LOW RATE DISTRIBUTED QUANTIZATION OF NOISY OBSERVATIONS

ROBERT M. GRAY, STEPHEN BOYD, & TOM LOOKABAUGH
Information Systems Laboratory, Stanford University, Stanford, CA

ABSTRACT

We consider the achievable performance of a network of nodes in which the nodes view a common random variable corrupted by a site specific noise and share low rate information about their observations in order to reproduce the common random variable.

INTRODUCTION

Distributed information systems consist of many isolated subsystems or nodes which measure their environment and take actions based on their own measurements and on information supplied to them by other subsystems in the network. A typical example is a distributed sensor network where isolated remote sensors and their associated computing facilities are connected by digital communication links; the separate sensors must use their own observations together with shared digital side information from other sensors in order to perform reliable estimation or detection on objects moving through the network.

Distributed compression and classification systems form a special case. An example is the following: Many separate nodes make measurements of a common phenomenon with different measurement noise and share digital information based on these measurements so that each node can combine its observation with the data from the other nodes in order to well reproduce (compression) or make a decision about (classification) the common phenomenon. Such a system can be viewed as a source coding analog of the broadcast network considered by Gallager [5]. Results providing Shannon theory bounds and simple code design techniques have been developed for the special case of only two nodes [1,2,3,4]. It is of interest, however, to consider the case where there are a large number of nodes, but the permitted communication rate is small, say one bit. The goal is to find the achievable performance for a fixed block size asymptotically as the number of nodes grows. This is in distinction to the usual asymptotic quantizer theory which considers asymptotically large rate for fixed block size or the Shannon theory which considers asymptotically large dimension for fixed rate. In this paper we consider a simple example of such a system.

DISTRIBUTED QUANTIZATION

Suppose that there are N nodes and that each node makes an observation $Y_i = X + W_i$, $i=1,2,\ldots,N$, where X is a random variable (the underlying phenomenon) and the W_i are noises. For example, X provides information on the location of a target and the W_i are measurement noises. In practice we will usually be interested in a sequence of such observations, but we here formulate the problem as a "one-shot" system to focus on the node action at a specific time. For simplicity we assume that the noise terms W_i are independent, e.g., the nodes are widely separated, and in addition, the noises are independent of the common random variable X. Each node sends low rate digital information about its measurement, say $q_i(Y_i)$, to all of the other nodes and attempts to either reproduce (compression) or classify X based on its own observation and the shared low rate information.For example, node #1 observes Y_1 and produces a digital signal $q_1(Y_1)$ which it sends to the other nodes. It then makes a decision based on Y_1 and $q_i(Y_i)$; $i=2,3,\cdots,N$. The $q_i(y)$ could be ordinary quantizers and simply attempt to send a good reproduction of y, but they might do something quite different, however, such as to extract some piece of information from the observations that is more meaningful to the other nodes when pooled with all of the digital data. In either case, however, we shall refer to the q_i as quantizers since they map analog variables into binary variables.

As a related problem, each node might only be able to use its own quantized output and hence must base its decision on $q_i(Y_i)$; $i=1,\cdots,N$. In this case all of the nodes use the same information to reach their final decision. Clearly this system provides a bound on the performance of the previous system since the decision makers have less information. If one has a distortion measure $d(x,\hat{x})$ which measures the distortion of reproducing x as $\hat{x}$, the common design goal of information theory is to minimize for each node the average distortion $E(d(X,\hat{X}(Y_i,q_j(Y_j);j\neq i)))$, where $\hat{X}$ is the reproduction of the original X based on the available information. By appropriate choice of the distortion measure, one can also treat the problem of classification of which of a finite collection of possible distributions produced X (this is essentially what vector quantization of linear predictive coded speech does).

Suppose that the quantizers are all constrained to have only 1 bit of information, that is, they are binary quantizers. One strategy would be to make all of the quantizers do as good a job at reproducing their observation with minimum average distortion (using, e.g., the Lloyd-Max algorithm), but this may not be good given the overall goal of sharing information to reproduce the underlying variable. A natural question is the following: Does there exist a strategy such that the average distortion between the guess based upon the shared information and the true value tends to 0 as $N\to\infty$? On one hand, one is getting an infinite number of bits in the limit and one would expect that that would yield ever smaller distortions, eventually converging to zero. On the other hand, each bit provides information about a random vector corrupted by continuous noise and hence in a sense there is more noise than 1 bit per node can handle. We shall argue below that at least in a special case, as $N \to \infty$ an appropriate choice of quantizers yields an arbitrarily small average distortion at each node, even though the rate of each node is only 1 bit per vector.

AN EXAMPLE

For simplicity we consider the scalar case so that the observation is a real random variable X. We also make the following simplifying assumptions:

1. The noises W_i are independent, identically distributed, and nonnegative, and EW is known (in particular, W has an expectation).
2. The random variable X is nonnegative, independent of the W_i, and $EX < \infty$.
3. The distortion measure is the squared error $d(x,\hat{x}) = (x-\hat{x})^2$.

Theorem: If 1. through 3. above hold, then there is a sequence of 1 bit quantizers $q_i^{(N)}$, $i=1,\cdots,N^2$ and decoding rules $\hat{X}_N = f(q_i^{(N)}(Y_i)$, $i=1,\cdots,N^2)$, such that

$$\lim_{N\to\infty} Ed(X,\hat{X}_N) = 0,$$

that is, by choosing the number of nodes N^2 large enough the average distortion for each node can be made arbitrarily small.

Sketch of Proof: For each N define for the ith node the quantizer

$$q_i^{(N)}(y) = 1_{(-\infty,i/N]}(y),\ i=1,\cdots,N^2,$$

where $1_F(y)$ is the indicator function for F, that is, is 1 if $y \in F$ and 0 otherwise. Define the decoder function

$$\hat{X}_N = \frac{1}{N}\sum_{i=1}^{N^2}(1-q_i^{(N)}(y_i))-EW. \quad (1)$$

For the moment consider conditional expectations with $X=x$ fixed. We have

$$E(\hat{X}_n \mid X=x) = \frac{1}{N}\sum_{i=1}^{N^2} Pr(W+x>i/N) - EW. \quad (2)$$

Using the fact that $Pr(W+x>z)$ is monotone in z, it can be shown that this sum converges as $N \to \infty$ to the integral

$$E(\hat{X}_N \mid X=x) \underset{N\to\infty}{\to} \int_0^\infty (1-F_{W+x}(w))dw - EW = x,$$

where F_{W+x} is the cumulative distribution function of $W+x$. Thus the estimate $\hat{X}_N$ is an asymptotically unbiased estimate of x.

Considering next the conditional variance, we have that

$$\sigma^2_{\hat{X}_N^2|X=x} = \frac{1}{N^2}\sum_{i=1}^{N^2} Pr(x+W>i/N)(1-Pr(x+W>i/N))$$

$$\le \frac{1}{N^2}\sum_{i=1}^{N^2} Pr(x+W>i/N) \underset{N\to\infty}{\to} 0.$$

since the sum in (2) converges to $EW+x$.
Considering conditional L_2 norms we have using the triangle inequality that

$$\|\hat{X}_N - x\| \le \|\hat{X}_N - E(\hat{X}_N \mid X=x)\| + |E(\hat{X}_N \mid X=x)-x| \underset{N\to\infty}{\to} 0$$

which proves that

$$\lim_{N\to\infty} E(d(X,\hat{X}_N) \mid X=x) = 0 \quad (3)$$

for all x. It can be shown that $E(\hat{X}_N \mid X=x) \le x$ for all x, so by dominated convergence we conclude from (3)

$$\lim_{N\to\infty} Ed(X,\hat{X}_N) = \lim_{N\to\infty} \int E(d(X,\hat{X}_N) \mid X=x)dPr(X<x) = 0$$

and the theorem is proved.

COMMENTS

The above simple example is new and makes several interesting points. The first point is that an intelligent quantizer strategy involves nodes cooperating to extract certain information from the observations and not acting independently to simply quantize their observations. The second point is that such a strategy can yield arbitrarily small average distortion in the limit of many nodes. The third point is that several of the nodes might fail to provide their bits without affecting the asymptotic properties since the nodes make overlapping tests. The estimate is simply modified to not include the missing information. Provided there are enough nodes, the overall approximation is still good. This robustness against missing nodes can be extremely valuable in a distributed system where nodes may fail.

There are many generalizations of possible interest: The nonnegative restriction can be removed in a straightforward manner. The generalization to noise values that are not independent is more important and less obvious. Intuitively one would expect that reasonable models would entail highly correlated noise in adjacent nodes with the correlation going to 0 with increased distance. It is also desirable to remove the assumption of identical distributions, although they must be related in order for the cooperative information to help. Generalizations to vector quantization in the presense of independent vector noises have been obtained using the ergodic theorem and quantizers that are indicator functions for sets that partition the space in an appropriate manner.

REFERENCES

1. T.J. Flynn, "Quantizer design for distributed sensing," Ph.D. Thesis, Stanford University, December 1984.
2. T.J. Flynn and R.M. Gray, "Quantizer Design for Correlated Observations", **Abstracts of Papers,** 1985 IEEE International Symposium on Information theory, Brighton, England, June 1985, p. 117.
3. T.J. Flynn and R.M. Gray, "Achievable Rates and Distortion for Correlated Observations", Ibid., p. 153.
4. T.J. Flynn and R.M. Gray, "Encoding of correlated observations," submitted for possible publication. (1985)
5. R.G. Gallager, "Finding parity in a broadcast network," **Abstracts of Papers**, 1985 IEEE International Symposium on Information Theory, June, 1985, p. 116.

ACKNOWLEDGEMENTS

This research was supported by the ARPA Distributed Information Theory Contract N00039-84-C-0211 Task 6.