Energy-Constrained Distributed Learning and Classification by Exploiting Relative Relevance of Sensors' Data

Majid Mahzoon, Christy Li, Xin Li, and Pulkit Grover

***Abstract*—We consider the problem of communicating data from energy-constrained distributed sensors. To reduce energy requirements, we go beyond the source reconstruction problem classically addressed, and focus on the problem where the recipient wants to perform supervised learning and classification on the data received from the sensors. Restricting our attention to a noiseless communication setting under simplistic Gaussian source assumptions, we study supervised learning and classification under total energy limitations. The energy constraints are modeled in two ways: 1) a linear scaling and 2) an exponential scaling of energy with number of bits used for compression at sensors. We first assume that the underlying parameters for Gaussian distributions have already been learned, and obtain (with linear scaling, reverse-waterfilling-type) strategies for allocating energy, and thus, bits, across different sensors under these two models. Intuitively, these strategies allocate larger rates and energies to sensors that are more "relevant" for the classification goal. These strategies are used to obtain an achievable bound on the tradeoff between energy and error-probability (classification risk). We then provide an algorithm for learning the distribution-parameters of the sensor-data under energy constraints to arrive at high-reliability energy-allocation strategies, while enabling the energy-allocation algorithm to backtrack when the underlying distributions change, or when there is noise in sensed data that can push the algorithm toward a local minimum. Finally, we provide numerical results on energy-savings for classification of simulated data as well as neural data acquired from electrocorticography (ECoG) experiments.**

***Index Terms*—Rate and energy-constrained learning, supervised learning, classification, statistical inference, distributed sensing, neural sensing.**

I. Introduction

ENERGY-EFFICIENT communication[1] for distributed sensors is gaining importance with the advent of wearable devices, structural health monitoring, body-area networks, and networked sensing. It is now widely acknowledged that energy will be a limiting bottleneck in the Internet of Things (IoT) revolution, where devices will sense and communicate enormous volumes of data which will be processed to make various inferences and decisions. Because of the tight coupling and tradeoffs between sensing, communication, and computing in these devices, there is an urgent requirement for cross-disciplinary approaches to reduce energy consumption that bridge concepts from information theory, computer science, hardware/circuits, and signal processing.

Towards this larger goal, this paper is motivated by the practitioner's observation that in many cases, different sensors have different relevance to the inference or signal processing task ("the goal") that the practitioner wants to perform. Often, there is a gentle gradation of relevance of sensors from immensely relevant to irrelevant. Thus, it might be possible to reduce required communication and sensing energy by reducing fidelity of observations from sensors that are less relevant to the goal[2]. In this work, we are motivated by problems where the *total* sensor network energy, summed over all sensors, is constrained. Are there applications where such constraints arise naturally? Indeed, one situation is when all sensors are wired, such as in health-monitoring and body-area networks[3]. Another situation is in wirelessly powered sensor networks, where the system can focus its power-beam towards sensors that are allocated more energy based on their relevance to the goal.

A particular application of our interest, that motivates our problem formulation as well, is high-electrode-count (thousands or tens of thousands of electrodes) brain-computer interfaces (BCIs) and neural recording systems. A parallel work of our team is exploring fundamental and practical challenges in development and utilization of such high-density systems [1]–[3]. Classification is a common tool used in BCIs for various applications such as neuroprostheses. It is commonly observed that only some of the electrodes are recording data that is relevant to the classification goal (as observed in neural data analysis on real data in [1]), and even for these electrodes, there is significant range of relative relevance. However, the most relevant electrodes can not be predicted in advance as they are different for different users, and even for the same user, they evolve over time due to neural plasticity. Thus, it is common to retrain the classifier frequently in order to maintain a low error-probability. The electrode count itself is limited by area

Manuscript received August 1, 2015; revised January 5, 2016; accepted February 24, 2016. Date of publication March 22, 2016; date of current version May 19, 2016. This work was supported by the NSF-ECCS-1343324, NSF-CCF-1350314 (NSF CAREER), NSF-CCF-1148778, and of Systems on Nanoscale Information fabriCs (SONIC), one of the six SRC STARnet Centers, sponsored by MARCO and DARPA.

The authors are with the Carnegie Mellon University, Pittsburgh, PA 15213 USA (e-mail: mmahzoon@andrew.cmu.edu; yli3@andrew.cmu.edu; xinli@andrew.cmu.edu; pgrover@andrew.cmu.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSAC.2016.2545381

[1] A part of this work (a subset of Section III-A) was presented at the Annual Allerton Conference in 2014 [1].

[2] Such techniques for reduction in energy would not work if the end-goal is not known. Often in such cases, collection from all sensors at high resolution is the only resort, and is common practice.

[3] In fact, it is not just a battery constraint here. If all of the sensors consume large energy in a dense sensing environment, the amount of heat generated can get large enough to burn the tissue.

and power of sensors: the power consumption of 1000s of electrodes can easily exceed tens of watts [2]! Such large power consumption can cause the user to perspire (which can cause the electrodes to short-circuit, compromising the recordings), or worse, burns in their tissue. The sensors are all wired together, and thus a total energy constraint is fairly realistic.

Thus, we seek strategies that allocate each sensor an amount of energy based on its relevance to the inference goal under a total energy constraint. We use two models of sensor-energy consumption as a function of the compression rate (measured in bits per observation) allocated to each sensor: a linear model, and an exponential model. The linear model is more suitable when communication energy dominates sensing energy, or when the sensing circuitry employs certain slow and low-resolution (e.g. "Successive Approximation Register (SAR)") Analog-to-Digital Converter (ADC) architectures. When sensing circuitry employs higher resolution and/or fast ADCs (e.g. flash ADCs or oversampled ADCs), and when sensing energy dominates communication energy, an exponential-type model is more appropriate [2].

Our goal is to use these models to obtain energy-efficient rate allocations that minimize the classification error-probability. Intuitively, an efficient rate allocation would allocate larger number of bits to more relevant sensors. The relevance of a sensor must in turn depend on what inference we desire from the measurements. In this work, our inference goal is supervised learning and classification from a parameterized Bayesian perspective. For simplicity, we limit ourselves to Gaussian priors on the data under different classes, and further restrict our attention to two classes (*i.e.*, binary classification). As a key simplifying step, we approximate the problem of minimizing error-probability with that of recovering the decision variable (prior to thresholding, see Section III-A). For this simplistic problem and for both models of energy consumption, we provide the following results on energy-constrained supervised learning:

(i) Energy-constrained algorithms for learning the parameters on labeled data (prior to testing) of the underlying distributions from quantized observations, and using them for relevance-based quantization. Importantly, our strategies learn parameters of more relevant sensors at a higher accuracy, incorporate backtracking in order to deal with changes in underlying distributions and/or burst noise in sensing and/or communication;

(ii) Energy-constrained testing/classification of unlabeled data assuming that the underlying distributions have been perfectly learned and fixed (*i.e.*, they do not evolve over time). We also provide an upper bound on the asymptotic "energy-risk" tradeoffs, that is the tradeoff between total available energy and the classification "risk," *i.e.*, the classification error-probability [4].

While (ii) above optimistically assumes that underlying distributions are known, and therefore appears practically less interesting in comparison with (i), it lays the necessary groundwork for (i): the strategies for (ii) are leveraged to obtain improved rate allocation in (i) even with uncertainty/errors in estimation of parameters. Because of energy constraints, we assume that sensors are not able to perform any sophisticated computations (such as parameter estimation) themselves, and instead rely on a data-fusion node (which collects the data sent by individual nodes) to estimate parameters and decide rate allocation.

This work is deeply connected with literature in information theory, signal processing, and distributed machine learning. Our strategies are related to the reverse-waterfilling strategy [5] in information theory that is the solution to the classical rate allocation problem for minimum distortion when communicating independent Gaussian sources. There is one important difference between our classification formulation and the classical communication problem. In the classical problem, sources with larger variances are assigned larger number of bits to be represented faithfully. This is reminiscent of Principle Component Analysis (PCA), where features with larger variances are retained through training [4]. The key observation is that even large-variance sources can be irrelevant when the goal is not communication, but classification. For instance, if the mean of a sensor's observations does not change significantly under the two hypotheses, then these observations are not very relevant. Instead of variance, a parameter called the "Fisher score" [6] of the sources — given by square of the difference of means under the two classes, divided by the (common) variance under each class — turns out to be the important parameter.

The fact that we naturally arrive at Fisher score to quantify relevance of each sensor is interesting, but hardly surprising: Fisher score is one of the key metrics used frequently in the problem of "sensor selection," which is a precursor to our problem (see, e.g., [7]–[10]). Motivated primarily by ease of computation through dimensionality reduction (once the data has been collected), sensor-selection algorithms select the sensors most relevant to the task at hand, and ignore data from other sensors (including using metric beyond the Fisher score, see e.g. [8]–[10]). More broadly, sensor-selection algorithms have been developed for many applications such as target tracking, distributed detection, field reconstruction, etc. (e.g., [11]). The main difference between our work and classical sensor selection is that while sensor-selection algorithms make a binary decision on whether to select a sensor, we seek a softer sensor selection, where we assign each sensor bits/energy based on the degree of relevance of a sensor to the inference goal.

A body of work in information theory implicitly addresses the problem of rate allocation based on relevance for a closely related problem of hypothesis testing, notably the works of Han and Amari [12] (see also the citations therein), Ahlswede and Csiszar [13], and Berger's work (and the follow-up work) on decentralized estimation [14] and the CEO problem [15]. Our main contributions that go beyond these works are:

- We do not allow our sensors to perform sophisticated computations. This is in part motivated by the BCI application, where sensors are largely just amplifiers followed by ADCs [2]. Thus, unlike e.g. in [12], our sensors are unable to carry out parameter estimation themselves, and can only communicate data at a flexible (but specified) resolution. This naturally leads us to a formulation where we reconstruct the decision variable which is also a formulation not considered in these works.

- We are interested in not merely the problem when the underlying distributions are static and known, but in the problem of learning itself. That is, the parameters of the underlying distributions also need to be learned under rate/energy constraints (and need to be relearned when the distribution changes).
- Finally, we address not just a sum-rate problem, but also a sum-energy problem using an exponential model of energy (Lemma 2 and ensuing results and algorithms).

We note that for conceptual simplicity, we allow the sensors to use vector quantization strategy in Section III-A. While we aim to provide solutions that work for scalars and small blocklengths, the current results fall short of that goal. However, we do note that the multiplicative difference in average distortion for scalars and infinite blocklengths can be small (e.g. for Gaussian sources [16], [17], the factor is for large rates $\pi\frac{\sqrt{3}}{2} \approx 2.72$, and empirically smaller for smaller rates [18]). Thus, we still use guidance obtained on rate allocation from the asymptotic analysis in Section III-A for scalar strategies in Section III-B.

More recently, within statistical learning, there has been an increasing interest in distributed learning algorithms, though most of the literature is motivated by connectivity, and not rate, constraints. Rate/energy-risk tradeoffs such as those proposed here, have also been examined in the recent work of Lafferty[4] [19], albeit not in a distributed sensing setting, but a distributed *computing* setting.

Another related body of work from signal processing perspective focuses on detecting edges in the sensors' field (e.g., [20], [21]), regression and clustering (e.g., [22]–[25]). In [22], [23], by adaptive learning, a constrained resource such as number of samples is assigned to more informative data region to reconstruct function with boundaries. This "active learning" is further studied in [26], [27] in the context of reconstructing sparse vectors by allocating more sampling points at non-zero values. Besides focusing on a different goal (classification) and using a different constraint (that on rate/energy), the active learning strategy proposed in this paper improves in the above aspects by providing a backtracking algorithm — "ambivalence-backtracking" — that backtracks towards "ambivalence," *i.e.*, chooses an allocation closer to uniform bit allocation. This allows our algorithm to retrain and classify even when the underlying distributions of sensor-observations are changing, or simply when a burst-noise pushes the classifier in an undesirable direction. The strategy is particularly useful in our reverse-waterfilling-type allocation (Lemma 1) where many of the sensors that are deemed less relevant may not receive any rate at all, and may therefore be turned off.

The rest of the paper is organized as follows. In Section II, we provide the problem statement, the notation, and the models of energy consumption (namely, linear and exponential in the number of sensor bits). In Section III, we provide energy-constrained algorithms for (i) classification/testing once learning has been performed (Section III-A), and (ii) energy-constrained learning (Section III-B). We explore the provided algorithms numerically in Section IV, where we also include an analysis on real-world neural data acquired from an Electrocorticographical (ECoG) system that is implanted on the surface of the brain and is used as a neuroprosthesis through a classification algorithm. We observe that while we treat features as individual sensors in our problem formulations, the application to real data brings out the fact that the same sensor can generate multiple features in this application. We discuss this disconnect further in the concluding section (Section V) along with other directions of future work.

II. Notation and Problem Statement

To understand the problem of rate- and energy-constrained supervised learning and classification, we will first (in Section III-A) investigate a problem in which the parameters of the underlying distributions have already been estimated. In Section III-B, we will use the insights obtained from this problem to provide strategies that adapt rate allocation to efficiently learn and use the knowledge of the parameters.

In our setup, there are M distributed sensors which have observations $X_i, i = 1, 2, \ldots, M$, that they communicate to a fusion center through noiseless but rate-limited channels that connect the sensors individually to the fusion center. No direct communication is allowed between sensors. The fusion center is allowed a small rate of feedback which it can use to tell each sensor its allocated rate (measured in bits per measurement). From a statistical inference perspective, for simplicity, each measurement is treated as a feature[5], and the vector $\mathbf{X} = [X_1, X_2, \ldots, X_M]$ of measurements is called a data point in the M-dimensional feature space.

At each time instant t, a data point $\mathbf{X}^{(t)}$ is drawn from class $C_t = 1$ with probability p_1 and from class $C_t = 2$ with probability p_2 *independently across time steps*.

If a data point $\mathbf{X}^{(t)}$ is drawn from class $C_t = j$, $j \in \{1, 2\}$, the sensors' observations are assumed to be distributed as jointly Gaussian, $\mathbf{X}^{(t)} \sim \mathcal{N}(\boldsymbol{\mu}_j, \Sigma_j)$. Thus, each sensor's observation is marginally (and unconditionally) a mixture Gaussian distribution. Without loss of generality, we assume that for the i-th sensor X_i, $\mu_{1i} = -\mu_{2i} = -\mu_i$, *i.e.*, the conditional means under the two classes are symmetric around zero. Further, for simplicity, we assume that $\Sigma_j = \Sigma$ for $j = 1, 2$, with diagonal elements σ_i^2. 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, $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$. Under our simplistic assumptions, the Bayes's optimal solution is to classify the data point $\mathbf{X}$ in the first class if the log-likelihood ratio is above some threshold constant ν', which is equivalent to,

$$(\mathbf{X} - \boldsymbol{\mu}_1)^T \, \Sigma_1^{-1} \, (\mathbf{X} - \boldsymbol{\mu}_1) + \ln |\Sigma_1| - (\mathbf{X} - \boldsymbol{\mu}_2)^T \, \Sigma_2^{-1} \, (\mathbf{X} - \boldsymbol{\mu}_2) - \ln |\Sigma_2| > \nu'. \quad (1)$$

Because we assume that $\Sigma_1 = \Sigma_2 = \Sigma$, the decision criterion in (1) simplifies to $\mathbf{w}^T \mathbf{X} < \nu$, where

$$\mathbf{w} = \Sigma^{-1} \left(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1 \right), \text{ and } \nu = \log \frac{p_1}{p_2}. \quad (2)$$

[4] Interestingly, reverse-waterfilling also shows up in Lafferty's problem [19].

[5] When each individual measurement is not a feature, but feature extraction can be performed at individual sensors, the results extend naturally.

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. This is called the Linear Discriminant Analysis (LDA) [28].

A. Energy Consumption Models for Given Rate of Compression

Two models of energy consumption are considered. In both models, the energy consumption E_{sensor} is a function of the number of bits, denoted by R_{sensor}, used to represent each sensor's source symbol:

Definition 1 (Linear Model): The energy consumed by a sensor is given by $E_{sensor} = R_{sensor}$. The proportionality constant is ignored for simplicity.

Definition 2 (Exponential Model): The energy consumed by a sensor is given by $E_{sensor} = 2^{R_{sensor}}$. The constant in the exponent is ignored for simplicity.

Our problem is therefore a rate allocation problem, which is naturally information theoretic in nature, with either sum-rate constraints (Linear Model) or a constraint on sum of exponentials of the rates (Exponential Model). Results will be obtained for both models.

B. Energy-Constrained Classification Assuming Underlying Parameters are Known and Fixed

In Section III, we address the problem of classification under total energy constraints assuming that the communication channels from sensors to the data-fusion center are noiseless. First, in Section III-A, we ask the question of how to quantize sensor data at each sensor (locally) based on the relevance of a sensor's data to the classification goal, *assuming that the underlying parameters of distributions* ($\boldsymbol{\mu}$ *and* Σ) *have already been perfectly estimated through learning*. Thus, the learning period using labeled data is assumed to be over, with perfect learning and parameter estimation accomplished.

The i-th sensor observation X_i is communicated to the fusion center at rate R_i. For analytical simplicity, we assume that the sensor can perform vector quantization across n time steps. Thus, the i-th sensor uses an encoder (quantization) function $\mathcal{E}_i : \mathcal{X}_i^n \rightarrow \{1, \ldots, 2^{nR_i}\}$ and sends the index $\mathcal{E}_i(X_i^n)$ to the fusion center. The goal is to design these strategies to minimize the classification error-probability averaged over data points realizations and normalized by time, subject to a constraint on the total energy of communication, *i.e.*,

$$\begin{aligned} &\inf_{\mathcal{E}_i, R_i} \frac{1}{n} \sum_{t=1}^{n} \Pr\left(C\left(\mathbf{X}^{(t)}\right) \neq \widehat{C}_t\left(\mathcal{E}_1(X_1^n), \ldots, \mathcal{E}_M(X_M^n)\right)\right) \\ &\text{subject to} \sum_{i=1}^{M} E_i \leq E_{total}, \end{aligned} \tag{3}$$

where $C\left(\mathbf{X}^{(t)}\right)$ is the true underlying class of $\mathbf{X}^{(t)}$ and $\widehat{C}_t\left(\cdot\right)$ is the predicted class of $\mathbf{X}^{(t)}$ based on the received indices.

Denote the result of the optimization in (3) as $P_{e,n}$, and the error-probability at any particular time t as $P_e^{(t)}$. In Section III-A, we will derive an asymptotic upper bound on $P_e^{(t)}$ (and thus on $P_{e,n}$ in the $n \rightarrow \infty$ limit) assuming the Gaussian parameters are perfectly estimated, thus providing an inner bound on the energy-risk tradeoff for the classification problem considered here. The resulting rate allocation will serve as guidance for algorithm-design for the learning problem in Section III-B.

Because our fusion center recovers $\widehat{X}_i$, an estimate of X_i, it uses a set of decoding functions $\mathcal{D}_i : \{1, \ldots, 2^{nR_i}\} \rightarrow \widehat{\mathcal{X}}_i^n$ to reproduce the sensor observations. With a slight abuse of notation, we continue to use $\widehat{C}_t$ as the classifier applied on received data at the fusion center, even though it is applied now on the decoded data point from the received indices.

C. Energy-Constrained Supervised Learning and Classification

In Section III-B, we then address the problem where underlying parameters of distributions of sensor-data are not known in advance, and have to be estimated at the data-fusion center. Two problems are considered: first, we consider a problem where the parameters are unknown but fixed, which allows a simple energy-constrained algorithm building on the strategies developed in Section III-A. Next, we consider the problem where the parameters are unknown and can slowly evolve.

Throughout the paper, vectors are generally denoted in bold font. $Q(.)$ (the "Q-function" [29]) is the tail probability of the standard Gaussian distribution, that is, $Q(x) = \frac{1}{\sqrt{2\pi}} \int_x^\infty e^{-\frac{u^2}{2}} du$.

III. Algorithms for Classification and Learning

We first (in Section III-A) provide a relevance-based data-compression technique that relies on perfectly estimated parameters of distributions at sensors to obtain energy-efficient rate allocations for the sensors. In Section III-B, we will then use the insights from these rate allocations to develop a learning strategy that learns more relevant distributions with higher precision, and thus tracks changes in the underlying distributions of more relevant sensors with higher precision.

A. Energy-Constrained Classification Assuming Underlying Parameters are Known and Fixed

We first provide a description of our energy-constrained classification strategies, and then (in Theorem 1) provide an inner bound on the "energy-risk tradeoff" by quantifying the tradeoff for our strategy.

At the fusion center, using a lossy reconstruction of the data, we cannot outperform the case with full observation of the data points in terms of classification error-probability. Therefore, we try to minimize the loss in classification accuracy due to use of the *estimates* of data points $\{\widehat{\mathbf{X}}^{(1)}, \widehat{\mathbf{X}}^{(2)}, \ldots, \widehat{\mathbf{X}}^{(n)}\}$ instead of the observed sensor observations themselves $\{\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \ldots, \mathbf{X}^{(n)}\}$, *i.e.*,

$$\begin{aligned} \inf_{\mathcal{E}_i, \mathcal{D}_i, E_i} \quad & \frac{1}{n} \sum_{t=1}^{n} \Pr\left(\widehat{C}\left(\widehat{\mathbf{X}}^{(t)}\right) \neq \widehat{C}\left(\mathbf{X}^{(t)}\right)\right) \\ \text{subject to} \quad & \sum_{i=1}^{M} E_i \leq E_{total}. \end{aligned} \tag{4}$$

Since LDA only depends on the decision variable $\mathbf{w}^T\mathbf{X}$, intuitively, a better approximation of $\mathbf{w}^T\mathbf{X}$ will lead to an improved classification accuracy. This suggests the following approximation of (4):

$$\begin{aligned} \inf_{\mathcal{E}_i,\mathcal{D}_i,E_i} \quad & \frac{1}{n}\sum_{t=1}^{n}\mathbb{E}\left[\left(\mathbf{w}^T\mathbf{X}^{(t)}-\mathbf{w}^T\widehat{\mathbf{X}}^{(t)}\right)^2\right] \\ \text{subject to} \quad & \sum_{i=1}^{M} E_i \leq E_{total}. \end{aligned} \tag{5}$$

This approximation, while suboptimal (because source reconstruction of even relevant sources may not be needed for classification), motivates a parameter estimation and tracking algorithm in Section III-B where we use the estimates of (parametrized) distributions of X_i to update the rate allocation. It allows us to track the distributions of more relevant sensors with higher precision.

Assuming that every sensor uses a blocklength n, we will use quantization strategies at each sensor that achieve a distortion $\overline{D}_n(E_{total})$, which denotes an achievable distortion-energy function (defined analogously to distortion-rate function for energy models in Section II-A) for blocklength n. We will use "equidistortion" quantization strategies at each sensor, defined as follows:

Definition 3 (Equidistortion quantization strategy): For an iid source vector $\mathbf{X} = [X^{(1)}, \ldots, X^{(n)}]$, an equidistortion quantization strategy is a vector quantization scheme for which the average reconstruction error for the i-th element $\mathbb{E}[(X^{(i)} - \widehat{X}^{(i)})^2]$ is the same for all i.

For equidistortion quantization strategies (at each sensor) that achieve an overall distortion-energy function $\overline{D}_n(E_{total})$, we now obtain the following result:

Theorem 1 (Energy-risk tradeoffs for classification): For the problem stated in Section II-B, using a blocklength n equidistortion quantization scheme at each sensor for quantizing each $X_i^{(t)}$ (where the average distortion is equal for all $t = 1, 2, \ldots, n$ for any fixed i) that achieves an average distortion-energy function $\overline{D}_n(E_{total})$, the classification error-probability $P_e^{(t)} = \Pr\left(C\left(\mathbf{X}^{(t)}\right) \neq \widehat{C}\left(\widehat{\mathbf{X}}^{(t)}\right)\right)$ of the t-th data-point is bounded as follows:

$$\begin{aligned} P_e^{(t)} \leq \min_{\xi>0} p_1 Q\left(\frac{\mu_w+\nu-\xi}{\sigma}\right) + p_2 Q\left(\frac{\mu_w-\nu-\xi}{\sigma}\right) \\ + \frac{\overline{D}_n(E_{total})}{\xi^2}, \end{aligned} \tag{6}$$

where p_j is the probability that the underlying class is $C_t = j$, $j \in \{1, 2\}$, $\mu_w = \mathbb{E}[\mathbf{w}^T\mathbf{X}^{(t)}|C_t = 2]$, and $\sigma^2 = Var[\mathbf{w}^T\mathbf{X}^{(t)}|C_t = 2]$.

Proof: Our decision rule for classifying $\mathbf{X}^{(t)}$ at time t is based on the reconstruction $\widehat{\mathbf{X}}^{(t)}$:

$$\widehat{C}_t = \begin{cases} 1, & \mathbf{w}^T\widehat{\mathbf{X}}^{(t)} < \nu, \\ 2, & \mathbf{w}^T\widehat{\mathbf{X}}^{(t)} > \nu, \end{cases} \tag{7}$$

where ν is the threshold defined in (2) and $\widehat{C}_t$ is the predicted class. Because $\mu_w = \mathbb{E}[\mathbf{w}^T\mathbf{X}^{(t)}|C_t = 2]$, by symmetry, $\mathbb{E}[\mathbf{w}^T\mathbf{X}^{(t)}|C_t = 1] = -\mu_w$. In the event that the underlying class at time t is $C_t = 2$, the error-probability $P_{e,2}^{(t)}$ can be bounded as follows:

$$\begin{aligned} P_{e,2}^{(t)} &= \Pr(\mathbf{w}^T\widehat{\mathbf{X}}^{(t)} < \nu|C_t = 2) \\ &= \Pr(\mathbf{w}^T\widehat{\mathbf{X}}^{(t)} < \nu, |\mathbf{w}^T\mathbf{X}^{(t)} - \mathbf{w}^T\widehat{\mathbf{X}}^{(t)}| < \xi|C_t = 2) \\ &\quad + \Pr(\mathbf{w}^T\widehat{\mathbf{X}}^{(t)} < \nu, |\mathbf{w}^T\mathbf{X}^{(t)} - \mathbf{w}^T\widehat{\mathbf{X}}^{(t)}| \geq \xi|C_t = 2) \\ &\leq \Pr(\mathbf{w}^T\mathbf{X}^{(t)} < \nu + \xi|C_t = 2) \\ &\quad + \Pr(|\mathbf{w}^T\mathbf{X}^{(t)} - \mathbf{w}^T\widehat{\mathbf{X}}^{(t)}| \geq \xi|C_t = 2) \\ &\leq Q\left(\frac{\mu_w-\nu-\xi}{\sigma}\right) \\ &\quad + \Pr(|\mathbf{w}^T\mathbf{X}^{(t)} - \mathbf{w}^T\widehat{\mathbf{X}}^{(t)}| \geq \xi|C_t = 2). \end{aligned} \tag{8}$$

From a similar equation for $P_{e,1}^{(t)}$, we get

$$\begin{aligned} P_e^{(t)} &= p_1 P_{e,1}^{(t)} + p_2 P_{e,2}^{(t)} \\ &\leq p_1 Q\left(\frac{\mu_w+\nu-\xi}{\sigma}\right) + p_2 Q\left(\frac{\mu_w-\nu-\xi}{\sigma}\right) \\ &\quad + \Pr(|\mathbf{w}^T\mathbf{X}^{(t)} - \mathbf{w}^T\widehat{\mathbf{X}}^{(t)}| \geq \xi). \end{aligned} \tag{9}$$

We now use[6] Markov's inequality to bound the third term in the RHS of (9):

$$\begin{aligned} \Pr(|\mathbf{w}^T\mathbf{X}^{(t)} - \mathbf{w}^T\widehat{\mathbf{X}}^{(t)}| \geq \xi) &\leq \frac{\mathbb{E}[|\mathbf{w}^T\mathbf{X}^{(t)} - \mathbf{w}^T\widehat{\mathbf{X}}^{(t)}|^2]}{\xi^2} \\ &\leq \frac{\overline{D}_n(E_{total})}{\xi^2}. \end{aligned} \tag{10}$$

Thus, from (9) and (10),

$$\begin{aligned} P_e^{(t)} \leq p_1 Q\left(\frac{\mu_w+\nu-\xi}{\sigma}\right) + p_2 Q\left(\frac{\mu_w-\nu-\xi}{\sigma}\right) \\ + \frac{\overline{D}_n(E_{total})}{\xi^2}. \end{aligned} \tag{11}$$

The theorem follows from the observation that (11) holds for all values of $\xi > 0$. ■

Remark 1: As a sanity-check, it is easy to observe that the bound in Theorem 1 converges to the error-probability of the unconstrained rate/energy problem in the limit of $E_{total} \to \infty$. To see this, note that for any n, as $E_{total} \to \infty$, $\overline{D}_n(E_{total}) \to 0$. Thus, setting $\xi = \sqrt{\overline{D}_n(E_{total})}$ in (6), and letting $\overline{D}_n(E_{total}) \to 0$, we get

$$P_e^{(t)} \leq p_1 Q\left(\frac{\mu_w+\nu}{\sigma}\right) + p_2 Q\left(\frac{\mu_w-\nu}{\sigma}\right), \tag{12}$$

which is the optimal error-probability with lossless data collection at the fusion center.

[6] Note that the event $|\mathbf{w}^T\mathbf{X}^{(t)} - \mathbf{w}^T\widehat{\mathbf{X}}^{(t)}| \geq \xi$ for $\xi > \sqrt{\overline{D}_n(E_{total})}$ is an "excess distortion" event [30], *i.e.*, the event in which the distortion at the t-th coordinate exceeds the average distortion. However, because we are dealing with coordinate-wise distortion, the probability of this event does not converge to zero for $\xi > \sqrt{\overline{D}_n(E_{total})}$, unlike for distortion averaged over time for any sensor. Thus, it does not seem straightforward to tighten this inequality.

In Lemmas 1 and 2 that follow, we provide upper bounds on $\overline{D}_n(E_{total})$ for the two models of energy consumption. These upper bounds are obtained by examining the asymptotic limit $n \to \infty$. Lemma 1, which is for the linear model, provides an intuitively pleasing reverse water-filling-type rate allocation. That is, a constant λ' is chosen appropriately (based on available total energy), and all features with parameter $w_i^2\left(\sigma_i^2+\mu_i^2\right)$ larger than λ' are described with equal distortion, but no rate is allocated to features with the parameter less than λ'. It turns out that the solution for the exponential model (Lemma 2) is not a reverse-waterfilling-type solution, and the sensors are always consuming energy simply because $R_i = 0$ still corresponds to $E_i = 1$. Even though these rate allocations are derived from the asymptotic case, in Section IV, we use them as guidance for rate allocations for finite-length (scalar) strategies. In Section V, we discuss the possibility of obtaining parallel results for the finite-length problem.

Lemma 1 (Linear Model): For the problem stated in Section II-B and the Linear Model of sensor energy consumption, for a given total energy E_{total}, an upper bound on the asymptotic distortion-energy function $D(E_{total})$ (asymptotically achievable distortion for a given total energy E_{total} in (5)) is given by:

$$D(E_{total}) \leq \sum_{i=1}^{M} D_i, \tag{13}$$

where

$$D_i = \begin{cases} \lambda', & \lambda' \leq \Lambda_i^2, \\ \Lambda_i^2, & \lambda' > \Lambda_i^2, \end{cases} \text{ and } \quad R_i = \frac{1}{2}\log\left(\frac{\Lambda_i^2}{D_i}\right), \tag{14}$$

where $\Lambda_i = w_i\sqrt{\sigma_i^2+\mu_i^2}$, and where λ' satisfies $\sum_{i=1}^{M} E_i = \sum_{i=1}^{M} R_i = E_{total}$.

Proof overview: We use a Gaussian upper bound for the mixture-Gaussian random variable. That is, we use the fact that for a given second moment, the distortion-rate function for any source – and hence our Gaussian mixture source – is upper bounded by that of the Gaussian source. To attain this rate-distortion tradeoff, however, requires the errors in reconstruction of different sensor observations to be uncorrelated. This is necessitated by the induced correlation in sensor observations by the underlying class, even when the sensor observations are conditionally independent (conditioned on the class). Therefore, we first use a lattice-quantization strategy with subtractive-dither [31] to ensure that the quantization errors are statistically independent of the sources being quantized [32]. Because the correlation is induced by the sources, and the errors are independent of the respective sources, the errors are mutually independent, and hence pairwise uncorrelated. Ensuring this uncorrelatedness, we carry out a derivation paralleling the classical reverse-waterfilling result. Detailed proof appears in Appendix A.

Lemma 2 (Exponential Model): For the problem stated in Section II-B, for a given total energy E_{total} in the Exponential Model, an upper bound on the asymptotic distortion-energy function $D(E_{total})$ (asymptotically achievable distortion for a given total energy E_{total} in (5)) is given by:

$$D(E_{total}) \leq \sum_{i=1}^{M} D_i, \tag{15}$$

where

$$D_i = \begin{cases} \sqrt[3]{\frac{\lambda^2\Lambda_i^2}{4}}, & \lambda \leq 2\Lambda_i^2, \\ \Lambda_i^2, & \lambda > 2\Lambda_i^2, \end{cases} \tag{16}$$

$$\text{and } R_i = \max\left\{\frac{1}{3}\log\left(\frac{2\Lambda_i^2}{\lambda}\right), 0\right\}, \tag{17}$$

where $\Lambda_i^2 = w_i^2(\mu_i^2+\sigma_i^2)$, and the value of λ (and hence also of R_i) is obtained by solving $\sum_{i=1}^{M} E_i = \sum_{i=1}^{M} 2^{R_i} = E_{total}$.

Proof: Notice that, unlike for the traditional reverse-waterfilling solution (and the solution to Lemma 1), the distortion is different for different sensors. The proof is analogous to that of Lemma 1, and is included in Appendix B. We remark that the expression on R_i is written in a slightly different form than in Lemma 1 to more explicitly bring out how the total energy constraint can be utilized to solve for λ. ■

Remark 2: We note that if a strategy is not an equidistortion strategy, one can easily obtain an equidistortion strategy by randomly permuting the strategy uniformly across all time (but not sensor) indices $\{1, 2, \ldots, n\}$. The resulting strategy is a probabilistic equidistortion strategy.

Remark 3: As seen in Lemmas 1 and 2, the rate allocated to the i-th feature is dependent on the parameter Λ_i^2. Using (2), we have $\mathbf{w} = \Sigma^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) = \Sigma^{-1}(2\boldsymbol{\mu})$, where we have used the assumption that $\boldsymbol{\mu_1} = -\boldsymbol{\mu_2}$. Therefore, $w_i = \frac{2\mu_i}{\sigma_i^2}$. In Lemmas 1 and 2, the rate allocated to each sensor is dependent on the parameter $\Lambda_i^2 = w_i^2\left(\sigma_i^2+\mu_i^2\right)$, which can be written as follows:

$$\begin{aligned} \Lambda_i^2 &= w_i^2\left(\sigma_i^2+\mu_i^2\right) = w_i^2\sigma_i^2\left(1+\frac{\mu_i^2}{\sigma_i^2}\right) \\ &= \frac{4\mu_i^2}{\sigma_i^2}\left(1+\frac{\mu_i^2}{\sigma_i^2}\right) = FS_i\left(1+\frac{FS_i}{4}\right), \end{aligned} \tag{18}$$

which is a function of Fisher score of the i-th feature, $FS_i = \frac{4\mu_i^2}{\sigma_i^2}$. Traditionally, Fisher score is used to determine the most discriminant features so that the features with the highest Fisher scores are deemed to be more relevant and are selected for further processing (this feature selection is a form of dimensionality reduction). Thus, under certain ("naive Bayes"-type [4]) assumptions, our rate allocation strategies are soft generalizations of feature selection using Fisher score where each feature, instead of receiving a hard decision (relevant or irrelevant), is assigned a soft "degree of relevance" quantified by the rate/energy allocated to it.

B. Energy-Constrained Supervised Learning for Classification

In Section III-A, we proposed a data-compression strategy that compresses the data streams based on their relevance.

There are two issues with this strategy: we assumed that the underlying distribution of data is (i) known in advance; and (ii) fixed over time. In general, neither of these assumptions is true: training is needed to estimate the parameters of the distribution, and the distribution keeps evolving over time (e.g., in BCI applications due to neuroplasticity [33]). Here we address these issues by proposing algorithms to learn the parameters that characterize marginal distributions at the sensors in a supervised manner, *i.e.*, using labeled data. Further, we provide algorithms to test when the learned distributions have evolved significantly so that the sensors can retrain themselves in that case. The novelty of our algorithms is that we learn the parameters in an energy-constrained manner, and they can thus be executed with lower energy requirements in both communication (through reduced rates) and sampling (through reduced ADC precision).

The core idea we bring in is that for estimations using a small sample size we need some form of "regularization" or ability to backtrack in the learning algorithm to adapt to large-deviation events or to evolving parameters. The algorithm should allow the rate allocation algorithm to backtrack when it settles into a local minimum (e.g., when most relevant sensors are turned off because of large-deviation events).

We now proceed to provide our algorithms. The algorithms have a "stabilization" aspect, and a "regularization" aspect. The stabilization aspect of each algorithm stabilizes it against large-deviation events by using a simple moving average-type filter. The regularization aspect reduces overfitting to data, and thus reduces the variance of the training algorithm, enabling better generalization. In the following, we detail the algorithms, emphasizing on the role these aspects play.

1) Energy-Constrained Learning Algorithm for Unknown but Fixed Distributions: The challenge lies in learning the rate allocation using data received at the fusion center that is (i) quantized; and (ii) finite sample size. Both aspects introduce errors in parameter estimation, and hence in classification.

In the first learning step, an "ambivalent rate vector" — that allocates equal number of bits to each sensor — is chosen. This initial rate vector is ambivalent in that it prefers no sensor to any other. Based on S samples of quantized data received from the sensors, the fusion center estimates parameters of these distributions, and computes a new rate allocation (using strategies in Section III-A) by assuming that the distributions are perfectly estimated. It computes this rate allocation, but does not send it to the sensors. Allocating rates based on this new rate vector would make the rate vector very sensitive to noise. So instead of sending this rate vector, it sends back a uniform average of the new "suggested" rate allocation and previous rate allocations (this average — and not the "suggested" rate allocation — is the actual rate allocation sent to the sensors). In this manner, the algorithm stabilizes itself while still modifying rate allocation to assign higher rates to more relevant sensors by obtaining improved resolution observations of these sensors. The process is detailed in Algorithm 1. Note that each learning step in the algorithm consists of multiple observations at each sensor (depending on the blocklength used by each sensor).

Algorithm 1 also compensates for burst noise: it is possible that due to noise in the data, some of the sensors receive low rates, or are simply turned off, even when their data is relevant to classification. If such relevant sensors are never turned on again, then the learned classifier would have a large variance and could result in overfitting. To address this issue, we introduce a form of regularization on the algorithm using what we call "ambivalence backtracking": the sensors backtrack to average the current rate vector with the ambivalent rate vector (see Algorithm 2 for how this is implemented). The key is to detect when to backtrack, and this can be done by simply computing the average classification error-probability on labeled data on a sliding time-window of one learning step. If the error-probability exceeds a threshold, the algorithm performs ambivalence backtracking. The advantage of stabilization and backtracking is explored further through numerical examples in Section IV. Because the selected rate-allocation is not changed during testing, in this problem with fixed distributions, we only need to backtrack during training (to reduce bias, we do not train on unlabeled data). However, backtracking while testing is a central issue in the case when the underlying distributions can evolve, which we discuss next.

Algorithm 1. Learning For Fixed Distributions On Labeled Data

```
procedure LEARNINGFORFIXEDDISTRIBUTIONS ($L$, $M$, $S$)  $\triangleright$ $L$ = # learning steps, $M$ = # sensors, $S$ = # data points per learning step
    $R \leftarrow$ zero matrix of size $(L, M)$  $\triangleright$ $R$ stores the rate vector of each learning step in one of its rows.
    $Ambivalence \leftarrow \left[\frac{R}{M}, \frac{R}{M}, \ldots, \frac{R}{M}\right]$
    $i \leftarrow 1, i_{prev} \leftarrow 1$
    $R(1,:) \leftarrow \left[\frac{R}{M}, \frac{R}{M}, \ldots, \frac{R}{M}\right]$  $\triangleright$ $R(1,:)$ is set to the ambivalent rate vector.
    for $i = 1, 2, \ldots, L$ do
        Fusion center sends $R(i,:)$ to sensors.
        Sensors transmit data points $\mathbf{X}^{((i-1)S+1)}, \ldots, \mathbf{X}^{(iS)}$ to fusion center quantized according to $R(i,:)$.
        Compute $\widetilde{R}(i+1,:)$  $\triangleright$ Fusion node computes preliminary $\widetilde{R}(i+1,:)$ based on parameter estimates from quantized data in the $i$-th step and using Lemma 1 or Lemma 2.
        $R(i+1,:) \leftarrow \frac{1}{i+2-i_{prev}} \sum_{j=i_{prev}}^{i+1} \widetilde{R}(j,:)$  $\triangleright$ Stabilize rate vector by uniform averaging starting from the previous backtrack.
        Compute error-rate $p_{err}$ during training.  $\triangleright$ Averaged over a sliding window of one learning step.
        if $i > 1$ and $\left(\frac{Metric(i)}{Metric(i-1)} \geq \rho \text{ or } \frac{Metric(i)}{Metric(i-1)} \leq \frac{1}{\rho}\right)$ then  $\triangleright$ Condition for backtracking.
            $R(i+1,:) \leftarrow \frac{1}{2}\left(R(i+1,:) + Ambivalence\right)$  $\triangleright$ Do ambivalence-backtracking.
            $i_{prev} \leftarrow i$
```

2) Energy-Constrained Learning Algorithm for Unknown and Evolving Distributions: Training can be performed as in Section III-B1, and the challenge lies in an unsupervised detection of change in the underlying parameters during testing. Once the change is detected, the data-fusion center can suggest a retraining to the control agent and the sensors, triggering

Algorithm 2. Learning With Backtracking For Evolving Distributions

```
procedure LearnRateVectorForEvolvingDistribution($L, M, S$)   ▷ $L$ = # learning steps, $M$ = # sensors, $S$ = # data points per learning step
    $R \leftarrow$ zero matrix of size $(L, M)$   ▷ $R$ stores the rate vector of each learning step in one of its rows.
    $Metric \leftarrow$ zero vector of size $L$   ▷ Initialize $Metric$ to zero.
    $Ambivalence \leftarrow \left[\frac{R}{M}, \frac{R}{M}, \ldots, \frac{R}{M}\right]$
    $i \leftarrow 1, i_{prev} \leftarrow 1$   ▷ $i_{prev}$ is the index of previous backtracking learning step.
    $R(1,:) \leftarrow \left[\frac{R}{M}, \frac{R}{M}, \ldots, \frac{R}{M}\right]$   ▷ $R(1,:)$ is initialized to the $Ambivalence$ vector.
    for $i = 1, 2, \ldots, L$ do   ▷ $i$ is the index of learning step.
        Fusion center sends $R(i,:)$ to sensors
        Sensors transmit data points $\mathbf{X}^{((i-1)S+1)}, \ldots, \mathbf{X}^{(iS)}$ to fusion center quantized according to $R(i,:)$
        Compute $\mathbf{w}$   ▷ Compute a new $\mathbf{w}$ based on newly received data points.
        Compute $\widetilde{R}(i+1,:)$   ▷ Fusion node computes preliminary $\widetilde{R}(i+1,:)$ based on parameter estimates from quantized data in the $i$-th step and using Lemma 1 or Lemma 2.
        $R(i+1,:) \leftarrow \frac{1}{i+2-i_{prev}} \sum_{j=i_{prev}}^{i+1} \widetilde{R}(j,:)$   ▷ Stabilize rate vector by uniform averaging starting from the previous backtrack.
        $Metric(i) \leftarrow \sum_{t=(i-1)S+1}^{iS} (\mathbf{w}^T \mathbf{X}^{(t)}$ - $\nu$   ▷ $t$ is the index of a data point during testing.
        if $i > 1$ and $\left(\frac{Metric(i)}{Metric(i-1)} \geq \rho\right.$ or $\left.\frac{Metric(i)}{Metric(i-1)} \leq \frac{1}{\rho}\right)$ then   ▷ Condition for change detection.
            $R(i+1,:) \leftarrow \frac{1}{2}\left(R(i+1,:) + Ambivalence\right)$   ▷ Do ambivalence-backtracking.
            $i_{prev} \leftarrow i$
```

a supervised training stage of the algorithm where the fusion center is supplied with labeled data. For simplicity of detection, and to reduce the amount of retraining, our algorithm requires retraining only when estimated classification error-probability becomes large. This error-probability is estimated using the distance $|\mathbf{w}^T\mathbf{X} - \nu|$ as a proxy. Intuitively, when this distance becomes small for successive periods of time (or has a large empirical variance with a small empirical mean), retraining is needed. The resulting algorithm is provided in Algorithm 2.

Numerical results that illustrate the utility of the stabilization and backtracking strategies proposed here are provided in Section IV-B.

IV. Numerical Results

In this section we numerically explore the performance of our algorithms for energy-constrained supervised learning and classification. The code for these numerical experiments is available online at [34].

A. Numerical Results on Energy-Constrained Classification When Underlying Parameters Have Been Perfectly Learned

In the following examples, while using the rate allocation strategies proposed in Lemma 1 and Lemma 2, we do not use long blocklengths (unlike in some of our results in Section III-A). In fact, we apply a scalar version of the strategy to investigate savings in energy. This provides our numerical results a greater degree of practicality[7]. Consequently, we round the rate allocated to each sensor to the nearest integer in order to obtain the corresponding number of bits for scalar quantization. The quantization points are chosen using the optimal Lloyd's algorithm [35].

In order to illustrate how inference-oriented communication can help reduce energy, we compare the classification results using two different rate allocation strategies: "inference-oriented" (using Lemma 1 or Lemma 2, depending on the energy model) and "classical" (based on classical rate-distortion theory). By classical rate allocation strategy, we mean allocating bits to represent the sources themselves as well as possible within the energy constraint, and not exploiting the fact that a weighted function $\mathbf{w}^T\mathbf{X}$ needs to be represented. Thus, features with higher (unconditional) variance receive the largest number of bits. Thus, this is a close analog of the Principal Component Analysis (PCA), and can be viewed as a soft PCA. Therefore, in the rest of the paper, we refer to this strategy as PCA-type rate allocation strategy. For instance, for the Linear Model, the PCA-type rate allocation strategy is given as below:

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

$$R_i = \frac{1}{2}\log\left(\frac{\sigma_i^2 + \mu_i^2}{D_i}\right), \tag{20}$$

where λ is chosen such that $\sum_{i=1}^{M} R_i = E_{total}$, and is easily obtained from Lemma 1. Similarly, the PCA-type rate allocation for the Exponential Model can be obtained from Lemma 2.

Simulated data: In the first example, we consider Gaussian 100-dimensional data points generated independently as follows: under the first class, the conditional mean vector is $\boldsymbol{\mu}_1 = [-100, -99, -98, \ldots, -1]$ and under the second class, the conditional mean vector is $\boldsymbol{\mu}_2 = -\boldsymbol{\mu}_1$. The conditional standard deviation vector under both classes is equal to $\boldsymbol{\sigma} = [100, 200, 300, \ldots, 10000]$ and the features are (conditioned on the class) mutually independent. Qualitatively, the relevance of the features gradually decreases as the index of the feature increases. Fig. 1 provides the tradeoff between classification accuracy and total energy (total number of bits) in the Linear Model using both inference-oriented (Lemma 1) and PCA-type rate allocation strategies. Similarly, Fig. 2 provides this tradeoff for the Exponential Model (Lemma 2). Note that in both cases, the number of bits allocated to each feature is rounded to the nearest integer.

[7] This scalar assumption also provides a better understanding of savings in number of bits of ADCs used in measurement process itself.

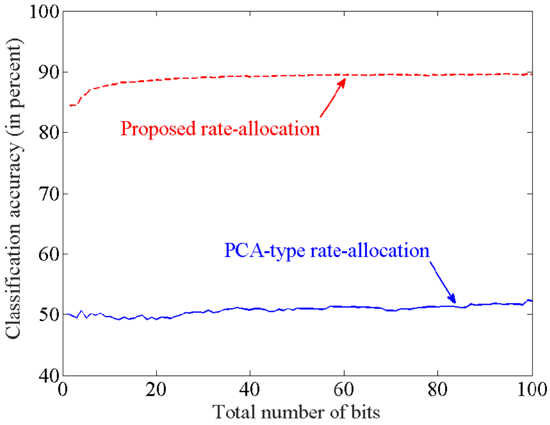

Fig. 1. Binary classification accuracy versus total number of bits achieved by "inference-oriented" strategy under Linear Model (Lemma 1; energy and number of bits are equivalent; red) and "PCA-type" (blue) rate allocation for simulated data of the first setup where $\mu_i = 100 - (i-1)$ and $\sigma_i = 100i, i = 1, \ldots, 100$ and the features are mutually (conditionally) independent. The number of bits is spread across 100 Gaussian features with decreasing relevance. The classification accuracy using unquantized features is about 89.6%. However, we can achieve almost the same accuracy using quantized features with just 25 total bits using our proposed rate allocation strategy.

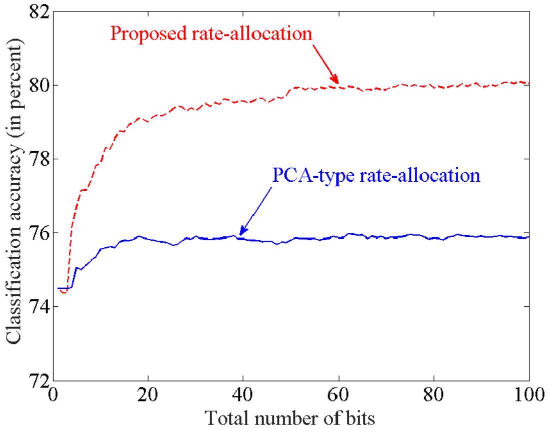

Fig. 3. Binary classification accuracy versus total number of bits achieved by "inference-oriented" strategy under Linear Model (Lemma 1; energy and number of bits are equivalent; red) and "PCA-type" (blue) rate allocation for simulated data of the second setup where $\mu_i = 100/i, \sigma_i = 150, i = 1, \ldots, 100$, and the features are mutually (conditionally) independent. The number of bits is spread across 100 Gaussian features with decreasing relevance. The classification accuracy using unquantized features is about 80.1%. However, we can achieve close-to-optimal accuracy using quantized features with just 35 total bits.

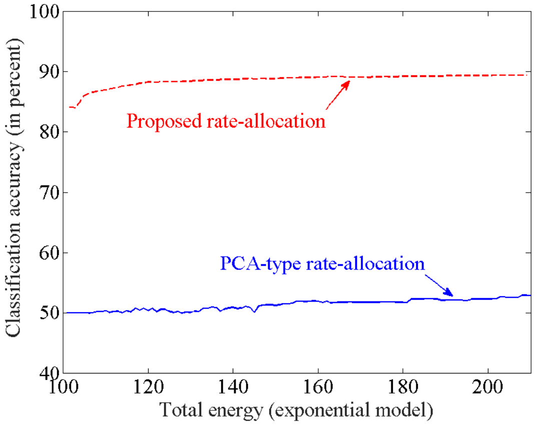

Fig. 2. Binary classification accuracy versus total energy achieved by "inference-oriented" strategy under Exponential Model (Lemma 2; red) and "PCA-type" (blue) rate allocation for simulated data of the first setup where $\mu_i = 100 - (i-1)$ and $\sigma_i = 100i, i = 1, \ldots, 100$ and the features are mutually (conditionally) independent. The number of bits is spread across 100 Gaussian features with decreasing relevance. The classification accuracy using unquantized features is about 89.6%. Again, the proposed rate-allocation outperforms the classical PCA-type allocation substantially.

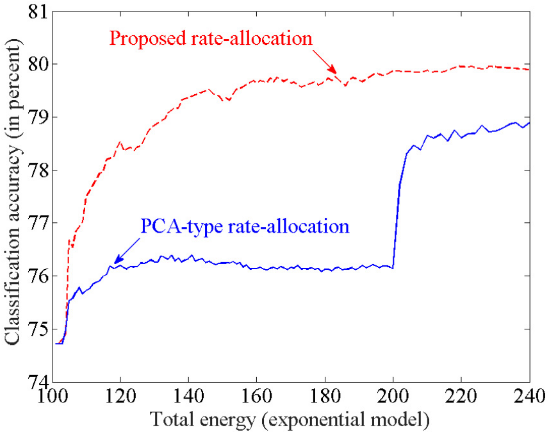

Fig. 4. Binary classification accuracy versus total energy achieved by "inference-oriented" strategy under Exponential Model (Lemma 2; red) and "PCA-type" (blue) rate allocation for simulated data of the second setup where $\mu_i = 100/i, \sigma_i = 150, i = 1, \ldots, 100$ and the features are mutually (conditionally) independent. The number of bits is spread across 100 Gaussian features with decreasing relevance. The classification accuracy using unquantized features of 80.1% can be approached with using much less energy in comparison with PCA-type rate allocation.

As a second example, we consider Gaussian 100-dimensional data points generated independently as follows: under the first class, the conditional mean vector is $\boldsymbol{\mu}_1 = [-100, -\frac{100}{2}, -\frac{100}{3}, \ldots, -1]$ and under the second class, the conditional mean vector is $\boldsymbol{\mu}_2 = -\boldsymbol{\mu}_1$. The common standard deviation under both classes is equal to 150 and the sensor observations are (conditioned on the class) mutually independent. Fig. 3 and 4 provide the tradeoff between classification accuracy and total number of bits/energy allocated to features using both inference-oriented (Lemmas 1 and 2) and PCA-type rate allocation strategies for the Linear and the Exponential Model.

For both examples and in both models, our inference-oriented strategies significantly outperform the PCA-type strategies. By judiciously allocating more bits to more relevant features, the classification accuracy quickly approaches the energy-unconstrained case. Fig. 4 shows an interesting aspect: there is a jump in accuracy in allocating the 101-st bit for the PCA-type allocation (100 bits correspond to an energy of 200 units). As it turns out, the first 100 bits are allocated equally among features in this model. The next (101-st) bit is allocated to the most relevant (first) feature, which brings a substantial improvement. In fact, there is a small dip in accuracy as energy is increased from about 130 units to 200 units, which we believe

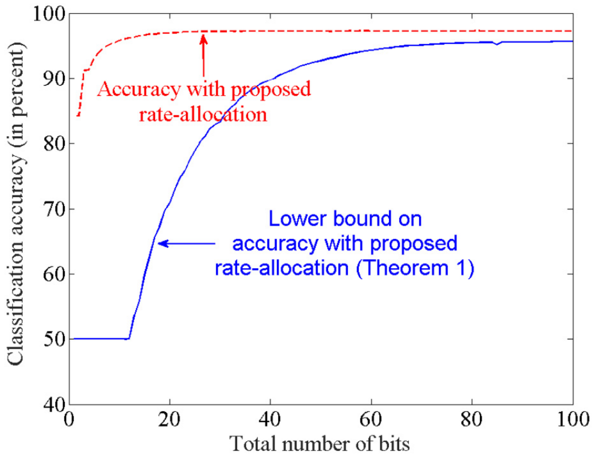

Fig. 5. Binary classification accuracy versus total number of bits achieved by "inference-oriented" strategy (Lemma 1; red) and the corresponding lower bound (blue), which is 1 − (upper bound) in Theorem 1, in the Linear Model for simulated data where $\mu_i = 110 - 10i$ and $\sigma_i = 100, i = 1, \ldots, 10$, and the features are mutually (conditionally) independent. The total number of bits is spread across 10 Gaussian features with decreasing relevance. Notice that classification accuracy using unquantized features is about 97.2%. Note that if the classification accuracy is less than 50% for a decision rule, we can rely on a fair coin toss instead. Thus, whenever the lower bound on the classification accuracy is less than 50%, it is replaced with 50%.

happens due to overfitting, and could be eliminated by proper regularization.

We noted earlier that we expect the bound in Theorem 1 to be loose. The theorem provides an upper bound on the achievable classification error-probability, or equivalently, a lower bound on the achievable classification accuracy. As an illustration, Fig. 5 compares the bound with numerically computed accuracy for a setup with 10-dimensional Gaussian data points generated independently as follows: under the first class, the conditional mean vector is $\boldsymbol{\mu}_1 = [-100, -90, \ldots, -10]$ and under the second class, the conditional mean vector is $\boldsymbol{\mu}_2 = -\boldsymbol{\mu}_1$. The common standard deviation under both classes is equal to 100, and the features are mutually conditionally independent (conditioned on the class). Qualitatively, the relevance of the features decreases as the index of the feature increases. The gap between numerical estimates using simulations and our bound is fairly large. Nevertheless, both the bound and the numerical classification accuracy approach the rate-unconstrained classification accuracy as the total rate increases to infinity (as noted in Remark 1).

Numerical results on recorded ECoG neural data

1) 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 300 ms long and sampled at 1.2 kHz frequency, resulting in 361 samples per trial. Given a trial, we are interested in decoding the movement direction.

2) Data Preprocessing: We use discrete cosine transform (DCT) as proposed in [36] 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. To make the problem consistent with our problem formulation, we view these 3872 features as different sensors, and hence quantize them according to their relevance. This view, and our models, are more applicable when computation of features consumes significant power, as is discussed in a noisy computing problem at the end of Section V.

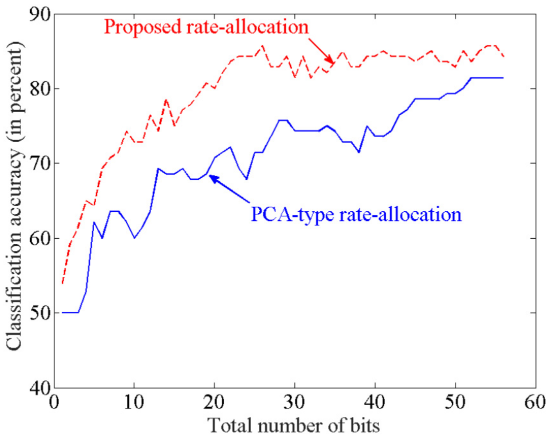

Fig. 6. Binary classification accuracy versus number of bits achieved by "inference-oriented" (red) and "PCA-type" (blue) rate allocation for neural data. The available bits are spread across 3872 features for neural data of binary movement decoding. Notice that an accuracy close to 84% is obtained with just 40 total bits. If relevance of features is disregarded, and the available bits are distributed uniformly across all features with one bit for each feature, it would still require 3872 bits, with barely any accuracy.

3) Evaluating Classification Accuracy: First, as a form of regularization, only the important features with Fisher scores above a threshold (in this case 0.25) are kept and the other features are removed. Then, in order to calculate the classification accuracy, we perform a 5-fold cross-validation: we partition the data points into 5 folds randomly where every fold consists of 14 trials of each class. For each fold, we train the LDA algorithm on the other 4 folds to obtain the vector **w** as in (2), and use the remaining fold for validation. Post-training, 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, and then performing the LDA algorithm on these quantized DCT features. Finally, we compute the average classification accuracy using the data labels.

The result, illustrated in Fig. 6, provides the tradeoff between classification accuracy and total number of bits allocated to features for both inference-oriented and PCA-type rate allocation under the assumption that the DCT values for different frequencies are independent. Even under this inaccurate assumption, the resulting classifier works with about 84% accuracy with just 40 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 (130x more energy) to be transmitted from the sensors, with barely any classification accuracy. Also, it can be seen from Fig. 6 that the inference-oriented rate allocation strategy outperforms the PCA-type strategy.

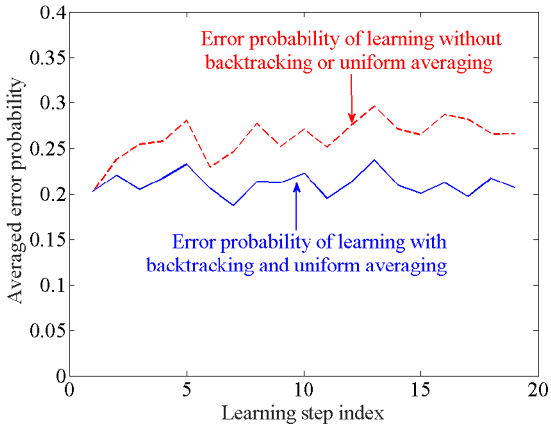

Fig. 7. Classification error-probability for fixed distribution of sensor-data with and without ambivalence-backtracking and uniform averaging. Each learning step consists of $S = 10$ data points. The setup is as follows: $\mu_i = \frac{100}{i}$ and $\sigma_i = 150$, $i = 1, 2, \ldots, 10$, and $E_{total} = 30$. Learning with ambivalence-backtracking and uniform averaging detects large fluctuations in the Metric and backtracks to a more ambivalent state.

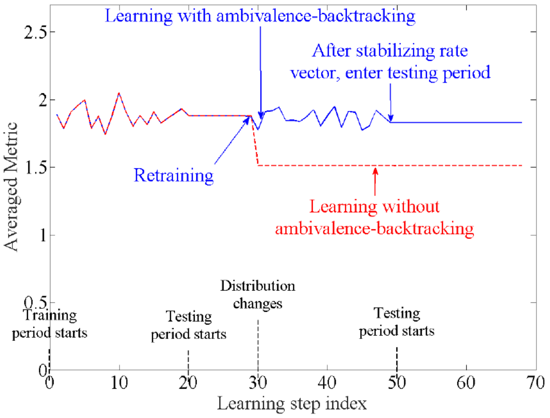

Fig. 8. Metric $\sum_{t=(i-1)S+1}^{iS}(\mathbf{w}^T\mathbf{X}^{(t)} - \nu)$ for evolving distributions. There are 100 data points for each class at each learning step. Total rate is 50. Means of data distribution change from $\mu_i = \frac{100}{i}$ to $\mu_i = \frac{100}{11-i}$ while the common standard deviation remains unchanged ($\sigma_i = 150$). Each training period consists of up to 20 learning steps, and training and testing periods are performed alternatively. Distribution changes at step 30 (10-th step of the testing period).

B. *Energy-Constrained Learning for Classification*

We limit our attention to the Linear Model of energy consumption for succinctness.

Parameters fixed but unknown: We first explore learning for fixed but unknown distribution of sensor-data. We consider 10-dimensional Gaussian data points generated independently as follows: under the first class, the conditional mean vector is $\mu_1 = [-100, -\frac{100}{2}, -\frac{100}{3}, \ldots, -10]$ and under the second class, the conditional mean vector is $\mu_2 = -\mu_1$. The common standard deviation under both classes is equal to 150 and the features are conditionally mutually independent. Qualitatively, the relevance of the features decreases as the index of the feature increases. We also consider Linear Model of energy consumption with total available energy (number of bits) $E_{total} = 30$.

When only a small number of data points is available, it is much more important to stabilize and allow for backtracking for regularization because the learning process becomes very sensitive to noise or large-deviation events. In Fig. 7, we consider 5 data points of each of the two classes in each learning step (and thus $S = 10$ data points total in each step), and a total of $L = 20$ learning steps with and without uniform averaging and backtracking. Choosing such a small value of S brings out the effects of variability. In Fig. 7, we plot the error-probability with ambivalence-backtracking and uniform averaging, and without them. When no ambivalence-backtracking or uniform averaging is used, the error-probability appears to get stuck in a local maximum because it turns off relevant sensors due to noise, quantization, and small number of data points.

Evolving parameters: Considering total available energy (number of bits) per sample as $E_{total} = 50$, we now explore an example where the distribution parameters are evolving over time. The simulation has a training period, followed by a testing period. In the training period, our algorithm learns a rate vector adaptively as provided in Algorithm 2. During the ensuing testing period, the classification weight $\mathbf{w}$, which is computed in training period, is kept constant. Each training period has $L = 20$ learning steps and 100 data points of each class are included in each learning step (a larger number of data points is allowed because we are more interested in examining how the algorithm performs when the underlying distribution changes). The focus here is on backtracking: it is assumed that when the fusion center decides to backtrack, it also asks the system-user as well as the circuit for a retraining so that it is supplied with labeled data for another training period. In a BCI, for e.g., this retraining request could be sent to the BCI user who could in response supply labeled data to retrain the system.

We consider the case when the distribution of sensor-data changes during the testing period. The distribution is assumed to remain unchanged during the training period, and change at the 10-th step during the testing period (after $L = 20$ steps of learning, a total of 30 steps). Specifically, the means of the distribution change from $\mu_i = 100 - 10i$ to $\mu_i = 10i$ (while the common standard deviation remains the same), reversing the relative-relevance of the sensors' data. Ambivalence-backtracking could be of help here: if sensors with low relevance initially have low resolution or are turned off, the only way to restore them is to have all sensors be turned on again.

Fig. 8 shows the normalized *Metric* $\sum_{t=(i-1)S+1}^{iS}(\mathbf{w}^T\mathbf{X}^{(t)} - \nu)$ for this example with ambivalence-backtracking and without. During the training period and first half of the testing period, when the distribution of data does not change, the performance with and without ambivalence-backtracking is similar. When the distribution changes (at step 30), initially the normalized *Metric* drops sharply in both cases. However, learning with ambivalence-backtracking asks for retraining in the very next testing step, and thus is able to update its rate vector through training on newly acquired labeled data. Learning without ambivalence-backtracking does not backtrack and thus cannot update rate vector for new distribution. Fig. 9 shows the corresponding error-probability for classification with and without ambivalence-backtracking.

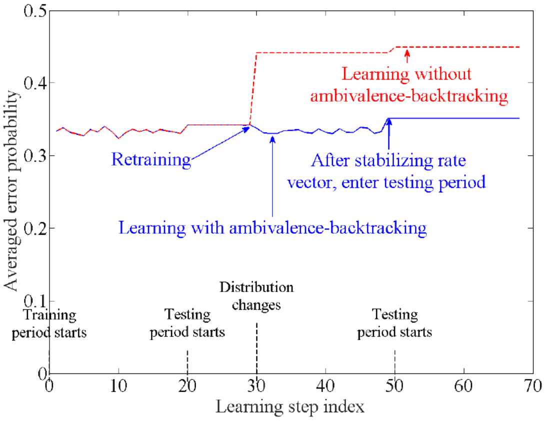

Fig. 9. Classification error-probability for the setup in Fig. 8. The algorithm with ambivalence-backtracking is able to retrain itself to reduce error-probability despite the distribution change at learning step 30. The testing period starts at "learning" step 20, and at that point, the rate-allocation is stabilized until detection of distribution change (which happens at learning step 31).

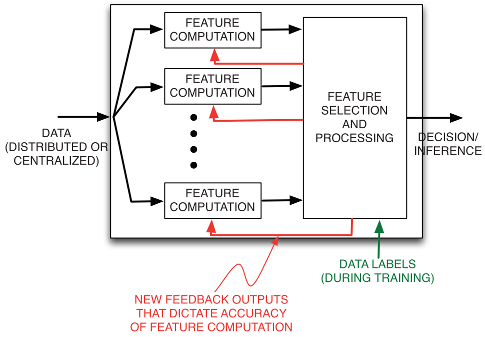

Fig. 10. Feature computation can be noisy, with noise levels adjusted by changing energy-level for each feature for a given total energy constraint.

V. Discussions and Conclusions

One of the strongest assumptions we make in the paper is that the acquired data is iid Gaussian conditioned on the class. How realistic is this assumption? The Gaussian aspect of the assumption, while not accurate [28], is a well known and widely used approximation for local field potentials (LFPs), Electroencephalography (EEG), and ECoG data used here. This assumption is thought to be a reasonable approximation because these recordings are measuring sums of smaller activations (e.g. nearby dendritic activity), suggesting that the central limit theorem is in action. What about the (conditional) independence assumption in our model? For ECoG data considered here, this assumption has often been used in neural decoding algorithms (see, e.g. [36] and references therein). In comparison, EEG signals recorded on the scalp (noninvasively) have significantly higher spatial correlations, especially in nearby electrodes [2]. This is because distance from the brain to the scalp (where the EEG sensors are located) acts as a spatial low-pass filter (as discussed in [2], [37]), introducing correlations in EEG signals recorded at nearby electrodes that are generated by uncorrelated brain sources. In contrast, ECoG signals are recorded on the brain surface and are influenced largely by activity close to the electrodes. Thus, the results are more directly applicable to ECoG data. However, even for EEG signals, our recent work has proposed differential-sensing strategies that – as a by-product of reducing sensing energy – reduce correlations in EEG signals (using a "hierarchical referencing" strategy) [2]. Lastly, we view these results as a step towards a broader understanding of energy-constrained distributed inference. Generalizing from this simplistic problem, it would be of interest to extend these results to cases with conditionally correlated measurements and multiclass classification.

Part of the novelty in this work lies in the assumption that the sensors are energy-limited, adaptive, and can not do sophisticated processing allowed in the information-theoretic works (e.g. [12]). To obtain fundamental limits on the problem, we need to bring in models of energy consumption in circuits as constraints on the problem (e.g., [2], [38], [39]), which would also bring our results closer to practice.

Our analysis on ECoG data also brings out a disconnect between our model and a technique used commonly (e.g. in [36]) for classifying ECoG data that relies on first taking a Fourier transform of the data, and then viewing different frequency components of different sensors as separate features. Since the features are not directly the sensor observations, but obtained from frequency analysis of time-series at each sensor, the same sensor can generate multiple features. We believe our results can extend to cover this case: intuitively, the more relevant features a sensor has, the more bits it should receive. Similarly, the results could extend to the case where data from multiple sensors is used to obtain a feature. Our future work will consider these problems as well.

The techniques and results obtained here could also be applied to the problem of computation in presence of noise (see Fig. 10) in the following fashion: consider the case where the data is available in a centralized fashion but different features need to be computed on the same data stream. This is often the case in neural data analysis and other applications. One can imagine that different features can be computed with different accuracy/noise levels, in particular using analog feature-computation and dot-product computation systems that are being built recently using emerging devices [40]. In these systems, the accuracy of computation can be controlled by energy expenditure on the fly: the larger the amount of energy spent, the smaller is the distortion in computation. Our algorithms, appropriately adapted to the energy models, can thus be used to compute more relevant features with higher accuracy. This is an exciting research direction that deserves further attention.

Appendix A
Proof of Lemma 1

We use a dithered-lattice codebook with optimal lattice quantizers [31] to ensure that the error in estimating X_i is independent of the source X_i, is zero mean [32], and is asymptotically Gaussian distributed. We note that the error is shown to be asymptotically Gaussian and white *regardless of the source distribution* (*i.e.*, the source need not be Gaussian). Thus the results of [31] are applicable in our problem where the source is mixture Gaussian.

We first derive an upper bound on the target function in (5) as follows:

$$\mathbb{E}\left[\left(\mathbf{w}^T\mathbf{X}^{(t)} - \mathbf{w}^T\widehat{\mathbf{X}}^{(t)}\right)^2\right] \overset{(a)}{=} \sum_{i=1}^{M} w_i^2 \mathbb{E}\left[\left(X_i^{(t)} - \widehat{X}_i^{(t)}\right)^2\right], \tag{21}$$

where (a) requires the errors in the two reconstructions to be uncorrelated and zero mean. The errors are uncorrelated because of independence of errors and source in subtractive dithered lattice quantization [32]. Thus, even though the sources themselves are correlated (unconditionally), the errors are not. Because errors in (subtractive dithered) lattice-quantization are distributed uniformly in a Voronoi region [32], they are zero mean as well. Using $D_i^{(t)} := \mathbb{E}\left[\left(\mathbf{w}^T\mathbf{X}^{(t)} - \mathbf{w}^T\widehat{\mathbf{X}}^{(t)}\right)^2\right]$, (5) becomes:

$$\min_{\mathcal{E}_i, \mathcal{D}_i, R_i} \quad \frac{1}{n}\sum_{t=1}^{n} D_i^{(t)} \text{ subject to } \sum_{i=1}^{M} R_i \leq E_{total}. \tag{22}$$

Allowing codebook to be chosen by randomly permuting its indices (to make it an equidistortion codebook), $D_i^{(t)}$, is the same for all t. Thus, we can drop superscript (t) and (22) can be simplified to

$$\min_{\mathcal{E}_i, \mathcal{D}_i, R_i} \quad \sum_{t=1}^{n} D_i \text{ subject to } \sum_{i=1}^{M} R_i \leq E_{total}. \tag{23}$$

For a given second moment, the Gaussian source has the largest distortion-rate function [41, Ch. 10], which provides a convenient (if loose) way to upper bound the distortion-rate function for our *mixture* Gaussian source. For a 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:

$$\mathbb{E}\left[X_i^2\right] = \sum_{c=1}^{2} p(C=c)\mathbb{E}\left[X_i^2|C=c\right] = \sigma_i^2 + \mu_i^2, \tag{24}$$

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

$$D_i \leq w_i^2 Var[X_i] 2^{-2R_i} = w_i^2\left(\sigma_i^2 + \mu_i^2\right) 2^{-2R_i}. \tag{25}$$

Further, this can be achieved using the lattice-quantization strategy proposed here, and the quantization noise will asymptotically be Gaussian distributed [31]. Thus, it remains to solve the following optimization problem:

$$\min_{R_i} \sum_{i=1}^{M} w_i^2(\sigma_i^2 + \mu_i^2) 2^{-2R_i} \text{ s.t.} \sum_{i=1}^{M} R_i = E_{total}, \tag{26}$$

which is precisely the source coding problem for parallel Gaussian sources considered, e.g., in [41, Ch. 10], which has the reverse-waterfilling solution given in Lemma 1.

Appendix B
Proof of Lemma 2

Following steps in the proof of Lemma 1, approximating the problem of minimizing the classification error-probability to the problem of minimizing the mean-squared error in reconstructing the decision variable $\mathbf{w}^T\mathbf{X}$, the optimization problem is:

$$\min_{R_i} \sum_{i=1}^{M} w_i^2(\sigma_i^2 + \mu_i^2) 2^{-2R_i} \; s.t. \sum_{i=1}^{M} 2^{R_i} \leq E_{total}, \, R_i \geq 0, \tag{27}$$

Denoting 2^{R_i} by x_i, the problem can be transformed to:

$$\min_{x_i} \sum_{i=1}^{M} \alpha_i x_i^{-2} \; s.t. \sum_{i=1}^{M} x_i \leq E_{total}, \, x_i \geq 1 \tag{28}$$

where $\alpha_i = \Lambda_i^2 = w_i^2(\sigma_i^2 + \mu_i^2)$, and $x_i = 2^{R_i}$. This is a convex optimization problem, and can be solved using Lagrange multipliers. This results in $x_i = \left(\frac{2\alpha_i}{\lambda - \theta_i}\right)^{\frac{1}{3}}$. Substituting $x_i = 2^{R_i}$ back, we have $R_i = \frac{1}{3}\log\frac{2\alpha_i}{\lambda - \theta_i}$. From complementary slackness conditions, $\theta_i(x_i - 1) = 0$ for all i. This means that either $x_i = 1$ (*i.e.*, $R_i = 0$), or $\theta_i = 0$. Thus, $R_i = \max\{\frac{1}{3}\log\frac{2\alpha_i}{\lambda}, 0\}$. The value of λ can now be obtained by solving $\sum_{i=1}^{M} 2^{R_i} = E_{total}$ resulting in $D_i = 2^{-\frac{2}{3}}\lambda^{\frac{2}{3}}[w_i^2(\mu_i^2 + \sigma_i^2)]^{\frac{1}{3}}$ when $R_i > 0$.

Acknowledgements

We thank Dr. Wei Wang from University of Pittsburgh for sharing his valuable experience and measurement data of ECoG-based BCI, and Vinod Prabhakaran, Aarti Singh, and Sanghamitra Dutta for thoughtful discussions.

References

[1] M. Mahzoon, H. Albalawi, X. Li, and P. Grover, "Using relative-relevance of data pieces for efficient communication, with an application to neural data acquisition," in *Proc. Annu. Allerton Conf. Commun. Control Comput.*, Monticello, IL, USA, Sep. 2014, pp. 160–166.

[2] P. Grover, J. Weldon, S. Kelly, P. Venkatesh, and H. Jeong, "A novel information-theoretic sensing strategy for ultra high-density EEG sensing," in *Proc. Allerton Conf. Commun. Control Comput.*, Monticello, IL, USA, Oct. 2015, to be published.

[3] P. Grover, "Ultra-high-density EEG: How many bits of resolution do the electrodes need?," in *Proc. Asilomar Conf. Signals Syst. Comput.*, Nov. 2015, pp. 943–947.

[4] L. Wasserman, *All of Statistics: A Concise Course in Statistical Inference*. New York, NY, USA: Springer, 2013.

[5] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Hoboken, NJ, USA: Wiley, 2006.

[6] T. Hastie, R. Tibshirani, J. Friedman, and J. Franklin, "The elements of statistical learning: Data mining, inference and prediction," *Math. Intell.*, vol. 27, no. 2, pp. 83–85, 2005.

[7] T. N. Lal *et al.*, "Support vector channel selection in BCI," *IEEE Trans. Biomed. Eng.*, vol. 51, no. 6, pp. 1003–1010, Jun. 2004.

[8] Y. Wang, S. Gao, and X. Gao, "Common spatial pattern method for channel selelction in motor imagery based brain-computer interface," in *Proc. IEEE 27th Annu. Int. Conf. Eng. Med. Biol. Soc.*, Jan. 2005, pp. 5392–5395.

[9] T. Lan, D. Erdogmus, A. Adami, M. Pavel, and S. Mathan, "Salient EEG channel selection in brain computer interfaces by mutual information maximization," in *Proc. IEEE 27th Annu. Int. Conf. Eng. Med. Biol. Soc.*, Jan. 2005, pp. 7064–7067.

[10] M. Schroder, M. Bogdan, T. Hinterberger, and N. Birbaumer, "Automated EEG feature selection for brain computer interfaces," in *Proc. IEEE 1st Int. EMBS Conf. Neural Eng.*, Mar. 2003, pp. 626–629.

[11] S. Liu, A. Vempaty, M. Fardad, E. Masazade, and P. K. Varshney, "Energy-aware sensor selection in field reconstruction," *IEEE Signal Process. Lett.*, vol. 21, no. 12, pp. 1476–1480, Dec. 2014.

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

[13] R. Ahlswede and I. Csiszár, "Hypothesis testing with communication constraints," *IEEE Trans. Inf. Theory*, vol. IT-32, no. 4, pp. 533–542, Jul. 1986.

[14] T. Berger, "Decentralized estimation and decision theory," in *Proc. IEEE Seven Springs Workshop Inf. Theory*, Mt. Kisco, NY, USA, 1979.

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

[16] R. Gray and D. Neuhoff, "Quantization," *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2325–2383, Oct. 1998.

[17] P. Panter and W. Dite, "Quantization distortion in pulse-count modulation with nonuniform spacing of levels," *Proc. IRE*, vol. 39, no. 1, pp. 44–48, 1951.

[18] S. Lloyd, "Least squares quantization in PCM," *IEEE Trans. Inf. Theory*, vol. IT-28, no. 2, pp. 129–137, Mar. 1982.

[19] Y. Zhu and J. Lafferty, "Quantized nonparametric estimation," arXiv preprint arXiv:1503.07368, 2015.

[20] R. Nowak and U. Mitra, "Boundary estimation in sensor networks: Theory and methods," in *Information Processing in Sensor Networks*. New York, NY, USA: Springer, 2003, pp. 80–95.

[21] G. Thatte *et al.*, "Knowme: An energy-efficient multimodal body area network for physical activity monitoring," *ACM Trans. Embedded Comput. Syst.*, vol. 11, no. S2, p. 48, 2012.

[22] R. Castro, R. Willett, and R. Nowak, "Faster rates in regression via active learning," University of Wisconsin, Madison, WI, USA, UW-Madison Tech. Rep. ECE-05-3, 2005.

[23] R. Castro and R. Nowak, "Active sensing and learning," *Found. Appl. Sensor Manage.*, 2009, pp. 177–200.

[24] R. Castro, J. Haupt, and R. Nowak, "Compressed sensing vs. active learning," in *Proc. Int. Conf. Acoust. Speech Signal Process.*, Toulouse, France, vol. 3, pp. 820–823, 2006.

[25] B. Eriksson, G. Dasarathy, A. Singh, and R. Nowak, "Active clustering: Robust and efficient hierarchical clustering using adaptively selected similarities," in *Proc. Int. Conf. Artif. Intell. Stat.*, vol. 15, 2011, pp. 260–268.

[26] J. Haupt, R. Castro, and R. Nowak, "Distilled sensing: Adaptive sampling for sparse detection and estimation," *IEEE Trans. Inf. Theory*, vol. 57, no. 9, pp. 6222–6235, Sep. 2011.

[27] J. D. Haupt, R. G. Baraniuk, R. M. Castro, and R. D. Nowak, "Compressive distilled sensing: Sparse recovery using adaptivity in compressive measurements," in *Proc. 43rd Asilomar Conf. Signals Syst. Comput.*, 2009, pp. 1551–1555.

[28] R. E. Kass, U. T. Eden, and E. N. Brown, *Analysis of Neural Data*. New York, NY, USA: Springer, 2014.

[29] Wikipedia, The Free Encyclopedia. (2016). *Q-Function* [Online]. Available: https://en.wikipedia.org/wiki/Q-function

[30] K. Marton, "Error exponent for source coding with a fidelity criterion," *IEEE Trans. Inf. Theory*, vol. 20, no. 2, pp. 197–199, Mar. 1974.

[31] R. Zamir and M. Feder, "On lattice quantization noise," *IEEE Trans. Inf. Theory*, vol. 42, no. 4, pp. 1152–1159, Jul. 1996.

[32] L. Schuchman, "A proof of the data compression theorem of Slepian and Wolf for ergodic sources (corresp.)," *IEEE Trans. Commun. Technol.*, vol. 12, no. 4, pp. 162–165, Mar. 1964.

[33] W. Wang *et al.*, "Neural interface technology for rehabilitation: Exploiting and promoting neuroplasticity," *Phys. Med. Rehabil. Clin. North Amer.*, vol. 21, no. 1, pp. 157–178, 2010.

[34] C. Yuan Li, M. Mahzoon, and P. Grover, *Code for 'rate and energy-constrained learning'*, 2016 [Online]. Available: http://www.tinyurl.com/pulkitgrover/RateConstrainedLearning.html

[35] S. P. Lloyd, "Least squares quantization in PCM," *IEEE Trans. Inf. Theory*, vol. IT-28, no. 2, pp. 129–137, Mar. 1982.

[36] 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," in *Proc. IEEE 36th Annu. Int. Conf. Eng. Med. Biol. Soc. (EMBC)*, 2014, pp. 1626–1629.

[37] P. L. Nunez and R. Srinivasan, *Electric Fields of the Brain: The Neurophysics of EEG*. London, U.K.: Oxford Univ. Press, 2006.

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

[39] P. Grover, K. Woyach, and A. Sahai, "Towards a communication-theoretic understanding of system-level power consumption," *IEEE J. Sel. Areas Commun.*, vol. 29, no. 8, pp. 1744–1755, Sep. 2011.

[40] I. Nahlus, E. P. Kim, N. R. Shanbhag, and D. Blaauw, "Energy-efficient dot product computation using a switched analog circuit architecture," in *Proc. Int. Symp. Low Power Electron. Des.*, 2014, pp. 315–318.

[41] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 1st ed. Hoboken, NJ, USA: Wiley, 1991.

Majid Mahzoon received the B.Sc. degree in electrical engineering from Sharif University of Technology, Tehran, Iran, in 2013. He is currently pursuing the Ph.D. degree at Carnegie Mellon University, Pittsburgh, PA, USA. His research interests include wireless communications, information theory, and machine learning.

Christy (Yuan) Li received the B.Sc. degree in information engineering from the Chinese University of Hong Kong, Hong Kong. She is currently pursuing the M.Sc. degree at Carnegie Mellon University, Pittsburgh, PA, USA. She has also worked with Applied Science and Technology Research Institution, Hong Kong, developing algorithms for converting 2-D images to 3-D representations, and with iAspec Hong Kong, as a Software Engineer. Her research interests include machine learning in sensing and with large datasets, with emphasis on energy-efficient machine learning.

Xin Li received the Ph.D. degree in electrical and computer engineering from Carnegie Mellon University, Pittsburgh, PA, USA, in 2005. He is currently an Associate Professor with the Department of Electrical and Computer Engineering, Carnegie Mellon University. His research interests include integrated circuit, signal processing, and data analytics. He was an Associate Editor of the IEEE TBME, TCAD, ACM TODAES, IEEE D&T, and JOLPE. He was the General Chair of ISVLSI and FAC. He was the recipient of the NSF CAREER Award in 2012 and five Best Paper Awards from the IEEE TCAD, DAC, ICCAD, and ISIC.

Pulkit Grover is an Assistant Professor with CMU, working on information theory, circuit design, and biomedical engineering. His research interests include developing a new theory of information for low-energy communication, sensing, and computing by incorporating novel circuit/processing-energy models to add to classical communication or sensing energy models. A common theme in his work is observing when optimal designs depart radically from classical theoretical intuition. He was the recipient of the 2010 Best Student Paper Award at the IEEE Conference on Decision and Control; the 2011 Eli Jury Dissertation Award from UC Berkeley; the 2012 Leonard G. Abraham Award from the IEEE Communications Society; the 2014 Best Paper Award at the International Symposium on Integrated Circuits; the 2014 NSF CAREER Award; and the 2015 Google Research Award.