

Joint Compression, Detection, and Routing in Capacity Constrained Wireless Sensor Networks

Onur G. Guleryuz and Ulaş C. Kozat
 DoCoMo USA Labs
 181 Metro Dr, Ste 300, San Jose, CA 95110
 {guleryuz,kozat}@docomolabs-usa.com

Abstract—This paper considers an important class of sensor networks where the ultimate goal is not necessarily to collect each individual measurement but rather a potentially smaller set of statistics. Considering link capacity constrained topologies, we derive results that optimally allocate rate/distortion to information collected by the sensors. As a key contribution, we determine how the flow of information emanating from the sensors should be managed, yielding optimal routing algorithms and jointly optimized networks. Our analysis encompasses the typical scenarios that are widely observed in sensor networks, and over these scenarios, we quantify the gains offered by sending the statistics rather than the measurement data itself.

I. INTRODUCTION

Wireless sensor networks have attracted significant attention in recent years. The rich investigation is due to the fact that these networks are application specific and hence they are amenable to highly engineered network design that removes layer separation. Nevertheless, most of literature independently treats transmission capacity, sensing capacity, energy efficiency, routing, compression, and detection issues [3].

In this paper, we look at an important class of sensor networks, where the ultimate goal is not to collect each individual measurement, but rather a smaller set of statistics. These statistics are often expressed as weighted sum of individual measurements and used for hypothesis testing or classification purposes. As stated such, we encounter several interesting problems that cross rate allocation, routing, and clustering problems under the gross picture of network detection. Starting from simpler yet tractable scenarios and moving to more involved ones, we investigate a fundamental subset of these problems.

In the sensor network model we consider, every node except for the remote control point is capable of: (i) data measurement, (ii) simple data processing, (iii) and relaying. Due to data processing capabilities, sensor nodes can decode, re-encode, and aggregate received information with the knowledge of local measurements.

We first analyze a relatively simple scenario (shown in figure 1), where we have a tree topology with depth two. Each sensor node transmits its own quantized measurements to a local aggregation node. Aggregation node, in return, relays the measurements to a remote location. As a natural constraint, each transmission is subject to the point-to-point throughput capacity constraints. We obtain optimal transmission strategies for two different cases: without and with signal processing. In the former case, aggregation node performs rate allocation to minimize the total distortion under the given capacity constraints. Whereas in the latter case, aggregation node has the knowledge of desired statistics (i.e. useful information) at the control point and can instead transmit a representation of these statistics. We provide analytical results for the total distortion seen at the controller node.

We then shift our attention to a more general topology, where we have L cluster heads that can transmit the data to the remote

central node. We pose the problem as optimal clustering under the capacity constraints. After some mild assumptions, such as identical observation variance and well-defined intra-cluster capacity scaling as a function of cluster size, we provide a polynomial-time optimum solution by reducing the question to a dynamic programming problem.

As our last contribution with this paper, we also look at the most general topology with arbitrary link capacities and observation variances. We present a greedy heuristic for our algorithmic solution, which returns a tree hierarchy solving the routing, scheduling and compression problems jointly. As before we compare distortion performance of both cases, i.e. when we apply and preclude signal processing at the aggregation points, via simulations.

The rest of the paper is organized as follows. In Section II, we present the notation that is used in the rest of the paper. In Section III, we provide our basic framework and corresponding analytical results. Section IV extends the basic framework and provides exact as well as approximate algorithmic solutions for different problem instances. We finally deliver some simulation results and conclude in Section V.

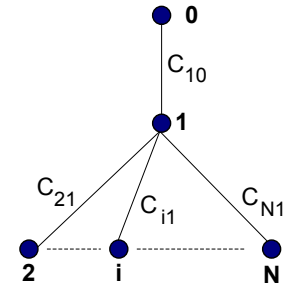


Fig. 1. Basic framework.

II. NOTATION

In the bulk of this paper we will be encoding scalar random variables using simple uniform quantizers and entropy coders. In order to obtain convenient formulas for the resulting average rate in bits, R , and the mean squared quantization error, D , let us consider a zero-mean random variable y having variance σ_y^2 . In general, R and D are complicated functions of the probability density function (pdf) of y . Regardless, for purposes of tractable optimization, we are interested in encoding strategies that produce $R = f_1(\sigma_y, \Delta)$, $D = f_2(\sigma_y, \Delta)$, where Δ is the uniform quantizer stepsize, and $f_1(\cdot, \cdot)$ and $f_2(\cdot, \cdot)$ are simple functions independent of the pdf of y .

Consider the entropy coding operation that takes the quantized version of y and losslessly encodes it as a sequence of bits [1]. By

using an entropy code that is optimized for a zero-mean Gaussian random variable of variance $\sigma^2 \geq \sigma_y^2$, we can obtain

$$R \leq \max(K_1, K_2 + \log(\sigma) - \log(\Delta)) \quad (1)$$

$$D \leq (\Delta/2)^2 \quad (2)$$

for positive constants K_1 and K_2 under modest assumptions on the regularity of the pdf of y ([2], section IV). Noting that $D = \sigma_y^2$ can be accomplished with $R = 0$ bits, and always choosing $\sigma = \sigma_y$, it is clear that we can define sufficiently large constants $K_2, K_3 > 0$ so that this simple encoding scheme obtains

$$R \leq K_2 + 1/2 \log\left(\frac{\sigma_y^2}{\Delta^2}\right) \quad (3)$$

$$D \leq K_3 \Delta^2 \quad (4)$$

for $D \leq \sigma_y^2$. Note also that in the important special case where y is a Gaussian random variable, the bound in Equation (3) becomes tighter, and in high rate regimes, we can find constants that replace inequalities with approximate equalities.

We will typically obtain y as a linear projection of a random vector x ($N \times 1$), via $y = h^T x$, for a given deterministic vector h . In this paper, x will be a conditionally Gaussian random vector conditioned on a state variable i . The state variable $i \in \{1, \dots, s\}$ determines the mean vector of x while the conditional covariance matrix of x is set to a fixed diagonal matrix, i.e., x becomes the observation of one of s deterministic signals under additive independent Gaussian noise. In much of the paper we will further assume that the noise is independent and identically distributed, i.e., the conditional covariance matrix is a multiple of identity.

A conditionally Gaussian x will in turn make y a mixture Gaussian random variable, i.e., given the mixture variable $i \in \{1, \dots, s\}$, y will be distributed conditionally Gaussian with mean $m_{y,i}$ and variance $\sigma_{y,i}^2$. In order to accommodate a wide range of means without a significant performance penalty, our encoder will first determine a state variable \hat{i} as the state that maximizes the a posteriori probability of the quantized $y - m_{y,\hat{i}}$, encode \hat{i} with $2\lceil \log(s) \rceil$ bits, and then encode the quantized $y - m_{y,\hat{i}}$ as above using $\sigma = \sigma_{y,\hat{i}}$. This will increase the constant K_2 in Equation (3) by $\sim \log(s)$ bits but it will not effect our main results to order N as we will have $s \ll N$. Observe that since any component of x can be obtained as a linear projection, we will be using the same encoding strategy when encoding the components of x . In addition as means are handled by the encoder/decoder, in what follows we will restrict ourselves to zero-mean analysis without affecting Equations (3) and (4).

Throughout, we will assume that relevant statistics (variances, means, etc.) are known at the encoder and decoder. In order to avoid technicalities with respect to cross correlations among quantized values and respective quantization errors we will assume that the uniform quantizer reproduction levels are at the centroids. For a given random variable y we will use the notation y^Δ to denote its quantized version using a quantizer of stepsize Δ . G will denote an $(M \times N)$ matrix whose application to a vector x results in M statistics that the central node is interested in.

III. BASIC FRAMEWORK AND COMPUTATION

In this section, we consider a relatively simple scenario where we have one local aggregation node that collects information from the sensors within its reach and combines with its own measurement to relay to a distant central node. The scenario is depicted in figure 1, where C_{ij} represents the channel capacity between nodes i and j . Indices 0 and 1 are reserved for the central and local aggregation nodes respectively. Node 1 can perform relaying in two

different ways: (i) Encode each node's information, i.e. each entry of $x = \{x_1, x_2^{\Delta_{21}}, \dots, x_N^{\Delta_{N1}}\}$, separately, or (ii) encode each desired statistic, i.e. each element in $y = Gx$, separately. We compare these two cases in terms of their optimally achievable total distortion.

Case 1: Optimal transmission strategy for this case can be expressed as a distortion (equivalently *rate*) allocation problem, which is formally defined as:

Problem P1:

$$\min_{D_1, \dots, D_N} \left\{ D_T = \sum_{i=1}^N D_i \right\} \quad (5)$$

$$\sum_{i=1}^N R_i = C_{10}, \quad (6)$$

$$D_i \leq \sigma_i^2, \quad i = 1, \dots, N, \quad (7)$$

$$D_i \geq \mathcal{D}(C_{i1}, \sigma_i^2), \quad i = 2, \dots, N, \quad (8)$$

where $\mathcal{D}(C_{i1}, \sigma_i^2) = K_3 \sigma_i^2 \exp(2K_2 - 2C_{i1})$, which is the distortion incurred by compressing a random variable with variance σ_i^2 using C_{i1} bits.

Solution of P1: When we consider the problem by assuming measurements from nodes 2 to N are available with no quantization, i.e. constraints in (8) are omitted, the solution is well-known and follows a "reverse water-filling" argument (see [1], pp. 348-349 for details). In a similar vein, by adding additional Kuhn-Tucker conditions, we can find the optimal distortion values (denoted as D_i^* 's) as:

$$D_i^* = \begin{cases} \mathcal{D}(C_{i1}, \sigma_i^2), & \text{if } \lambda \leq \mathcal{D}(C_{i1}, \sigma_i^2), \\ \lambda, & \text{if } \mathcal{D}(C_{i1}, \sigma_i^2) < \lambda < \sigma_i^2, \\ \sigma_i^2, & \text{if } \lambda \geq \sigma_i^2. \end{cases} \quad (9)$$

Define the sets $P = \{i : \mathcal{D}(C_{i1}, \sigma_i^2) < \lambda < \sigma_i^2\}$, $Q = \{i : D_i^* = \sigma_i^2\}$, and $Z = \{i : D_i^* = \mathcal{D}(C_{i1}, \sigma_i^2)\}$. Then,

$$\lambda = \frac{(\prod_{i \in P} \sigma_i^2)^{1/|P|} \exp(2\beta)}{\exp\left\{\frac{2}{|P|} [C_{10} - \sum_{i \in Z} C_{i1}]\right\}}, \quad (10)$$

$$R_i^* = \frac{1}{2} \log\left(\frac{\sigma_i^2}{(\prod_{i \in P} \sigma_i^2)^{1/|P|}}\right) + \frac{C_{10} - \sum_{i \in Z} C_{i1}}{|P|}, \quad (11)$$

where $\beta = \log(\sqrt{K_3}) + K_2$.

Case 2: Intuition suggests that we should be able to do much better in the sense of total distortion by directly working on the desired statistics. Therefore, as a simple scheme, we are interested in optimally transmitting each statistics $y_j = \sum_{i=1}^N g_{ji} x_i^{\Delta_{i1}}$ with $x_1^{\Delta_{11}} = x_1$. Under the independence of $x_i^{\Delta_{i1}}$'s, we obtain $E[y_j^2] = \sum_{i=1}^N g_{ji}^2 E[(x_i^{\Delta_{i1}})^2]$. Denote quantization error at the i th node as w_i , i.e. $x_i = x_i^{\Delta_{i1}} + w_i$. Since $x_i^{\Delta_{i1}}$ are at the centroids $E[x_i^{\Delta_{i1}} w_i] = 0$, and we have $\sigma_i^2 = E[x_i^2] = E[(x_i^{\Delta_{i1}})^2] + E[w_i^2]$ or $E[(x_i^{\Delta_{i1}})^2] \leq \sigma_i^2$. Hence,

$$\sigma_{y_j}^2 = E[y_j^2] \leq \sum_{i=1}^N g_{ji}^2 \sigma_i^2. \quad (12)$$

Using the upper-bound (12) and the results from case 1, we can compute the optimal transmission rates R_j^* for each statistics y_j . Then, by (3), we can find the quantization step size Δ_j for each statistics as:

$$\Delta_j = \sigma_{y_j} \exp(K_2 - R_j^*). \quad (13)$$

Since we are actually interested in $\sum_{i=1}^N g_{ji} x_i$, we need to compute

the overall distortion by:

$$D_T^{(j)} = E \left[\left(\sum_{i=1}^N g_{ji} x_i - y_j^{\Delta_j} \right)^2 \right] \quad (14)$$

$$= E \left[\left(\sum_{i=1}^N g_{ji} x_i - y_j \right) + (y_j - y_j^{\Delta_j}) \right]^2 \quad (15)$$

$$\leq K_3 \left(\sum_{i=2}^N g_{ji}^2 \Delta_{i1}^2 + \Delta_j^2 \right), \quad (16)$$

where cross terms in Equation (15) are zero since the correlation of quantized values and quantization errors is zero. Note that in case 1, we may throw away some of the measurements due to insufficient capacity at node 1 following reverse water filling arguments, whereas in case 2, we effectively compress each statistics of interest by utilizing all the received information provided $g_{ji} > 0, \forall i$.

A. Comparison of Case 1 and Case 2 in Special Scenarios:

Let's limit ourselves to the scenario where we are only interested in a single statistic that is the sum of all measurements, i.e. $g_{1i} = 1$.

1) $C_{i1} = C > C_{10}/N$ and $\sigma_i^2 = \sigma^2 > \lambda$: In this scenario, we have a communication bottleneck between the local aggregation and central nodes. Since measurement variances are same, in case 1, all measurements are transmitted at rate $R_i^* = C_{10}/N$. Therefore, we can express the ratio of total distortion of case 1 to that of case 2 as:

$$\rho = \frac{(D_T)^I}{(D_T)^{II}} = \frac{\exp(-2C_{10}/N)}{\frac{N-1}{N} \exp(-2C) + \exp(-2C_{10})}. \quad (17)$$

When $C \approx C_{10}$, we have an exponential improvement in distortion in terms of C_{10} :

$$\rho = \frac{N}{2N-1} \exp \left\{ 2 \frac{(N-1)}{N} C_{10} \right\}.$$

2) $C_{i1} = C < C_{10}/N$ and $\sigma_i^2 = \sigma^2 > \lambda$: When $C < C_{10}/N$, we observe that transmissions to local aggregation node constitute the bottleneck and there is no incentive in case 1 to encode at a higher rate to the central node. In return, ρ becomes:

$$\rho = \frac{\exp(-2C)}{\frac{(N-1)}{N} \exp(-2C) + \exp(-2C_{10})} \approx \frac{N}{(N-1)}, \quad (18)$$

for sufficiently large C_{10} . Hence, when transmissions to local aggregation point form a bottleneck, both schemes perform similar for moderate number of nodes.

IV. GENERALIZED FRAMEWORK

In the following two sections, we investigate two different topological settings. In section IV-A, we limit ourselves to a routing tree structure that is of depth two or less and we jointly optimize it under the specified conditions. In section IV-B, we look at the most general scenario, which allows arbitrary topologies, and we search for the optimal routing tree structure.

A. Optimal Clustering Under a Simplified Scenario

In this section we are interested in the the N nodes forming L clusters, with every cluster of the particular form shown in Figure 1, i.e., the information flow from the nodes to the central node is via a tree of depth two. We would like to find the optimal tree that conforms to a network/transmission model, i.e., we are interested in finding the tree of depth less than or equal to two that minimizes the total distortion observed at the central node, given capacity constraints enforced by the transmission model. In order to obtain optimal

solutions we restrict discussion to the **Case 1** scenario identified in Section III, and $\sigma_i^2 = \sigma^2, i = 1, \dots, N$.

Assume that the capacity between node i and the central node 0 is given by $C_{i0}, i = 1, \dots, N$. We will start with the case where the internode capacities $C_{ij} = C, i, j = 1, \dots, N$, in order to define an algorithm that finds the optimal tree, and then generalize this algorithm to conform to the more general transmission model where C varies for each cluster as a function of the number of nodes aggregated by that cluster.

Let node $\gamma(l)$ be the aggregator of cluster l and let $\xi(l)$ be the set of nodes that are aggregated by $\gamma(l)$ so that the information coming out of cluster l is due to the nodes $\xi(l) \cup \{\gamma(l)\}$. Let $\alpha_l = |\xi(l)|$ denote the cardinality of $\xi(l)$ with $\sum_{l=1}^L \alpha_l = N - L$. Using **Case 1** results, we determine that nodes in $\xi(l) \cup \{\gamma(l)\}$ will encode $x_j, j \in \xi(l) \cup \{\gamma(l)\}$ to achieve the distortions

$$\begin{aligned} \text{if } \frac{c_{k,0}}{\alpha_l + 1} \geq c &\rightarrow \begin{cases} D_j = \mathcal{D}(c_{k,0} - \alpha_l c, \sigma^2), & j = \gamma(l) \\ D_j = \mathcal{D}(c, \sigma^2), & j \in \xi(l) \end{cases} \\ \text{otherwise} &\rightarrow D_j = \mathcal{D}(c, \sigma^2), \quad j \in \xi(l) \cup \{\gamma(l)\}, \end{aligned} \quad (19)$$

with the total distortion incurred at the central node due to cluster l given by

$$\sum_{j \in \xi(l) \cup \{\gamma(l)\}} D_j. \quad (20)$$

We are now ready to specify a dynamic programming algorithm that optimally determines the cluster number $L \leq N$, the number of nodes α_l that are aggregated by cluster $l, l = 1, \dots, L$, and the nodes that act as the cluster aggregators. Due to symmetry, observe that L, α_l , and the indices of the cluster aggregator nodes are sufficient to determine the optimal tree.

Consider the $N + 1$ step trellis, where the states at step i are determined by the possible values of $\mathcal{N}(i)$, the number of nodes left available for clustering to steps $i + 1$ through N . We have $0 \leq \mathcal{N}(i) \leq N, \mathcal{N}(N) = 0$, and we define $\mathcal{N}(0) = N$.

Consider a path through this trellis given by the sequence of values $(N, \mathcal{N}(1), \dots, \mathcal{N}(N-1), 0)$, with $\mathcal{N}(m) \geq \mathcal{N}(n)$ for all $m \leq n$. We will say that node $i, i = 1, \dots, N$, is *aggregated* if $\mathcal{N}(i-1) = \mathcal{N}(i)$, and *aggregating* otherwise. If node i is aggregating, then the number of nodes it aggregates is given by $\mathcal{N}(i-1) - \mathcal{N}(i)$.

It is clear that any tree of depth less than or equal to two can be mapped to a sequence of $\mathcal{N}(\cdot)$'s and vice versa. For example, suppose we have $N = 5, L = 2$, and the two aggregating nodes are given by node 1 with $\alpha(1) = 2$, and node 4 with $\alpha(2) = 1$. This corresponds to the sequence $(5, 2, 2, 2, 0, 0)$.

It is also clear that, since the total distortion at the central node is additive over clusters and the optimal cluster distortions can be allocated independently, we can choose the optimal first $i - 1$ steps among all \mathcal{N} sequences that pass through $\mathcal{N}(i) = j$, simply by comparing the distortions they induce at the central node. Hence to carry out step i , we need to only consider the $j + 1$ paths that emanate from $\mathcal{N}(i) = j$ to $\mathcal{N}(i+1) = j, \dots, \mathcal{N}(i+1) = 0$, and repeat this for all $N + 1$ possible values of $\mathcal{N}(i)$. With N steps, each requiring such $\sim N^2$ path evaluations, we can determine the optimal sequence and thus the optimal tree with a dynamic programming algorithm of complexity $\sim N^3$.

Now, suppose we generalize to the case where the node capacities to the aggregator of cluster l are given by a cluster specific constant $C_l = C(\alpha_l)$, i.e., $C_{i\gamma(l)} = C(\alpha_l)$ if $i \in \xi(l)$. For example in the transmission scenario where bandwidth inside each cluster is allocated as a function of α_l we can have $C(\alpha_l) = W/\alpha_l$. This changes the distortion allocation specifics in Equation (19) but the

cluster additive nature of the total central node distortion and the independent allocation of optimal cluster distortions do not change. Hence the above outlined dynamic program will again find the optimal tree. Note that the algorithm can also be utilized to solve a specific case of **Case 2** with a single statistic that is the summation of all measurements, i.e., $g_{1i} = 1$. However, the solution for general G remain difficult. Section V includes simulation results that compare **Case 1** solutions with the specific **Case 2** scenario.

B. Efficient topology formation with tree-growing heuristic

Topology formation has been an active research area in sensor and ad hoc networks from the power and capacity efficiency perspectives. In this section, we utilize our earlier results to construct an efficient topology in the distortion sense under an arbitrary setting, where capacity constraints C_{ij} 's between node pairs $\{i, j\}$ and σ_i 's are arbitrary. Our objective is to deliver the desired statistics to node 0 with minimal total distortion D_T^* by altering the routing tree, which in return determines a hierarchy of aggregation nodes with node 0 at the root.

We start from an initial tree T_0 with depth one, i.e. all nodes 1 through N directly send their individual measurements to node 0. Our algorithm iterates over the trees: at k th iteration, it inputs T_k and outputs T_{k+1} . We also define the total distortion due to tree T_k as $D(T_k)$ and denote the subtree rooted at i as $ST(i)$. Below, we provide the pseudo-code of our algorithm.

```

while  $T_{k+1} \neq T_k$ 
  for all  $i \in \{1, \dots, N\}$ 
     $p = \arg \min_{j \notin ST(i)} D(f(T_k, i, j));$ 
     $T_{k+1}(i) = f(T_k, i, p);$ 
  end
   $T_{k+1} = \min_i T_{k+1}(i);$ 
   $k = k + 1;$ 
end

```

Function $f(T_k, i, j)$ in this procedure constructs a new tree by changing the *parent* node of i to j by keeping the subtree of i intact. We can compute $D(T_k)$ for both transmission schemes as examined in case1 and case2.

V. SIMULATION RESULTS AND CONCLUSION

In this paper we considered sensor networks with link capacity constraints and we proposed techniques that jointly optimize compression and routing under two cases of information flow. In the first case (**Case 1**), the network is optimized to convey all of the sensor measurements under a fidelity criterion, optimally aggregating (and if it is to the benefit of the overall network, sometimes discarding) sensor measurements, compressing them, and routing them. In the second case, where the ultimate application is detection related (**Case 2**), our results show that optimized networks significantly outperform their first case counterparts (which are optimized without regard to the final application), provided that in-network bandwidth is sufficient. The second case optimization is also much less willing to discard individual sensor measurements for nontrivial statistics. Our work provides a locally optimal algorithm for the solution of the most general cases, while identifying specific subcases which can be solved optimally using proposed algorithms.

REFERENCES

- [1] Thomas M. Cover and Joy A. Thomas, "Elements of Information Theory", Wiley-Interscience, 1991.
- [2] Albert Cohen, Ingrid Daubechies, Onur G. Guleryuz, and Michael T. Orchard, "On the importance of combining wavelet-based nonlinear approximation with coding strategies," IEEE Transactions on Information Theory, vol. 48, no. 7 pp. 1895-1921, July 2002.
- [3] Enrique J. Duarte-Melo and Mingyan Liu, "Data-Gathering Wireless Sensor Networks: Organization and Capacity," *Computer Networks (COMNET)*, Special Issue on Wireless Sensor Networks, vol. 43, no. 4, pp. 519-537, November 2003.