Sparse Sensing for Distributed Detection

Sundeep Prabhakar Chepuri, *Member, IEEE*, and Geert Leus, *Fellow, IEEE*

***Abstract*—An offline sampling design problem for distributed detection is considered in this paper. To reduce the sensing, storage, transmission, and processing costs, the natural choice for the sampler is the sparsest one that results in a desired global error probability. Since the numerical optimization of the error probabilities is difficult, we adopt simpler costs related to distance measures between the conditional distributions of the sensor observations. We design sparse samplers for the Bayesian as well as the Neyman-Pearson setting. The developed theory can be applied to sensor placement/selection, sample selection, and fully decentralized data compression. For conditionally independent observations, we give an explicit solution, which is optimal in terms of the error exponents. More specifically, the best subset of sensors is the one with the smallest local average root-likelihood ratio and largest local average log-likelihood ratio in the Bayesian and Neyman-Pearson setting, respectively. We supplement the proposed framework with a thorough analysis for Gaussian observations, including the case when the sensors are conditionally dependent, and also provide examples for other observation distributions. One of the results shows that, for nonidentical Gaussian sensor observations with uncommon means and common covariances under both hypotheses, the number of sensors required to achieve a desired detection performance reduces significantly as the sensors become more coherent.**

***Index Terms*—Bhattacharyya distance, convex optimization, distributed detection, energy-efficiency, J-divergence, Kullback-Leibler distance, sensor placement, sensor selection, sparse sensing.**

I. Introduction

SENSORS are widely used in a variety of applications such as environmental monitoring, surveillance, social networks, and power networks, to list a few. The datasets generated by the sensors have to be optimally processed to extract relevant information, which typically involves solving a statistical inference problem like detection, estimation, and filtering, for instance. Often, the data generated by the ubiquitous sensors are excessively large. Thus, it is of paramount importance to parsimoniously acquire the data keeping in mind the inference problem to be solved.

The number of sensors available is often limited either by economical constraints including hardware costs, the availability of physical or data storage space, or restrictions on the available energy and other resources. In order to reduce these costs, the sensors have to be frugally deployed and processed. Naturally, this limits the performance and results in a trade-off between the performance, number of sensors, and sensing pattern. In this paper, we are interested in studying this trade-off, and choose only informative sensors (or sampling locations) that offer a tolerable detection fidelity, that is, we study sensor (or sample) selection or placement in the context of distributed detection. More generally, we design a sparse compression matrix to perform sensor selection, sensor placement, or data compression, hence the name *sparse sensing*. We now describe a scenario to illustrate the considered sensing problem.

Example (Field Detection): Consider a multi-core processor with a hot spot. A historical question of interest is to estimate the thermal distribution, for instance, by interpolating noisy measurements. In some applications though, a precise estimation of the temperature field might not be required, instead, detecting the hot spots (i.e., the areas where the temperature exceeds a certain threshold) would be sufficient for subsequent control actions. An important question of interest for such detection problems then is, how to design spatial samplers (i.e., sensor placement) by exploiting the knowledge of the underlying model, physical space and processing limitations.

In this paper, the focus is on distributed detection pertinent to applications in sensor networks, radar and sonar systems, wireless cognitive radio networks, biometrics, social networks, imaging platforms, to name a few. We assume that the field is sampled by distributed sensors, and these samples are delivered to a central unit. The central unit then makes a single global decision as to the true hypothesis using binary hypothesis testing. More specifically, the observations at each sensor are related to the state of nature $\mathcal{H}$, where the random variable $\mathcal{H}$ is drawn from a binary alphabet set $\{\mathcal{H}_0, \mathcal{H}_1\}$. In the Bayesian setting, we assume that the prior probabilities $\pi_0 = \Pr(\mathcal{H}_0)$ and $\pi_1 = \Pr(\mathcal{H}_1)$ are known, whereas in the Neyman-Pearson setting, the prior probabilities are not known.

A. Related Earlier Works

The problem of choosing the best subset of sensors that guarantees a desired inference performance is referred to as sensor selection. Sensor selection for estimation and filtering is a well-studied problem [2]–[6] (and references therein), where the best subset of sensors that results in a prescribed estimation accuracy is chosen. For selecting the sensors, a scalar function of the error covariance matrix (more generally, the inverse Fisher information matrix) is optimized to obtain the required information gain.

Manuscript received March 22, 2015; revised August 10, 2015; accepted September 17, 2015. Date of publication October 16, 2015; date of current version February 10, 2016. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Marco Lops. This work was supported in part by STW under FASTCOM project (10551), and in part by NWO-STW under the VICI program (10382). A conference precursor of this manuscript appears in the 40th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), May 2015.

The authors are with the Faculty of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology, Delft 2628CD, The Netherlands (e-mail: s.p.chepuri@tudelft.nl; g.j.t.leus@tudelft.nl).

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/TSP.2015.2491890

The minimum error probability criterion is a standard performance measure for design problems related to statistical detection such as signal design [7]–[9], censoring [10], sampling design [11], and so on. However, in most cases, optimizing the error probabilities is very difficult. This may be because these error probabilities do not admit a known closed form or their expression is too complicated for numerical optimization. Therefore, weaker performance criteria that are easier to evaluate and optimize are often used. A number of measures related to the distance between the conditional probabilities are widely used in the design of experiments as proxies for the error probability [9]–[13]. Some of the prominent distance measures that are often used are the Kullback-Leibler distance, J-divergence, Chernoff information, and Bhattacharyya distance.

A related topic in the context of energy-efficient distributed detection is *data censoring*, wherein the uninformative sensor observations are not transmitted to the central unit [10], [14], [15]. However, in censoring, data still has to be acquired in order to choose informative sensors, thus, it incurs a sensing cost. That is, censoring schemes are data dependent as opposed to the proposed data-independent sparse sensing schemes that can be designed offline. In other words, the actual measurements are not needed and only model information is used.

B. Overview and Main Results

We focus on both the Bayesian as well as the Neyman-Pearson setting for distributed detection. The sparse sensing operation is designed based on a number of distance measures that belong to the general class of Ali-Silvey distances [16].

The main question addressed in the paper is similar to that of [11], [12], [17]–[20], but with the following differences. Firstly, the proposed framework is *general*, that is, it is not limited to Gaussian observations, especially for conditionally independent observations. Secondly, we propose a *sparsity-promoting* cost function to design structured samplers to achieve the lowest sensing cost as compared to the previously adopted periodic, regular, or random samplers. The main contributions of this paper that broaden the existing literature are listed below.

- For conditionally independent observations, the best subset of sensors is the one with the *smallest local average root-likelihood ratio* and *largest local average log-likelihood ratio* in the Bayesian and Neyman-Pearson setting, respectively. This leads to an explicit solution for the sensing design problem that is optimal in terms of the error exponents. As a special case, for Gaussian observations with common covariances and uncommon means under both hypotheses, the selected sensors are also optimal in terms of the error probabilities (initial results for the Gaussian case were reported in [1]). The computational complexity of the proposed solvers is independent of the number of candidate sensors, and is as low as $O(K)$, where K is the number of selected sensors (or sampling locations).
- For conditionally dependent observations, we focus on the Gaussian setting. When the mean vectors are uncommon and the covariance structure is common under both hypotheses, the sensing design problem can be relaxed to a convex optimization problem. Although this leads to a suboptimal solution, we propose a randomized rounding technique that further improves the solution. Moreover, in this case, for non-identical sensor observations, we show that the number of sensors required to achieve a prescribed detection performance decreases significantly as the correlation among them increases (i.e., when the sensors become more coherent), which is in complete contrast to the case of identical sensor observations. When the covariances are uncommon and the means are common under both hypotheses, the sensing design problem remains nonconvex, except for the J-divergence optimization (this also holds for a more general case with uncommon means).

C. Outline and Notation

The remainder of the paper is organized as follows. The sampling design problem is stated in Section II. The performance measures that determine the sparsity order of the samplers are discussed in Section III. The solution to the sparse sensing problem for conditionally independent observations is discussed in Section IV. A few examples to illustrate the proposed framework are provided in Section V. The solution to the sparse sensing problem for conditionally dependent Gaussian observations is discussed in Section VI. Finally, the paper concludes with Section VII.

The notation used in this paper can be described as follows. Upper (lower) bold face letters are used for matrices (column vectors). $(\cdot)^T$ denotes transposition. $\mathrm{diag}(\cdot)$ refers to a diagonal matrix with its argument on the main diagonal. $\mathrm{diag_r}(\cdot)$ represents a diagonal matrix with the argument on its diagonal but with the all-zero rows removed. $\mathbf{1}$ $(\mathbf{0})$ denotes the vector of all ones (zeros). $\boldsymbol{I}$ is an identity matrix. $\mathbb{E}\{\cdot\}$ denotes the expectation operation. $\mathrm{tr}\{\cdot\}$ is the matrix trace operator. $\det\{\cdot\}$ is the matrix determinant. $\lambda_{\max}\{\boldsymbol{A}\}(\lambda_{\min}\{\boldsymbol{A}\})$ denotes the maximum (minimum) eigenvalue of a symmetric matrix $\boldsymbol{A}$. The ℓ_0-(quasi) norm refers to the number of non-zero entries in $\boldsymbol{w}$, i.e., $\|\boldsymbol{w}\|_0 := |\{m : w_m \neq 0\}|$. The ℓ_1-norm of an $N \times 1$ vector $\boldsymbol{w}$ is denoted by $\|\boldsymbol{w}\|_1 = \sum_{n=1}^{N} |w_n|$. The notation $\sim$ is read as "is distributed according to". Unless and otherwise noted, logarithms are natural.

II. Problem Statement

Consider a network with M candidate sensors. These candidate sensors might represent temporal, spatial, or even spatio-temporal samples. The observations are related to the following model

$$\mathcal{H}_0 : x_m \sim p_m(x|\mathcal{H}_0), \quad m = 1, 2, \ldots, M, \tag{1a}$$

$$\mathcal{H}_1 : x_m \sim p_m(x|\mathcal{H}_1), \quad m = 1, 2, \ldots, M, \tag{1b}$$

where the probability density function (pdf) of the observation at the mth sensor, x_m, conditioned on the state of nature $\mathcal{H}$ is denoted by $p_m(x|\mathcal{H}_i)$ for $i = 0, 1$. Further, the observations are collected in $\boldsymbol{x} = [x_1, x_2, \ldots, x_M]^T \in \mathbb{R}^M$. The pdf of $\boldsymbol{x}$ under $\mathcal{H}_0$ and $\mathcal{H}_1$ is denoted by $p(\boldsymbol{x}|\mathcal{H}_0)$ and $p(\boldsymbol{x}|\mathcal{H}_1)$, respectively.

We acquire the data $\boldsymbol{x}$ via a linear sensing operation, where the sensing task is modeled through a vector whose entries belong to a binary alphabet, i.e., through

$$\boldsymbol{w} = [w_1, w_2, \ldots, w_M]^T \in \{0, 1\}^M,$$

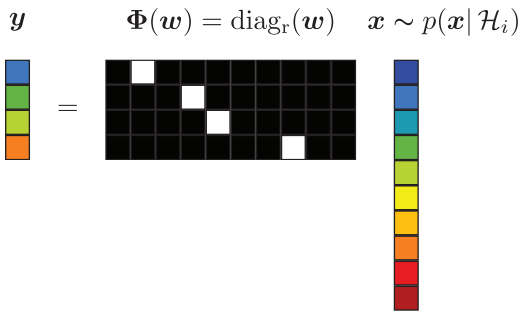

$\boldsymbol{y}$
$\boldsymbol{\Phi}(\boldsymbol{w}) = \mathrm{diag_r}(\boldsymbol{w})$
$\boldsymbol{x} \sim p(\boldsymbol{x}|\mathcal{H}_i)$
=

Fig. 1. Discrete sparse sensing scheme for distributed detection. Here, a white (black) and colored square represents a one (zero) and an arbitrary value, respectively.

where the variable $w_m = (0)1$ indicates whether the mth sensor is (not) selected. More specifically, we define the sensing matrix $\boldsymbol{\Phi}(\boldsymbol{w}) = \mathrm{diag_r}(\boldsymbol{w}) \in \{0,1\}^{K\times M}$, to acquire the data as

$$\boldsymbol{y} = \mathrm{diag_r}(\boldsymbol{w})\boldsymbol{x} = \boldsymbol{\Phi}(\boldsymbol{w})\boldsymbol{x},$$

where K is not assumed to be known. Note that we are interested in cases where $K \ll M$. The reduced dimension data vector $\boldsymbol{y} \in \mathbb{R}^K$ is used instead of $\boldsymbol{x} \in \mathbb{R}^M$ to solve the detection problem. In this paper we seek a sparsest $\boldsymbol{w}$, i.e., a vector with many zeros and just a few non-zero entries, such that a prescribed global detection performance is achieved. Due to the construction of $\boldsymbol{\Phi}(\boldsymbol{w})$, we label the resulting deterministic and structured sensing scheme as *sparse sensing*; see the illustration in Fig. 1. Such a sparse sensing matrix enables a completely distributed compression and sampling, which are instrumental to distributed detection. Furthermore, it leads to possible reductions in the hardware costs, as well as processing and communications overhead.

Sparse sensing differs from the broad research area of compressive sensing—state of the art in the field of sensing cost reduction [21]. In compressive sensing, the underlying signal is always considered sparse in some domain and the goal is sparse signal reconstruction. On the other hand for sparse sensing, the underlying signal does not necessarily have to be sparse and other signal processing tasks (including sparse signal reconstruction [22]) can be considered. Furthermore, in compressive sensing, the compression is generally random, which introduces robustness, but might limit the maximum amount of compression if a specific signal processing task needs to be carried out. Sparse sensing, on the other hand, is a deterministic type of data compression, where the sparse vector inside the sensing function gives a handle on the compression and it can be used for optimally designing the sensing process.

Let $\widehat{\mathcal{H}}$ denote an estimate of the state of nature $\mathcal{H}$, based on a certain decision rule. In the Neyman-Pearson setting, the optimal detector minimizes the probability of miss detection (type II error),

$$P_m = \Pr(\widehat{\mathcal{H}} \neq \mathcal{H}_1|\mathcal{H}_1)$$

for a fixed probability of false alarm (type I error),

$$P_f = \Pr(\widehat{\mathcal{H}} = \mathcal{H}_1|\mathcal{H}_0).$$

This is the well-known Neyman-Pearson detector. In the Bayesian setting, given the prior probabilities $\pi_i = \Pr(\mathcal{H}_i)$ for $i = 0, 1$, the optimal detector minimizes the Bayesian error probability,

$$P_e = \Pr(\widehat{\mathcal{H}} \neq \mathcal{H}) = \pi_0 P_f + \pi_1 P_m,$$

or more generally, the detector minimizes the Bayes' risk. Having introduced the data model, we now formally state the design problem of interest.

Problem 1 (Sparse Sampler Design): Given the data model (1), design a sparsest Boolean vector $\boldsymbol{w}$ that results in a prescribed

i) Bayesian probability of error, P_e, in the Bayesian setting, or

ii) probability of miss detection, P_m, for a fixed probability of false alarm, P_f, in the Neyman-Pearson setting.

Mathematically, the sparse sensing problem for distributed detection can be formulated as

$$\begin{aligned} \text{P-B}: \ & \underset{\boldsymbol{w}\in\{0,1\}^M}{\arg\min} \ \|\boldsymbol{w}\|_0 \\ & \text{s.to } P_e(\boldsymbol{w}) \leq e; \end{aligned} \tag{2a}$$

$$\begin{aligned} \text{P-N}: \ & \underset{\boldsymbol{w}\in\{0,1\}^M}{\arg\min} \ \|\boldsymbol{w}\|_0 \\ & \text{s.to } P_f(\boldsymbol{w}) \leq \alpha, \text{ and } P_m(\boldsymbol{w}) \leq \beta, \end{aligned} \tag{2b}$$

where e, α and β are, respectively, the desired Bayesian probability of error, maximum false-alarm rate and maximum miss-detection rate. Here, $P_e(\boldsymbol{w})$, $P_f(\boldsymbol{w})$, and $P_m(\boldsymbol{w})$ denote the error probabilities due to the selected sensor subset indicated by the non-zero entries of $\boldsymbol{w}$. When prior probabilities are available, we solve P-B (P denotes problem and B denotes Bayesian), otherwise in the Neyman-Pearson setting we solve P-N (N denotes Neyman-Pearson).

In order to ease the design, we next discuss some performance measures that can substitute the error probabilities in the above optimization problems.

III. Optimality Criteria

The error probabilities P_e, P_m or P_f might not admit a known closed-form expression or their expressions might not be favorable for numerical optimization. In this section, we will discuss several weaker and simpler substitutes, which can be optimized instead of the error probabilities. These substitutes are based on the notion of distance (closeness or divergence) between the two distributions of the observations under test. They lead to tractable, if not always optimal (in terms of the error probabilities) design procedures for sampler design. Nevertheless, optimizing the distance measures improves the performance of any practical system.

Let the likelihood ratio of the two hypotheses under test be defined as

$$l(\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\mathcal{H}_1)}{p(\boldsymbol{y}|\mathcal{H}_0)}.$$

In what follows, we consider a number of distance measures that belong to the general class of Ali-Silvey distances [16], which are of the form

$$\psi\left(\mathbb{E}_{|\mathcal{H}_i}\left\{\phi\left[l(\boldsymbol{y})\right]\right\}\right),$$

where $\psi(\cdot)$ is an increasing real-valued function, $\phi[\cdot]$ is a continuous convex function on $(0, \infty)$, and the notation $\mathbb{E}_{|\mathcal{H}_i}\{\phi[l(\boldsymbol{y})]\}$ indicates that $\phi[l(\boldsymbol{y})]$ is averaged under the pdf $p(\boldsymbol{y}|\mathcal{H}_i)$ for either $i = 0$ or $i = 1$.

A. The Bayesian Setting

The Bayes detector minimizes P_e, and makes a decision based on comparing the optimal statistic to a threshold:

$$\log l(\boldsymbol{y}) = \log \frac{p(\boldsymbol{y}|\mathcal{H}_1)}{p(\boldsymbol{y}|\mathcal{H}_0)} \overset{\mathcal{H}_0}{\underset{\mathcal{H}_1}{\lessgtr}} \log \frac{\pi_0}{\pi_1}.$$

In the Bayesian setting, our goal is to choose the best subset of sensors that results in a prescribed Bayesian probability of error P_e. The best achievable exponent in the Bayesian probability of error is parameterized by the Chernoff information (sometimes also referred to as the Chernoff distance) [23, Chernoff's theorem], and it is given by

$$\begin{aligned}\mathcal{C}(\mathcal{H}_1\|\mathcal{H}_0) &= -\log \min_{0\le n\le 1} \int [p(\boldsymbol{y}|\mathcal{H}_1)]^n [p(\boldsymbol{y}|\mathcal{H}_0)]^{1-n} d\boldsymbol{y} \\ &= -\log \min_{0\le n\le 1} \mathbb{E}_{|\mathcal{H}_0} \left\{[l(\boldsymbol{y})]^n\right\}. \qquad (3)\end{aligned}$$

Due to the involved minimization over n, the Chernoff information in (3) is difficult to optimize over $\boldsymbol{w}$. Therefore, we use a special case of the Chernoff information called the *Bhattacharyya distance* as the optimization criterion, where the Bhattacharyya distance is obtained by fixing $n = 0.5$ in (3). The Bhattacharyya distance is given by

$$\mathcal{B}(\mathcal{H}_1\|\mathcal{H}_0) = -\log \rho, \qquad (4)$$

where the *Bhattacharyya coefficient* [9] or the *Hellinger integral* [8], ρ, is given by

$$\begin{aligned}\rho &= \int \sqrt{p(\boldsymbol{y}|\mathcal{H}_1)p(\boldsymbol{y}|\mathcal{H}_0)}d\boldsymbol{y} = \int p(\boldsymbol{y}|\mathcal{H}_0)\sqrt{\frac{p(\boldsymbol{y}|\mathcal{H}_1)}{p(\boldsymbol{y}|\mathcal{H}_0)}}d\boldsymbol{y} \\ &= \mathbb{E}_{|\mathcal{H}_0}\left\{\sqrt{l(\boldsymbol{y})}\right\}. \qquad (5)\end{aligned}$$

It is easy to verify from (5) that the Bhattacharya distance is symmetric, which means $\mathcal{B}(\mathcal{H}_1\|\mathcal{H}_0) = \mathcal{B}(\mathcal{H}_0\|\mathcal{H}_1)$. More importantly, the upper and lower bounds for the Bayesian probability of error can be obtained using the Bhattacharyya coefficient. The bounds are given as follows [8, Appendix A], [9]:

$$\frac{1}{2}\min(\pi_0, \pi_1)\rho^2 \le P_e \le \sqrt{\pi_0\pi_1}\rho. \qquad (6)$$

Therefore, in place of the Bayesian error probability, we minimize the Hellinger integral, or equivalently, maximize the Bhattacharyya distance. Furthermore, when $\int [p(\boldsymbol{y}|\mathcal{H}_1)]^n [p(\boldsymbol{y}|\mathcal{H}_0)]^{1-n} d\boldsymbol{y}$ is symmetric in n and the observations are independent and identically distributed, the Bhattacharyya distance is exponentially the best [9], i.e.,

$$P_e \overset{\text{as.}}{=} \exp\left(-\mathcal{B}(\mathcal{H}_1\|\mathcal{H}_0)\right) \text{ for } P_e \to 0.$$

We now introduce the following assumption:

Assumption 1 (Conditional Independence): The sensor observations are statistically independent, conditioned on the hypothesis $\mathcal{H}$.

Under Assumption 1, the likelihood ratio simplifies to

$$l(\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\mathcal{H}_1)}{p(\boldsymbol{y}|\mathcal{H}_0)} = \prod_{m=1}^{M} [l_m(x)]^{w_m}$$

where $l_m(x) = p_m(x|\mathcal{H}_1)/p_m(x|\mathcal{H}_0)$ is the local likelihood ratio related to the mth sensor, and $p_m(x|\mathcal{H}_i)$ for $i = 0, 1$ are the conditional pdfs of x for the mth sensor. Here, the conditional pdf of the selected sensors is of reduced dimension, i.e., it does not include the measurements that are set to zero.

Besides being a reasonable measure, the Bhattacharya distance is much simpler to optimize under Assumption 1 because of the following result:

Proposition 1 (Linearity of the Bhattacharyya Distance): The considered sparse sensing mechanism preserves the additivity of the Bhattarcharyya distance under Assumption 1, i.e., we can express

$$f_{\mathrm{B}}(\boldsymbol{w}) := \mathcal{B}(\mathcal{H}_1\|\mathcal{H}_0) = \sum_{m=1}^{M} w_m \mathcal{B}_m(\mathcal{H}_1\|\mathcal{H}_0), \qquad (7)$$

where

$$\mathcal{B}_m(\mathcal{H}_1\|\mathcal{H}_0) = -\log \mathbb{E}_{|\mathcal{H}_0}\left\{\sqrt{l_m(x)}\right\}. \qquad (8)$$

Proof: See Appendix A. □

Thus, Proposition 1 enables us to optimize $f_{\mathrm{B}}(\boldsymbol{w})$ over $\boldsymbol{w}$ (subscript B denotes Bayesian). We underline here that $f_{\mathrm{B}}(\boldsymbol{w})$ assumes only the knowledge of the data model and does not need actual measurements, hence the sensing operation can be designed offline. We also remark that the Chernoff information (3) is not additive for conditionally independent observations, unlike the Bhattacharyya distance.

Before discussing the optimization criterion for the Neyman-Pearson setting, we end this subsection with the following remark that generalizes the sampling design in the Bayesian setting.

Remark 1 (Bayes Risk): Let C_{ij} be the cost if we decide $\mathcal{H}_i$ when $\mathcal{H}_j$ is true. A generalization of the minimum P_e detector, is to minimize the Bayes risk

$$\mathcal{R} = \sum_{i=0}^{1}\sum_{j=0}^{1} C_{ij}\mathrm{Pr}(\mathcal{H}_i|\mathcal{H}_j)\mathrm{Pr}(\mathcal{H}_j),$$

where we arrive at a special case of $\mathcal{R} = P_e$ for $C_{00} = C_{11} = 0, C_{10} = C_{01} = 1$. This results in the sensing design problem

$$\underset{\boldsymbol{w}\in\{0,1\}^M}{\arg\min} \ \|\boldsymbol{w}\|_0 \quad \text{s.to } \mathcal{R}(\boldsymbol{w}) \le e_r,$$

where $\mathcal{R}(\boldsymbol{w})$ denotes the Bayes risk due to the selected sensor subset indicated by the non-zero entries of $\boldsymbol{w}$, and e_r is the desired Bayes risk.

The bounds in (6) can be generalized to [24]

$$\mathcal{R}_0 + \mathcal{R}_2\rho^2 \le \mathcal{R} \le \mathcal{R}_0 + \sqrt{\mathcal{R}_1}\rho,$$

where $\mathcal{R}_0 = \pi_0 C_{00} + \pi_1 C_{11}, \mathcal{R}_1 = \pi_0\pi_1(C_{11} - C_{01})(C_{00} - C_{10})$, and $\mathcal{R}_2 = \mathcal{R}_1/(\pi_0(C_{00} - C_{10}) + \pi_1(C_{11} - C_{01}))$. Therefore, maximizing the Bhattacharyya distance (or minimizing the

Hellinger integral) is a reasonable optimality criterion also for a more general minimum Bayes risk detector.

B. The Neyman-Pearson Setting

When the prior probabilities are not known, we solve the Neyman-Pearson problem, where one of the error probabilities (P_f, for example) is fixed while the second error probability, P_m is minimized. More specifically, the decision is based upon the log-likelihood ratio test

$$\log l(\boldsymbol{y}) = \log \frac{p(\boldsymbol{y}|\mathcal{H}_1)}{p(\boldsymbol{y}|\mathcal{H}_0)} \underset{\mathcal{H}_1}{\overset{\mathcal{H}_0}{\lesseqgtr}} \gamma, \tag{9}$$

where γ is the threshold obtained by setting $P_f = \alpha$. In what follows, we discuss two distance measures that we can optimize in the Neyman-Pearson setting.

1) Kullback-Leibler Distance: For a Neyman-Pearson problem, the best achievable error exponent in the probability of error (P_m, for example) is given by the relative entropy or *Kullback-Leibler distance* $\mathcal{K}(\mathcal{H}_1\|\mathcal{H}_0)$ [23, Stein's lemma]. That is, for a fixed value of P_f,

$$\log P_m \overset{\text{as.}}{=} -\mathcal{K}(\mathcal{H}_1\|\mathcal{H}_0) \text{ for } P_m \to 0.$$

The Kullback-Leibler distance is the average log-likelihood ratio, and is given by [25]

$$\begin{aligned}\mathcal{K}(\mathcal{H}_1\|\mathcal{H}_0) &= \mathbb{E}_{|\mathcal{H}_1}\{\log l(\boldsymbol{y})\}\\ &= \int \log l(\boldsymbol{y}) p(\boldsymbol{y}|\mathcal{H}_1) d\boldsymbol{y}.\end{aligned} \tag{10}$$

A lower bound on P_m for a fixed P_f, say α ($0 \le \alpha \le 1$) can be obtained using [25, pp. 74-75 and tables in pp. 378-379]

$$\begin{aligned}\mathcal{K}(\mathcal{H}_1\|\mathcal{H}_0) \geq \alpha \log\left(\frac{\alpha}{1-P_m}\right)&\\ +(1-\alpha)\log\left(\frac{1-\alpha}{P_m}\right) &= g(P_m).\end{aligned} \tag{11}$$

Since $g(P_m)$ is a strictly monotonic function (for values of α that are of practical interest), we can write

$$P_m \geq g^{-1}\left(\mathcal{K}(\mathcal{H}_1\|\mathcal{H}_0)\right). \tag{12}$$

For example, a very small (close to zero) α simplifies (12) to $P_m \geq \exp(-\mathcal{K}(\mathcal{H}_1\|\mathcal{H}_0))$. The following theorem gives an upper bound on P_m.

Theorem 1 (Upper Bound on P_m): If the variance of the log-likehood ratio is v^2, then

$$P_m \leq \frac{1}{1 + \frac{(\mathcal{K}(\mathcal{H}_1\|\mathcal{H}_0) - \log\gamma)^2}{v^2}}, \tag{13}$$

where the threshold γ corresponds to a desired $P_f = \alpha$.

Proof: See Appendix B. □

Therefore, the bounds in (12) and (13) make the maximization of $\mathcal{K}(\mathcal{H}_1\|\mathcal{H}_0)$ a reasonable optimality criterion. We stress here that the above bounds (12) and (13) are valid even when Assumption 1 is not true.

The following property of the Kullback-Leibler distance further allows its easy numerical optimization.

Proposition 2 (Linearity of the Kullback-Leibler Distance): The considered sparse sensing mechanism preserves the additivity of the Kullback-Leibler distance under Assumption 1, i.e., we can express

$$f_{\text{N},1}(\boldsymbol{w}) := \mathcal{K}(\mathcal{H}_1\|\mathcal{H}_0) = \sum_{m=1}^{M} w_m \mathcal{K}_m(\mathcal{H}_1\|\mathcal{H}_0), \tag{14}$$

where

$$\begin{aligned}\mathcal{K}_m(\mathcal{H}_1\|\mathcal{H}_0) &= \int \log l_m(x) p_m(x|\mathcal{H}_1) dx\\ &= \mathbb{E}_{|\mathcal{H}_1}\{\log l_m(x)\}\end{aligned} \tag{15}$$

with $l_m(x) = p_m(x|\mathcal{H}_1)/p_m(x|\mathcal{H}_0)$ being the local likelihood ratio that was defined earlier.

Proof: See Appendix C. □

Therefore, Proposition 2 allows us to maximize $f_{\text{N},1}(\boldsymbol{w})$ over $\boldsymbol{w}$ (subscript N denotes Neyman-Pearson).

Remark 2: For the problem that minimizes the probability of false alarm P_f for a fixed probability of miss detection P_m, the Kullback-Leibler distance

$$\begin{aligned}\mathcal{K}(\mathcal{H}_0\|\mathcal{H}_1) &= -\mathbb{E}_{|\mathcal{H}_0}\{\log l(\boldsymbol{y})\}\\ &= -\int \log l(\boldsymbol{y}) p(\boldsymbol{y}|\mathcal{H}_0) d\boldsymbol{y}\end{aligned} \tag{16}$$

has to be optimized. Note that the Kullback-Leibler distance is not symmetric, i.e., $\mathcal{K}(\mathcal{H}_0\|\mathcal{H}_1) \neq \mathcal{K}(\mathcal{H}_1\|\mathcal{H}_0)$. Furthermore, Proposition 2 holds with the 0 and 1 subscripts interchanged in (14) and (15), which leads to the objective function

$$f_{\text{N},2}(\boldsymbol{w}) := \sum_{m=1}^{M} w_m \mathcal{K}_m(\mathcal{H}_0\|\mathcal{H}_1). \tag{17}$$

2) J-Divergence: The symmetric form of the Kullback-Leibler distance, *J-divergence*, is another frequently used criterion in the design of experiments. The J-divergence is defined as

$$\mathcal{D}(\mathcal{H}_1\|\mathcal{H}_0) = \mathcal{K}(\mathcal{H}_1\|\mathcal{H}_0) + \mathcal{K}(\mathcal{H}_0\|\mathcal{H}_1). \tag{18}$$

A lower bound on $(P_f + P_m)/2$ can be obtained using [25]

$$\begin{aligned}\mathcal{D}(\mathcal{H}_1\|\mathcal{H}_0) \geq 2\Bigg[&\frac{P_f + P_m}{2}\log\left(\frac{(P_f + P_m)/2}{1 - (P_f + P_m)/2}\right)\\ &+ \left(1 - \frac{P_f + P_m}{2}\right)\log\left(\frac{1 - (P_f + P_m)/2}{(P_f + P_m)/2}\right)\Bigg],\end{aligned}$$

and the results from Theorem 1 can be generalized to arrive at an upper bound.

Remark 3: The J-divergence is also a reasonable measure in the Bayesian setting with $\pi_0 = 0.5$ as the Bayesian error probability $P_e = (P_f + P_m)/2$ can be both upper and lower bounded by $\mathcal{D}(\mathcal{H}_1\|\mathcal{H}_0)$. However, for other prior probabilities an upper bound on P_e can be obtained in terms of $\mathcal{D}(\mathcal{H}_1\|\mathcal{H}_0)$ only for Gaussian observations [8].

The J-divergence is also additive for conditionally independent observations, i.e.,

$$f_{\text{N},3}(\boldsymbol{w}) := \mathcal{D}(\mathcal{H}_1\|\mathcal{H}_0) = \sum_{m=1}^{M} w_m \mathcal{D}_m(\mathcal{H}_1\|\mathcal{H}_0),$$

where

$$\mathcal{D}_m(\mathcal{H}_1\|\mathcal{H}_0) = \mathcal{K}_m(\mathcal{H}_1\|\mathcal{H}_0) + \mathcal{K}_m(\mathcal{H}_0\|\mathcal{H}_1). \tag{19}$$

The additive property of the J-divergence is straightforward to verify, and it follows directly from Proposition 2.

Note that all the distance measures introduced in this section admit a closed-form expression irrespective of the observation distributions. The solvers for designing the sensing operation based on the developed performance measures are presented next.

IV. Sparse Sampler Design

The performance measures derived in Section III greatly simplify the sensing design problems P-B and P-N, which are otherwise difficult to solve. The simplified problem is stated as follows.

Problem 2 (Simplified Sparse Sensing Design): Under Assumption 1, given M candidate sensors characterized by the conditional pdfs $\{p_m(x|\mathcal{H}_i)\}_{m=1}^M$ for $i = 0, 1$, design a sparsest vector $\boldsymbol{w}$ such that a desired

i) Bhattacharyya distance in the Bayesian setting, or
ii) Kullback-Leibler distance (or J-divergence) in the Neyman-Pearson setting,

is achieved.

These sampling design problems are, respectively, expressed as the following cardinality minimization problems

$$\text{S-B}: \underset{\boldsymbol{w}\in\{0,1\}^M}{\arg\min}\ \|\boldsymbol{w}\|_0 \quad \text{s.to}\ f_{\text{B}}(\boldsymbol{w}) \geq \lambda_{\text{B}}; \tag{20a}$$

$$\text{S-N}: \underset{\boldsymbol{w}\in\{0,1\}^M}{\arg\min}\ \|\boldsymbol{w}\|_0 \quad \text{s.to}\ f_{\text{N}}(\boldsymbol{w}) \geq \lambda_{\text{N}}, \tag{20b}$$

where λ_{B} and λ_{N} specify the required Bhattacharyya distance and Kullback-Leibler distance (or J-divergence), respectively. The optimization problems S-B and S-N (S denotes simplified problem) are Boolean linear programming problems. In place of $f_{\text{N}}(\boldsymbol{w})$ in (20b), either one of the three performance measures $f_{\text{N},1}(\boldsymbol{w})$, $f_{\text{N},2}(\boldsymbol{w})$, or $f_{\text{N},3}(\boldsymbol{w})$ can be used; however, there is no general answer to the question of how does one performance metric compare with the other.

For the sake of brevity, we collect $\{\mathcal{B}_m(\mathcal{H}_1\|\mathcal{H}_0)\}$, $\{\mathcal{K}_m(\mathcal{H}_1\|\mathcal{H}_0)\}$, $\{\mathcal{K}_m(\mathcal{H}_0\|\mathcal{H}_1)\}$, or $\{\mathcal{D}_m(\mathcal{H}_1\|\mathcal{H}_0)\}$ in a common distance vector denoted by $\boldsymbol{d} \in \mathbb{R}^M$. The optimization problems in (20) can then be expressed in a general form as

$$\underset{\boldsymbol{w}\in\{0,1\}^M}{\arg\min}\ \|\boldsymbol{w}\|_0 \quad \text{s.to}\ \boldsymbol{d}^T\boldsymbol{w} \geq \lambda, \tag{21}$$

where the threshold corresponds to $\lambda := \lambda_{\text{B}}$ or $\lambda := \lambda_{\text{N}}$ for the Bayesian or Neyman-Pearson setting, respectively, with $0 \leq \lambda \leq \mathbf{1}^T\boldsymbol{d}$. Boolean linear programming problems are in general hard to solve. However, S-B and S-N are some of the few special cases of a Boolean linear program that have an explicit solution. We give the solution to the considered offline sampling design problem in the following theorem.

Theorem 2 (Sparse Sampler for Distributed Detection): Assuming the entries of $\boldsymbol{d}$ are (pre-)sorted in descending order and the entries of $\boldsymbol{w}$ are sorted accordingly. The optimal solution $\boldsymbol{w}$ to (21) has entries equal to 1 at the first $\widehat{K}$ entries corresponding to the largest entries in $\boldsymbol{d}$, where

$$\widehat{K} = \min\{i \in \{1, 2, \ldots, M\} | d_1 + d_2 + \ldots d_i \geq \lambda\}. \tag{22}$$

Proof: The proof is straightforward, thus, not detailed. □

In essence, the integer program (21) has an explicit solution and it is optimal for (21). The solution can be interpreted as follows: recalling $\widehat{K}$ from (22), the best subset of sensors out of the M candidate sensors are those $\widehat{K}$ sensors having the smallest local average root-likelihood ratio and largest local average log-likelihood ratio in the Bayesian and Neyman-Pearson setting, respectively.

The appeal of the proposed solution lies in its simplicity. Computationally, the proposed solver is very attractive, for example, with a complexity of $\mathcal{O}(M \log M)$, which is essentially the complexity of the involved sorting algorithm [26]. A parallel implementation on different processors (i.e., still in an offline centralized setting) of the ordering algorithm further reduces the complexity to $\mathcal{O}(\widehat{K})$ using a back-off mechanism as detailed next: The distance measure d_m is made available to the central unit after a time c/d_m, where c is a known positive constant, and the central unit computes the sum of the received values. If the accumulated sum exceeds the desired threshold λ, the central unit declares a transmission stop[1]. Thus, only the $\widehat{K}$ largest distance values are gathered at the central unit.

In many applications, we might know the number of sensors to select (e.g., we might have already purchased the hardware and we want to use all of them). That is, for a fixed sampler size K, the sensing design problem can be expressed as

$$\text{E-B}: \underset{\boldsymbol{w}\in\{0,1\}^M}{\arg\min}\ P_e(\boldsymbol{w}) \quad \text{s.to}\ \|\boldsymbol{w}\|_0 = K; \tag{23a}$$

$$\text{E-N}: \begin{array}{l} \underset{\boldsymbol{w}\in\{0,1\}^M}{\arg\min}\ P_m(\boldsymbol{w}) \\ \quad \text{s.to}\ P_f(\boldsymbol{w}) \leq \alpha, \\ \qquad \|\boldsymbol{w}\|_0 = K, \end{array} \text{ or } \begin{array}{l} \underset{\boldsymbol{w}\in\{0,1\}^M}{\arg\min}\ P_f(\boldsymbol{w}) \\ \quad \text{s.to}\ P_m(\boldsymbol{w}) \leq \beta, \\ \qquad \|\boldsymbol{w}\|_0 = K, \end{array} \tag{23b}$$

where E-B (E-N) represents the equivalent Bayesian (equivalent Neyman-Pearson) problem, and α and β are, respectively, the maximum false-alarm rate and maximum miss-detection rate to be satisfied. By appropriately choosing the thresholds e, α and β in (2), we can obtain the optimal objective value of (2) equal to K, for which P-B (P-N) and E-B (E-N) are equivalent.

We can also simplify E-B and E-N using the Bhattacharyya and Kullback-Leibler distance (or J-divergence) as proxies for the error probabilities, respectively, to arrive at a general form given by

$$\underset{\boldsymbol{w}\in\{0,1\}^M}{\arg\max}\ \boldsymbol{d}^T\boldsymbol{w} \quad \text{s.to}\ \|\boldsymbol{w}\|_0 = K, \tag{24}$$

where it is straightforward to verify that the optimal objective value is given by the sum of the K largest entries of $\boldsymbol{d}$.

We underline that the proposed solver is valid as long as Assumption 1 holds, and the observations need not necessarily be Gaussian distributed.

[1] If more than one distance is made available at the same time, we randomly pick as many as we need.

V. Illustrative Examples

In this section, we illustrate the developed theory of offline sampling design for binary hypothesis testing with a few examples. The sensing operation is designed such that a desired detection performance determined by the Bhattacharyya distance, Kullback-Leibler distance, or J-divergence is achieved. We begin with some examples of Gaussian observations and later on extend it to exponential observation distributions.

A. Gaussian Observations

1) Uncommon Means and Common Covariances: Detecting signals in Gaussian noise is a well-studied problem in detection theory. In particular, it finds applications in spectrum sensing, target detection, and communications, to list a few. For binary signals in Gaussian noise, that is, observations with uncommon means and common covariance structure under both hypotheses, the conditional distributions are given by

$$\begin{aligned}\mathcal{H}_0 &: \boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\theta}_0, \boldsymbol{\Sigma})\\ \mathcal{H}_1 &: \boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\theta}_1, \boldsymbol{\Sigma}),\end{aligned} \tag{25}$$

where $\mathcal{N}(\boldsymbol{\theta}, \boldsymbol{\Sigma})$ denotes a Gaussian distribution with mean vector $\boldsymbol{\theta}$ and covariance matrix $\boldsymbol{\Sigma}$, the mean vectors $\boldsymbol{\theta}_i = [\theta_{i,1}, \theta_{i,2}, \ldots, \theta_{i,M}]^T \in \mathbb{R}^M$ for $i = 0, 1$ as well as the covariance matrix $\boldsymbol{\Sigma} = \operatorname{diag}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_M^2) \in \mathbb{R}^{M \times M}$ are assumed to be perfectly known. The error probabilities admit the following expressions [27, p. 475]

$$\begin{aligned}P_f(\boldsymbol{w}) &= \mathcal{Q}\left(\frac{\gamma + s(\boldsymbol{w})/2}{\sqrt{s(\boldsymbol{w})}}\right);\\ P_m(\boldsymbol{w}) &= 1 - \mathcal{Q}\left(\frac{\gamma - s(\boldsymbol{w})/2}{\sqrt{s(\boldsymbol{w})}}\right),\end{aligned} \tag{26}$$

where γ is the threshold defined in (9),

$$s(\boldsymbol{w}) = (\boldsymbol{\theta}_1 - \boldsymbol{\theta}_0)^T \operatorname{diag}(\boldsymbol{w}) \boldsymbol{\Sigma}^{-1} (\boldsymbol{\theta}_1 - \boldsymbol{\theta}_0) \tag{27}$$

is the *signal-to-noise ratio* (sometimes referred to as the deflection coefficient), and $\mathcal{Q}$ is the complementary Gaussian cumulative distribution function

$$\mathcal{Q}(x) = \int_x^\infty \frac{1}{\sqrt{2\pi}} \exp\left(-y^2/2\right) dy.$$

Note that the signal-to-noise ratio (27) is also linear in $\boldsymbol{w}$. The Bayesian error probability is given by [27, p. 494]

$$P_e(\boldsymbol{w}) = \pi_0 \mathcal{Q}\left(\frac{\gamma' + s(\boldsymbol{w})/2}{\sqrt{s(\boldsymbol{w})}}\right) + \pi_1 \left[1 - \mathcal{Q}\left(\frac{\gamma' - s(\boldsymbol{w})/2}{\sqrt{s(\boldsymbol{w})}}\right)\right], \tag{28}$$

where $\gamma' = \log(\pi_0/\pi_1)$ is the threshold in the Bayesian setting.

For the detection problem (25), the local Bhattacharrya distance, Kullback-Leibler distance, and J-Divergence can be computed respectively as

$$\begin{aligned}\mathcal{B}_m(\mathcal{H}_1 \| \mathcal{H}_0) &= \frac{1}{8\sigma_m^2}(\theta_{0,m} - \theta_{1,m})^2,\\ \mathcal{K}_m(\mathcal{H}_1 \| \mathcal{H}_0) = \mathcal{K}_m(\mathcal{H}_0 \| \mathcal{H}_1) &= \frac{1}{2\sigma_m^2}(\theta_{0,m} - \theta_{1,m})^2,\\ \mathcal{D}_m(\mathcal{H}_1 \| \mathcal{H}_0) &= \frac{1}{\sigma_m^2}(\theta_{0,m} - \theta_{1,m})^2.\end{aligned}$$

We next remark the following interesting observation. All the three distance measures are equal to the signal-to-noise ratio up to a constant. That is, $\mathcal{B}(\mathcal{H}_1 \| \mathcal{H}_0) = s(\boldsymbol{w})/8, \mathcal{K}(\mathcal{H}_1 \| \mathcal{H}_0) = s(\boldsymbol{w})/2$, and $\mathcal{D}(\mathcal{H}_1 \| \mathcal{H}_0) = s(\boldsymbol{w})$. However, these relations are not universal (e.g., they do not hold for non-Gaussian observations). This fact allows us to state the following fundamental result in sampling design for Gaussian observations with common covariance.

Theorem 3: For Gaussian observations with uncommon means and common covariance structure under both hypotheses, maximizing the signal-to-noise ratio over all the possible sampler choices is optimal for P-B and P-N.

Proof: The proof is straightforward. It can be derived based on results from [17] and the monotonicity of the $\mathcal{Q}$ function. Thus, it is omitted. □

As an example, consider the sinusoidal detection problem with $M = 15$ candidate sensors. The means are $\theta_{0,m} = 0$ and $\theta_{1,m} = \cos 2\pi f m$ with $f := 0.33$ for $m = 1, 2, \ldots, M$. Furthermore, we use $\boldsymbol{\Sigma} = \boldsymbol{I}, \pi_0 = 0.3, \pi_1 = 0.7$, and $\alpha = 0.01$. In this example, we use a smaller dimension for M to compare the results with the optimal solution of (2). Nevertheless, the proposed solvers based on ordering easily scale to higher dimensional problems. We solve (23) using exhaustive search over all the $\binom{M}{K}$ combinations for different values of K such that the error probabilities (26) and (28) are optimized. This is labelled as "Neyman-Pearson/Bayesian optimal" in Fig. 2. For this particular example, due to Theorem 3, the simplified sensing design problem can be solved optimally also in terms of error probabilities. This is evident from Fig. 2, where the solution based on ordering the distance measures (labelled as "Neyman-Pearson/Bayesian simplified, sorting") is on top of the optimal solution obtained from exhaustive search. The shaded regions in Fig. 2 indicate the error probabilities with the worst to best subset of K sensors (including any possible subset of K sensors) for different numbers of selected sensors. In particular, the error probabilities with *random sampling* (or any other sub-optimal sampling), for example, [17], [20], would span the shaded region.

Remark 4 (Choosing λ): For a desired P_m, say β, and fixed P_f, say α, the threshold $\lambda := \lambda_{\rm N}$ (for a desired signal-to-noise ratio) can be computed using (26). Specifically,

$$\lambda_{\rm N} = \left(\mathcal{Q}^{-1}(\alpha) - \mathcal{Q}^{-1}(1 - \beta)\right)^2.$$

When λ does not admit a closed form (e.g., with other distributions), the solution path can be used as a guideline to choose λ that results in a desired error probability (often needs to be computed numerically); for example, see Fig. 3 to compute $\lambda := \lambda_{\rm B}$, where we solve (21) with the same simulation parameters as before.

2) Uncommon Covariances and Common Means: Detecting a change in variance is also frequently encountered in practice, for example, while measuring a physical phenomenon with different sensors each characterized with different noise levels both

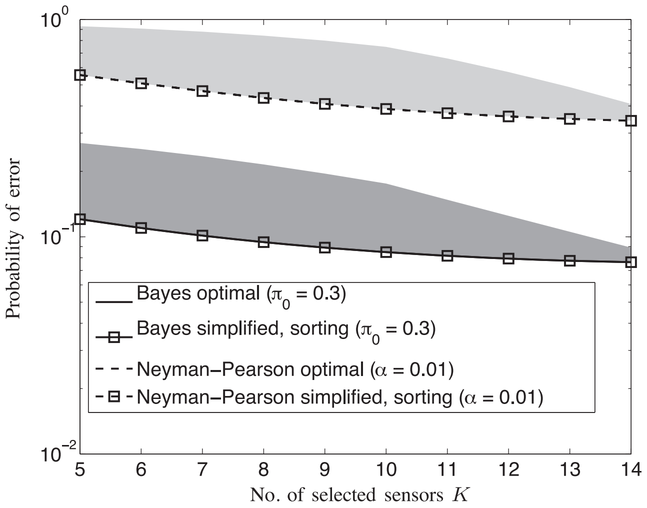


Fig. 2. The (Bayesian/Neyman-Pearson) probability of error for (25) with different numbers of selected sensors K out of $M = 15$ sensors for independent observations. The shaded regions indicate the performance with the worst to best subset of K sensors.

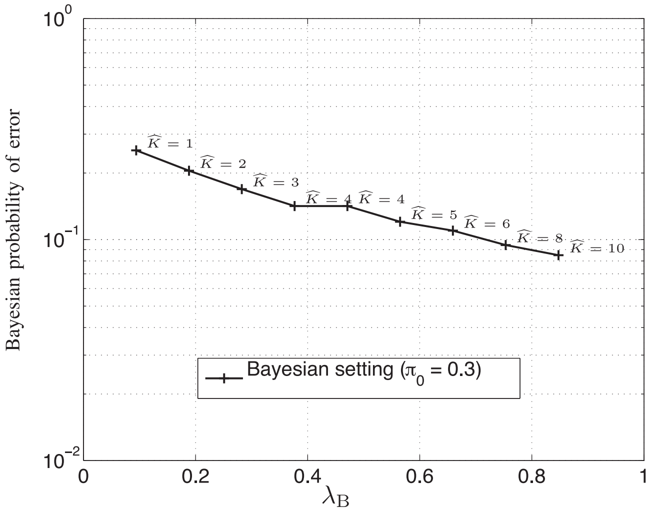


Fig. 3. The solution path illustrates the Bayesian error probability for different values of the threshold $\lambda_{\rm B}$. We use $M = 15$. The number of selected sensors $\widehat{K}$ for a specific value of the threshold is also shown.

across the sensors and under both hypotheses. The conditional distributions in this case are given by

$$\mathcal{H}_0 : \boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\theta}, \boldsymbol{\Sigma}_0)$$
$$\mathcal{H}_1 : \boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\theta}, \boldsymbol{\Sigma}_1), \tag{29}$$

where $\boldsymbol{\theta}$ is the known mean vector and the diagonal matrix $\boldsymbol{\Sigma}_i = \text{diag}(\sigma_{i,1}^2, \sigma_{i,2}^2, \ldots, \sigma_{i,M}^2)$ for $i = 0, 1$ is known. The local log-likelihood ratio is

$$\log l_m(x) = \frac{1}{2}\log\frac{\sigma_{0,m}^2}{\sigma_{1,m}^2} + x^2\left(\frac{1}{2\sigma_{0,m}^2} - \frac{1}{2\sigma_{1,m}^2}\right).$$

Quantifying the performance of the detector, i.e., expressing P_m, P_f, and P_e in a known closed form is more difficult than before, as the pdf of $l(\boldsymbol{x})$ can be obtained only by numerical integration [27]. However, the proposed performance measures admit known expressions as given next. The local Bhattacharyya distance between the conditional distributions in (29) is given by

$$\mathcal{B}_m(\mathcal{H}_1\|\mathcal{H}_0) = \frac{1}{2}\log\left(\frac{\sigma_{0,m}^2 + \sigma_{1,m}^2}{2\sigma_{0,m}\sigma_{1,m}}\right), \tag{30}$$

the local Kullback-Leibler distance is given by

$$\mathcal{K}_m(\mathcal{H}_1\|\mathcal{H}_0) = \frac{1}{2}\left(\frac{\sigma_{1,m}^2}{\sigma_{0,m}^2} - 1 - \log\frac{\sigma_{1,m}^2}{\sigma_{0,m}^2}\right), \tag{31}$$

and $\mathcal{K}_m(\mathcal{H}_0\|\mathcal{H}_1)$ is obtained by interchanging the subscripts 0 and 1 in the above equation. Finally, the J-divergence is given by

$$\mathcal{D}_m(\mathcal{H}_1\|\mathcal{H}_0) = \frac{1}{2}\left(\frac{\sigma_{1,m}^2}{\sigma_{0,m}^2} + \frac{\sigma_{0,m}^2}{\sigma_{1,m}^2} - 2\right). \tag{32}$$

Assume that

$$\boldsymbol{\Sigma}_0 = \begin{bmatrix} 0.1 & 0 \\ 0 & 0.01 \end{bmatrix} \quad \text{and} \quad \boldsymbol{\Sigma}_1 = \begin{bmatrix} 0.5 & 0 \\ 0 & 0.25 \end{bmatrix}$$

and that we want to find the best sensor out of $M = 2$ candidate sensors ($K = 1$). A quick calculation shows that $d_2 > d_1$ for all distances (i.e., the local distance measure of the second sensor is larger than that of the first sensor). Thus, the solution to the S-B (and S-N) will be $\boldsymbol{w} = [0, 1]^T$. This is intuitive as the conditional variance of the second sensor has a larger gap as compared to that of the first sensor, hence the second sensor is more informative.

B. Exponential Observations

Exponentially distributed observations occur while detecting a complex Gaussian signal at the output of a non-coherent receiver. The conditional distributions for exponentially distributed observations for $m = 1, 2, \ldots, M$, are given by

$$\mathcal{H}_0 : x_m \sim \mu_{0,m}\exp(-\mu_{0,m}x)$$
$$\mathcal{H}_1 : x_m \sim \mu_{1,m}\exp(-\mu_{1,m}x), \tag{33}$$

where $x \in [0, \infty)$. The local log-likelihood ratio is

$$\log l_m(x) = \log\frac{\mu_{1,m}}{\mu_{0,m}} + x(\mu_{0,m} - \mu_{1,m}).$$

Using (8), the local Bhattacharyya distance can be computed as

$$\mathcal{B}_m(\mathcal{H}_1\|\mathcal{H}_0) = -\log\frac{\sqrt{4\mu_{0,m}\mu_{1,m}}}{\mu_{0,m} + \mu_{1,m}}.$$

Similarly, the local Kullback-Leibler distance can be computed as

$$\mathcal{K}_m(\mathcal{H}_1\|\mathcal{H}_0) = \log\frac{\mu_{0,m}}{\mu_{1,m}} + \frac{\mu_{1,m}}{\mu_{0,m}} - 1,$$

the local Kullback-Leibler distance $\mathcal{K}_m(\mathcal{H}_0\|\mathcal{H}_1)$ is obtained by interchanging the subscripts 0 and 1 in the above equation, and the local J-divergence is given as

$$\mathcal{D}_m(\mathcal{H}_1\|\mathcal{H}_0) = 2\log\frac{\mu_{1,m}}{\mu_{0,m}} + \frac{\mu_{0,m}^2 - \mu_{1,m}^2}{\mu_{0,m}\mu_{1,m}}.$$

These measures can be directly used in the proposed solvers to design sparse samplers.

VI. Dependent Observations

Throughout most of the paper so far, we have assumed that the observations are conditionally independent. This assumption is generally valid if the sensors are responsible for the noise in the observations (i.e., receiver noise). However, if the sensors are subject to external noise or if the signal itself is stochastic in nature, then Assumption 1 might not be reasonable anymore. Consequently, the additive property of the considered distance measures is also no more valid.

The simplified design problem for this general case (i.e., without any independence assumption), again consists of finding a sparsest $\boldsymbol{w}$ that results in a prescribed distance measure, where we express the Bhattacharyya distance, Kullback-Leibler distance, or J-divergence in terms of $\boldsymbol{w}$. The solution to the above generic problem is hard, nevertheless, we can solve it using standard nonlinear and often nonconvex optimization techniques for a given problem instance (see the example in Section VI.B). However, in some cases, a solution can be computed efficiently. As an example, the Gaussian observation case with uncommon means is detailed next.

A. Gaussian Observations With Uncommon Means

Let us consider the case of binary signal detection in Gaussian noise, and assume the related conditional distributions are given by

$$\begin{aligned}\mathcal{H}_0 &: \boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\theta}_0, \boldsymbol{\Sigma})\\ \mathcal{H}_1 &: \boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\theta}_1, \boldsymbol{\Sigma}),\end{aligned} \tag{34}$$

where the mean vectors $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_1$ as well as the covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{N\times N}$ are assumed to be perfectly known. Note that this model is a generalization of (25) with a nondiagonal covariance matrix. The results from Theorem 3 generalize to dependent observations. Thus, the error probabilities in (2) can without loss of optimality be replaced with the signal-to-noise ratio (which is also related to the considered distance measures up to a constant)

$$s(\boldsymbol{w}) := [\boldsymbol{\Phi}(\boldsymbol{w})\boldsymbol{m}]^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{w}) [\boldsymbol{\Phi}(\boldsymbol{w})\boldsymbol{m}], \tag{35}$$

where we use $\boldsymbol{m} = \boldsymbol{\theta}_1 - \boldsymbol{\theta}_0$ and

$$\boldsymbol{\Sigma}(\boldsymbol{w}) = \boldsymbol{\Phi}(\boldsymbol{w})\boldsymbol{\Sigma}\boldsymbol{\Phi}^T(\boldsymbol{w}) \in \mathbb{R}^{K\times K}$$

is a submatrix of $\boldsymbol{\Sigma}$ that includes only the entries corresponding to the selected measurements. More specifically, we want to solve the problem

$$\underset{\boldsymbol{w}\in\{0,1\}^M}{\arg\min}\ \|\boldsymbol{w}\|_0 \quad \text{s.to}\ s(\boldsymbol{w}) \geq \lambda, \tag{36}$$

where λ is the desired signal-to-noise ratio (or distance measure, or error probability). However, in this case, the simplified problem does not admit an explicit solution. The optimal sampling scheme maximizes $s(\boldsymbol{w})$ in (35) over all possible $\boldsymbol{w} \in \{0,1\}^M$ such that $\boldsymbol{w}$ is as sparse as possible. This incurs a combinatorial search over all the 2^M possible combinations. For example, with $M = 100$ candidate sensors, a performance evaluation of about 10^{30} possible choices is needed whose direct enumeration is clearly impossible.

The sampling design $\boldsymbol{w}$ for (34) depends on the first and second order moments of the observations. In particular, it depends on $\boldsymbol{\theta}_0, \boldsymbol{\theta}_1$, and $\boldsymbol{\Sigma}$.

We next propose some simplifications to solve this problem sub-optimally in polynomial time, yet with a performance that is comparable to the optimal one. Firstly, we write the covariance matrix $\boldsymbol{\Sigma}$ as

$$\boldsymbol{\Sigma} = a\boldsymbol{I} + \boldsymbol{S}, \tag{37}$$

where a non-zero $a \in \mathbb{R}$ is chosen such that $\boldsymbol{S} \in \mathbb{R}^{M\times M}$ is invertible and well-conditioned. Using (37) in (35), we obtain

$$s(\boldsymbol{w}) = \boldsymbol{m}^T\boldsymbol{\Phi}^T(\boldsymbol{w}) \left[a\boldsymbol{I} + \boldsymbol{\Phi}(\boldsymbol{w})\boldsymbol{S}\boldsymbol{\Phi}^T(\boldsymbol{w})\right]^{-1} \boldsymbol{\Phi}(\boldsymbol{w})\boldsymbol{m}. \tag{38}$$

We now state the following property.

Property 1: Using the fact that $\boldsymbol{\Phi}^T\boldsymbol{\Phi} = \operatorname{diag}(\boldsymbol{w})$, we have

$$\begin{aligned}\boldsymbol{\Phi}^T \left(a\boldsymbol{I} + \boldsymbol{\Phi}\boldsymbol{S}\boldsymbol{\Phi}^T\right)^{-1} \boldsymbol{\Phi} = \boldsymbol{S}^{-1}\\ -\boldsymbol{S}^{-1}\left[\boldsymbol{S}^{-1} + a^{-1}\operatorname{diag}(\boldsymbol{w})\right]^{-1}\boldsymbol{S}^{-1}.\end{aligned} \tag{39}$$

Proof: Applying the matrix inversion lemma [28]:

$$\boldsymbol{C}(\boldsymbol{B}^{-1} + \boldsymbol{C}^T\boldsymbol{A}^{-1}\boldsymbol{C})^{-1}\boldsymbol{C}^T = \boldsymbol{A} - \boldsymbol{A}(\boldsymbol{A} + \boldsymbol{C}\boldsymbol{B}\boldsymbol{C}^T)^{-1}\boldsymbol{A},$$

with $\boldsymbol{C} = \boldsymbol{\Phi}^T, \boldsymbol{B}^{-1} = a\boldsymbol{I}$, and $\boldsymbol{A} = \boldsymbol{S}^{-1}$, it is easy to verify (39). □

From Property 1, we can express $s(\boldsymbol{w})$ as

$$\begin{aligned}s(\boldsymbol{w}) = \boldsymbol{m}^T\boldsymbol{S}^{-1}\boldsymbol{m}\\ -\boldsymbol{m}^T\boldsymbol{S}^{-1}\left[\boldsymbol{S}^{-1} + a^{-1}\operatorname{diag}(\boldsymbol{w})\right]^{-1}\boldsymbol{S}^{-1}\boldsymbol{m}.\end{aligned} \tag{40}$$

Note that in contrast to (35), the design parameter $\boldsymbol{w}$ only shows up at one place in (40), which makes the problem much easier. Using the Schur complement, the performance constraint in (36), i.e.,

$$\boldsymbol{m}^T\boldsymbol{S}^{-1}\left[\boldsymbol{S}^{-1} + a^{-1}\operatorname{diag}(\boldsymbol{w})\right]^{-1}\boldsymbol{S}^{-1}\boldsymbol{m} \leq \lambda'$$

with $\lambda' := \lambda - \boldsymbol{m}^T\boldsymbol{S}^{-1}\boldsymbol{m}$ can be equivalently expressed as a linear matrix inequality in $\boldsymbol{w}$, i.e.,

$$\begin{bmatrix}\boldsymbol{S}^{-1} + a^{-1}\operatorname{diag}(\boldsymbol{w}) & \boldsymbol{S}^{-1}\boldsymbol{m}\\ \boldsymbol{m}^T\boldsymbol{S}^{-1} & \lambda'\end{bmatrix} \succeq \boldsymbol{0}, \tag{41}$$

and therefore, it is convex in $\boldsymbol{w}$. The parameter a should be chosen such that $\boldsymbol{S}$ is invertible and well-conditioned. Furthermore, because of (41) the matrix $\boldsymbol{S}^{-1} + a^{-1}\operatorname{diag}(\boldsymbol{w})$ should be positive definite. This can be achieved, for example, by choosing a such that it satisfies the condition $0 < a < \lambda_{\min}\{\boldsymbol{\Sigma}\}$, since $w_m \geq 0$ for $m = 1, 2, \ldots, M$. Although the constraint (41) is convex on $\boldsymbol{w}$, the optimization problem (36) is still not a convex problem due to the ℓ_0-(quasi) norm cost function and the Boolean constraint.

1) Convex Relaxation: The Boolean constraint set is relaxed to its convex hull, i.e., $0 \leq w_m \leq 1, m = 1, 2, \ldots, M$, and we also relax the $\|\boldsymbol{w}\|_0$ constraint in (36) to its best convex approximate $\boldsymbol{1}^T\boldsymbol{w}$. Thus, the relaxed convex problem, more specifically, a semi-definite programming problem, is given as

$$\begin{aligned}&\underset{\boldsymbol{w}}{\arg\min}\ \boldsymbol{1}^T\boldsymbol{w}\\ &\text{s.to}\ \begin{bmatrix}\boldsymbol{S}^{-1} + a^{-1}\operatorname{diag}(\boldsymbol{w}) & \boldsymbol{S}^{-1}\boldsymbol{m}\\ \boldsymbol{m}^T\boldsymbol{S}^{-1} & \lambda'\end{bmatrix} \succeq \boldsymbol{0},\\ &\qquad 0 \leq w_m \leq 1, m = 1, 2, \ldots, M.\end{aligned} \tag{42}$$

For a fixed K, the equivalent problem of the form (24) can be relaxed to

$$\begin{aligned}&\arg\max_{\boldsymbol{w}}\ s(\boldsymbol{w})\\ &\text{s.to } \mathbf{1}^T\boldsymbol{w} = K,\\ &\qquad 0 \le w_m \le 1, m = 1,2,\ldots,M,\end{aligned}$$

which simplifies to

$$\begin{aligned}&\arg\min_{\boldsymbol{w}}\ \boldsymbol{m}^T\boldsymbol{S}^{-1}\left[\boldsymbol{S}^{-1} + a^{-1}\mathrm{diag}(\boldsymbol{w})\right]^{-1}\boldsymbol{S}^{-1}\boldsymbol{m}\\ &\text{s.to } \mathbf{1}^T\boldsymbol{w} = K,\\ &\qquad 0 \le w_m \le 1, m = 1,2,\ldots,M.\end{aligned} \tag{43}$$

Here, only the second term of (40), which depends on $\boldsymbol{w}$ is optimized (minimization is due to its negative sign). Writing (43) in the epigraph form [29], we obtain

$$\begin{aligned}&\arg\min_{\boldsymbol{w},t}\ t\\ &\text{s.to } \mathbf{1}^T\boldsymbol{w} = K,\\ &\quad \begin{bmatrix}\boldsymbol{S}^{-1} + a^{-1}\mathrm{diag}(\boldsymbol{w}) & \boldsymbol{S}^{-1}\boldsymbol{m}\\ \boldsymbol{m}^T\boldsymbol{S}^{-1} & t\end{bmatrix} \succeq \mathbf{0},\\ &\quad 0 \le w_m \le 1, m = 1,2,\ldots,M,\end{aligned} \tag{44}$$

with auxiliary variable $t \in \mathbb{R}$.

Subsequently, the selected sensors (i.e., an approximate Boolean solution) can be computed using randomization techniques based on the solution from (42) or (44) as described in [4]. For the sake of completeness, we summarize the randomized rounding as Algorithm 1. The relaxed convex problem can be solved using off-the-shelf software, for example, SeDuMi [30].

Algorithm 1: Randomized Rounding

```
1. Given the solution $\boldsymbol{w}^\star$ of (42) or (44) and a number of randomizations $L$.
2. for $l = 1$ to $L$
3.     generate $w_{m,l} = 1$ with a probability $w_m^\star$
       (or $w_{m,l} = 0$ with a probability $1 - w_m^\star$)
       for $m = 1,2,\ldots,M$, where $w_m^\star = [\boldsymbol{w}^\star]_m$.
4. end
5. define $\boldsymbol{w}_l = [w_{1,l},\ldots,w_{M,l}]^T$ and the index set of the candidate estimates satisfying the constraints as
   $\Omega \triangleq \begin{cases}\{l \mid s(\boldsymbol{w}_l) \ge \lambda, l = 1,2,\ldots,L\}, & \text{for (42)}\\ \{l \mid \|\boldsymbol{w}_l\|_0 = K, l = 1,2,\ldots,L\}, & \text{for (44)}.\end{cases}$
6. If the set $\Omega$ is empty, go back to step 2.
7. output approximate solution $\boldsymbol{w}^\star_{\mathrm{round}} = \boldsymbol{w}_{l^\star}$, where
   $l^\star = \begin{cases}\arg\min_{l\in\Omega} \|\boldsymbol{w}_l\|_0, & \text{for (42)}\\ \arg\max_{l\in\Omega} s(\boldsymbol{w}_l), & \text{for(44)}.\end{cases}$
```

2) Numerical Example: To illustrate sparse sensing with dependent observations, we recall the simulation parameters from Section V.A1, but instead of independent noise, we use an autoregressive correlation matrix $\boldsymbol{\Sigma}$, which is a Toeplitz matrix of the form

$$\boldsymbol{\Sigma} = \sigma^2\begin{bmatrix}1 & \rho & \rho^2 & \cdots & \rho^{M-1}\\ \rho & 1 & \rho & & \\ \rho^2 & \rho & 1 & & \vdots\\ \vdots & & & \ddots & \\ \rho^{M-1} & & & \cdots & 1\end{bmatrix}, \tag{45}$$

with a known correlation coefficient $\rho \in [0,1]$ and variance $\sigma^2 = 1$. Such a $\boldsymbol{\Sigma}$ is useful for modeling correlations between distributed sensors; for example, it can represent a spatially decaying correlation function. The convex relaxed problem (44) is solved using SeDuMi [30].

The probability of error, i.e., P_m in the Neyman-Pearson setting and P_e in the Bayesian setting for different numbers of selected sensors is shown in Fig. 4. We underline the following observations. The solution with randomized rounding ($L = 50$) is shown in Fig. 4 for $\rho = \{0.25, 0.75\}$ with $a = 0.11$ in (37). For low values of the correlation coefficient, ρ, the convex relaxation with deterministic rounding is very close to optimal. For larger values of ρ, the solution of the relaxed problem with randomized rounding is still very close to optimal for large values of K, but less optimal for small values of K. As observed in the simulations, for $L = 50 \ll 2^{15}$, the sensing design with randomization is near-optimal in terms of the error probability.

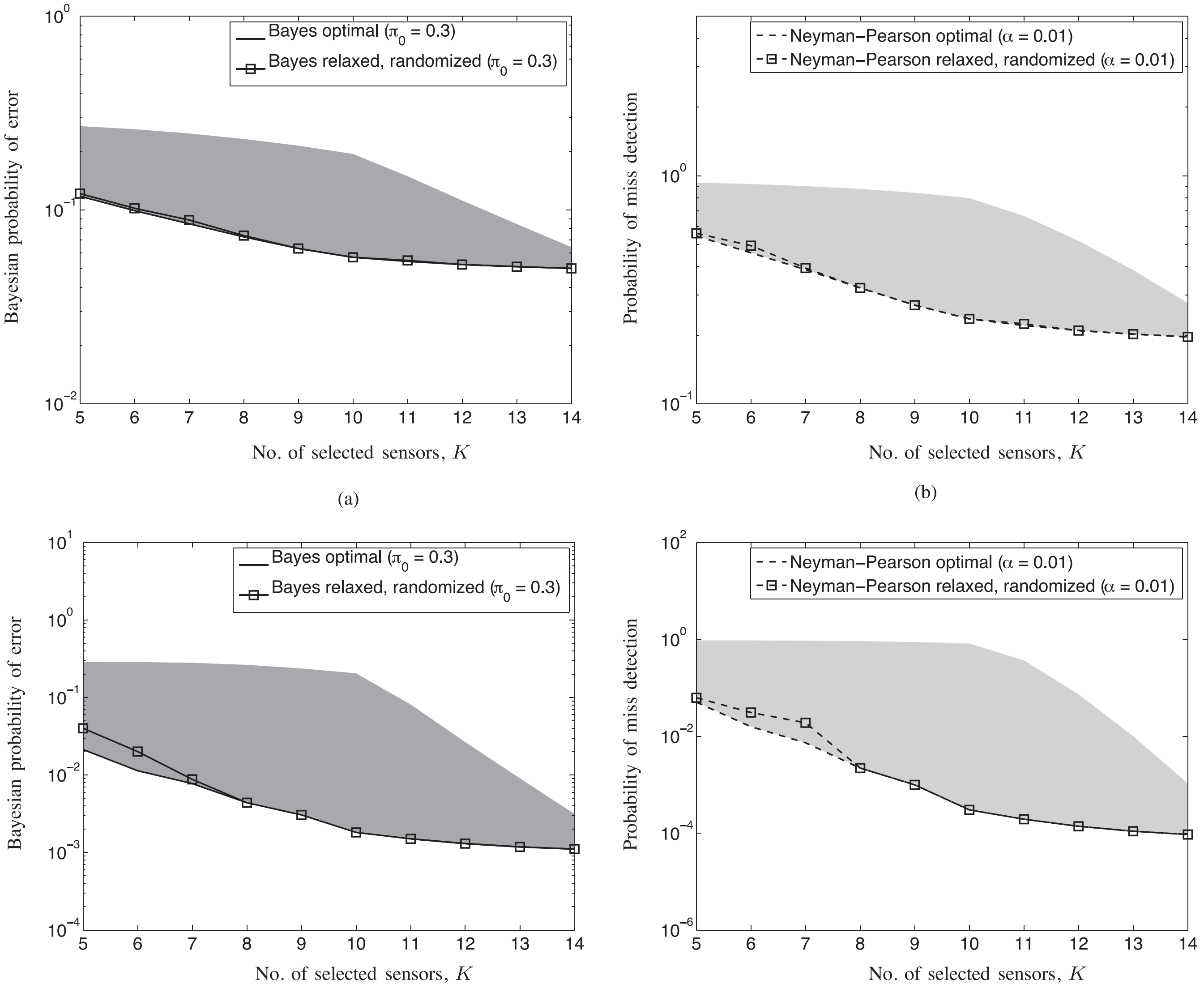
3) Correlation Versus Number of Selected Sensors: In this subsection, we focus on the number of sensors required to achieve a certain detection performance when the sensors become more *coherent*, i.e., as the correlation coefficient ρ approaches 1. To illustrate this, let us consider the numerical example introduced in Section V.A1 with $f \in \{0, 0.33\}$, but with an equi-correlated covariance matrix of the form

$$\boldsymbol{\Sigma} := \begin{bmatrix}1 & \rho & \cdots & \rho\\ \rho & 1 & \cdots & \rho\\ \vdots & \vdots & \ddots & \rho\\ \rho & \rho & \cdots & 1\end{bmatrix} = (1-\rho)\boldsymbol{I}_M + \rho\mathbf{1}_M\mathbf{1}_M^T, \tag{46}$$

with a known correlation coefficient $\rho \in [0,1]$. Note that for such a covariance matrix, any $a \ne 1-\rho$ leads to an invertible $\boldsymbol{S}$ in (37) that can be used in the solver (42).

We first consider the case when $f = 0$, where all the M sensors have the same mean value, i.e., $\boldsymbol{m}$ is the all-one vector up to a constant scaling. We refer to them as *identical sensors*. In this case, any subset of sensors is also the best subset of sensors, hence, random sensing is optimal. As the correlation coefficient ρ approaches 1, the amount of information (Kullback-Liebler distance/Bhattacharyya distance/J-divergence/signal-to-noise ratio) contributed by any random subset of $K > 1$ sensors is the same as that of the contribution from $K = 1$ sensor; see Fig. 5(a). Thus, even with all the sensors selected the detection performance is limited to that of the performance with one sensor. This is a well-known result from distributed detection that extends to sampling design problems [13].

A more interesting case, in particular for sensing design problems, is when the sensors are not identical ($f = 0.33$), i.e., $\boldsymbol{m}$ has all different entries. When the sensors are not identical, as the correlation coefficient ρ approaches 1, the amount of information contained in the best subset of $K > 1$ sensors increases



(a)

(b)

(c)

(d)

Fig. 4. The (Bayesian/Neyman-Pearson) probability of error for (34) with different numbers of selected sensors K out of $M = 15$ sensors. The shaded regions indicate the performance with the worst to best subset of K sensors. (a) $\rho = 0.25$, (b) $\rho = 0.25$, (c) $\rho = 0.75$, and (d) $\rho = 0.75$.

significantly; see Fig. 5(b). More specifically, with equi-correlated yet different observations, to achieve a certain detection performance, the number of sensors required decreases significantly as the correlation coefficient ρ increases. The maximum achievable signal-to-noise ratio is proportional to the inverse of the minimum eigenvalue of $\boldsymbol{\Sigma}(\boldsymbol{w})$, which is $\lambda_{\min}^{-1}\{\boldsymbol{\Sigma}(\boldsymbol{w})\} = 1/(1-\rho)$, for any sampler size $K \neq 0$. The optimal sparse sampler would choose the entries of $\boldsymbol{m}$ that are most aligned to the eigenvector corresponding to the minimum eigenvalue of $\boldsymbol{\Sigma}(\boldsymbol{w})$ (hence, as $\rho \to 1$ the signal-to-noise ratio is large). Similarly, if the entries of $\boldsymbol{m}$ are parallel to the eigenvector corresponding to the maximum eigenvalue of $\boldsymbol{\Sigma}(\boldsymbol{w})$, that is, the all-one vector, then the signal-to-noise ratio is minimized; this is the case in Fig. 5(a).

B. Gaussian Observation With Uncommon Covariances

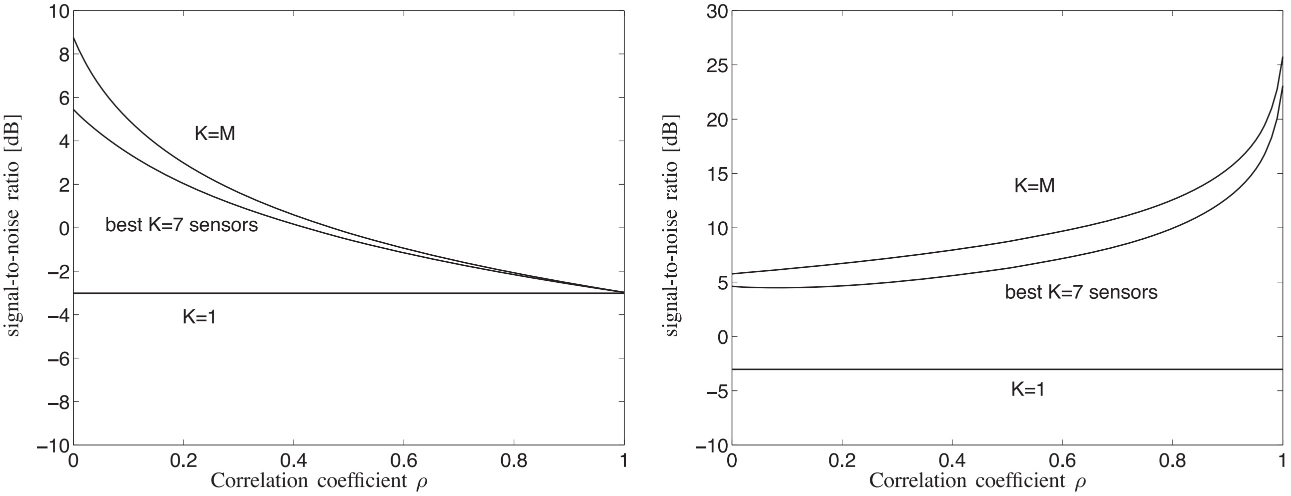
We now provide some extensions and offer guidelines for determining sparse sensing mechanisms for testing between two covariance matrices. That is, when the covariance structures are different under both hypotheses. Suppose the conditional distributions are given by

$$\begin{aligned}\mathcal{H}_0 &: \boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\theta}, \boldsymbol{\Sigma}_0)\\ \mathcal{H}_1 &: \boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\theta}, \boldsymbol{\Sigma}_1),\end{aligned} \tag{47}$$

where the mean vector $\boldsymbol{\theta} \in \mathbb{R}^N$ as well as the $N \times N$ covariance matrices $\boldsymbol{\Sigma}_0$ and $\boldsymbol{\Sigma}_1$ are assumed to be perfectly known. This model is a generalization of (29) with nondiagonal covariance matrices.

As with (29), the distance measures are not equal to each other. Using (5), the Bhattacharyya distance for the observations of the form $\boldsymbol{y} = \boldsymbol{\Phi}\boldsymbol{x}$ can be computed as

$$\begin{aligned}\mathcal{B}(\mathcal{H}_1 \| \mathcal{H}_0) = &\frac{1}{2}\log\det\{\boldsymbol{\Sigma}_{01}(\boldsymbol{w})\}\\ &-\frac{1}{4}\left(\log\det\{\boldsymbol{\Sigma}_0(\boldsymbol{w})\} + \log\det\{\boldsymbol{\Sigma}_1(\boldsymbol{w})\}\right),\end{aligned} \tag{48}$$



(a)

(b)

Fig. 5. The signal-to-noise ratio for different values of the correlation coefficient ρ. (a) Identical sensors ($f = 0$). (b) Non-identical sensors ($f = 0.33$).

where $\boldsymbol{\Sigma}_i(\boldsymbol{w}) = \boldsymbol{\Phi}\boldsymbol{\Sigma}_i\boldsymbol{\Phi}^T$ for $i = 01, 0, 1$, with $2\boldsymbol{\Sigma}_{01} = \boldsymbol{\Sigma}_0 + \boldsymbol{\Sigma}_1$. Similarly, using (10), we can show that the Kullback-Leibler distance is given by

$$\mathcal{K}(\mathcal{H}_1 \| \mathcal{H}_0) = \frac{1}{2}\left(\mathrm{tr}\{\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{w})\boldsymbol{\Sigma}_1(\boldsymbol{w})\} - \|\boldsymbol{w}\|_0 \right.$$
$$\left. - \log\det\{\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{w})\boldsymbol{\Sigma}_1(\boldsymbol{w})\}\right). \quad (49)$$

Here, $\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{w})\boldsymbol{\Sigma}_1(\boldsymbol{w})$ is the *signal-to-noise ratio* matrix.

We can express the Bhattacharyya and Kullback-Leibler distance as a difference of concave functions by relaxing $\boldsymbol{w} \in \{0,1\}^M$ to $[0,1]^M$. That is, we can express (48) and (49) as

$$f(\boldsymbol{w}) = f_0(\boldsymbol{w}) - f_1(\boldsymbol{w}),$$

where $f_0(\boldsymbol{w})$ and $f_1(\boldsymbol{w})$ are concave functions of its arguments; see Appendix D for the explicit expressions of $f_0(\boldsymbol{w})$ and $f_1(\boldsymbol{w})$. As a consequence, the relaxed problem (for fixed K)

$$\underset{\boldsymbol{w}\in[0,1]^M}{\arg\min} \; f_1(\boldsymbol{w}) - f_0(\boldsymbol{w}) \quad \text{s.to } \mathbf{1}^T\boldsymbol{w} = K,$$

is not a convex problem as the cost is not a convex function of its argument and has to be solved using nonconvex optimization techniques.

One such heuristic to solve the difference of convex problems is the convex-concave procedure [31] where the concave term, i.e., $f_1(\boldsymbol{w})$ is replaced with its affine approximation (more generally, any reasonable convex approximation) while the convex portion, i.e., $-f_0(\boldsymbol{w})$ is retained. The resulting convex problem is iteratively solved to obtain a local optimum.

The J-divergence can be computed using (18) as

$$\mathcal{D}(\mathcal{H}_1 \| \mathcal{H}_0) = \frac{1}{2}\mathrm{tr}\{\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{w})\boldsymbol{\Sigma}_1(\boldsymbol{w})\}$$
$$+ \frac{1}{2}\mathrm{tr}\{\boldsymbol{\Sigma}_1^{-1}(\boldsymbol{w})\boldsymbol{\Sigma}_0(\boldsymbol{w})\} - \|\boldsymbol{w}\|_0. \quad (50)$$

We next show that maximizing the J-divergence over $\boldsymbol{w}$ can be cast as a convex problem.

Let the covariance matrices $\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}_0^{1/2}\boldsymbol{\Sigma}_0^{T/2}$ and $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_1^{1/2}\boldsymbol{\Sigma}_1^{T/2}$, respectively, admit the decomposition

$$\boldsymbol{\Sigma}_0 = a_0\boldsymbol{I} + \boldsymbol{S}_0,$$

and

$$\boldsymbol{\Sigma}_1 = a_1\boldsymbol{I} + \boldsymbol{S}_1,$$

with scalars a_0 and a_1 chosen such that $\boldsymbol{S}_0$ and $\boldsymbol{S}_1$ are invertible. Using Property 1, we can show that the J-divergence (50) is equivalent to

$$\begin{aligned}\mathcal{D}(\mathcal{H}_1 \| \mathcal{H}_0) = \; & \frac{1}{2}\mathrm{tr}\left\{\boldsymbol{S}_0^{-1}\boldsymbol{\Sigma}_1 \right. \\ & \left. -\boldsymbol{S}_0^{-1}\left[\boldsymbol{S}_0^{-1} + a_0^{-1}\mathrm{diag}(\boldsymbol{w})\right]^{-1}\boldsymbol{S}_0^{-1}\boldsymbol{\Sigma}_1\right\} \\ & + \frac{1}{2}\mathrm{tr}\left\{\boldsymbol{S}_1^{-1}\boldsymbol{\Sigma}_0 \right. \\ & \left. -\boldsymbol{S}_1^{-1}\left[\boldsymbol{S}_1^{-1} + a_1^{-1}\mathrm{diag}(\boldsymbol{w})\right]^{-1}\boldsymbol{S}_1^{-1}\boldsymbol{\Sigma}_0\right\} - \|\mathbf{w}\|_0.\end{aligned}$$

Thus, maximizing the J-divergence over $\boldsymbol{w}$ for a fixed K is the same as minimizing

$$\frac{1}{2}\mathrm{tr}\left\{\boldsymbol{S}_0^{-1}\left[\boldsymbol{S}_0^{-1} + a_0^{-1}\mathrm{diag}(\boldsymbol{w})\right]^{-1}\boldsymbol{S}_0^{-1}\boldsymbol{\Sigma}_1\right\}$$
$$+ \frac{1}{2}\mathrm{tr}\left\{\boldsymbol{S}_1^{-1}\left[\boldsymbol{S}_1^{-1} + a_1^{-1}\mathrm{diag}(\boldsymbol{w})\right]^{-1}\boldsymbol{S}_1^{-1}\boldsymbol{\Sigma}_0\right\}$$

over $\boldsymbol{w}$. To cast this as a convex problem, we introduce two variables

$$\boldsymbol{Z}_0 = \boldsymbol{\Sigma}_1^{T/2}\boldsymbol{S}_0^{-1}\left[\boldsymbol{S}_0^{-1} + a_0^{-1}\mathrm{diag}(\boldsymbol{w})\right]^{-1}\boldsymbol{S}_0^{-1}\boldsymbol{\Sigma}_1^{1/2};$$
$$\boldsymbol{Z}_1 = \boldsymbol{\Sigma}_0^{T/2}\boldsymbol{S}_1^{-1}\left[\boldsymbol{S}_1^{-1} + a_1^{-1}\mathrm{diag}(\boldsymbol{w})\right]^{-1}\boldsymbol{S}_1^{-1}\boldsymbol{\Sigma}_0^{1/2},$$

and obtain

$$\begin{aligned} & \underset{\boldsymbol{w},\boldsymbol{Z}_0,\boldsymbol{Z}_1}{\arg\min} \; \frac{1}{2}\mathrm{tr}\{\boldsymbol{Z}_0\} + \frac{1}{2}\mathrm{tr}\{\boldsymbol{Z}_1\} \\ & \text{s.to } \mathbf{1}^T\boldsymbol{w} = K, \\ & \quad \boldsymbol{\Sigma}_1^{T/2}\boldsymbol{S}_0^{-1}\left[\boldsymbol{S}_0^{-1} + a_0^{-1}\mathrm{diag}(\boldsymbol{w})\right]^{-1}\boldsymbol{S}_0^{-1}\boldsymbol{\Sigma}_1^{1/2} \preceq \boldsymbol{Z}_0 \\ & \quad \boldsymbol{\Sigma}_0^{T/2}\boldsymbol{S}_1^{-1}\left[\boldsymbol{S}_1^{-1} + a_1^{-1}\mathrm{diag}(\boldsymbol{w})\right]^{-1}\boldsymbol{S}_1^{-1}\boldsymbol{\Sigma}_0^{1/2} \preceq \boldsymbol{Z}_1 \\ & \quad 0 \leq w_m \leq 1, \quad m = 1, 2, \ldots, M. \end{aligned} \quad (51)$$

TABLE I
SUMMARY OF RESULTS

	Neyman-Pearson setting	**Bayesian Setting**
Optimization criterion	Kullback-Leibler distance or J-divergence	Bhattacharyya distance
Independent observations	ordering distances	ordering distances
Dependent Gaussian observations (uncommon means)	signal-to-noise (convex) optimization	signal-to-noise (convex) optimization
Dependent Gaussian observations (uncommon covariances)	nonconvex (Kullback-Leibler distance) or convex (J-divergence) optimization	nonconvex (Bhattacharyya distance) or convex (J-divergence) optimization

The second and the third constraint can be, respectively, expressed as an LMI in $\boldsymbol{w}$, i.e.,

$$\begin{bmatrix} \boldsymbol{Z}_0 & \boldsymbol{S}_0^{-1}\boldsymbol{\Sigma}_1^{1/2} \\ \boldsymbol{\Sigma}_1^{T/2}\boldsymbol{S}_0^{-1} & \boldsymbol{S}_0^{-1} + a_0^{-1}\mathrm{diag}(\boldsymbol{w}) \end{bmatrix} \succeq \mathbf{0},$$

$$\begin{bmatrix} \boldsymbol{Z}_1 & \boldsymbol{S}_1^{-1}\boldsymbol{\Sigma}_0^{1/2} \\ \boldsymbol{\Sigma}_0^{T/2}\boldsymbol{S}_1^{-1} & \boldsymbol{S}_1^{-1} + a_1^{-1}\mathrm{diag}(\boldsymbol{w}) \end{bmatrix} \succeq \mathbf{0}.$$

An approximate Boolean solution has to be subsequently computed using randomized rounding.

The optimization problem of the form (20) with unknown K can be derived along similar lines by relaxing the $\|\boldsymbol{w}\|_0$ in the cost function. Before we end this section, we make the following remarks.

- For Gaussian observations, we recall that an upper bound on P_e and P_m can be obtained in terms of J-divergence. Hence, optimizing J-divergence is reasonable under the Bayesian and Neyman-Pearson setting.
- For general Gaussian dependent observations (with uncommon means and uncommon covariances under both hypotheses), the design problems are straightforward combinations of the problems derived in Sections VI.A and VI.B.

We have summarized the results in Table I.

VII. CONCLUSION

In this paper, we have developed a framework for structured and sparse sampler design for distributed detection problems. In particular, we have addressed binary hypothesis testing in both the Bayesian and Neyman-Pearson setting. The proposed framework can be directly applied to sensor placement/selection, sample selection, and fully-decentralized data compression, where we seek the best subset of sensor/sampling locations or data samples that results in a desired detection probability. To simplify the design problem, we have used a number of distance measures that quantify the closeness or divergence between the conditional distributions of the observations. We give an explicit solution for the sampling design problem with conditionally independent observations and the results are summarized as follows. The best sensors are the ones with the smallest local average root-likelihood ratio and largest local average log-likelihood ratio in the Bayesian and Neyman-Pearson setting, respectively. The framework has also been generalized to conditionally dependent observations with a thorough analysis for the Gaussian case. In that context, we have shown that, for uncommon means and common covariances under both hypotheses, the number of non-identical Gaussian sensors required to achieve a desired detection performance reduces significantly as the sensors become more coherent.

APPENDIX A
PROOF OF PROPOSITION 1

In this section, we prove that the additivity of the Bhattacharyya distance is preserved with compression using $\boldsymbol{\Phi}(\boldsymbol{w})$. Using the conditional independence across sensors, i.e., Assumption 1, the Bhattacharyya distance in (5) can be expressed as

$$\begin{aligned} \mathcal{B}(\mathcal{H}_1\|\mathcal{H}_0) &= -\log \mathbb{E}_{|\mathcal{H}_0}\{\sqrt{l(\boldsymbol{y})}\} \\ &= -\log \mathbb{E}_{|\mathcal{H}_0}\left\{\prod_{m=1}^{M} [l_m(x)]^{w_m/2}\right\} \\ &= -\log \prod_{m=1}^{M} \mathbb{E}_{|\mathcal{H}_0}\left\{[l_m(x)]^{w_m/2}\right\}, \end{aligned}$$

where $l_m(x)$ is the local likelihood ratio at the mth sensor. Since $w_m \in \{0,1\}$, we can further simplify $\mathcal{B}(\mathcal{H}_1\|\mathcal{H}_0)$ to

$$\begin{aligned} \mathcal{B}(\mathcal{H}_1\|\mathcal{H}_0) &= -\log \prod_{m=1}^{M} \left(\mathbb{E}_{|\mathcal{H}_0}\left\{\sqrt{l_m(x)}\right\}\right)^{w_m} \\ &= \sum_{m=1}^{M} -w_m \log \mathbb{E}_{|\mathcal{H}_0}\left\{\sqrt{l_m(x)}\right\} \\ &= \sum_{m=1}^{M} w_m \mathcal{B}_m(\mathcal{H}_1\|\mathcal{H}_0). \end{aligned}$$

APPENDIX B
UPPER BOUND ON P_m

To derive the upper bound on P_m stated in Theorem 1, we use Chebyshev's inequality [32]

$$\Pr(X - \mathbb{E}\{X\} \geq t) \leq \frac{1}{1 + \frac{t^2}{v^2}}, \tag{52}$$

where X is a random variable with variance v^2 and t is a constant. Then, P_m simplifies to

$$\begin{aligned} P_m &= \Pr\left(\log l(\boldsymbol{y}) \leq \log\gamma|\mathcal{H}_1\right) \\ &= \Pr\big(\log l(\boldsymbol{y}) - \mathbb{E}_{|\mathcal{H}_1}\{\log l(\boldsymbol{y})\} \\ &\qquad \leq \log\gamma - \mathbb{E}_{|\mathcal{H}_1}\{\log l(\boldsymbol{y})\}|\mathcal{H}_1\big) \\ &= \Pr\left(\log l(\boldsymbol{y}) - \mathcal{K}(\mathcal{H}_1\|\mathcal{H}_0) \geq \mathcal{K}(\mathcal{H}_1\|\mathcal{H}_0) - \log\gamma|\mathcal{H}_1\right), \end{aligned}$$

where the last equation has the same form as the inequality (52) with $t = \log\gamma - \mathcal{K}(\mathcal{H}_1\|\mathcal{H}_0)$.

If the variance of $\log l(\boldsymbol{y})$ is v^2, then, from (52), we have

$$P_m \leq \frac{1}{1+\frac{(\mathcal{K}(\mathcal{H}_1\|\mathcal{H}_0)-\log\gamma)^2}{v^2}}$$

This completes the proof.

Appendix C
Proof of Proposition 2

In this section, we prove that the additivity of the Kullback-Leibler distance for independent observations is preserved with compression using $\boldsymbol{\Phi}(\boldsymbol{w})$. Assuming Assumption 1 holds, then the Kullback-Leibler distance in (10) can be expressed as

$$\begin{aligned}\mathcal{K}(\mathcal{H}_1\|\mathcal{H}_0) &= \mathbb{E}_{|\mathcal{H}_1}\{\log l(\boldsymbol{y})\}\\ &= \mathbb{E}_{|\mathcal{H}_1}\left\{\log\prod_{m=1}^{M}[l_m(x)]^{w_m}\right\}\\ &= \mathbb{E}_{|\mathcal{H}_1}\left\{\sum_{m=1}^{M} w_m\log l_m(x)\right\}\\ &= \sum_{m=1}^{M} w_m\mathbb{E}_{|\mathcal{H}_1}\{\log l_m(x)\}\\ &= \sum_{m=1}^{M} w_m\mathcal{K}_m(\mathcal{H}_1\|\mathcal{H}_0),\end{aligned}$$

where $l_m(x)$ is the local likelihood ratio at the mth sensor.

Appendix D
Expressions for $f_0(\boldsymbol{w})$ and $f_1(\boldsymbol{w})$

Let the covariance matrices $\boldsymbol{\Sigma}_{01}, \boldsymbol{\Sigma}_0$ and $\boldsymbol{\Sigma}_1$, respectively, admit a decomposition of the form $\boldsymbol{\Sigma}_{01} = a_{01}\boldsymbol{I} + \boldsymbol{S}_{01}, \boldsymbol{\Sigma}_0 = a_0\boldsymbol{I} + \boldsymbol{S}_0$, and $\boldsymbol{\Sigma}_1 = a_0\boldsymbol{I} + \boldsymbol{S}_0$. Here, the scalars a_{01}, a_0, and a_1 are, respectively, chosen such that the matrices $\boldsymbol{S}_{01}, \boldsymbol{S}_0$, and $\boldsymbol{S}_1$ are invertible.

Using the Sylvester's determinant identity

$$\det\{\boldsymbol{A}+\boldsymbol{B}\boldsymbol{C}\} = \det\{\boldsymbol{A}\}\det\{\boldsymbol{I}+\boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{B}\}, \qquad (53)$$

we can express, for example,

$$\begin{aligned}\det\{\boldsymbol{\Phi}\boldsymbol{\Sigma}_0\boldsymbol{\Phi}^T\} &= \det\{a_0\boldsymbol{I}+\boldsymbol{\Phi}\boldsymbol{S}_0\boldsymbol{\Phi}^T\}\\ &= a_0^M\det\{\boldsymbol{I}+a_0^{-1}\boldsymbol{\Phi}^T\boldsymbol{\Phi}\boldsymbol{S}_0\}\\ &= a_0^M\det\{\boldsymbol{I}+a_0^{-1}\mathrm{diag}(\boldsymbol{w})\boldsymbol{S}_0\}.\end{aligned}$$

Bhattacharyya Distance

Ignoring the terms that are independent of the optimization variable $\boldsymbol{w}$, we can express the Bhattacharyya distance (48) as

$$f(\boldsymbol{w}) = f_0(\boldsymbol{w}) - f_1(\boldsymbol{w}),$$

where

$$f_0(\boldsymbol{w}) := \frac{1}{2}\log\det\{\boldsymbol{I}+a_{01}^{-1}\mathrm{diag}(\boldsymbol{w})\boldsymbol{S}_{01}\}$$

and

$$\begin{aligned}f_1(\boldsymbol{w}) := \frac{1}{4}\big(&\log\det\{\boldsymbol{I}+a_0^{-1}\mathrm{diag}(\boldsymbol{w})\boldsymbol{S}_0\}\\ &+\log\det\left\{\boldsymbol{I}+a_1^{-1}\mathrm{diag}(\boldsymbol{w})\boldsymbol{S}_1\right\}\big),\end{aligned}$$

are concave functions on $\boldsymbol{w}\in[0,1]^M$.

Kullback-Leibler Distance

Recalling Property 1, where we had shown that the matrix of the form $\boldsymbol{\Phi}^T\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{w})\boldsymbol{\Phi}$ is equivalent to

$$\begin{aligned}&\boldsymbol{\Phi}^T\left(a_0\boldsymbol{I}+\boldsymbol{\Phi}\boldsymbol{S}_0\boldsymbol{\Phi}^T\right)^{-1}\boldsymbol{\Phi}\\ &\qquad = \boldsymbol{S}_0^{-1} - \boldsymbol{S}_0^{-1}\left[\boldsymbol{S}_0^{-1}+a_0^{-1}\mathrm{diag}(\boldsymbol{w})\right]^{-1}\boldsymbol{S}_0^{-1},\end{aligned}$$

we can write the first term of (49), that is,

$$\mathrm{tr}\left\{\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{w})\boldsymbol{\Phi}\boldsymbol{\Sigma}_1\boldsymbol{\Phi}^T\right\} = \mathrm{tr}\left\{\boldsymbol{\Phi}^T\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{w})\boldsymbol{\Phi}\boldsymbol{\Sigma}_1\right\}$$

as

$$\mathrm{tr}\left\{\boldsymbol{S}_0^{-1}\boldsymbol{\Sigma}_1 - \boldsymbol{S}_0^{-1}\left[\boldsymbol{S}_0^{-1}+a_0^{-1}\mathrm{diag}(\boldsymbol{w})\right]^{-1}\boldsymbol{S}_0^{-1}\boldsymbol{\Sigma}_1\right\}.$$

The above function can be expressed as a convex function in $\boldsymbol{w}$ (e.g., using the epigraph form). The second term of (49) can be relaxed to a convex function $\boldsymbol{1}^T\boldsymbol{w}$. The last term of (49), that is, $\log\det\{\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{w})\boldsymbol{\Sigma}_1(\boldsymbol{w})\}$ is equivalent to

$$\begin{aligned}&\log\det\{\boldsymbol{\Phi}\boldsymbol{\Sigma}_1\boldsymbol{\Phi}^T\} - \log\det\{\boldsymbol{\Phi}\boldsymbol{\Sigma}_0\boldsymbol{\Phi}^T\}\\ &= \log\det\left\{\boldsymbol{I}+a_1^{-1}\mathrm{diag}(\boldsymbol{w})\boldsymbol{S}_1\right\} - \log\det\left\{\boldsymbol{I}+a_0^{-1}\mathrm{diag}(\boldsymbol{w})\boldsymbol{S}_0\right\}.\end{aligned}$$

Thus, we can equivalently express (49) as $f(\boldsymbol{w}) = f_0(\boldsymbol{w}) - f_1(\boldsymbol{w})$ with

$$\begin{aligned}f_0(\boldsymbol{w}) := &-\mathrm{tr}\left\{\boldsymbol{S}_0^{-1}\left[\boldsymbol{S}_0^{-1}+a_0^{-1}\mathrm{diag}(\boldsymbol{w})\right]^{-1}\boldsymbol{S}_0^{-1}\boldsymbol{\Sigma}_1\right\}\\ &-\boldsymbol{1}^T\boldsymbol{w}+\log\det\left\{\boldsymbol{I}+a_1^{-1}\mathrm{diag}(\boldsymbol{w})\boldsymbol{S}_1\right\}\end{aligned}$$

and

$$f_1(\boldsymbol{w}) := \log\det\left\{\boldsymbol{I}+a_0^{-1}\mathrm{diag}(\boldsymbol{w})\boldsymbol{S}_0\right\},$$

which are concave in $\boldsymbol{w}$.

References

[1] S. P. Chepuri and G. Leus, "Sparse sensing for distributed Gaussian detection," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2015.

[2] A. Krause, A. Singh, and C. Guestrin, "Near-optimal sensor placements in Gaussian processes: Theory, efficient algorithms and empirical studies," *J. Mach. Learn. Res.*, vol. 9, pp. 235–284, Feb. 2008.

[3] S. Joshi and S. Boyd, "Sensor selection via convex optimization," *IEEE Trans. Signal Process.*, vol. 57, no. 2, pp. 451–462, Feb. 2009.

[4] S. P. Chepuri and G. Leus, "Sparsity-promoting sensor selection for non-linear measurement models," *IEEE Trans. Signal Process.*, vol. 63, no. 3, pp. 684–698, Feb. 2015.

[5] S. P. Chepuri and G. Leus, "Sparsity-promoting adaptive sensor selection for non-linear filtering," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2014, pp. 5100–5104.

[6] S. P. Chepuri and G. Leus, "Continuous sensor placement," *IEEE Signal Process. Lett.*, vol. 22, no. 5, pp. 544–548, May 2015.

[7] T. Grettenberg, "Signal selection in communication and radar systems," *IEEE Trans. Inf. Theory*, vol. 9, no. 4, pp. 265–275, 1963.

[8] T. Kadota and L. A. Shepp, "On the best finite set of linear observables for discriminating two Gaussian signals," *IEEE Trans. Inf. Theory*, vol. 13, no. 2, pp. 278–284, 1967.

[9] T. Kailath, "The divergence and Bhattacharyya distance measures in signal selection," *IEEE Trans. Commun. Technol.*, vol. 15, no. 1, pp. 52–60, Feb. 1967.

[10] C. Rago, P. Willett, and Y. Bar-Shalom, "Censoring sensors: A low-communication-rate scheme for distributed detection," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 32, no. 2, pp. 554–568, 1996.

[11] C.-T. Yu and P. K. Varshney, "Sampling design for Gaussian detection problems," *IEEE Trans. Signal Process.*, vol. 45, no. 9, pp. 2328–2337, 1997.

[12] D. Bajovic, B. Sinopoli, and J. Xavier, "Sensor selection for event detection in wireless sensor networks," *IEEE Trans. Signal Process.*, vol. 59, no. 10, pp. 4938–4953, Oct. 2011.

[13] J. Chamberland and V. Veeravalli, "Wireless sensors in distributed detection applications," *IEEE Signal Process. Mag.*, vol. 24, no. 3, pp. 16–25, May 2007.

[14] S. Appadwedula, V. V. Veeravalli, and D. L. Jones, "Decentralized detection with censoring sensors," *IEEE Trans. Signal Process.*, vol. 56, no. 4, pp. 1362–1373, 2008.

[15] R. S. Blum and B. M. Sadler, "Energy efficient signal detection in sensor networks using ordered transmissions," *IEEE Trans. Signal Process.*, vol. 56, no. 7, pp. 3229–3235, 2008.

[16] S. Ali and S. D. Silvey, "A general class of coefficients of divergence of one distribution from another," *J. Royal Stat. Soc. Ser. B (Methodolog.)*, pp. 131–142, 1966.

[17] S. Cambanis and E. Masry, "Sampling designs for the detection of signals in noise," *IEEE Trans. Inf. Theory*, vol. 29, no. 1, pp. 83–104, Jan. 1983.

[18] R. K. Bahr and J. A. Bucklew, "Optimal sampling schemes for the Gaussian hypothesis testing problem," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 38, no. 10, pp. 1677–1686, 1990.

[19] Z. Quan, W. J. Kaiser, and A. H. Sayed, "Innovations diffusion: A spatial sampling scheme for distributed estimation and detection," *IEEE Trans. Signal Process.*, vol. 57, no. 2, pp. 738–751, 2009.

[20] Y. Sung, L. Tong, and H. V. Poor, "A large deviations approach to sensor scheduling for detection of correlated random fields," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2005, vol. 3, pp. iii–649, IEEE.

[21] D. L. Donoho, "Compressed sensing," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[22] S. P. Chepuri and G. Leus, "Compression schemes for time-varying sparse signals," in *Conf. Rec. 46th Asilomar Conf. Signals, Syst. Comput. (ASILOMAR)*, Nov. 2014.

[23] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York, NY, USA: Wiley, 2012.

[24] H. Kobayashi and J. B. Thomas, "Distance measures and related criteria," in *Proc. 5th Ann. Allerton Conf. Circuit Syst. Theory*, 1967, pp. 491–500.

[25] S. Kullback, *Information Theory and Statistics*. New York, NY, USA: Courier Dover , 2012.

[26] C. H. Papadimitriou, *Computational Complexity*. New York, NY, USA: Wiley, 2003.

[27] T. K. Moon and W. C. Stirling, *Mathematical Methods and Algorithms for Signal Processing*. New York, NY, USA: Prentice-Hall, 2000, vol. 1.

[28] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1993.

[29] S. Boyd and L. Vandenberghe, *Convex Optimization*. New York, NY, USA: Cambridge Univ. Press, 2004.

[30] J. F. Sturm, "Using SeDuMi 1.02, a MATLAB toolbox for optimization over symmetric cones," *Optimiz. Methods Software*, vol. 11, no. 1–4, pp. 625–653, 1999.

[31] A. L. Yuille and A. Rangarajan, "The concave-convex procedure," *Neural Computat.*, vol. 15, no. 4, pp. 915–936, 2003.

[32] W. Hoeffding, "Probability inequalities for sums of bounded random variables," *J. Amer. Statist. Assoc.*, vol. 58, no. 301, pp. 13–30, 1963.

Sundeep Prabhakar Chepuri (S'11–M'16) was born in India in 1986. He received his M.Sc. degree (*cum laude*) in electrical engineering and Ph.D. degree (*cum laude*) from the Delft University of Technology, The Netherlands, in July 2011 and January 2016, respectively. He has held positions at Robert Bosch, India, during 2007–2009, and Holst Centre/imec-nl, The Netherlands, during 2010–2011. He is currently a postdoctoral scholar with the Circuits and Systems group at the Faculty of Electrical Engineering, Mathematics and Computer Science of the Delft University of Technology, The Netherlands. His general research interest lies in the field of mathematical signal processing, statistical inference, sensor networks, and wireless communications.

Dr. Chepuri received the Best Student Paper Award for his publication at the ICASSP 2015 conference in Australia.

Geert Leus (M'01–SM'05–F'12) received the electrical engineering degree and the Ph.D. degree in applied sciences from the Katholieke Universiteit Leuven, Belgium, in 1996 and 2000, respectively.

Currently, he is an "Antoni van Leeuwenhoek" Full Professor at the Faculty of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology, Delft, The Netherlands. His research interests are in the area of signal processing for communications.

Prof. Leus received the 2002 IEEE Signal Processing Society Young Author Best Paper Award and the 2005 IEEE Signal Processing Society Best Paper Award. He was the Chair of the IEEE Signal Processing for Communications and Networking Technical Committee, and an Associate Editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING, the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, and the IEEE SIGNAL PROCESSING LETTERS. Currently, he is a member of the IEEE Sensor Array and Multichannel Technical Committee and serves as the Editor-in-Chief of the EURASIP *Journal on Advances in Signal Processing*.