network communication constraints. A few recent works have analysed finite field compressive sensing from information theory perspective. For example, the work in [8] studies the error exponent value for recovery of finite field sparse signals from their linear measurements performed in Galois field. This expression is developed under the assumption that the sampling matrix elements are i.i.d. and uniformly sampled from the Galois field. The problem of the recovery of a low rank matrix from its linear measurements in a Galois field is studied in [9], where the authors provide fundamental limits on sampling requirements and show that the decoding error bound is small for arbitrary matrices of given rank. In this work we rather consider a generic setup where processing is performed in a finite field of arbitrary size and signal and sampling matrices have arbitrary probability mass functions. In this context, our study is the first work that addresses the problem of data reconstruction from underdetermined systems at decoder and for general signal models.

The rest of this paper is organised as follows. Section 2 describes the proposed data gathering framework. Section 3 studies the analysis of the decoding error. Simulation results and discussions are then given in Section 4.

2. DATA GATHERING FRAMEWORK

In this section, we present in detail our data gathering system for adhoc sensor networks. We consider that the network is represented by a directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ as illustrated in Fig. 1. The set of vertices $\mathcal{V}$ represents sensors and the edge set $\mathcal{E}$ represents the connections between sensors. We consider that the network is connected. We represent the set of directed edges $\mathcal{E}$ by an asymmetric adjacency matrix, where element $e_{i,j}$ has a nonzero value if the sen-

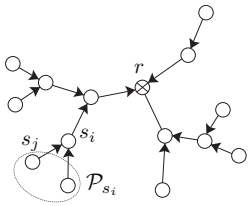

Fig. 1. Directed sensor network. Predecessors $\mathcal{P}_{s_i}$ of sensor s_i are sensors positioned farther than s_i from the receiver r.

sor s_i communicates data towards s_j. We call as predecessors of s_i the set of nodes $\mathcal{P}_{s_i}$ that are the direct neighbors of s_i but with a larger hop-distance to the receiver r.

The sensors $\{s_i\}$, with $1 \leq i \leq S$, measure values of a scalar function that lives on the sensor graph. For minimal communication costs, the sensors should ideally transmit to the receiver only data that are innovative with respect to the information form other sensors. This is however hard to achieve in realistic settings where sensors are of low complexity and only have a local view of the network. We therefore propose that the network conveys messages that are random combinations of sensor data, in order to encode innovative information with high probability. In more details, we assume that the sensor measurements x_i, $1 \leq i \leq S$, represent uniformly quantized values that are mapped by injection to the values $(0, \ldots, q-1)$ of a finite field F_q of size q. The data collection is performed as follows. We first assign *weights* $w_i \in F_q$ to every sensor. Then, sensors at the border of the network (i.e., sensors that do not have any incoming link) initiate the data gathering. Messages are collected synchronously along the directed collection paths where each node performs weighted combinations of sensor measurements using modulo q operations. For instance, the message $y_i \in F_q$ created at the sensor s_i is the result of the weighted combination of its current measurement and the messages y_j received from its predecessors $\mathcal{P}_{s_i}$. It reads

$$y_i = (w_i \odot x_i) \oplus \sum_{j \in P_{s_i}} y_j. \tag{1}$$

The message y_j has been similarly constructed by weighted combination of the measurement of sensor s_j and data received from its predecessors $\mathcal{P}_{s_j}$. Note that the operators $\odot$ and $\oplus$ represent element-wise modulo-q multiplication and addition. The data gathering terminates when M messages have reached the receiver. Finally, the messages at receiver can be represented in a matrix form as

$$Y = \mathbf{W} \otimes X, \tag{2}$$

where $X \in F_q^{S \times 1}$ represents a vector of sensor measurement values x_i, $1 \leq i \leq S$ and the vector $Y \in F_q^{M \times 1}$ contains the received messages. The coding matrix $\mathbf{W} \in F_q^{M \times S}$ describes the network coding operations and the operator $\otimes$ describes the modulo-q matrix multiplication. Remark that the i-th row of $\mathbf{W}$ represents the coding vector used for building the i-th message y_i.

In general, the system of Eq. (2) is undetermined. The reconstruction problem at decoder thus consists in reconstructing the sensor data from a small number of messages with help of priors on the signal under observation. In other words, the data has to determine the value X that both satisfies the constraints from Eq. (2), and fits the data model. Such problem can be solved with algorithms such as message-passing decoders [10]. We are interested in this paper in analysing the performance bounds of the proposed system.

3. DECODER PERFORMANCE BOUNDS

We compute performance bounds for our data gathering framework by studying the decoding error probability. We assume a general signal model where the finite field representation of the sensor signal $X \in F_q^{S \times 1}$ and the decoded signal $\hat{X} \in F_q^{S \times 1}$ belong to a class of signals denoted by $\mathcal{F}$. This class represents all the possible signals that match the data model. Under this assumption, a reconstruction error occurs when the decoder selects a signal from $\mathcal{F}$ that matches the coding conditions, i.e., $Y = \mathbf{W} \otimes X = \mathbf{W} \otimes \hat{X}$, but that is different from X. The decoding error probability thus reads

$$\begin{aligned} p(\hat{X}|X) &= p\Big(\hat{X} \in \mathcal{F}, \text{ s.t. } Y = \mathbf{W} \otimes \hat{X} \text{ and } \hat{X} \neq X\Big) \\ &\leq \sum_{\hat{X} \in \mathcal{F}} p\Big(\hat{X} \text{ s.t. } Y = \mathbf{W} \otimes \hat{X}\Big). \end{aligned} \tag{3}$$

The work in [8] uses a similar setup for expressing the decoding error probability when the coding is performed in a Galois field with uniformly distributed coefficients. We study here the decoding performance for a more generic framework in finite fields of arbitrary size and with an arbitrary coding matrix. The computation of the performance bounds is quite different in such a generic setup.

We can derive the bound in Eq. (3) by computing the probability for a signal in $\hat{X} \in \mathcal{F}$ to satisfy the coding equations at decoder. It reads

$$p(Y = \mathbf{W} \otimes \hat{X}) = \prod_{i=1}^{M} p\Big(y_i = \mathbf{W}_{i,:} \otimes \hat{X}\Big), \tag{4}$$

where $\mathbf{W}_{i,:}$ represents the i-th row of the coding matrix $\mathbf{W}$, with the coding coefficients in each row chosen independently. We observe now that, if $\mathbf{W}_{i,:} \otimes \hat{X} = Y = \mathbf{W}_{i,:} \otimes X$, then $\mathbf{W}_{i,:} \otimes (\hat{X} \ominus X) = 0$. Thus, we have

$$p\Big(y_i = \mathbf{W}_{i,:} \otimes \hat{X}\Big) = p\Big(\sum_{s=1}^{S} W_{i,s} \odot (\hat{x}_s \ominus x_s) = 0\Big). \tag{5}$$

Therefore, the set of error events E is given as the collection of M events e_i for which

$$\sum_{s=1}^{S} W_{i,s} \odot (\hat{x}_s \ominus x_s) = 0 \tag{6}$$

holds. Computing the decoder error probability boils down to building systematically the error event set E. Once the full set of events is known, the decoder error probability is computed as the sum of the probabilities of events in this set that generates a decoding error.

We first identify the events that simplify the systematic construction of the set of error events E. As a illustrative example, we consider the case where the vectors X and $\hat{X}$ differ only in the first three positions. Let u_s for $s \in \{1, 2, 3\}$ be $u_s = (W_{i,s} \odot (\hat{x}_s \ominus x_s))$. Then, we observe that the expression in Eq. (6) is equal to $(u_1 \oplus u_2 \oplus u_3) = 0$ in a finite field, when the sum of the u_s would take the values $\mathcal{I} = \{0, q, 2q\}$ in a (non-finite) integer field. In particular, this sum would be 0 if all three coefficients $W_{i,s}$ in our example are equal to zero; it would possibly sum up to q if more than one coefficient value is non-zero, and would possibly sum up to $2q$ if all

three coefficients are non-zero. With this example that illustrates the cyclic properties of additions in finite fields, we see that the error events could be separated into different cases that reflects the number of coefficients $W_{i,s}$ that have a zero value, where the index s defines the positions where the vectors X and $\hat{X}$ are different.

We now extend these development to more general cases. We denote by $A \in \{1, \ldots, 2K\}$ the number of non-zero coding matrix coefficients at positions where X and $\hat{X}$ differ. Alternatively, let $B \in \{0, \ldots, 2K\}$ denote the number of zero elements among coding matrix coefficients at positions where X and $\hat{X}$ differ. Since X and $\hat{X}$ differ in $k \in \{1, \ldots, K\}$ positions by our initial assumption, we have $A + B = 2k$; in other words the signals are multiplied by zero coefficients at $2k - A$ positions. We further denote by e_A and $\bar{e}_B$ the sets corresponding to the cases with A non-zero values in the coding matrix coefficients at positions where X and $\hat{X}$ differ, and respectively B zero coefficients at these positions. The size of these sets is denoted by $\mathcal{S}(e_A)$ and $\mathcal{S}(\bar{e}_B)$, respectively. Then, the probability of an error event e_i to happen can thus be computed over all values A as follows

$$p(e_i) = \sum_{k=1}^{K} \sum_{A=1}^{2k} \left\{ \Big(p(c_W \neq 0)\Big)^A \mathcal{S}(e_A) \Big(p(c_W = 0)\Big)^{2k-A} \mathcal{S}(\bar{e}_B) \right\}, \tag{7}$$

where c_W is one of the coefficients in the coding matrix, which follow an i.i.d. distribution in the design of W. We now compute the expected size of the sets e_A and e_B.

Because of the circular property of multiplication in arbitrary fields, the same coding vector multiplied with two different signals may give the same result. The event e_A denotes such cases amongst the coding vector with A nonzero elements at positions where X and $\hat{X}$ differ. Then we can write the probability of expected size of the set e_A as

$$\mathcal{S}(e_A) \leq |\hat{\mathcal{X}}_A| \sum_{m=1}^{A} \sum_{l=1}^{A} \sum_{n=1}^{q-1} \frac{p\Big(\mathcal{C}(P((m-1)q, n, l))\Big)}{p\Big(\mathcal{C}(P((m-1)q, n))\Big)}, \tag{8}$$

where $|\hat{\mathcal{X}}_A|$ here counts the total number of vectors in $\mathcal{F}$ with at least A nonzero values at positions where X and $\hat{X}$ differ. It multiplies the expected number of error event realizations for each particular $\hat{\mathcal{X}}_A \subset \mathcal{F}$ to give an bound on the size of e_A. In particular, the function $\mathcal{C}(P)$ lists all the possible combinations of summands from the realisation of a partition function P. The partition function $P(a, b)$ of a number a defines a partition of the different representations of the number a with the biggest summand being b. The function $P(a, b, c)$ represents a partition of the possible representations of the number a with c summands, where the biggest element in the sum is b. In our case, we compute partition functions considering up to c nonzero summands. The reason for this is that modulo multiplication result can be equal to zero, even when both multiplied values are nonzero. We therefore here modify the classical partition function such that it includes these special cases.

As consequence of the circularity of the modulo product in the arbitrary field, certain realizations are more probable than other. In the previous equation p denotes the probability mass function of the product of the two random variables $Z = Z_1 Z_2$, that stand for distributions of the coding vector and signal vector values. The pmf of the random variable Z is first computed in the real field from the pmfs of Z_1 and Z_2 using transform techniques and the values of Z are quantized to the range $\{0, \ldots, q-1\}$. The probabilities of the random variable realizations that have the same congruent modulo q (same reminders) are summed up together. Remark that the data model $\mathcal{F}$ influence is not explicitly visible in the error term. It however drives the number of partitions and it constraints values in $\hat{\mathcal{X}}_A$.

Finally, the size of the set $\mathcal{S}(\bar{e}_B)$ is simply given as

$$\mathcal{S}(\bar{e}_B) = |\hat{\mathcal{X}}_B|, \tag{9}$$

which is the cardinality of the set of possible vectors in $\mathcal{F}$ with B arbitrary values at positions multiplied by zero coefficients in the coding vector. Remark that the size of the set $\bar{e}_B$ is easier to compute than the size of e_A; indeed, the actual values of the signals to be consider are unimportant since they are multiplied by zero coefficients.

By combining Eq. (8) into (7), we obtain the probability $p(e_i)$ of the error event e_i. Finally, by inserting this result into Eq. (5), we can bound the error probability of Eq. (3) as

$$p(\hat{X}|X) \leq p(e_i)^M \tag{10}$$

Recall that, even if the influence of the data model is not explicit in this last relation, the form of the data in the set $\mathcal{F}$ drives the probability $p(e_i)$ through Eqs. (8) and (9).

4. EXPERIMENTAL RESULTS AND DISCUSSION

We now study the decoder error bound and analyse the performance of the proposed framework in different settings. We assume that sensors are randomly distributed and that the data gathering paths are built with a shortest path algorithm. The results are given for two types of networks, namely line array and tree-like networks.

Signals in sensor networks used for environmental monitoring very often have a small number of non-zero values locally positioned, while the rest of the values are zeros. Therefore, we adopt this data model in our simulations and we assume that the class of signals of interest $\mathcal{F}$ is formed by signals with up to K non-zero values that are grouped locally on the graph. Furthermore, we consider that these non-zeros values follow Uniform or Discrete Laplacian distributions. Finally, we assume that the values of the coding matrix are chosen uniformly at random, similar to the coding matrix design used in [8].

First, we study the evolution of the performance bound as a function of the number of messages, for different sparsity levels (i.e., different values of K). In order to help the evaluation, we also show the bound given in [8]

$$p(\hat{X}|X) \leq K 2^{S H_B(K/S)} (q-1)^K q^{-M}, \tag{11}$$

where $H_B(K/S)$ is a binary entropy. This decoding error has however been developed for a framework where linear combinations are performed in $GF(q)$ for prime values of the field size q, where the elements of the matrix $\mathbf{W}$ that are chosen uniformly at random in a Galois field $GF(q)$ and the signals are chosen uniformly from the set of sparse signals. Note that our bound in Eq. (10) can be seen as a generalization of the bound in Eq. (11), since we allow network combinations to be performed in a finite field or arbitrary size q using modulo q operations. Remark that for the same set of the assumptions (parameters, linear combinations in Galois field and data model) both bounds actually match.

Figure 2 gives the error decoder performance vs. number of received messages for (DR) in [8] and for our setup (PM), for the parameter set $(S, K, q, \text{signal model})$. The lower error values for PM is a consequence of set cardinality: the cardinality of the set $\mathcal{F}$ is by construction smaller than for the set of K-sparse signals whose non-zero values are arbitrarily distributed in DR.

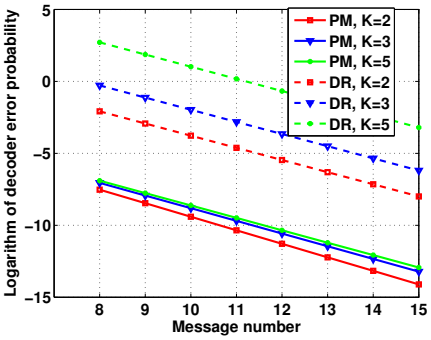

Fig. 2. Error decoding probability (log scale) for line array sensor network with parameters $(S, K, q, \text{sparse signal distribution})$. PM: proposed method bound calculated for parameters $(20, [2, 3, 5], 7, \text{Uniform})$. DR: Draper et al. bound for parameters $(20, [2, 3, 5], 7, \text{Uniform})$.

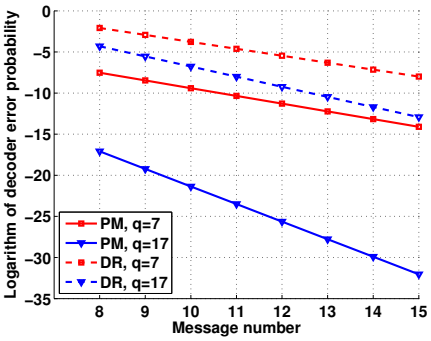

Fig. 3. Error decoding probability (log scale) for line array sensor network with parameters $(S, K, q, \text{sparse signal distribution})$. PM: proposed method bound calculated for parameters $(20, 2, [7, 17], \text{Uniform})$. DR: Draper et al. bound for parameters $(20, 2, [7, 17], \text{Uniform})$.

In addition, we repeat similar comparisons but for different sizes of the finite fields where processing is performed. We clearly see that the decoding error decreases for larger values q of the field size. Note that we chose prime values for the field size q for the sake of comparison with the framework in [8]; however, our framework could use any value for q and the results actually follow the same tendency as the one shown in Figure 3.

Finally, we want to study the influence of the data model on the performance bounds. We consider a second class of signals, where the number of non-zero values is exactly K. These non-zeros values are again grouped locally on the graph. The computations of our performance bounds is adapted in this case by putting $k = K$ exactly, and by choosing A even in $\{2, \ldots, 2K\}$ (i.e., two different vectors with fixed K may differ only in an even number of positions). Fig. 4 illustrates the decoding error probability for both signal models and for line array (LA) and tree (TR) network. The tree network has roughly 30% more connections between sensors than the line array network. As expected, we see that the error is smaller for signals when the sparsity is fixed, as the set of possible signals is smaller in this case, hence the decoding error is reduced. Due to the same reasons, smaller values of sparsity (i.e., K) leads to smaller decoding error probabilities. Finally, we observe that in all cases the error bounds decreases exponentially with the number of messages, which is very important for building effective data gathering solutions.

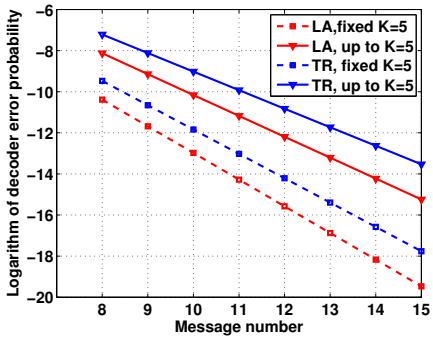

Fig. 4. Error decoding probability (log scale) for line array and tree-based sensor network with parameters: $(S, K, q, \text{sparse signal model}) = (20, 5, 8, \text{DiscreteLaplacian})$. The error is calculated for sparsity of K and up to K.

5. CONCLUSIONS

In this work, we have proposed a new data gathering system for sensor networks, where a small number of network messages are created by combination of quantized sensor measurements in a finite field of arbitrary size. These messages are communicated through the network from the leaf sensors towards the receiver. We have developed bounds on the decoding error probability as a function of the number of received messages, the design of the coding matrix and the signal class. We finally illustrate these dependencies by simulations in different network settings, which confirm that the framework offers an interesting solution for data gathering with a small number of messages in communication-constrained networks.

6. REFERENCES

[1] R. Ahlswede, N. Cai, S-Y. R. Li, and R. W. Yeung, "Network Information Flow," *IEEE Transactions on Information Theory*, vol. 46, no. 4, pp. 1204–1216, 2000.

[2] R. Rajagopalan and P. K. Varshney, "Data aggregation techniques in sensor networks: A survey," *IEEE Communications Surveys and Tutorials*, vol. 8, no. 4, pp. 48–63, Oct. 2006.

[3] R. Critescu, B. Beferull-Lozano, and M. Vetterli, "On network correlated data gathering," *IEEE INFOCOM*, pp. 2571–2582, 2004.

[4] S. Pradhan and K. Ramchandran, "Distributed Sounce Coding Using Syndromes (DISCUS): Design and Construction," *IEEE Trans. of Inf. Theory*, vol. 49, no. 3, pp. 626–642, March 2003.

[5] Z. Xiong, A. D. Liveris, and S. Cheng, "Distributed Sounce Coding for Sensor Networks," *IEEE Signal Processing Magazine*, vol. 21, no. 5, pp. 80–94, Sept. 2004.

[6] S. B. Cruz, G. Maierbacher, and J. Barros, "Joint source-network coding for large-scale sensor networks," *IEEE Int. Symp. on Inf. Theory Proceedings (ISIT)*, 2011.

[7] R. G. Baraniuk, "Compressive sensing," *Lecture Notes in IEEE Signal Processing Magazine*, vol. 24, no. 4, pp. 118–120, Jul. 2007.

[8] S. C. Draper and S. Malekpour, "Compressed sensing over finite fields," *Int. Symp. on Inf. Theory*, 2009.

[9] V. Y. F. Tan, L. Balzano, and S. C. Draper, "Rank minimization over finite fields: Fundamental limits and coding-theoretic interpretations," *IEEE Trans. on Inf. Theory*, vol. 58, no. 4, pp. 2018–2039, 2012.

[10] F. R. Kschischang, B. J. Frey, and H-A. Loeliger, "Factor graphs and the sum-product algorithm," *IEEE Transactions on Information Theory*, vol. 47, no. 2, pp. 498–519, Feb 2001.