**Fifty-second Annual Allerton Conference**
**Allerton House, UIUC, Illinois, USA**
**October 1 - 3, 2014**

# Using Relative-Relevance of Data Pieces for Efficient Communication, with an Application to Neural Data Acquisition

Majid Mahzoon, Hassan Albalawi, Xin Li, and Pulkit Grover
Electrical and Computer Engineering Department, Carnegie Mellon University, Pittsburgh, PA, USA 15213
{mmahzoon, halbalaw, xinli, pgrover}@andrew.cmu.edu

***Abstract*—In this paper, we consider the problem of communicating data from distributed sensors for the goal of inference. Two inference problems of linear regression and binary linear classification are investigated. Assuming perfect training of the classifier, an approximation of the problem of minimizing classification error-probability under Gaussianity assumptions leads us to recover Fisher score: a metric that is commonly used for feature selection in machine learning. Further, this allows us to soften the notion of feature selection by assigning a degree of relevance to each feature based on the number of bits assigned to it. This relative relevance is used to obtain numerical results on savings on number of bits acquired and communicated for classification of neural data obtained from Electrocorticography (ECoG) experiments. The results demonstrate that significant savings on costs of communication can be achieved by compressing Big Data at the source.**

## I. Introduction

In this work, we investigate the problem of compressing data from different data streams. These data streams could be generated by different sources (e.g., sensors), some of which are more "relevant" to the goal at hand than others. To motivate these problems, consider the practical example of Brain-Computer Interfaces (BCIs) [1]. Such interfaces typically record large amounts of data [2], and communicate them at high rates (and consequently, with large amounts of energy, see e.g., [3]). With the demand to make these interfaces wireless (e.g., [4]), it is important to reduce the communication requirements.

Standard information-theoretic compression techniques tend to be difficult to apply in these problems. For instance, the sensors often sense correlated data, which can make the possibility of using distributed source-coding techniques (see e.g., [5]–[8], etc.) very appealing. However, the techniques rely on estimation of joint distribution. Even when with additional assumptions the underlying distribution can be parameterized, the problem is hard: (i) it requires $O(n^2)$ data to estimate correlations between $n$ sources, where $n$ can typically be very large; and (ii) it often requires moving data to a central node prior to compression so that correlations can be computed. In contrast, parameter estimation at individual sensors can require less data and can be performed in isolation. One observation that can help reduce the required data rates comes from the literature of "channel selection" (see, e.g., [9]–[17]) in brain-machine interfaces. Motivated primarily by ease of computation through dimensionality reduction (once the data has been collected), channel-selection algorithms select the channels most relevant to the task at hand, and ignore data from other channels. When the goal is one of classifying data streams into one of many classes (e.g., for neuroprosthetic applications [18]), techniques based on computing Fisher scores [9], recursive channel selection [9], common spatial pattern [10]–[13], mutual information maximization [14], [15], genetic algorithms [16], [17], etc., have been proposed.

The core idea examined in this work is the following: instead of hard selection of each data stream (as is done in channel selection), our goal is to quantize each stream using minimal number of bits in order to accomplish the inference goal with a target accuracy. The allocation of number of bits to each data stream depends on the relevance of that stream to the goal. Further, we want to preserve, to some degree, the data in the more relevant data streams themselves in order to enable improved estimation of the underlying joint distribution at the receiving end. The receiver can then use sophisticated inference algorithms that could exploit correlations. Thus, our problem (posed formally in Section II) is one where we are forced to reconstruct the relevant data streams themselves for distribution estimation. We propose strategies (in Section III) that assign more weight to the reconstruction error of more relevant data streams, which would enable improved estimates of the marginal distributions of more relevant streams, and improved understanding of redundancy in these streams. We do believe that this is merely an approximation of the actual underlying problem, and that a system-level approach is needed in order to understand the costs of learning the system and classification together.

Towards understanding how bits should be allocated to different data streams that reflect their relevance, two problems of inference are considered in this work: linear regression for prediction, and linear classification. For the problem of linear classification, the proposed strategy is tested on neural data (in Section IV). The obtained results show substantial reductions in data rates (over uniform quantization over all sensors) suggesting that the proposed compression strategies can enable significant lowering of energy requirements if the

**978-1-4799-8009-3/14/$31.00 ©2014 IEEE** 

needed intelligence for computing relevance can be built into the sensing system.

Channel-selection approaches often rely on feature-selection techniques from machine learning [19], [20]. A common feature-selection technique is one that uses "Fisher score" [21], selecting only features that have sufficiently high Fisher scores (see Section III and Section IV). For an underlying Gaussian model and the goal of binary linear classification, interestingly, our results in Section III show that Fisher score falls out naturally from our proposed strategies.

Within information theory, in the problems of distributed data compression, perhaps the most closely related works are those of Berger (the "CEO problem" [22]) and Han [23] (and the follow-up work on these papers). From the perspective of the CEO problem, the underlying class can be viewed as the source of information. Corrupted versions of this source are observed at various sensors. The problem becomes one of rate-constrained minimization of error in estimating the source. Han [23] surveys the literature in distributed compression for inference, and hence is more closely related. The core difference here from both of these works is that we need reconstruction of not only the underlying class, but also of the data streams themselves, with increasing resolution for the more relevant data streams. Nevertheless, the problems are intimately related, and our ongoing work seeks to understand how strategies proposed in this literature can prove helpful in reducing the data and energy overhead faced by brain-machine interfaces, and sensor networks in general.

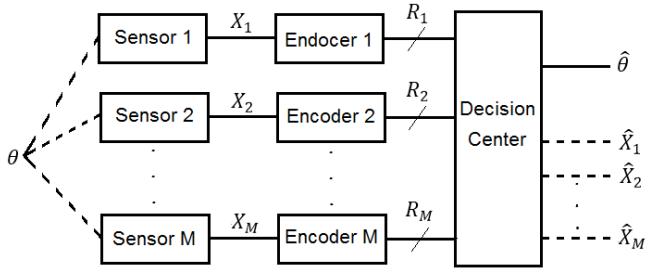


Fig. 1. Model of a distributed sensor network. In the CEO problem, the parameter $\theta$ is the unobservable source $X$ while in the statistical inference problem, the parameter $\theta$ is the underlying class (or hypothesis) as the problem posed in this paper. The main difference between our problem formulation and the other two is that while in the CEO and statistical inference problems, only reconstruction of the parameter $\theta$ is required, we also need to obtain estimations of the data streams.

This paper is organized as follows. Section II contains the formulation of two inference problems: linear regression and binary linear classification. Section III develops the main results of the paper and draws the connection between the proposed rate-allocation strategies for classification and Fisher score. In section IV, the performance of the proposed rate-allocation method is evaluated using numerical results, and section V concludes the paper.

## II. Problem Formulation

In this section, the inference problems of linear regression for prediction and binary linear classification based on measurements obtained from a set of distributed sensors under a sum-rate constraint are presented.

Consider the problem of inference based on data obtained from sensors on which training has already been performed and thus, the distributions on the data have already been estimated. Suppose there are $M$ distributed sensors which are sending their measurements $X_i, i = 1, 2, ..., M$ to a decision center for the goal of inference through noiseless channels between each sensor and the decision center. Considering each of these measurements as a feature, the vector $\mathbf{X} = [X_1, X_2, ..., X_M]$ of measurements is called a data point in the $M$-dimensional feature space.

Assuming $R_i$ bits of resolution are dedicated to the $i$-th feature $X_i$, its quantized representation using $R_i$ bits is denoted by $\widehat{X}_i$. More precisely, the $i$-th sensor uses an encoder (quantization) function $\mathcal{E}_i : \mathcal{X}_i^n \rightarrow \{1, ..., 2^{nR_i}\}$ and sends $\mathcal{E}_i(X_i^n)$ to the decision center. The decision center uses a set of decoding functions $\mathcal{D}_i : \{1, ..., 2^{nR_i}\} \rightarrow \widehat{\mathcal{X}}_i^n$ to reproduce the features. Given a set of data points $\{\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, ..., \mathbf{X}^{(n)}\}$ in the feature space, consider the problem of inference based on a reconstruction of those data points $\{\widehat{\mathbf{X}}^{(1)}, \widehat{\mathbf{X}}^{(2)}, ..., \widehat{\mathbf{X}}^{(n)}\}$ subject to a constraint on the total rate.

### A. Linear Regression for Prediction

Suppose that the data points in the $M$-dimensional feature space are such that the $i$-th feature of each data point is distributed according to a Gaussian distribution, i.e., $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$. Consider the problem of predicting the value of the dot product $\mathbf{w}^T\mathbf{X}$ subject to a constraint on the sum-rate of communication, where the vector $\mathbf{w}$ has been estimated beforehand from the estimated data distribution. In other words, the problem is how to divide the total rate among different features so that the mean-squared error in predicting $\mathbf{w}^T\mathbf{X}$, i.e., $E[(\mathbf{w}^T\mathbf{X} - \mathbf{w}^T\widehat{\mathbf{X}})^2]$ is minimized. More formally, the inference problem can be formulated as follows:

$$\begin{aligned} \underset{\mathcal{E}_i, \mathcal{D}_i, R_i}{\text{minimize}} \quad & \frac{1}{n}\sum_{t=1}^{n} E\left[\left(\mathbf{w}^T\mathbf{X}^{(t)} - \mathbf{w}^T\widehat{\mathbf{X}}^{(t)}\right)^2\right]. \\ \text{subject to} \quad & \sum_{i=1}^{M} R_i \leq R \end{aligned} \tag{1}$$

We will derive an inner bound for the above optimization problem in section III.A.

### B. Binary Linear Classification

Assume that each data point belongs to one of the two existing classes with equal probabilities. If a data point $\mathbf{X}$ is of class $j$, $j = 1, 2$, the $i$-th, $i = 1, 2, ..., M$ feature of that data point is distributed according to a Gaussian distribution, i.e., $X_i \sim \mathcal{N}(\mu_{ji}, \sigma_{ji}^2)$; in other words, each feature is a random variable distributed according to a Gaussian mixture model with two equiprobable components.

Without loss of generality, we can assume that for each feature $X_i$, $\mu_{1i} = -\mu_{2i} = -\mu_i$. Also, for simplicity, we assume that for each feature $X_i$, $\sigma_{1i} = \sigma_{2i} = \sigma_i$. Thus, in the remainder of the paper, we will assume that under class 1, $X_i \sim \mathcal{N}(-\mu_i, \sigma_i^2)$ and under class 2, $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$.

Given a data point $\mathbf{X}$ in the feature space, consider the problem of classification by the linear discriminant analysis (LDA) algorithm using a reconstruction of that data point $\widehat{\mathbf{X}}$ subject to a constraint on the total rate. The LDA algorithm for classification is based on two assumptions. First, it assumes that the conditional probability distribution of a data point given its class is normal, i.e., under class 1, $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and under class 2, $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$. Under this assumption, the Bayes optimal solution is to classify the data point in the first class if the log-likelihood ratio is above some threshold constant $d$, i.e.,

$$\begin{aligned}(\mathbf{X}-\boldsymbol{\mu}_1)^T \boldsymbol{\Sigma}_1^{-1} (\mathbf{X}-\boldsymbol{\mu}_1) + \ln|\boldsymbol{\Sigma}_1| \\ - (\mathbf{X}-\boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}_2^{-1} (\mathbf{X}-\boldsymbol{\mu}_2) - \ln|\boldsymbol{\Sigma}_2| > d.\end{aligned} \tag{2}$$

Second, it assumes that the conditional covariance matrices under the two classes are identical, i.e., $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$. Under this assumption, the decision criterion in (2) simplifies to $\mathbf{w}^T\mathbf{X} < d'$, for some threshold constant $d'$, where

$$\mathbf{w} \propto \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1). \tag{3}$$

This means that the decision rule which classifies a data point $\mathbf{X}$ in class $j, j = 1, 2$ is a function of this linear combination of the known measurements.

Assuming that the vector $\mathbf{w}$ has been estimated beforehand from the estimated data distribution, consider the problem of minimizing the classification error-probability averaged over time subject to a constraint on the sum-rate of communication, i.e.,

$$\begin{aligned}\underset{\mathcal{E}_i,\mathcal{D}_i,R_i}{\text{minimize}} \quad & \frac{1}{n}\sum_{t=1}^{n} Pr\left(\text{Class}\left(\widehat{\mathbf{X}}^{(t)}\right) \neq \text{Class}\left(\mathbf{X}^{(t)}\right)\right). \\ \text{subject to} \quad & \sum_{i=1}^{M} R_i \leq R\end{aligned} \tag{4}$$

In other words, the problem is how to divide the total rate among different features so that the probability of misclassification is minimized. We will derive two inner bounds for the above optimization problem in section III.B.

## III. Main Results

This section includes the main results of the paper, specifically an inner bound for the linear regression problem in (1) and two inner bounds for the binary linear classification problem in (4). We also draw a connection between our proposed methods of rate-allocation for (4) and Fisher score.

### A. Linear Regression for Prediction

We first derive an upper bound on the target function in (1) as follows:

$$E\left[\left(\mathbf{w}^T\mathbf{X}^{(t)} - \mathbf{w}^T\widehat{\mathbf{X}}^{(t)}\right)^2\right] \tag{5}$$

$$= E\left[\left(\mathbf{w}^T\left(\mathbf{X}^{(t)} - \widehat{\mathbf{X}}^{(t)}\right)\right)^2\right] \tag{6}$$

$$= E\left[\left(\sum_{i=1}^{M} w_i\left(X_i^{(t)} - \widehat{X}_i^{(t)}\right)\right)^2\right] \tag{7}$$

$$= E\left[\sum_{i=1}^{M}\sum_{j=1}^{M} w_i w_j\left(X_i^{(t)} - \widehat{X}_i^{(t)}\right)\left(X_j^{(t)} - \widehat{X}_j^{(t)}\right)\right] \tag{8}$$

$$= \sum_{i=1}^{M}\sum_{j=1}^{M} w_i w_j E\left[\left(X_i^{(t)} - \widehat{X}_i^{(t)}\right)\left(X_j^{(t)} - \widehat{X}_j^{(t)}\right)\right] \tag{9}$$

$$\leq \sum_{i=1}^{M}\sum_{j=1}^{M} w_i w_j \sqrt{E\left[\left(X_i^{(t)} - \widehat{X}_i^{(t)}\right)^2\right] E\left[\left(X_j^{(t)} - \widehat{X}_j^{(t)}\right)^2\right]} \tag{10}$$

$$= \left(\sum_{i=1}^{M} w_i \sqrt{E\left[\left(X_i^{(t)} - \widehat{X}_i^{(t)}\right)^2\right]}\right)^2 \tag{11}$$

$$= \left(\sum_{i=1}^{M} \sqrt{w_i^2 E\left[\left(X_i^{(t)} - \widehat{X}_i^{(t)}\right)^2\right]}\right)^2 \tag{12}$$

$$= \left(\sum_{i=1}^{M} \sqrt{E\left[\left(w_i X_i^{(t)} - w_i \widehat{X}_i^{(t)}\right)^2\right]}\right)^2 \tag{13}$$

$$= \left(\sum_{i=1}^{M} \sqrt{D_i^{(t)}}\right)^2, \tag{14}$$

where (10) follows from the Cauchy-Schwarz inequality and (14) follows from the definition $D_i^{(t)} \triangleq E[(w_i X_i^{(t)} - w_i \widehat{X}_i^{(t)})^2]$. Therefore, the solution to the following problem is an inner bound to that of (1):

$$\begin{aligned}\underset{\mathcal{E}_i,\mathcal{D}_i,R_i}{\text{minimize}} \quad & \frac{1}{n}\sum_{t=1}^{n}\left(\sum_{i=1}^{M}\sqrt{D_i^{(t)}}\right)^2. \\ \text{subject to} \quad & \sum_{i=1}^{M} R_i \leq R\end{aligned} \tag{15}$$

Since codebooks are generated randomly in the achievability proof of the distortion-rate function, averaging over the randomly chosen codebook, the expected distortion in recovering the $i$-th feature of the $t$-th data point, and thus, $D_i^{(t)}$, is the same for all data points. Thus, we can drop $(t)$ in the problem definition and (15) can be simplified to

$$\begin{aligned}\underset{\mathcal{E}_i,\mathcal{D}_i,R_i}{\text{minimize}} \quad & \left(\sum_{i=1}^{M}\sqrt{D_i}\right)^2, \\ \text{subject to} \quad & \sum_{i=1}^{M} R_i \leq R\end{aligned} \tag{16}$$

which is equivalent to the following problem:

$$\begin{aligned} \underset{\mathcal{E}_i,\mathcal{D}_i,R_i}{\text{minimize}} \quad & \sum_{i=1}^{M} \sqrt{D_i}. \\ \text{subject to} \quad & \sum_{i=1}^{M} R_i \leq R \end{aligned} \tag{17}$$

Rather than deriving the rate-allocation strategy which minimizes the target function in (1), we are trying to determine the strategy that minimizes the target function in (17) which is an upper bound to the former one.

**Theorem 1.** *Let $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$, $i = 1, 2, \ldots, M$ be independent Gaussian random variables. Then, an inner bound for (17) is given by*

$$R_i = \frac{1}{2} \log \left( \frac{w_i^2 \sigma_i^2}{D_i} \right), \tag{18}$$

*where*

$$D_i = \begin{cases} \lambda', & \lambda' \leq w_i^2 \sigma_i^2, \\ w_i^2 \sigma_i^2, & \lambda' > w_i^2 \sigma_i^2, \end{cases} \tag{19}$$

*where $\lambda'$ is chosen such that $\sum_{i=1}^{M} R_i = R$.*

*Proof.* From the distortion-rate function of a single Gaussian source $\left(D(R) = \sigma^2 2^{-2R}\right)$, we know that $D_i = w_i^2 \sigma_i^2 2^{-2R_i}$. Thus, it remains to solve the following optimization problem:

$$\begin{aligned} \underset{R_i}{\text{minimize}} \quad & \sum_{i=1}^{M} w_i \sigma_i 2^{-R_i}. \\ \text{subject to} \quad & \sum_{i=1}^{M} R_i = R \end{aligned} \tag{20}$$

Using Lagrange multipliers, the Lagrange function is

$$J(R) = \sum_{i=1}^{M} w_i \sigma_i 2^{-R_i} + \lambda \left( \sum_{i=1}^{M} R_i - R \right). \tag{21}$$

Differentiating with respect to $R_i$ and setting to zero, we have

$$\frac{\partial J}{\partial R_i} = -w_i \sigma_i 2^{-R_i} \ln 2 + \lambda = 0, \tag{22}$$

which results in $R_i = \log \frac{w_i \sigma_i \ln 2}{\lambda}$ which can be translated to $D_i = \frac{\lambda^2}{\ln^2 2} = \lambda'$. Thus, we arrive at the same reverse water-filling result for the rate-allocation strategy as that in minimizing the sum-distortion (as opposed to sum of square roots of distortions) of independent Gaussian sources. ■

### *B. Binary Linear Classification*

Since classification using the LDA algorithm only depends on $\mathbf{w}^T\mathbf{X}$, intuitively, better approximation of $\mathbf{w}^T\mathbf{X}$ leads to lower classification error-probability. This suggests the approximation of (4) by

$$\begin{aligned} \underset{\mathcal{E}_i,\mathcal{D}_i,R_i}{\text{minimize}} \quad & \frac{1}{n} \sum_{t=1}^{n} E\left[ \left( \mathbf{w}^T \mathbf{X}^{(t)} - \mathbf{w}^T \widehat{\mathbf{X}}^{(t)} \right)^2 \right]. \\ \text{subject to} \quad & \sum_{i=1}^{M} R_i \leq R \end{aligned} \tag{23}$$

In other words, the problem is how to divide the total rate among different features so that the mean-squared error in recovering the dot product $\mathbf{w}^T\mathbf{X}$, i.e., $E[(\mathbf{w}^T\mathbf{X} - \mathbf{w}^T\widehat{\mathbf{X}})^2]$ is minimized. The above approximation is also important because in inference problems, the estimated distributions of $X_i$'s often need to be updated.

Using the same procedure and reasoning as in section III.A, instead of deriving the rate-allocation strategy which minimizes the target function in (23), we are trying to determine the strategy that minimizes the target function in (17) which is an upper bound to the former one.

**Theorem 2.** *Let $X_i, i = 1, 2, \ldots, M$ be independent random variables each distributed as a mixture of two equiprobable Gaussian densities, i.e., $p(X_i) = \frac{1}{2}\mathcal{N}(-\mu_i, \sigma_i^2) + \frac{1}{2}\mathcal{N}(\mu_i, \sigma_i^2)$. Then, an inner bound for (17) is given by*

$$R_i = \frac{1}{2} \log \left( \frac{\Lambda_i^2}{D_i} \right), \tag{24}$$

*where*

$$D_i = \begin{cases} \lambda', & \lambda' \leq \Lambda_i^2, \\ \Lambda_i^2, & \lambda' > \Lambda_i^2, \end{cases} \tag{25}$$

*where $\Lambda_i = w_i \sqrt{\sigma_i^2 + \mu_i^2}$ and $\lambda'$ is chosen such that $\sum_{i=1}^{M} R_i = R$.*

*Proof.* It is known that for a given second moment and a given distortion, the Gaussian source has the largest rate-distortion function; equivalently, for a given second moment and a given rate, the Gaussian source has the largest distortion-rate function. Thus, we can derive an upper bound on the distortion-rate function of a Gaussian mixture source by considering the distortion-rate function of the Gaussian source with the same second moment. For the Gaussian mixture random variable $X_i \sim \frac{1}{2}\mathcal{N}(-\mu_i, \sigma_i^2) + \frac{1}{2}\mathcal{N}(\mu_i, \sigma_i^2)$, the second moment is obtained as follows:

$$E\left[X_i^2\right] = \sum_{c=1}^{2} p(C = c) E\left[X_i^2 | C = c\right] = \sigma_i^2 + \mu_i^2, \tag{26}$$

where $C$ represents whether $X_i$ is taken from the first or the second Gaussian component. Therefore, considering the Gaussian random variable $Y_i \sim \mathcal{N}(0, E\left[X_i^2\right])$, the distortion-rate function of the Gaussian mixture source $X_i$ is bounded by

$$D_i \leq w_i^2 Var[X_i] 2^{-2R_i} = w_i^2 \left(\sigma_i^2 + \mu_i^2\right) 2^{-2R_i}. \tag{27}$$

Thus, it remains to solve the following optimization problem:

$$\begin{aligned} \underset{R_i}{\text{minimize}} \quad & \sum_{i=1}^{M} w_i \sqrt{\sigma_i^2 + \mu_i^2} 2^{-R_i}. \\ \text{subject to} \quad & \sum_{i=1}^{M} R_i = R \end{aligned} \tag{28}$$

Substituting $w_i^2 \sigma_i^2$ in Theorem 1 with $w_i^2 \left(\sigma_i^2 + \mu_i^2\right)$, the rest of the proof is along the same lines as that of Theorem 1. Thus, the achievable rate-allocation strategy in (24) leads to

a kind of reverse water-filling for the rate-allocation strategy, as illustrated in Fig. 2. ■

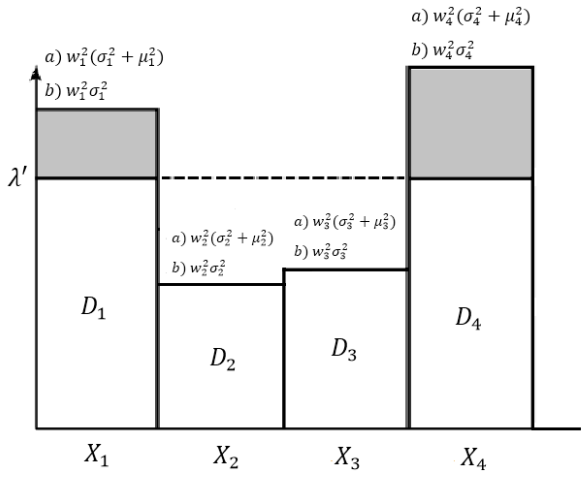


Fig. 2. Reverse water-filling. a) In the rate-allocation strategy corresponding to the first inner bound, only features with parameter $w_i^2\left(\sigma_i^2+\mu_i^2\right)$ larger than some constant $\lambda'$ are quantized. All these features suffer the same individual distortion. b) In the rate-allocation strategy corresponding to the second inner bound, features with parameter $w_i^2\sigma_i^2$ larger than some constant $\lambda'$ are quantized such that all of them suffer the same individual distortion. Each of the other features is quantized such that its individual distortion is equal to its parameter $w_i^2\sigma_i^2$. (Figure is a modified version of Fig. 10.7 in [24].)

The rate-distortion function of a Gaussian mixture random variable with two equiprobable components for distortions $D \leq \sigma^2$ satisfies [25]

$$R(D) = \frac{1}{2}\log\left(\frac{\sigma^2}{D}\right) + 1 - \epsilon\left(\frac{\mu}{\sigma}\right), \tag{29}$$

where $\epsilon(x)$ goes to zero as $x \to \infty$. Using the fact that the distortion-rate function $D(R)$ is a non-increasing function of the rate $R$, the distortion-rate function of a Gaussian mixture random variable with two equiprobable components for $R \geq 1-\epsilon\left(\frac{\mu}{\sigma}\right)$ is given by

$$D(R) = \sigma^2 2^{-2\left(R-1+\epsilon\left(\frac{\mu}{\sigma}\right)\right)}. \tag{30}$$

**Theorem 3.** *Let $X_i, i = 1,2,\ldots,M$ be independent random variables each distributed as a mixture of two equiprobable Gaussian densities, i.e., $p(X_i) = \frac{1}{2}\mathcal{N}(-\mu_i,\sigma_i^2) + \frac{1}{2}\mathcal{N}(\mu_i,\sigma_i^2)$. Then, assuming $R \geq \sum_{i=1}^{M}\left(1-\epsilon\left(\frac{\mu_i}{\sigma_i}\right)\right)$, an inner bound for (17) is given by*

$$R_i = \frac{1}{2}\log\left(\frac{w_i^2\sigma_i^2}{D_i}\right) + 1 - \epsilon\left(\frac{\mu_i}{\sigma_i}\right), \tag{31}$$

*where*

$$D_i = \begin{cases} \lambda', & \lambda' \leq w_i^2\sigma_i^2, \\ w_i^2\sigma_i^2, & \lambda' > w_i^2\sigma_i^2, \end{cases} \tag{32}$$

*where $\lambda'$ is chosen such that $\sum_{i=1}^{M} R_i = R$.*

*Proof.* First, we assign $1-\epsilon\left(\frac{\mu_i}{\sigma_i}\right)$ bits to the $i$-th feature for $i = 1,\ldots,M$. Then, using (30), we have $D_i = w_i^2\sigma_i^2 2^{-2\left(R_i-1+\epsilon\left(\frac{\mu_i}{\sigma_i}\right)\right)}$. Thus, it remains to solve the following optimization problem:

$$\begin{aligned} \underset{R_i}{\text{minimize}} \quad & \sum_{i=1}^{M} w_i\sigma_i 2^{-R_i+1-\epsilon\left(\frac{\mu_i}{\sigma_i}\right)}. \\ \text{subject to} \quad & \sum_{i=1}^{M} R_i = R \end{aligned} \tag{33}$$

Following the same lines as those in the proof of Theorem 1, we get $D_i = \frac{\lambda^2}{\ln^2 2} = \lambda'$. Thus, the achievable rate-allocation strategy in (31) also leads to a kind of reverse water-filling for the rate-allocation strategy, as illustrated in Fig. 2. ■

**Remark 1.** As seen in Theorem 2 and Theorem 3, the rate allocated to the $i$-th feature is dependent on the parameters $\Lambda_i^2$ and $w_i^2\sigma_i^2$, respectively. Traditionally, Fisher score is used to determine the most discriminant features so that the features with the highest Fisher scores (larger than a threshold) are assumed to be more relevant and the others are assumed less relevant (and thereby ignored). Now, we show that under certain assumptions, our rate-allocation strategies are soft generalizations of feature selection using Fisher score where the relevance of each feature is not measured in a *hard* manner, i.e., whether the feature is relevant or irrelevant. Instead, each feature is assigned a degree of relevance quantified by the rate allocated to it. For this, assume that the features of each data point are statistically mutually independent and also $\sigma_{1i} = \sigma_{2i} = \sigma_i$ for the $i$-th feature, which imply that the covariance matrix $\mathbf{\Sigma}$ is a diagonal matrix whose diagonal elements are $\sigma_i^2$, $i = 1,2,\ldots,M$. Obtaining the vector $\mathbf{w}$ as in (3) with factor of proportionality equal to 1, we have

$$\mathbf{w} = \mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) = \mathbf{\Sigma}^{-1}(2\boldsymbol{\mu}), \tag{34}$$

where we have used the assumption that $\boldsymbol{\mu_1} = -\boldsymbol{\mu_2}$. Therefore,

$$w_i = \frac{2\mu_i}{\sigma_i^2}. \tag{35}$$

In Theorem 2, the rate allocated to each feature is dependent on the parameter $\Lambda_i^2 = w_i^2\left(\sigma_i^2+\mu_i^2\right)$. Using (35), we have

$$\Lambda_i^2 = w_i^2\sigma_i^2\left(1+\frac{\mu_i^2}{\sigma_i^2}\right) \tag{36}$$

$$= \frac{4\mu_i^2}{\sigma_i^2}\left(1+\frac{\mu_i^2}{\sigma_i^2}\right) \tag{37}$$

$$= FS_i\left(1+\frac{FS_i}{4}\right), \tag{38}$$

which is a function of Fisher score of the $i$-th feature, $FS_i$.

In Theorem 3, the rate allocated to the $i$-th feature depends on the variance of $w_iX_i$, i.e., $w_i^2\sigma_i^2$. From (35), we have

$$w_i^2\sigma_i^2 = \frac{4\mu_i^2}{\sigma_i^2}, \tag{39}$$

which is equal to the Fisher score of the $i$-th feature, $FS_i$. Thus, the rate-allocation strategy corresponding to the first

inner bound can be considered as a generalization of feature selection using Fisher score.

## IV. Numerical Results: An Application to Recorded Neural Data

In this section, we test our proposed method of inference-oriented compression on data obtained from real-world neural measurement systems to evaluate its effect on classification accuracy.

### A. Data Description

In this study, the ECoG signals recorded with a high-density 32-electrode grid over the hand and arm area of the left sensorimotor cortex of a paralyzed individual are used. The individual can activate his sensorimotor cortex using attempted movements to the left or right. The ECoG data set used, consists of $140$ trials, $70$ trials for each of the movement directions. Each trial is $300ms$ long and sampled at $1.2kHz$ frequency, resulting in $361$ samples per trial. Given a trial, we are interested in decoding the movement direction.

### B. Data Pre-processing

In this experiment, instead of common feature extraction techniques which rely on spectral density estimation [26] or band-pass filters [27], we use discrete cosine transform (DCT) as proposed in [3] which reduces the power consumption in extracting brain-computer interface (BCI) features substantially. Taking the DCT of the signals recorded by each of the 32 channels for integer frequencies from $0$ to $120$, we obtain a $3872$-dimensional feature vector ($121$ frequencies for $32$ channels) for each trial. Linear classification using the LDA algorithm is performed on these $3872$-dimensional data points.

### C. Evaluating Classification Accuracy

First, only the important features with Fisher scores above a threshold, in this case $0.25$, are kept and the other features are removed. Then, we select $35$ trials of each class randomly to train the LDA algorithm, i.e., to obtain the vector $\mathbf{w}$ as in (3) with factor of proportionality equal to $1$, and use the other trials for validation. After training is done, the classification is performed by first quantizing each of the remaining $28$ features of a validation data point with the number of bits allocated to that feature according to Theorem 2 (note that the total number of bits are divided only between $28$ features), and then performing the LDA algorithm on these quantized DCT features. The result, illustrated in Fig. 3, provides the tradeoff between classification accuracy and total number of bits allocated to features under the assumption that the DCT values for different frequencies are independent. Even under this inaccurate assumption, the resulting classifier works with about $90\%$ accuracy with just $30$ total bits allocated across $3872$ features. This illustrates the dramatic potential for energy savings: allocating only one bit to represent each of $3872$ features would need $3872$ bits ($130$x more energy) to be transmitted from the sensors, with barely any classification accuracy. For another comparison, allocating $8$ bits to represent each of $28$ important features as suggested in [3], would need $224$ bits ($8$x more energy) and the classification accuracy would not be any better (about $82.5\%$).

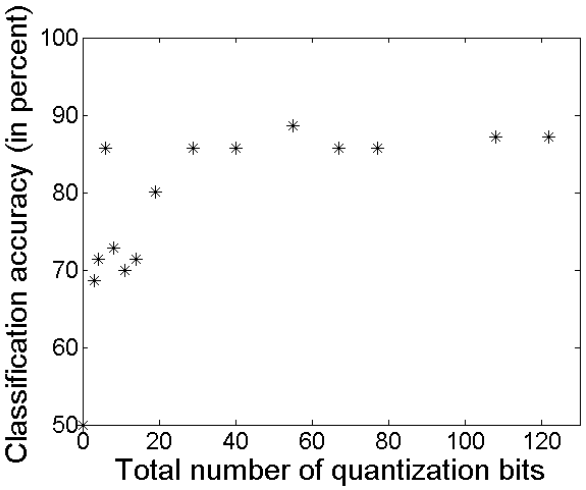


Fig. 3. Binary classification accuracy vs. number of bits with "inference-oriented" communication. The number of bits is spread across 3872 features for neural data of binary movement decoding. Notice that accuracy close to $90\%$ is obtained with just 30 total bits. If relevance is disregarded, and bits are distributed uniformly across all features with one bit for each feature, it would still require 3872 bits, with barely any accuracy.

## V. Conclusions

We investigated the problem of allocating bits to data streams as a function of their relevance for inference obtained from the data. The proposed techniques go beyond standard compression and compress more relevant data with higher resolution. These techniques could save a substantial amount of energy in communication for neural data acquisition, and for Big Data acquisition in general. Under simplifying assumptions of Gaussianity and independence of features, our techniques for classification can be interpreted as soft feature selection techniques. A metric widely used in practice for feature selection in machine learning, namely, Fisher score, falls out of the Gaussian formulation. Numerical experiments on neural data indicate that our proposed method achieves substantial savings on the number of bits communicated while the gains in classification saturate with increasing number of bits.

Many problems remain to be addressed. Besides extension to multiclass classification and stronger connections with existing works in distributed source coding, to bring our results closer to practice, we need more realistic models of energy consumption in circuits (e.g., [28], [29]).

## VI. Acknowledgements

The authors wish to thank Wei Wang from University of Pittsburgh for sharing his valuable experience and measurement data of ECoG-based BCI. We thank Vinod Prabhakaran for commenting on an early version of the paper and

most useful pointers. We also acknowledge the support of NSF-ECCS-1343324 (NSF EARS program) and NSF-CCF-1350314 (NSF CAREER).

## REFERENCES


[1] M. A. Lebedev and M. A. Nicolelis, "Brain–machine interfaces: past, present and future," *TRENDS in Neurosciences*, vol. 29, no. 9, pp. 536–546, 2006.

[2] A. H. Marblestone, B. M. Zamft, Y. G. Maguire, M. G. Shapiro, T. R. Cybulski, J. I. Glaser, D. Amodei, P. B. Stranges, R. Kalhor, D. A. Dalrymple *et al.*, "Physical principles for scalable neural recording," *Frontiers in Computational Neuroscience*, vol. 7, 2013.

[3] M. Won, H. Albalawi, X. Li, and D. E. Thomas, "Low-power hardware implementation of movement decoding for brain computer interface with reduced-resolution discrete cosine transform," *Accepted by IEEE EMBC 2014*.

[4] D. Seo, J. M. Carmena, J. M. Rabaey, E. Alon, and M. M. Maharbiz, "Neural dust: An ultrasonic, low power solution for chronic brain-machine interfaces," *arXiv preprint arXiv:1307.2196*, 2013.

[5] T. Berger, "Decentralized estimation and decision theory," *presented at the IEEE 7th Workshop on Information Theory, NY*, Sept 1979.

[6] D. Slepian and J. Wolf, "Noiseless coding of correlated information sources," *Information Theory, IEEE Transactions on*, vol. 19, no. 4, pp. 471–480, Jul 1973.

[7] T. Cover, "A proof of the data compression theorem of Slepian and Wolf for ergodic sources (corresp.)," *Information Theory, IEEE Transactions on*, vol. 21, no. 2, pp. 226–228, Mar 1975.

[8] A. Wyner and J. Ziv, "The rate-distortion function for source coding with side information at the decoder," *Information Theory, IEEE Transactions on*, vol. 22, no. 1, pp. 1–10, Jan 1976.

[9] T. N. Lal, M. Schroder, T. Hinterberger, J. Weston, M. Bogdan, N. Birbaumer, and B. Scholkopf, "Support vector channel selection in BCI," *Biomedical Engineering, IEEE Transactions on*, vol. 51, no. 6, pp. 1003–1010, 2004.

[10] Y. Wang, S. Gao, and X. Gao, "Common spatial pattern method for channel selelction in motor imagery based brain-computer interface," in *Engineering in Medicine and Biology Society, 2005. IEEE-EMBS 2005. 27th Annual International Conference of the*, Jan 2005, pp. 5392–5395.

[11] J. Farquhar, J. Hill, T. N. Lal, and B. Schölkopf, "Regularised CSP for sensor selection in BCI," 2006.

[12] J. Meng, G. Liu, G. Huang, and X. Zhu, "Automated selecting subset of channels based on CSP in motor imagery brain-computer interface system," in *Robotics and Biomimetics (ROBIO), 2009 IEEE International Conference on*, Dec 2009, pp. 2290–2294.

[13] M. Arvaneh, C. Guan, K. K. Ang, and C. Quek, "Optimizing the channel selection and classification accuracy in EEG-based BCI," *Biomedical Engineering, IEEE Transactions on*, vol. 58, no. 6, pp. 1865–1873, June 2011.

[14] T. Lan, D. Erdogmus, A. Adami, M. Pavel, and S. Mathan, "Salient EEG channel selection in brain computer interfaces by mutual information maximization," in *Engineering in Medicine and Biology Society, 2005. IEEE-EMBS 2005. 27th Annual International Conference of the*, Jan 2005, pp. 7064–7067.

[15] T. Lan, D. Erdogmus, A. Adami, S. Mathan, and M. Pavel, "Channel selection and feature projection for cognitive load estimation using ambulatory EEG," *Computational intelligence and neuroscience*, vol. 2007, pp. 8–8, 2007.

[16] M. Schroder, M. Bogdan, T. Hinterberger, and N. Birbaumer, "Automated EEG feature selection for brain computer interfaces," in *Neural Engineering, 2003. Conference Proceedings. First International IEEE EMBS Conference on*, March 2003, pp. 626–629.

[17] L. Citi, R. Poli, C. Cinel, and F. Sepulveda, "Feature selection and classification in brain computer interfaces by a genetic algorithm," in *Late-breaking papers of the Genetic and Evolutionary Computation Conference (GECCO-2004)*, vol. 400, 2004.

[18] G. Santhanam, M. Y. Byron, V. Gilja, S. I. Ryu, A. Afshar, M. Sahani, and K. V. Shenoy, "Factor-analysis methods for higher-performance neural prostheses," *Journal of neurophysiology*, vol. 102, no. 2, pp. 1315–1330, 2009.

[19] Y. Yang and J. O. Pedersen, "A comparative study on feature selection in text categorization," in *Proceedings of the Fourteenth International Conference on Machine Learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1997, pp. 412–420.

[20] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, mar 2003.

[21] Q. Gu, Z. Li, and J. Han, "Generalized Fisher score for feature selection," *arXiv preprint arXiv:1202.3725*, 2012.

[22] T. Berger, Z. Zhang, and H. Viswanathan, "The CEO problem," *IEEE Trans. Information Theory*, vol. 42, no. 3, pp. 887–902, May 1996.

[23] T. Han and S.-I. Amari, "Statistical inference under multiterminal data compression," *Information Theory, IEEE Transactions on*, vol. 44, no. 6, pp. 2300–2324, Oct 1998.

[24] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York, NY, USA: Wiley-Interscience, 2006.

[25] Z. Reznic, R. Zamir, and M. Feder, "Joint source-channel coding of a gaussian mixture source over the gaussian broadcast channel," *IEEE TRANS. INFORM. THEORY*, vol. 48, pp. 776–781, 2002.

[26] W. Wang, J. L. Collinger, A. D. Degenhart, E. C. Tyler-Kabara, A. B. Schwartz, and et al., "An electrocorticographic brain interface in an individual with tetraplegia," *PLoS One*, vol. 8, no. 2, pp. 1–8, Feb 2013.

[27] J. Yoo, L. Yan, D. El-Damak, M. Altaf, A. Shoeb, and A. Chandrakasan, "An 8-channel scalable EEG acquisition SoC with patient-specific seizure classification and recording processor," *Solid-State Circuits, IEEE Journal of*, vol. 48, no. 1, pp. 214–228, Jan 2013.

[28] K. Ganesan, P. Grover, and J. Rabaey, "The power cost of over-designing codes," in *Signal Processing Systems (SiPS), 2011 IEEE Workshop on*, Oct 2011, pp. 128–133.

[29] P. Grover, K. Woyach, and A. Sahai, "Towards a communication-theoretic understanding of system-level power consumption," *Selected Areas in Communications, IEEE Journal on*, vol. 29, no. 8, pp. 1744–1755, September 2011.