Optimal aggregation of noisy observations: A large deviations approach

To cite this article: Tatsuto Murayama and Peter Davis 2010 J. Phys.: Conf. Ser. 233 012007

View the article online for updates and enhancements.
Optimal aggregation of noisy observations:  
A large deviations approach

Tatsuto Murayama and Peter Davis  
NTT Communication Science Laboratories, NTT Corporation, 2-4, Hikaridai, Seika-cho, Keihanna, Kyoto 619-0237, Japan  
E-mail: murayama@cslab.kecl.ntt.co.jp, davis@cslab.kecl.ntt.co.jp

Abstract. Sensing and data aggregation tasks in distributed systems should not be considered as separate issues. The quality of collective estimation involves a fundamental tradeoff between sensing quality, which can be increased by increasing the number of sensors, and aggregation quality under a given capacity of the network, which decreases if the number of sensors is too large. In this paper, we examine a system level strategy for optimal aggregation of data from an ensemble of independent sensors. In particular, we consider large scale aggregation from very many sensors, in which case the network capacity diverges to infinity. Then, by applying the large deviations techniques, we conclude the following significant result: larger scale aggregation always outperforms smaller scale aggregation at higher noise levels, while below a critical value of noise, there exist moderate scale aggregation levels at which optimal estimation is realized. At a critical value of noise, there is an abrupt change in the behavior of a parameter characterizing the aggregation strategy, similar to a phase transition in statistical physics.

1. Introduction
The purpose of this paper is to present results which give a new perspective on the growing field of networked sensors. We focus on the issue of reducing sensing error by collecting sensing data from many sensors. In particular, we work on the simple idea that the collective sensing error can be reduced by collecting data from more sensors. However, in distributed sensor networks, collecting data from many sensors usually involves some cost in terms of network resources; total bandwidth, power consumptions, computational capabilities, and so forth [1]. When the network resources and the amount of data from each sensor are fixed, then there will be a limit on the number of sensors, and thus a limit on the collective sensing quality. If it is possible to reduce the amount of sensor data by data compression then the number of sensors, and thus the sensing quality, will depend on the amount of data compression which can be achieved.

Now, there is a fundamental limit on the amount of data reduction which is possible. The limit is called the entropy rate and is determined by the statistical property of the source [2]. However, if it is possible to tolerate some error introduced by the reduction process, then data can be coded and compressed beyond this fundamental limit [3]. This is a key idea in information theory. When we extend our consideration of the collective sensing problem to include the possibility of arbitrary data reduction, then there arises a new type of tradeoff, i.e., the tradeoff between the effect of reducing collective sensing error with increasing number of sensors, and the effect of increasing aggregation error with increasing number of sensors due to the need to reduce the amount of data by data compression below the entropy rate. This tradeoff determines the
optimal number of sensors for minimal collective sensing error. This new type of tradeoff and its implications for optimal size of sensor networks is the new problem that we focus on in this paper. We see this as a fundamental and universal problem in sensor networks. Moreover with the growing feasibility of deploying huge numbers of smart sensors with advanced coding and communication abilities, this is a problem of rapidly increasing relevance and significance [4].

In this paper, we consider a fundamental formulation of the problem with only one information source and suppose that all sensors are symmetrical, i.e., exchangeable with respect to their contributions to the final result of aggregation. This allows us to treat the problem in terms of the theory of large deviations. The paper is divided into 5 sections. Section 2 presents our system model which is first introduced in the context of multi terminal information theory. Section 3 briefly summarizes our main result. The detailed proof for the proposition, however, is postponed until the following section 4. Conclusions are given in section 5.

2. System model

2.1. Ensemble of independent sensors

We consider that an observer is interested in observing the Bernoulli(1/2) source $X$, the state of which can be represented by a series of Ising variables $X_\mu$ and their realizations are explicitly denoted by the lower case letters $x_\mu = \pm 1$. We assume that this observer can not directly observe the source. Instead, he deploys a collection of $L$ sensors, labeled by an index $a$, to independently observe the source and report the results of their observations over a communication network. Assuming a certain level of environmental noise, the individual observations $Y_\mu(a)$ could be different for different sensors. We define a common level of noise $p \in [0, 1/2]$ for our observations

$$\langle \delta(X_\mu, -Y_\mu(a)) \rangle = p$$  (1)

with Kronecker’s delta $\delta$, where the braket $\langle \cdot \rangle$ denotes the expectation of an argument. Here we also assume that $y_\mu(a) = \pm 1$.

Then we suppose that each sensor can compress (i.e., lossy encode), if necessary, its sensor readings $y(a) = (y_1(a), \cdots, y_M(a))$ into a codeword $z(a) = (z_1(a), \cdots, z_N(a))$ independently. In this paper, we assume that the codeword is represented by a series of Ising variables $Z_\nu(a)$ and thus their realizations are restricted to $z_\nu(a) = \pm 1$ as well. We further assume that the sensors themselves can not share any information about their observations. That is, they are not permitted to communicate with each other to decide what to send beforehand. As a result, the observer must collect the $L$ codewords from all the sensors, each of which separately encodes its own observations $y_\mu(a)$, and use them to estimate the original $x_\mu$ for $\mu = 1, \cdots, M$. We assumed here that the lengths of the codewords are the same $N$, so that all the sensors are identical with respect to the theoretical ability of encoding their observations. That is, regardless of the sensor label, the rate for the lossy encoding is given by $R = N/M$. Therefore, the load level of our network can be measured by the sum rate $LR$, which should not be greater than the network capacity given by, say, $C$. We assume that $C$ is a given integer, not a real, in which case our argument will be greatly simplified.

If the sensors were able to share information about their observations before reporting to the observer, then they would be able to smooth out their independent environmental noises entirely as the number of sensors $L$ diverges. Then the observer can figure out all the realizations of $X_\mu$ if the network capacity $C$ exceeds 1, which is the entropy rate of the source $X$. However, if the mutual communications are prohibited, there does not exist any finite value of $C$ for which even infinitely many sensors can transmit all the information [5]. Therefore, our goal should be the semifaithful reconstruction of the original $x_\mu$ given the codewords $z_\nu(a)$ under a certain fidelity criterion.
2.2. Exchangeable sensor ansatz

Suppose that $\hat{y}(a) = (\hat{y}_1(a), \cdots, \hat{y}_M(a))$ be best reproductions for the observations obtained by using the codewords, respectively. Assume that the distortion between two sequences are always measured by the Hamming distance per symbol. Then it is easy to see that the distortion is given by, in this case,

$$d(y(a), \hat{y}(a)) = \frac{1}{M} \sum_{\mu=1}^{M} \delta(y_{\mu}(a), \hat{y}_{\mu}(a))$$

for $a = 1, \cdots, L$. Since we have exchangeable sensors as stated, we can impose that

$$\langle d(Y(a), \hat{Y}(a)) \rangle = D$$

for any given pairs. With this Hamming distortion constraint, the lower bound on the rate $R(D)$ required to describe a variable $Y_{\mu}(a)$ is given by

$$R(D) = 1 - H_2(D) ,$$

where $H_2(D)$ denotes the binary entropy function [6]. This is called the rate distortion function for the Bernoulli(1/2) source.

The observer then collects all the transmitted information $z_{\nu}(a)$ to calculate the estimate $\hat{x}_{\mu}$ for the $\mu$th symbol of the unknown $x = (x_1, \cdots, x_M)$. To go further, we now restrict ourselves to the case where

$$\langle \delta(Y_{\mu}(a), \hat{Y}_{\mu}(a)) \rangle = D \ .$$

That is, every variable $\hat{Y}_{\mu}(a)$ in the reproductions is expected to have the same error probability $D$. Since the observation and communication processes are independent, the above three variables $X_{\mu}, Y_{\mu}(a)$, and $\hat{Y}_{\mu}(a)$ form a Markov chain. Here the best estimator for $X_{\mu}$ is $\hat{Y}_{\mu}(a)$ if $0 \leq p, D < 1/2$ holds. Then it is straightforward to get, independently,

$$\langle \delta(X_{\mu}, -\hat{Y}_{\mu}(a)) \rangle = \rho ,$$

where

$$\rho = p(1 - D) + (1 - p)D$$

represents the combined error probability for replacing the original $x_{\mu}$ by the available symbols $\hat{y}_{\mu}(a)$. In other words, the error indicator function $\delta(X_{\mu}, -\hat{Y}_{\mu}(a))$ reduces to the Bernoulli random variable that takes the value 1 with probability $\rho$ for $a = 1, \cdots, L$.

2.3. Bayes optimal estimator

Now let us consider the most probable realization of $X_{\mu}$ given a set of evidences $\tilde{y}_{\mu} = (\tilde{y}_{\mu}(1), \cdots, \tilde{y}_{\mu}(L))$. Since $\delta(X_{\mu}, -\hat{Y}_{\mu}(a))$ obeys the Bernoulli statistics, it is easy to see that the majority vote procedure gives the best strategy [7]. That is, the optimal estimator should be a mapping

$$\hat{X}_{\mu} = \text{sgn}\left\{ \sum_{a=1}^{L} \hat{Y}_{\mu}(a) \right\} .$$
Then overall error probability for the estimate \( \hat{x}_\mu \) is minimized. The probability of getting more errors than \( L/2 \) out of \( L \) Bernoulli trials is given by

\[
P(X_\mu \neq \hat{X}_\mu) = \langle \delta(X_\mu, \hat{X}_\mu(a)) \rangle
\]

\[
= \begin{cases} 
\sum_{l=0}^{L} Q_\rho(l|L), & \text{(L is odd)} \\
\sum_{l=\frac{L}{2}+1}^{L} Q_\rho(l|L) + \frac{1}{2} Q_\rho(L/2|L) & \text{(L is even)}
\end{cases}
\]

where

\[
Q_\rho(l|L) = \binom{L}{l} \rho^l (1 - \rho)^{L-l}
\]

denotes the binomial distribution. In principle, we may choose whatever value of \( L \) which is compatible with the sum rate constraint of \( LR \leq C \). To minimize the error probability for the estimator \( \hat{X}_\mu \), however, we should use the largest possible value. Hereafter we assume that \( L \) denotes the largest possible value. In particular, suppose that the sensors do not encode their observations. Instead, each sensor simply sends the whole information of the noisy \( Y_\mu(a) \). Then, the error probability for the estimator \( \hat{X}_\mu \) reduces to

\[
P(X_\mu \neq \hat{X}_\mu) = \begin{cases} 
\sum_{l=\frac{L}{2}+1}^{L} Q_\rho(l|C), & \text{(C is odd)} \\
\sum_{l=\frac{L}{2}+1}^{L} Q_\rho(l|C) + \frac{1}{2} Q_\rho(C/2|C) & \text{(C is even)}
\end{cases}
\]

3. Statement of results
Assume that a network capacity \( C \) is given. Consider that the common data rate \( R \) is first allocated to all the sensors. The number of sensors \( L \) is thus determined as the maximum value of \( L \) satisfying the sum rate constraint \( RL \leq C \). In our system model, it is obvious to say that \( P(X_\mu \neq \hat{X}_\mu) \to 0 \) as \( C \to \infty \). As is shown in section 4, it is not hard to refine the above statement of convergence and to prove that \( P(X_\mu \neq \hat{X}_\mu) \) decays to 0 exponentially fast as \( C \to \infty \). By analogy with large deviation theory [8], we define the exponential rate of decay by

\[
I_p(R) = -\lim_{C \to \infty} \frac{1}{C} \ln P(X_\mu \neq \hat{X}_\mu) \quad (0 < R \leq 1)
\]

The decay rate \( I_p(R) \) describes the limiting behavior of the system from a macroscopic level, on which the rate \( R \) could be used as a control parameter. The case of \( R = 1 \) reduces to a naive aggregation scheme in which the sensors just send their noisy observations to the observer. For this smallest aggregation, we aggregate data from only \( L = C \) sensors. Hereafter, we call this scheme the level-1 aggregation. For a given \( R > 0 \), the level-\( R \) aggregation is defined in which every sensor encodes its observations at the rate of \( R \) independently. As an extension of the definition of \( I_p(R) \) for \( R > 0 \), we could naturally define the level-0 decay rate as

\[
I_p(0) = -\lim_{C \to \infty} \frac{1}{C} \lim_{R \to 0} \ln P(X_\mu \neq \hat{X}_\mu).
\]
Figure 1. Optimal aggregation levels for ensemble of independent sensors in noisy environment. The $p$ is a given noise level, while $R^*$ is the optimal value of $R$, the data rate per sensor, which maximizes the decay rate $I_p(R)$ of error probability with increase of network capacity $C$.

Assume that $D(R)$ denotes the distortion rate function, which is the inverse function of $R(D)$. Suppose that $\alpha(p, R) = (1 - 2p)(1 - 2D(R))$ for $0 \leq p < 1/2$ and $0 < R \leq 1$. Then, the main result of this paper is given below.

**Proposition 3.1 (Decay rate)** We have

$$I_p(R) = \begin{cases} 
(1 - 2p)^2 \ln 2 & (R=0) \\
\frac{\alpha(p, R)^2}{2R(1 - \alpha(p, R))(1 + \alpha(p, R))} & (0 < R \leq 1)
\end{cases} \quad (5)$$

The maximum of $I_p(R)$ is of great interest from an engineering point of view. That is, we prefer larger values of $I_p(R)$. Therefore, we examine the optimal levels defined by $R^* = \arg\max_{0 \leq R \leq 1} I_p(R)$. The optimal aggregation, for a given $p$, is called the level-$R^*$ aggregation.

We now examine the behavior of formula (5) which gives the optimal levels $R^*$ for the noise $p$. As is seen in figure 1, the optimal aggregation scale diverges, i.e., the optimal data rate $R^*$ per sensor vanishes for noise levels larger than the critical point $p_0 = 0.295$. In this noisy region, we want the system to be as large as possible. The larger the system we have, the smaller the error probability. By definition, the optimal aggregation is said to be level-0. In contrast, we can always find the non-zero optimal levels below $p_0$. In particular, if the noise level is below $p_1 = 0.066$, our investigations indicate that the level-1 aggregation is optimal. Moderate aggregation levels could be optimal in the intermediate noise levels between the two critical points. It is also worth noticing that the behavior of $R^*$ of $p$ is reminiscent of that of order parameters at a continuous phase transition in statistical mechanics [9]. The analytical results presented here are also consistent with numerical simulations for the system size $C = 10^3$.

Since the optimal levels $R^*$ are unique values for each noise $p$, we can plot the optimal decay rate $I_p(R^*)$ as is given in figure 2. The optimal rate $I_p(R^*)$ describes the limiting behavior of
Figure 2. Maximum decay rates for error probability of final decision. The $I_p(R^*)$ is the largest decay rate at the noise level $p$, which is given by the optimal data rate $R^*$ per sensor.

Figure 3. The asymptotics of $I_p(0)$ and $I_p(1)$. The analysis indicates that the combination of these two options can provide a suboptimal aggregation strategy.

the smallest error probability $P(X_\mu \neq \hat{X}_\mu)$ in terms of macroscopic variables. Clearly, it is a strongly decreasing function of the noise $p$. An approximation of $I_p(R^*)$ by the combination of $I_p(0)$ and $I_p(1)$ is also presented in figure 3. However the crossover point of $I_p(0)$ and $I_p(1)$ as observed in figure 3 is quite different from the value obtained by numerics [10]. We think that this inconsistency should be due to the use of Gaussian approximation and that more rigorous approach should be taken toward the quantitative argument.
4. Large deviations analysis

This section is devoted to presenting the large deviations analysis which gives proposition 3.1. For sufficiently large \( L \), the binomial distribution \( Q_\rho(l|L) \) is well approximated by the Gaussian distribution \( N(L\rho, L\rho(1-\rho)) \) with mean \( L\rho \) and variance \( L\rho(1-\rho) \) \([11]\). Changing the variable

\[
s = \frac{l - L\rho}{\sqrt{L\rho(1-\rho)}}
\]

enables us to use the central limit theorem to get

\[
P(X_\mu \neq \hat{X}_\mu) \sim \int_{\lambda_1}^{\lambda_2} \frac{ds}{\sqrt{2\pi}} e^{-s^2/2},
\]

where we denote, respectively,

\[
\lambda_1 = \frac{1/2 - \rho}{\sqrt{\rho(1-\rho)}} \sqrt{L}, \quad \lambda_2 = \frac{1 - \rho}{\sqrt{\rho(1-\rho)}} \sqrt{L}.
\]

Since every sensor can achieve the optimal rate \( R(D) \), we may evaluate the number of sensors \( L \) as \( C/R(D) \). Write \( \beta(p, D) = (1-2p)(1-2D) \) for simplicity. Together with an identity

\[
\frac{1}{2} - \rho = (1-2p) \left( \frac{1}{2} - D \right),
\]

we have

\[
\lambda_1 = \frac{\beta(p, D) \sqrt{C}}{\sqrt{R(D)(1-\beta(p, D))(1+\beta(p, D))}}, \quad \lambda_2 = \frac{(1+\beta(p, D)) \sqrt{C}}{\sqrt{R(D)(1-\beta(p, D))(1+\beta(p, D))}}.
\]

In particular, the asymptotics of \( R(D) \) yields, respectively,

\[
\lambda_1 \sim \frac{(1-2p) \sqrt{2C \ln 2}}{\sqrt{(1-\beta(p, D))(1+\beta(p, D))}}, \quad \lambda_2 \sim \frac{(1+\beta(p, D)) \sqrt{C \ln 2}}{\sqrt{(1-\beta(p, D))(1+\beta(p, D))}} \left( \frac{1}{2} - D \right)^{-1}.
\]

Therefore \( \lambda_1 \) takes a finite value for every \( C \) given, while \( \lambda_2 \) could diverge if we consider the level-0 aggregation scheme. These formulas assure us that \( I_\rho(0) \) is well defined.

Now consider

\[
\Phi(\lambda) = \int_\lambda^\infty \frac{ds}{\sqrt{2\pi}} e^{-s^2/2}
\]

and write

\[
\int_{\lambda_1}^{\lambda_2} \frac{ds}{\sqrt{2\pi}} e^{-s^2/2} = \Phi(\lambda_1) - \Phi(\lambda_2).
\]

The following lemma tells us that the asymptotic behavior of the definite integral is governed by the asymptotics of \( \Phi(\lambda_1) \). The proof is given in the appendix.

**Lemma 4.1 (Asymptotic expansion)** We have

\[
\Phi(\lambda) = \frac{1}{\sqrt{2\pi}} e^{-\lambda^2/2} \left\{ \frac{1}{\lambda} - F_2(\lambda) \right\},
\]

where the second term is given by

\[
F_2(\lambda) = e^{\lambda^2/2} \int_{\lambda}^{\infty} \frac{ds}{s^2} e^{-s^2/2}.
\]
Here we have $1/\lambda \gg F_2(\lambda)$ as $\lambda \to \infty$ [12]. Then simple algebra gives

$$\Phi(\lambda_1) - \Phi(\lambda_2) \sim \frac{e^{-\lambda_2^2/2}}{\sqrt{2\pi\lambda_2}} \left(1 - \frac{\lambda_1}{\lambda_2} e^{-\frac{(\lambda_2^2 - \lambda_1^2)/2}{2}}\right).$$

Taking the logarithm results in

$$\ln P(X_\mu \neq \hat{X}_\mu) \sim -\frac{\lambda_1^2}{2}$$

which proves the proposition 3.1.

5. Conclusions
It has been shown that the optimal aggregation strategy for a system of sensors exhibits a kind of phase transition with respect to the noise level. We expect this result to be useful when engineering a large sensor network. We described the critical behavior of the ratio $R = C/L$, the data rate per sensor which is the ratio of the network communication capacity $C$ and the number of sensors $L$.

The analysis shows that in the high noise region beyond a critical value of noise $p_0$, the ratio $R$ should converge to zero in order to reduce collective estimation error. This means that we should deploy very many sensors $L \gg C$ in the large $C$ limit. In contrast, if the noise level is lower than the critical point, the ratio $R$ should take a positive value. In this case, the number of sensors scales as $L = \mathcal{O}(C)$. The authors hope that this brief paper will trigger interest in this kind of asymptotic property of the sensory aggregation and applications of the large deviations approach in the field of systems science.

References
[1] Akyildiz I F, Su W, Sankarasubramaniam Y and Cayirci E 2002 IEEE Communications Magazine 40 102–14
[2] Shannon C E 1948 Bell Syst. Tech. J 27 379–423
[3] Shannon C E 1959 IRE National Convention Record 4 142–63
[4] Culler D E and Muller H 2004 Scientific American 290 84–91
[5] Berger T, Zhang Z and Viswanathan H 1996 IEEE Transactions on Information Theory 42 887–902
[6] Cover T M and Thomas J A 1991 Elements of Information Theory (New York: Wiley)
[7] MacKay D J C 2003 Information Theory, Inference, and Learning Algorithms (Cambridge: Cambridge University Press)
[8] Ellis R S 1985 Entropy, Large Deviations, and Statistical Mechanics (Berlin: Springer)
[9] Monasson R, Zecchina R, Kirkpatrick S, Selman B and Troyansky L 1999 Nature 400 133–7
[10] Murayama T and Davis P 2008 Journal of Physics: Conference Series 95 012010
[11] Hays W L 1981 Statistics (New York: Holt, Rinehart and Winston)
[12] Copson E T 2004 Asymptotic Expansions (Cambridge: Cambridge University Press)

Proof of lemma 4.1
Define a series of integrals,

$$F_n(\lambda) = e^{\lambda^2/2} \int_{\lambda}^{\infty} \frac{ds}{s^n} e^{-s^2/2},$$

the notations of which is also consistent with the preceding $F_2(\lambda)$. By definitions, it is an easy matter to check that

$$\Phi(\lambda) = \frac{F_0(\lambda)}{\sqrt{2\pi}} e^{-\lambda^2 / 2}.$$
Since elementary calculus gives
\[ F_n(\lambda) = -e^{\lambda^2/2} \int_{\lambda}^{\infty} \frac{ds}{s^{n+1}} \frac{d}{ds} e^{-s^2/2}, \]
the integration by parts results in the recurrence formula
\[ F_n(\lambda) = \frac{1}{\lambda^{n+1}} - (n + 1)F_{n+2}(\lambda). \]
Therefore, we find the asymptotic expansion
\[
\Phi(\lambda) = \frac{1}{\sqrt{2\pi}} e^{-\lambda^2/2} \left\{ \frac{1}{\lambda} - \frac{1}{\lambda^3} + \frac{3!}{\lambda^5} + \cdots + \frac{(-1)^n (2n - 1)!!}{\lambda^{2n+1}} \right. \\
+ \left. (-1)^{n+1} (2n + 1)!! F_{2n+2}(\lambda) \right\}
\]
which is also exact for \( n \) equals 0. The conclusion follows.