Reliability vs. Efficiency in Distributed Source Coding for Field-Gathering Sensor Networks *

Daniel Marco
University of Michigan
EECS Department
1301 Beal Ave.
Ann Arbor, MI 48109
idaniel@umich.edu

David L. Neuhoff
University of Michigan
EECS Department
1301 Beal Ave.
Ann Arbor, MI 48109
neuhoff@umich.edu

ABSTRACT

The tradeoff between reliability and efficiency in distributed source coding for field-gathering sensor networks is examined. In the considered networks, sensors measure some underlying random field, quantize their measurements, encode the quantized values into bits and transmit these directly, or via relays, to a collector that reconstructs the field. The bits from one sensor's encoder are regarded as a packet. The minimum achievable coding rate can be attained if the sensors are ordered and each applies Slepian-Wolf distributed coding to its data assuming the decoder knows the data from all prior sensors. However, with such a coding scheme, losing even one sensor's packet would cause decoding failure for all subsequent sensors' values. Therefore, one might consider other ways of applying Slepian-Wolf coding, where in trade for increased coding rate, fewer sensor values are lost when a packet is lost. In this paper, the tradeoff between efficiency, i.e. coding rate, and reliability, characterized by a loss factor, is considered for several different Slepian-Wolf based coding schemes as a function of the packet error probability and the size of the network.

Categories and Subject Descriptors

H.1.1 [**Models and Principles**]: Systems and Information Theory—*Sensor Networks*

General Terms

Performance, Design, Reliability, Theory

Keywords

Sensor Networks, Reliability, Efficiency, Distributed source coding

*This work was supported by NSF Grant ANI-0112801.

1. INTRODUCTION

Consider a sensor network whose purpose is to measure some random field. Each sensor samples the field at its location, quantizes the samples values, encodes the quantized values into bits and transmits these bits to a central processing unit, which we name the collector. At the collector a reconstruction of the field is created. We assume that all sensors transmit their encoded bits directly to the collector, or that there exists a predetermined set of relay sensors, which do not sample the field, do not generate new bits and do not aggregate encoded packets from various sensors.

In order to minimize the number of generated bits, sensors utilize Slepian Wolf coding [1]. The idea behind which is that lossless encoders can separately encode data from correlated sources (such as the data produced by neighboring sensors) as efficiently as if each encoder could see the values produced by the other data sources, under the assumption that these values are perfectly known at the decoder. It follows that when the encoded bits from one sensor are lost, the effect might be significant in the sense that the collector may not be able to reconstruct several sensor values. Let p denote the probability that the encoded bits of a sensor are lost (either due to relay failure, or due to any other reason). One should think of the encoded bits from one sensor as a packet. Thus, p is the probability of losing a packet, and the assumption is that packets are lost independently. For each sensor the decoder may require some number of packets from additional sensors in order to decode its value. If even one of these packets is not received, then the sensor value cannot be reconstructed at the decoder, i.e. that sensor value is lost. This motivates the following

Definition 1. *The* loss factor *of a wireless sensor network (denoted by L) is the expected fraction of sensor values that cannot be reconstructed at the collector.*

Clearly, the larger the loss factor the larger the distortion of the reconstructed field. The precise relationship between these two quantities depends on the distortion measure used and on the reconstruction method applied at the collector. In this paper we are not concerned with a particular distortion measure or reconstruction method, and thus all discussion will be with respect to loss factor rather than distortion. We do note, however, that when the distortion measure is mean squared error (MSE), and when the loss factor goes to one, the distortion of the reconstructed field goes to the variance for any reasonable reconstruction method (an unreasonable reconstruction method could attain arbitrarily large distortion, of course). We will refer to this fact again in Subsection 4.1.

Our goal is to characterize the tradeoff between the rate (i.e. the average number of newly generated bits per sensor) and the loss factor of various encoding schemes. A desirable scheme would have small rate and small loss factor.

We note that there has been some work with regard to packet distribution reliability in sensor networks. For example, in [2] reliability of small and large message deliveries was examined; in [3] the connectivity of sensor networks, in which sensors have some failure probability, was considered; and in [4] an information-theoretic rate-distortion region was derived, for a sensor network in which sensor measurement statistics are symmetric, and where k out of n packets are guaranteed to arrive at the collector.

The remainder of this paper is organized in the following way. In Section 2 we describe four encoding schemes and find formulas for their rates and loss factors. When the measured field is one-dimensional, schemes 1, 2 and 3 can be improved. Section 3 discusses these improved schemes. In Section 4 scheme comparisons are made. Finally, our concluding remarks are offered in Section 5.

2. FOUR ENCODING SCHEMES

In this section we describe four encoding schemes. For each one we will evaluate the loss factor and the rate. We assume that the measured field has some type of spatial stationarity so that if sensors are divided into congruent clusters (as will be discussed shortly), these clusters behave the same in terms of rate and loss factor.

Schemes 0 and 1 represent benchmark schemes, in the sense that they each operate on one extreme of the loss factor vs. rate curve (to be discussed in Section 4). Let N denote the number of sensors. Let $X_1, ..., X_N$ and $Y_1, ..., Y_N$ denote the N sensor values and their respective quantized versions. Let p denote the packet loss probability. Following is a description of the four schemes, which are illustrated in Figure 1.

Scheme 0 - No Slepian-Wolf:

Each sensor encodes its quantized value independently of all other sensors (so no Slepian-Wolf coding is performed). The rate of this scheme, i.e. the number of encoded bits per quantized sensor value, is

$$R_0 = H(Y_1) \ , \tag{1}$$

where $H(Y_1)$ is the entropy of a quantized sensor value and we have assumed $H(Y_i)$ is same for all i (which is a kind of stationarity).

As for the loss factor, whenever a packet is lost, it affects only the sensor that sent that packet. Therefore,

$$L_0(p) = \frac{1}{N}\sum_{i=1}^{N} p = p \ . \tag{2}$$

Notice that neither the rate nor the loss factor depend on the number of sensors, N.

Scheme 1 - Sequential Slepian-Wolf:

Sensors utilize sequential Slepian-Wolf coding. Namely, Sensor 1 encodes Y_1 using rate $H(Y_1)$. Sensor 2 encodes Y_2 using rate $H(Y_2|Y_1)$, assuming the decoder knows Y_1. Continuing in this fashion, Sensor i encodes Y_i using rate $H(Y_i|Y_1, Y_2..., Y_{i-1})$, assuming that the decoder knows $(Y_1, Y_2..., Y_{i-1})$. The rate of this scheme is the N^{th} order entropy of the quantized sensors values, i.e.

$$R_1(N) = \frac{H(Y_1, ..., Y_N)}{N} = H_N \ . \tag{3}$$

Consider next the loss factor. We observe that there is a lack of symmetry between the sensors in terms of the importance of their packets. Specifically, if the first packet is lost (i.e. the encoded bits of Sensor 1), then no other sensor can be decoded at the collector. Thus, a loss of the first packet induces an overall reconstruction failure of all N sensors. However, loss of the N^{th} packet would cause no other sensor to suffer reconstruction failure. To evaluate the expected number of failed sensors (i.e. sensors that cannot be reconstructed at the collector), we observe that in order for exactly $N-k$ sensors to fail, packets 1 through k must arrive at the collector and packet $k+1$ must not arrive at the collector. Thus,

$$L_1(p, N) = \frac{1}{N}\sum_{k=0}^{N}(N-k)(1-p)^k p = 1 - \frac{1}{N}\frac{1-p}{p}\left[1-(1-p)^N\right] \ . \tag{4}$$

Remark: If $N \to \infty$, then $\frac{1}{N}\frac{1-p}{p}\left[1-(1-p)^N\right] \to 0$ and so $\lim_{N\to\infty} L_1(p, N) = 1$, which means that asymptotically the loss factor approaches 1.

Scheme 2 - Clustered Slepian-Wolf:

Sensors are partitioned into congruent clusters of equal size K. We assume that K divides N so as to avoid edge affects. Each cluster performs its encoding separately and independently of all other clusters. The encoding in each cluster is sequential Slepian-Wolf as in Scheme 1. Consider the rate as function of the cluster size K. Each cluster has the same rate (due to spatial stationarity of the field) $\frac{1}{K}H(Y_1, ..., Y_K)$, which is the rate of the scheme in general. Thus,

$$R_2(K) = \frac{1}{K}H(Y_1, ..., Y_K) \ . \tag{5}$$

Next, let us consider the loss factor. We observe that when a packet is lost it can influence at most the K sensors of its cluster in the reconstruction phase. Since all clusters have the same size and the same encoding scheme, it follows that all clusters have the same cluster loss factor. It follows from (4) that the cluster loss factor is $1-\frac{1}{K}\frac{1-p}{p}[1-(1-p)^K]$. Since the overall loss factor is the average of these loss factors, it follows that the loss factor of the entire network equals the cluster loss factor. Thus,

$$L_2(p, K) = 1 - \frac{1}{K}\frac{1-p}{p}\left[1-(1-p)^K\right] \ . \tag{6}$$

Remarks:

1. If $K = 1$, i.e. no Slepian-Wolf coding is utilized, then Scheme 2 reduces to Scheme 0, and indeed $L_2(p, 1) = 1 - \frac{1-p}{p}[1-(1-p)] = p$, as expected.
2. If $K = N$, then Scheme 2 reduces to Scheme 1, and indeed $L_2(p, N) = 1 - \frac{1}{N}\frac{1-p}{p}[1-(1-p)^N] = L_1(p, N)$.
3. If $N/K = c$, $c \geq 1$ and N is increased (this corresponds to clustering together sensors in a certain geographical

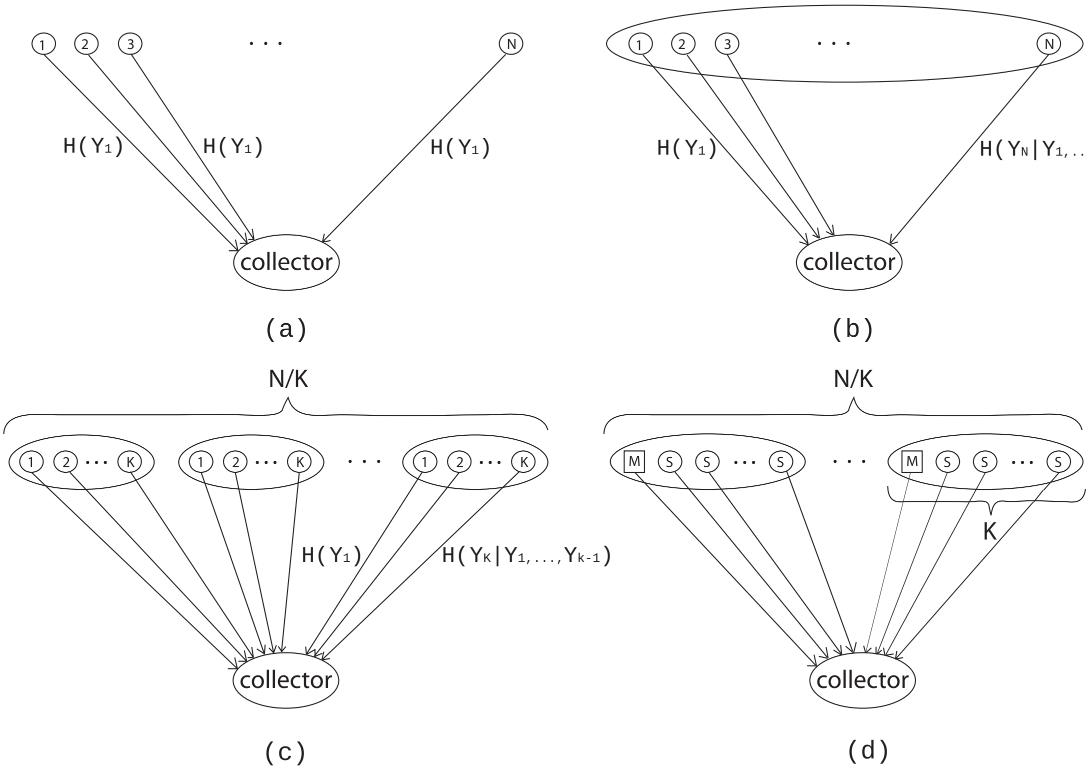

Figure 1: Four encoding schemes (general case). (a) No Slepian-Wolf coding. (b) Sequential Slepian-Wolf. (c) Clustered Slepian-Wolf. There are N/K clusters of size K. (d) Master–Slave. There are N/K clusters of size K.

area and increasing their density), then as $N \to \infty$ also $K \to \infty$ and so $\frac{1}{K}\frac{1-p}{p}[1-(1-p)^K] \to 0$. Thus, $\lim_{N\to\infty} L_2(p, \frac{N}{c}) = 1$. This is not surprising since each cluster behaves much like Scheme 1, where the loss factor goes to 1 as N tends to infinity.

4. The loss factor depends on K but not on N.

Scheme 3 - Master-Slave:

Sensors are partitioned into congruent clusters of equal size K. As before, we assume K divides N. Each cluster performs its encoding separately and independently of all other clusters. In each cluster one sensor is the *master* and all other sensors are *slaves*. The master encodes its quantized value using rate $H(Y_1)$. Each slave encodes its quantized value using Slepian-Wolf coding with respect to the master. Thus, the i^{th} slave encodes its quantized value using rate $H(Y_i|Y_1)$. The rate of all clusters is the same (due to spatial stationarity of the field) and equals the rate of the whole network. This rate is given by

$$R_3(K) = \frac{1}{K}[H(Y_1) + \sum_{i=2}^{K} H(Y_i|Y_1)] \ . \tag{7}$$

Next, consider the loss factor. As in the case of Scheme 2, the network loss factor equals the cluster loss factor. In order to evaluate the cluster loss factor we observe that when the packet corresponding to the master sensor is lost, all sensors in the cluster cannot be reconstructed. When, however, a packet corresponding to a slave is lost, it causes no additional sensors to suffer reconstruction failure. We consider one cluster and define an indicator variable

$$I_i = \begin{cases} 1, & \text{if sensor } i \text{ cannot be reconstructed} \\ 0, & \text{else} \end{cases} \ , \tag{8}$$

for all $i \in \{1, ..., K\}$. Next, observe that $\Pr(I_i = 1) = \Pr(\text{packet 1 is lost and/or packet } i \text{ is lost}) = 2p - p^2$, when $i \in \{2, ..., K\}$. $\Pr(I_1 = 1) = p$ (recalling that Sensor 1 is the master of the cluster). Now, the loss factor can be evaluated as follows,

$$\begin{aligned} L_3(p, K) &= \frac{1}{K}\sum_{i=1}^{K} E[I_i] = \frac{1}{K}\left[p + \sum_{i=2}^{K}(2p - p^2)\right] \\ &= p(2-p) - \frac{1}{K}p(1-p) \ . \end{aligned} \tag{9}$$

Remarks:

1. If $K = 1$, i.e. no Slepian-Wolf coding is utilized, then Scheme 3 reduces to Scheme 0, and indeed $L_3(p, 1) = p(2-p) - p(1-p) = p$, as expected.
2. If $N/K = c$, $c \geq 1$ and N is increased, then as $N \to \infty$ also $K \to \infty$. Thus, $\lim_{N\to\infty} \frac{1}{K}p(1-p) = 0$, and so $\lim_{N\to\infty} L_3(p, \frac{N}{c}) = p(2-p)$.
3. If $K = N$, then $\lim_{N\to\infty} L_3(p, N) = p(2-p)$. This is not surprising since this case may be viewed as the previous case with $c = 1$.

4. As in Scheme 2, the loss factor depends on K but not on N.
5. For any K, $p \leq L_3(p, K) < 2p - p^2$.

3. IMPROVED ENCODING SCHEMES FOR ONE-DIMENSIONAL FIELD

Consider a one-dimensional network, measuring a stationary Markov random process such that we may approximate the stationary sampled quantized process as being Markov as well. Alternatively, let us assume that the encoders/decoders are such that encoding and decoding can be performed conditioned only on one other quantized value. Under these new settings, the schemes described can be improved. These improved schemes are described next.

Scheme $1'$ - Sequential Slepian-Wolf:

The scheme is the same as Scheme 1, except for having the sensor that encodes its quantized value at rate $H(Y_1)$ (which we call the *head*) be in the center. Due to the Markov property, the rate is the same as that of Scheme 1 (or more precisely, the rate is the same as that of Scheme 1 when conditioning is allowed with respect to only one other sensor). The rate is given by

$$R_1'(N) = \frac{1}{N} H(Y_1) + \frac{N-1}{N} H(Y_2|Y_1) \ . \tag{10}$$

As for the loss factor, it becomes smaller than that of Scheme 1. Specifically, we let I_i be as defined in (8), where the head is referred to as sensor 0, and the sensors to its right are numbered 1 through $(N-1)/2$, and the sensors to its left are numbered -1 through $-(N-1)/2$. We assume N is odd. It follows that $\Pr(I_i = 1) = 1 - (1-p)^{|i|+1}$. Consequently,

$$\begin{aligned} L_1'(p, N) &= \frac{1}{N} \sum_{i=-\frac{N-1}{2}}^{\frac{N-1}{2}} E[I_i] \\ &= 1 - \frac{1-p}{N} - \frac{2}{N} \frac{(1-p)^2}{p} \left[1 - (1-p)^{\frac{N-1}{2}} \right] \ , \end{aligned} \tag{11}$$

where the second equality follows using some algebraic steps.

Scheme $2'$ - Clustered Slepian-Wolf:

As in the case of Scheme $1'$, the improvement lies in having the cluster heads lie in the center of their clusters. The rate and loss factors are the same as for Scheme $1'$, except for changing N's to K's, where K is the cluster size. As in the case of schemes 2 and 3, we assume that K is odd and divides N. It follows that

$$R_2'(K) = \frac{1}{K} H(Y_1) + \frac{K-1}{K} H(Y_2|Y_1) \ , \tag{12}$$

and

$$L_2'(p, K) = 1 - \frac{1-p}{K} - \frac{2}{K} \frac{(1-p)^2}{p} \left[1 - (1-p)^{\frac{K-1}{2}} \right] \ . \tag{13}$$

Schemes $3'$ and $3''$ - Master-Slave:

We consider two improvements to the master slave scheme. The first improvement is as for the other schemes, i.e. letting the master be in the center. We dub this scheme as Scheme $3'$. In such a case the loss factor remains the same as for Scheme 3, namely,

$$L_3'(p, K) = p(2-p) - \frac{1}{K} p(1-p) \ , \tag{14}$$

where K is the cluster size and is assumed to divide N.

The rate, however, decreases since sensors are no longer as far away from the master as previously. Specifically,

$$R_3'(K) = \frac{1}{K} [H(Y_0) + 2 \sum_{i=1}^{\frac{K-1}{2}} H(Y_i|Y_0)] \ , \tag{15}$$

where K is assumed to be odd and where sensor 0 is the master, and sensors 1 through $(K-1)/2$ lie to its right, and sensors -1 through $-(K-1)/2$ lie to its left.

As it is readily seen, in the one-dimensional case considered, Scheme $3'$ is strictly superior to Scheme 3, since for any loss factor it has better rate. Next, we consider a modified Scheme $3'$, which we refer to as Scheme $3''$. In this scheme the master transmits its packet twice. Since every transmission fails independently of all other transmissions, the probability of losing the master's packet becomes p^2. As we shall see shortly this leads to major reduction in the loss factor. We should point out that this comes at a cost of higher rate. However, when N is large (which is typically the case in a dense sensor network), the excess rate is marginal.

The rate for Scheme $3''$ is simply the rate of Scheme $3'$ plus $H(Y_1)/K$, thus,

$$R_3''(K) = \frac{1}{K} [2H(Y_0) + 2 \sum_{i=1}^{\frac{K-1}{2}} H(Y_i|Y_0)] \ . \tag{16}$$

As for the loss factor, we define the indicator variable I_i as in (8), where the sensors are numbered as in Scheme $3'$. Next, we observe that $\Pr(I_i = 1) = \Pr(\text{packet 0 is lost and/or packet } i \text{ is lost}) = p + p^2 - p^3$, when $i \neq 0$, and $\Pr(I_0 = 1) = p^2$ (recalling that Sensor 0 is the master of the cluster). The loss factor is obtained as follows,

$$L_3''(p, K) = \frac{1}{K} \sum_{i=-\frac{K-1}{2}}^{\frac{K-1}{2}} E[I_i] = p(1 + p - p^2) - \frac{p(1-p^2)}{K} \ , \tag{17}$$

where the last equality follows using some algebraic steps.

Remarks:

1. When p is small and K is large schemes 3 and $3'$ have a loss factor approximately $2p$, while Scheme $3''$ has loss factor approximately p.
2. It is not hard to see that having the master send its packet more than twice will have negligible affect on the loss factor.
3. The idea behind Scheme $3''$ (i.e. sending the master's packet twice) can be implemented in the general case as well.

4. COMPARISON OF SCHEMES

In this section we compare the various schemes first in the general case and then for the one-dimensional Markov case.

Table 1: Rates and loss factors for the general case.

Scheme	Rate	Loss Factor
0	$H(Y_1)$	p
1	$\frac{1}{N}H(Y_1,...,Y_N)$	$1-\frac{1}{N}\frac{1-p}{p}[1-(1-p)^N]$
2	$\frac{1}{K}H(Y_1,...,Y_K)$	$1-\frac{1}{K}\frac{1-p}{p}[1-(1-p)^K]$
3	$\frac{1}{K}\left[H(Y_1)+\sum_{i=2}^{K}H(Y_i\|Y_1)\right]$	$p(2-p)-\frac{1}{K}p(1-p)$

4.1 General Case

Table 1 Summarizes the rates and loss factors of the four schemes in the general case. First let us recall that schemes 0 and 1 are special cases of Scheme 2. They will serve as benchmarks. We compare the schemes under two settings.

Case 1: Number of clusters $N/K = c$, $c \geq 1$ is constant.

It follows from the remarks made in Section 2 that as $N \to \infty$ the loss factors of schemes 1, 2 and 3 become $L_1(p,N) \approx 1$, $L_2(p, \frac{N}{c}) \approx 1$ and $L_3(p, \frac{N}{c}) \approx p(2-p)$, respectively. Translating this in terms of MSE distortion, schemes 1 and 2 attain the variance of the measured field (since no sensors can be reconstructed). Therefore, regardless of any rate benefit these schemes might have with respect to the Scheme 3, it is clear that these two schemes are inferior to Scheme 3. As for Scheme 0, it is not affected by enlarging N. Its rate and loss factor remain constant. While, its loss factor, is indeed favorable, its rate is very high.

Case 2: Cluster size K is constant.

As before, Scheme 0 is not affected by N. Furthermore, as before, letting $N \to \infty$, makes $L_1(p,N) \to 1$. As for schemes 2 and 3, their loss factors are given in Table 1. As mentioned earlier when $K = 1$, $L_2(p,1) = L_3(p,1) = p$ (of course the rates of these two schemes are the same as well). One should point out that cluster size should naturally be determined by the distribution function and correlation function of the measured field. Thus, a cluster should cover a certain geographical area. Consequently, increasing the number of sensors in the network (while letting the network be deployed over the same fixed geographical area) will make each cluster have more sensors. Therefore, it is reasonable to assume large values of K, in which case, the previous case applies again.

4.2 One-Dimensional Case

Table 2 Summarizes the rates and loss factors of schemes 0, $1'$, $2'$, $3'$ and $3''$.

We computed the performance of these schemes for particular settings. Specifically, we considered a zero mean one-dimensional continuous-space-parameter stationary Gaussian random process with variance σ^2 and exponential autocorrelation $R_x(\tau) = \sigma^2 e^{-|\tau|}$. (Note, that a Gaussian random process with exponential autocorrelation is a Markov process.) We let the sensors utilize infinite level uniform scalar quantizers with step size Δ and a level at the origin. The network is deployed over one-dimensional region of length 1, and so the spacing between sensors is $1/N$. When N is large this spacing is small. Therefore, we may use the asymptotic (for small spacing between sensors) formula from [5] that evaluates $H(Y_i|Y_j)$ for $i \neq j$ and sufficiently small values of $\tau_{i,j} \triangleq |i-j|/N$. For the given autocorrelation function, the

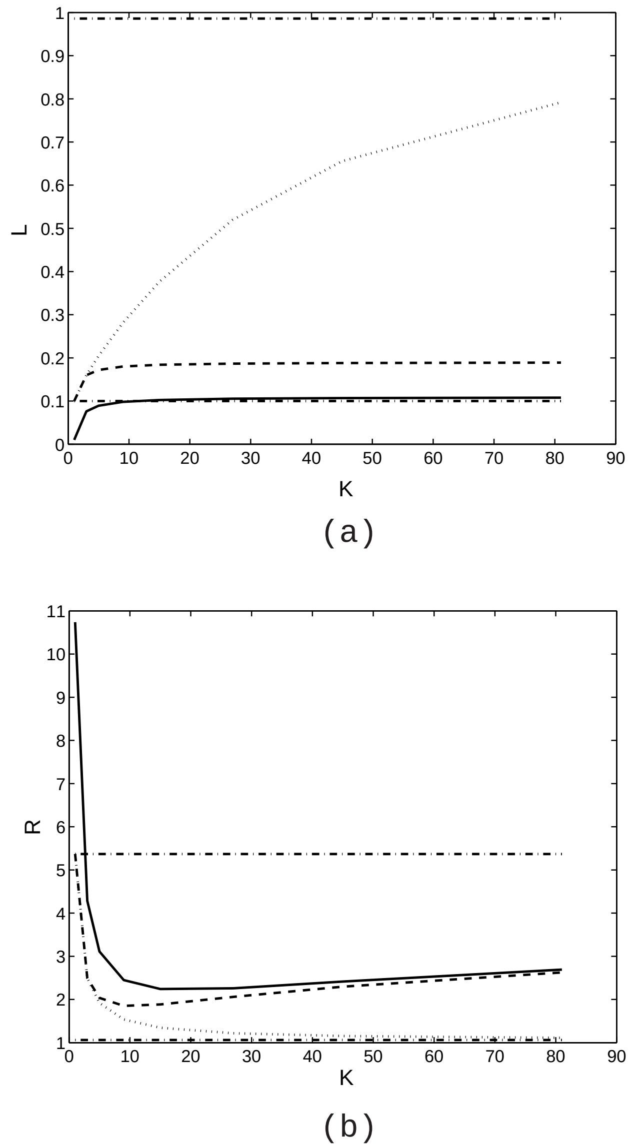

Figure 2: (a) Loss Factor vs. Cluster Size, for $N = 1215$ and $p = 0.1$ (b) Rate vs. Cluster Size K, for $N = 1215$. In both (a) and (b) the dashdot line denotes schemes 0 and $1'$, the dotted line denotes Scheme $2'$, the dashed line denotes Scheme $3'$ and the solid line denotes Scheme $3''$.

Table 2: Rates and loss factors for the one-dimensional Markov case.

Scheme	Rate	Loss Factor
0	$H(Y_1)$	p
$1'$	$\frac{1}{N}H(Y_1) + \frac{N-1}{N}H(Y_2\|Y_1)$	$1 - \frac{1-p}{N} - \frac{2}{N}\frac{(1-p)^2}{p}\left[1-(1-p)^{\frac{N-1}{2}}\right]$
$2'$	$\frac{1}{K}H(Y_1) + \frac{K-1}{K}H(Y_2\|Y_1)$	$1 - \frac{1-p}{K} - \frac{2}{K}\frac{(1-p)^2}{p}\left[1-(1-p)^{\frac{K-1}{2}}\right]$
$3'$	$\frac{1}{K}[H(Y_0) + 2\sum_{i=1}^{\frac{K-1}{2}} H(Y_i\|Y_0)]$	$p(2-p) - \frac{1}{K}p(1-p)$
$3''$	$\frac{1}{K}[2H(Y_0) + 2\sum_{i=1}^{\frac{K-1}{2}} H(Y_i\|Y_0)]$	$p(1+p-p^2) - \frac{p(1-p^2)}{K}$

formula has the form

$$H(Y_i|Y_j) = -M_{\sigma,\Delta}\sqrt{\tau_{i,j}}\log\sqrt{\tau_{i,j}} \quad (18)$$

where $M_{\sigma,\Delta}$is a constant that depends on σ and Δ. In all our computeations we used $\sigma = 1$ and $\Delta = 0.1$.

Figure 2 shows the effect of the cluster size on the loss factor and on the rate of schemes 0, $1'$, $2'$, $3'$ and $3''$. Specifically, Figure 2a shows how the loss factor depends on the cluster size, when $N = 1215$ and $p = 0.1$. Figure 2b shows the dependence of rate on cluster size, when $N = 1215$. Before elaborating some more, we note that not every $K < N$, can serve as a cluster size. An *admissible* cluster size is one that divides N. For our computation purposes, we chose the cluster size to never exceed $N/15$, so as to ensure that the two most distant sensors in a cluster are sufficiently close that (18) may be used.

In Figure 2a, schemes 0 and $1'$ are denoted by the horizontal lines. The bottom line represents Scheme 0 with constant loss factor of p. The top line represents Scheme $1'$ with loss factor $1 - \frac{1-p}{N} - \frac{2}{N}\frac{(1-p)^2}{p}\left[1-(1-p)^{\frac{N-1}{2}}\right]$, which depends only on N, and is close to 1 for large N. Scheme $2'$ has a loss factor that increases monotonically with cluster size to the loss factor of Scheme $1'$. Lastly, schemes $3'$ and $3''$ converge very quickly to loss factors slightly larger than $2p$ and p, respectively.

Figure 2b shows that schemes 0 and $1'$ serve again as benchmarks (these are the horizontal top (Scheme 0) and bottom (Scheme $1'$) lines). The R value of the top horizontal line is $H(Y_1)$, which in our example equals approximately 5.37. The rate of Scheme $2'$ is seen to monotonically decrease to that of Scheme $1'$. Finally, we observe that the rate curves of schemes $3'$ and $3''$ are not monotonic and have minima. We will discuss this phenomenon shortly, but first let us consider Figure 3, which combines Figures 2a and 2b.

Figure 3 illustrates the loss factor as a function of rate for schemes 0, $1'$, $2'$, $3'$ and $3''$. Figures 3a and 3b were obtained by fixing N and p and varying K so as to obtain (L, R) pairs. The p values that were used for each N are $0.2, 0.1$ and 0.01. The curves that are further up the L axis correspond to higher p values, where we compare dotted curves to dotted curves, dashed curves to dashed curves and solid curves to solid curves.

The circle ('o') denotes the operation point of Scheme 0. This point is at $R = H(Y_1)$ and $L = p$ and does not depend on N. The asterisk ('*') denotes the operation point of Scheme $1'$. While it depends on N and p, it has no dependence on K. The dotted line represents Scheme $2'$. It is clear from the figures that when N is large, the loss factor tends to 1. This of course was to be expected. Finally, the dashed line and the solid line represent Scheme $3'$ and Scheme $3''$, respectively. As discussed earlier, when K is

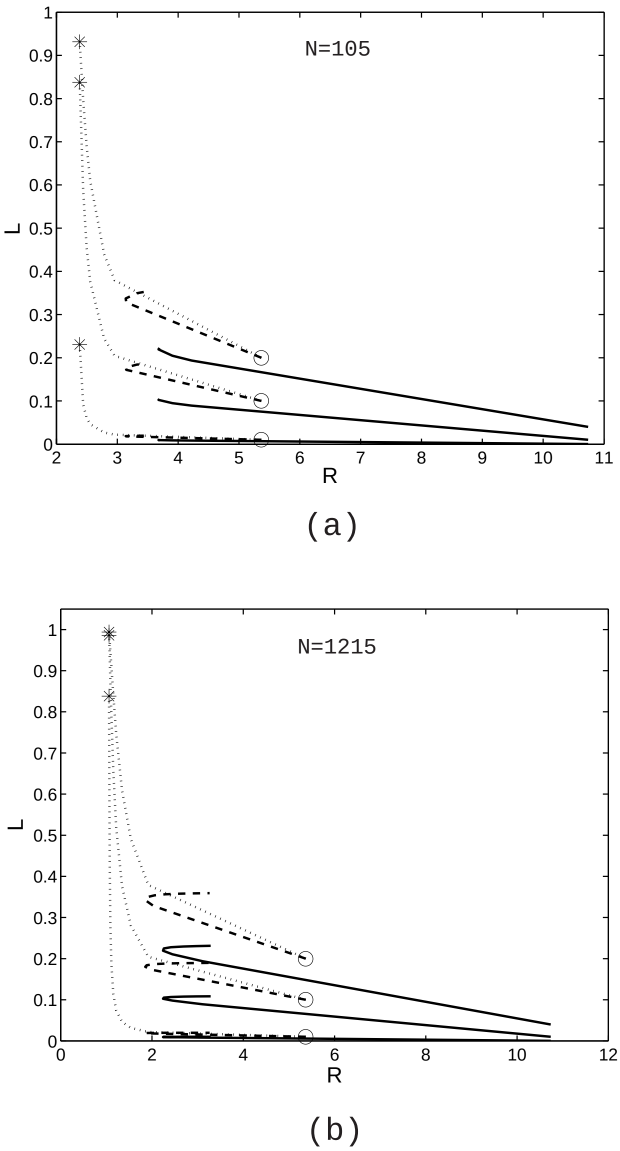

Figure 3: Loss Factor vs. Rate for two values of N and for $p = 0.2, p = 0.1$ and $p = 0.01$. 'o' denotes Scheme 0, '*' denotes Scheme $1'$, the dotted line denotes Scheme $2'$, the dashed line denotes Scheme $3'$ and the solid line denotes Scheme $3''$.

large the loss factor of Scheme $3'$ tends to $2p$ and the loss factor of Scheme $3''$ tends to p.

We observe that the curves of schemes $3'$ and $3''$ are multi-valued (this can be seen clearly in Figure 3b). The reason for this is that the curves of Scheme $3'$ and $3''$ in Figure 2b are not monotonic and they each attain a minimum. These minimum values correspond to the values of R in Figure 3b, for which the curves of schemes $3'$ and $3''$ begin to wrap back on themselves. It is natural to have schemes $3'$ and $3''$ operate at that value of K where wrapping begins. Clearly, a larger value of K results in worse performance both in loss factor and rate. Smaller values of K induce lower loss factors and higher rates. However, it appears that the rate increases much faster than the loss factor decreases. This motivates us to dub this value of K the "optimal K value". More formally,

Definition 2. *The optimal cluster size K is the K that minimizes the rate of Scheme $3'$ or $3''$ (depending on which scheme is used).*

The process by which the optimal K can be found is by increasing the cluster size of schemes $3'$ and $3''$ up to that point where the average conditional entropies of the edge sensors (i.e. left and right edges) is larger than the average entropy per sensor when the edge sensors are not included. Otherwise stated, the optimal K can be found by observing that it is the first K for which

$$R(K) < R(\widetilde{K}) \;, \tag{19}$$

where $\widetilde{K} > K$ is the next admissible value of K, and R represents the rate of Scheme $3'$ or $3''$. Notice, that since not all odd values of K are admissible, i.e. they do not all divide N (recall that N is always chosen to be odd), it follows that there could be several sensors on each edge, whose conditional entropies we average and compare to the previous rate.

Notice further that the optimal value of K does not depend on the probability of lost packet p. This is clear since p does not play a role in evaluating rates.

Figure 4 shows optimal values of K as function of N. As we can see, schemes $3'$ and $3''$ have different optimal K values. Specifically, the optimal K for Scheme $3''$ is larger than that of Scheme $3'$, for the same N. This follows from the fact that Scheme $3''$ has a higher initial rate (the extra $H(Y_0)$ term). At the same time, the conditional entropy of all sensors is the same for both schemes. Thus, the K that satisfies the condition in (19) needs to be larger for Scheme $3''$ than for Scheme $3'$.

We observe further that as N increases so does the optimal K. This follows from the fact that as N increases the spacing between the sensors decreases and thus the conditional entropies decrease. Consequently, one can add more conditional entropies before they have an ill-affect on the overall rate.

We also note that schemes $3'$ and $3''$ are sensitive to the value K in that they incur significant rate increase when K is greater than its optimal value (for the respective scheme). On the other hand, Scheme $2'$, is not as sensitive to K. We observe from Figure 3 that schemes $3'$ and $3''$ are superior to Scheme $2'$, when they operate with K values that are not greater than the optimal values.

Finally, we postulate that the optimal K for schemes $3'$ and $3''$ depends on the correlation of the measured random process, so that the size (in inches) of a master-slave cluster is constant.

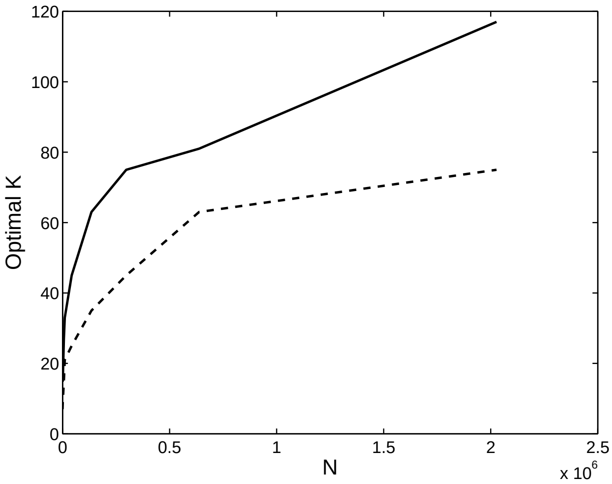

Figure 4: The optimal cluster size K of Scheme $3'$ (dashed line) and Scheme $3''$ (solid line) as a function of the network size N.

5. CONCLUSIONS

We examined several encoding schemes for sensor networks, under the premise that packets transmitted by sensors are independently lost with probability p. We first considered the problem in general, i.e. with no restrictions on the dimensionality of the network or on the measured random field. It was shown that the master-slave scheme (Scheme 3) is superior to the clustered Slepian-Wolf scheme (Scheme 2), when the cluster size is large. Specifically, the clustered Slepian-Wolf scheme has a loss factor that tends to 1 as the cluster size increases. Therefore, regardless of any rate benefit this scheme might have, it becomes useless for large cluster size.

Next, we examined the special case of a one-dimensional sensor network measuring a Gaussian random process with exponential autocorrelation. For this case, we introduced improved versions of clustered Slepian-Wolf scheme and of the master-slave scheme. We provided computations of the performance of each of the improved schemes. As in the general case, as the cluster size increases the loss factor of the improved clustered Slepian-Wolf tends to 1. We further saw that for a given number of sensors N, each of the various master-slave schemes has an optimal cluster size, which increases with N.

We should point out that these results strengthen the result in [6], where it was shown that under conditions similar to those in this paper, as $N \to \infty$ the efficiency of the network degrades to zero, where efficiency was measured by the number of cycles/times slots needed to transport the bits to the collector. The result in [6] assumed that sequential Slepian-Wolf coding is performed so as to minimize the rate (no packet loss was assumed). The results of the present paper show that in a scenario where packets can be lost, sequential Slepian-Wolf coding (or any other type of joint Slepian-Wolf coding) is of little use, when N is large, due to high loss factor. Consequently, in such a scenario one would have to use a scheme such as Scheme 3 (or an improved version of it), whose rate is higher than the rate attained by sequential Slepian-Wolf coding. Thus, reducing the efficiency, as described in [6], even further.

Finally, we make two additional remarks. First, we note that similar strategies to those used in the improved schemes for the one-dimensional field can be used for the two dimensional field as well. Secondly, we refer to the work done by Ishwar et al. [4], where a more efficient Slepian-Wolf based encoding scheme was presented, in which using rate $\frac{1}{k}H(Y_1, ..., Y_k)$ is enough to guarantee successful reconstruction of any k or more packets. However, this is under the condition that the distribution of any m quantized values is the same. This condition does not hold in the one-dimensional case we examined. If this condition could be relaxed, it would lead to a scheme that is superior to the schemes described in this paper.

6. REFERENCES

[1] D. Slepian and J. Wolf, "Noiseless coding of correlated information sources," *IEEE Transactions on Information Theory*, vol. 19, pp. 471-480, Jul. 1973.

[2] S. J. Park and R. Sivakumar, "Sink-to-Sensors Reliability in Sensor Networks," *Poster Presentation, Mobile Add Hoc Networking and Computing (MobiHoc), Annapolis, MD.*, pp. 27-28, June 2003.

[3] S. Shakkottai, R. Srikant, and N. B. Shroff, "Unreliable Sensor Grids: Coverage, Connectivity and Diameter," *CSL Technical Report.*, June 2002.

[4] P. Ishwar, R. Puri, S. S. Pradhan and K. Ramchandran, "On Rate-Constrained Estimation in Unreliable Sensor Networks," *Workshop on Information Processing in Sensor Networks (IPSN), Palo Alto, CA.*, pp. 178-192, Apr. 2003.

[5] D. Marco and D. L. Neuhoff, "On the Entropy of Quantized Data at High Sampling Rates," *In preparation.*

[6] D. Marco, E. 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," *Workshop on Information Processing in Sensor Networks (IPSN), Palo Alto, CA.*, pp. 1-16, Apr. 2003.