Can we break the dependency in distributed detection?

Osama A. Hanna†, Xinlin Li†, Christina Fragouli† and Suhas Diggavi†
†University of California, Los Angeles
Email:{ohanna, xinlinli, christina.fragouli, suhasdiggavi}@ucla.edu

2022 IEEE International Symposium on Information Theory (ISIT) | 978-1-6654-2159-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/ISIT50566.2022.9834790

***Abstract*—We consider a distributed detection problem where sensors observe dependent observations. We ask, if we can allow the sensors to locally exchange a few bits with each other, whether we can use these bits to "break" the dependency of the sensor observations, and thus reduce the dependent detection problem to the much better-studied and understood case of conditionally independent observations. To this end, we propose an optimization problem that we prove is equivalent to minimizing the dependency between the sensor observations. This problem is in general NP-hard, however, we show that for at least some cases of Gaussian distributions it can be solved efficiently. For general distributions, we propose to use alternating minimization and derive a constant factor approximation algorithm. Numerical evaluations indicate that our approach can offer significant improvement in detection accuracy over alternative schemes.**

I. Introduction

Consider a set of distributed sensor nodes that collect real value observations $\mathbf{x}_i$ and communicate some discrete function of their observations to a central node (fusion center). We assume that the $\mathbf{x}_i$ are drawn from arbitrary distributions determined by an environment state C that can take one out of a finite number of values, and that these distributions are known in advance to the sensors and to the fusion center. The central node aims to decide the state C of the environment that generates these observations.

For the case where the observations $\mathbf{x}_i$ are independent (conditioned on the state value C), there already exists a rich literature that has explored a number of variations of this problem, and optimal or near-optimal performance bounds and algorithms have been derived. For two states, our setup can be viewed as a binary hypothesis testing problem in distributed detection, see for instance the seminal work by Tsitsiklis [1] and numerous followup work and references, such as [2]–[10]; it can also be viewed as a distributed quantization problem for classification, see for instance [11].

However, in many practical cases, the sensors would collect dependent observations. For example, when road sensors are used for vehicle identification, the size, weight, and noise a vehicle makes are correlated; similarly, when classifying animal or plant species, the features can be correlated. This has been recognized in the literature, and some schemes for dependent observations have been proposed [2], [7], [12]–[17].

This work was supported in part by NSF grant 2007714 and by Army Research Laboratory under Cooperative Agreement W911NF-17-2-0196.

In this paper, we ask the following question: if we allow the sensors to locally exchange a few bits with each other, can we use these bits to "break" the dependency of the sensor observations, and thus reduce our problem to the well-studied and understood case of independent observations? If we can do so, we can then re-use all the existing (and future) algorithms for the independent case in the literature. Note that sensors that observe the same environment would be located in a proximal geographical location and thus directly communicating with each other might not be a hard to meet requirement.

Contributions. Our main contributions include:

- We formulate an optimization problem and prove an equivalence result: it is possible to break the dependency if and only if our objective function achieves zero value.
- We prove that solving this problem (and as a result, finding what bits to exchange to break the dependency between distributed features) is in general NP-hard.
- We prove what we think is a surprising result: for (at least some) cases of Gaussian distributions this problem can be solved efficiently.
- For general distributions, we propose using alternating minimization over the functions each sensor implements; we prove that the resulting simplified problem is still NP-hard, however, we can design an approximation algorithm that offers constant approximation guarantees.
- We numerically evaluate our proposed algorithms and show significant accuracy improvements over alternative approaches.

Related Work. There is a large body of literature devoted to the design of distributed detection systems, see for example [1]–[10], [12]–[26]. For conditionally independent observations and binary hypothesis testing, quantizing the likelihood ratio is proved to be optimal [1]. However, this is no longer the case when there are dependencies among the observations [27]. As shown by Tsitsiklis and Athans [28], the distributed detection problem with dependent observations is NP-complete. The work in [7] modifies the detection rule at the fusion center to account for correlation. However, these works do not consider modifying the quantization scheme at the sensor observation to take correlation into account. The work in [29], [30] assumes that a known hidden variable induces conditional independence among sensor observations, based on which they propose a new framework that unifies distributed detection with dependent and independent observations, in both parallel fusion [29] and tandem networks [30]. However, finding such a hidden variable

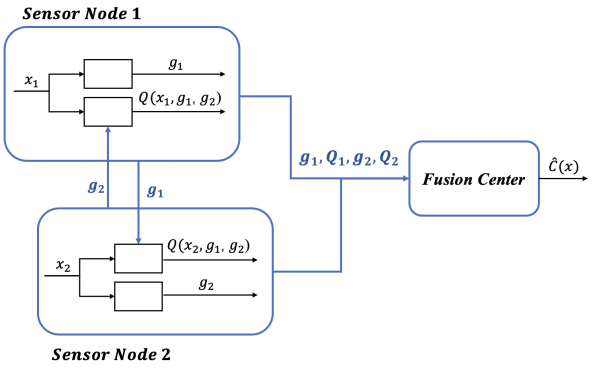

Fig. 1. System overview for the case of two sensors.

remains challenging.

To the best of our knowledge, ours is the first work to explore whether cooperation, by sharing a function of the local observations using a few bits between the sensor nodes, can introduce conditional independence.

II. System Model and Objective

In the following, we denote the Probability Density Function (PDF) of a random vector $\mathbf{x}$ by $f_{\mathbf{X}}$[1].

Classical system model. We start from the classical communication-constrained distributed hypothesis testing setup [1]: a fusion center needs to make a decision on one of M hypothesis $C_1, ..., C_M$, leveraging the observations of N distributed sensors $\mathbf{x} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N]$, where $\mathbf{x}_i \in \mathbb{R}^{n_i}$ is the observation of node i, as depicted in Fig. 1. The true hypothesis C is a random variable that takes one of the values $C_1, ..., C_M$, with probability $f(C_1), ..., f(C_M)$, respectively. The distribution of $(\mathbf{x}, C)$ is assumed to be known to all nodes in the system. The goal is, to design compression schemes at the sensor nodes that only need a few bits to convey the $\mathbf{x}_i$'s while maintaining a good testing accuracy (probability of making a correct decision) at the fusion center. For the case where the observations $\mathbf{x}_1, ..., \mathbf{x}_N$ are conditionally independent given the hypothesis C, it has been shown that it is optimal to quantize a function of $\{f_{\mathbf{X}_i|C}(\mathbf{x}_i|C_j)\}_{j=1}^{M}$ [1].

Our system model. In this paper, we consider the case where the conditional independence assumption does not hold. Moreover, we change the classical system model as follows. We assume that the sensor nodes are allowed to have a *single message* interaction. Namely, each sensor node i can calculate a low bit representation $g_i(\mathbf{x}_i)$ and share it with the other sensor nodes and the fusion center, as depicted in Fig. 1.

Objective. The main question we want to explore is, whether and how we can use the functions $g(\mathbf{x}) = [g_1(\mathbf{x}_1), ..., g_N(\mathbf{x}_N)]$ to break the dependency between the features collected at the N sensors. Clearly, if $g_i(\mathbf{x}_i) = \mathbf{x}_i$, this can trivially be achieved. We are interested in the communication constrained case, where each g_i needs to be represented with (a small number of) R_i bits. That is, we ask whether there exists g with range $|ran(g_i)| \leq 2^{R_i}$, such that conditioned on C and $g(\mathbf{x})$, the observations $\mathbf{x}_1, ..., \mathbf{x}_N$ are conditionally independent. If we can achieve this, we can then re-use any of the existing algorithms for the uncorrelated case in the literature, to quantize at the sensor nodes and take a decision in the fusion center.

In this paper we mostly focus on binary hypothesis testing with N sensor observations (yet our approach can be extended to more general cases). For the purpose of evaluation, beyond the communication using g, we will use for the conditionally independent parts of the message the same quantization as in [1]. Namely, for two hypothesis, each sensor i quantizes and shares with the fusion center the likelihood ratio $\frac{f(\mathbf{x}_i|g(\mathbf{x}),C_1)}{f(\mathbf{x}_i|g(\mathbf{x}),C_2)}$. The quantization is performed by using simulations to tune a set of thresholds that divide $[0, \infty)$ into quantization regions. The fusion center uses the maximum a posteriori probability rule to make the hypothesis decision

$$\hat{C} = \arg\max_{j=1,2} f(C_j|\{Q_i(\frac{f(\mathbf{x}_i|g(\mathbf{x}), C_1)}{f(\mathbf{x}_i|g(\mathbf{x}), C_2)})\}_{i=1}^{N}, g(\mathbf{x})). \quad (1)$$

We emphasize that our goal is the design of the functions g; we simply evaluate the achieved performance using g together with the above off-the-shelf functions designed for independent observations.

III. Optimization Problem and Equivalence

Optimization problem formulation. To translate our objective to an optimization problem, we will use a function $P(g)$ that measures the conditional dependency between $\mathbf{x}_1, ..., \mathbf{x}_N$ given $\{C, g\}$. In particular, we want that $P(g) = 0$ if and only if $\mathbf{x}_1, ..., \mathbf{x}_N$ are conditionally independent given $\{C, g\}$; and $P(g)$ takes large values if the dependence between $\mathbf{x}_1, ..., \mathbf{x}_N$ given $\{C, g\}$ is high. We will discuss how to design such a P next. Then, as we will formally prove, achieving our objective reduces to solving the following optimization problem

$$\mathcal{P}_1 : \min_{g:|ran(g_i)|\leq 2^{R_i}, i=1,...,N} P(g), \quad (2)$$

where R_i is the number of bits used to represent $g_i(\mathbf{x}_i)$, and $ran(g_i)$ is the range of the function g_i.

Choice of P. We here propose a method to select the objective function $P(g)$ for the optimization problem $\mathcal{P}_1$ in (2).

Recall that, for each of the M values of the hypothesis C, we know the conditional distribution of $\mathbf{x}$; what we propose to do is, for each such distribution solve a different optimization problem that selects for that particular hypothesis, a function g; note that these M optimization problems can be solved offline in advance (e.g., at the fusion center). Then, at the time of operation, since each sensor will not know the actual hypothesis, it will apply the corresponding g for all M hypothesis and convey these values to the other sensor nodes and the fusion center; conditioned on the observed g and a hypothesis C_j, the fusion center will see nearly conditionally independent feature observartions. In the following we discuss how to select the objective function *given a specific hypothesis* (distribution). To simplify notation, we will continue talking about selecting values for g, with the understanding that these values are for a particular hypothesis.

[1]We abuse notation and use f_X to denote the Probability Mass Function for discrete random variables and the PDF for continuous random variables. We also omit the subscript when it can be implicitly understood.

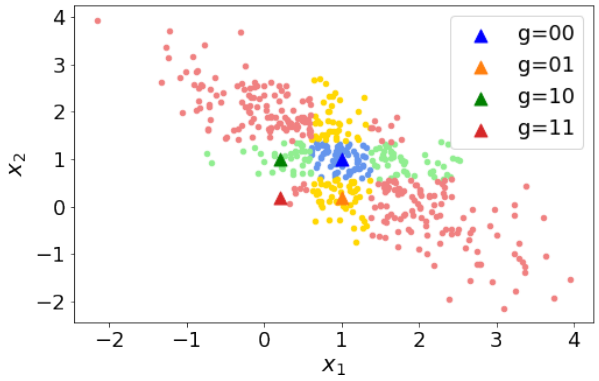

Fig. 2. Example of representatives points for g using 2 bits.

For a given distribution, the function $g(\mathbf{x}) = [g_1(\mathbf{x}_1), ..., g_N(\mathbf{x}_N)]$ would act as a type of quantizer for $\mathbf{x}$. In the example of Fig. 2 we plot the distribution for a specific hypothesis (we assume two sensors each observing a scalar value x_i); all observed points with the same color are mapped to the same g value.

We now select a function $d(\mathbf{x}, \mathbf{y})$, that we will use to "measure the dependency" between two different observations $\mathbf{x}$ and $\mathbf{y}$. We would like this function to satisfy the "dependency symmetrical" properties we define next.

Definition 1. *We say that* $d : (\mathbf{R}^{n_1} \times ... \times \mathbf{R}^{n_N})^2 \to \mathbf{R}$ *is* ***Dependency Symmetrical (DS)*** *if and only if for any* $\mathbf{x}$ *and* $\mathbf{y}$

1) $d(\mathbf{x}, \mathbf{y}) \geq 0$.
2) $d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x})$.
3) $d(\mathbf{x}, \mathbf{y}) = 0$ *if and only if*

$$\begin{aligned} &f(\mathbf{x}_1,\mathbf{x}_2,...,\mathbf{x}_N)f(\mathbf{y}_1,\mathbf{y}_2,...,\mathbf{y}_N) \\ &=f(b_1\mathbf{x}_1+(1-b_1)\mathbf{y}_1,...,b_N\mathbf{x}_N+(1-b_N)\mathbf{y}_N) \\ &\quad f(b_1\mathbf{y}_1+(1-b_1)\mathbf{x}_1,...,b_N\mathbf{y}_N+(1-b_N)\mathbf{x}_N) \\ &\qquad \forall \mathbf{b}=(b_1,...,b_N)\in\{0,1\}^N. \quad (3) \end{aligned}$$

For $N = 2$, examples of such functions include $|f(\mathbf{x}_1, \mathbf{x}_2)f(\mathbf{y}_1, \mathbf{y}_2) - f(\mathbf{x}_1, \mathbf{y}_2)f(\mathbf{y}_1, \mathbf{x}_2)|$, $|\frac{f(\mathbf{x}_1|\mathbf{x}_2)}{f(\mathbf{x}_1|\mathbf{y}_2)} - \frac{f(\mathbf{y}_1|\mathbf{x}_2)}{f(\mathbf{y}_1|\mathbf{y}_2)}| + |\frac{f(\mathbf{x}_1|\mathbf{y}_2)}{f(\mathbf{x}_1|\mathbf{x}_2)} - \frac{f(\mathbf{y}_1|\mathbf{y}_2)}{f(\mathbf{y}_1|\mathbf{x}_2)}|$; these choices can be generalized for $N > 2$.

Given a choice for d, we define $P(g)$ as

$$P(g) = \max_{\mathbf{x},\mathbf{y}:g(\mathbf{x})=g(\mathbf{y})} d(\mathbf{x}, \mathbf{y}). \quad (4)$$

The intuition for this choice is as follows. As mentioned earlier, if g mapped each $\mathbf{x}$ to a different point (thus perfectly conveyed $\mathbf{x}$) we would have conditional independence; however, this can require a large, (in theory infinite), number of bits. To use a small number of bits we want g to map as many different observations as possible to the same value while still achieving conditional independence. As we show in Theorem 1, we approach conditional independence as $d(\mathbf{x}, \mathbf{y})$ is made small for each pair of points $\mathbf{x}, \mathbf{y}$ that are mapped to the same value by g. Thus, the optimization problem we propose to solve is:

$$\mathcal{P}_2 : \min_{\substack{g:|ran(g_i)|\leq 2^{R_i}, \\ i=1,...N}} \max_{\mathbf{x},\mathbf{y}:g(\mathbf{x})=g(\mathbf{y})} d(\mathbf{x}, \mathbf{y}), \quad (5)$$

for any DS (see Definition 1) function d.

Equivalence. The next theorem formally proves the equivalence between finding g to break the dependency between $\mathbf{x}_1, ..., \mathbf{x}_N$ given g, C and solving (5). We use the notation $\mathbf{x}_{-i} = [\mathbf{x}_1, ..., \mathbf{x}_{i-1}, \mathbf{x}_{i+1}, ..., \mathbf{x}_N]$ and $g_{-i}(\mathbf{x}_{-i}) = [g_1(\mathbf{x}_1), ..., g_{i-1}(\mathbf{x}_{i-1}), g_{i+1}(\mathbf{x}_{i+1}), ..., g_N(\mathbf{x}_N)]$.

Theorem 1. *Suppose that* $p_2^\star$ *is the optimal value of the optimization problem* $\mathcal{P}_2$*, then* $p_2^\star = 0$ *and* g *is an optimal solution of* $\mathcal{P}_2$ *if and only if:*
$(*)$ $|ran(g_i)| \leq 2^{R_i}$ *and* $\mathbf{x}_1, ..., \mathbf{x}_N$ *are conditionally independent given* $g(\mathbf{x}), C$.

Proof. We here provide a proof outline; the proof details are in [31]. Suppose that $(*)$ holds. From the conditional independence, after a few steps of algebra, we have that

$$\frac{f(\mathbf{x}_{-i}|\mathbf{x}_i)}{f(\mathbf{x}_{-i}|\mathbf{y}_i)} = \frac{f(g_{-i}(\mathbf{x}_{-i})|\mathbf{x}_i)}{f(g_{-i}(\mathbf{x}_{-i})|\mathbf{y}_i)} \forall \mathbf{x}_{-i}, \mathbf{x}_i, \mathbf{y}_i : g_i(\mathbf{x}_i) = g_i(\mathbf{y}_i).$$

Since the right hand side only depends on $g(\mathbf{x_{-i}}), \mathbf{x}_i, \mathbf{y}_i$, we have that $\forall\ \mathbf{x}, \mathbf{y} : g(\mathbf{x}) = g(\mathbf{y})$,

$$\frac{f(\mathbf{x}_i)}{f(\mathbf{y}_i)} \cdot \frac{f(\mathbf{x}_{-i}|\mathbf{x}_i)}{f(\mathbf{x}_{-i}|\mathbf{y}_i)} = \frac{f(\mathbf{y}_{-i}|\mathbf{x}_i)}{f(\mathbf{y}_{-i}|\mathbf{y}_i)} \frac{f(\mathbf{x}_i)}{f(\mathbf{y}_i)},$$

and thus $f(\mathbf{x}_{-i}, \mathbf{x}_i)f(\mathbf{y}_{-i}, \mathbf{y}_i) = f(\mathbf{x}_{-i}, \mathbf{y}_i)f(\mathbf{y}_{-i}, \mathbf{x}_i)$. As i was arbitrary, we get that $\forall\ \mathbf{x}, \mathbf{y} : g(\mathbf{x}) = g(\mathbf{y})$, (3) holds. Then, from (2), we have that $d(\mathbf{x}, \mathbf{y}) = 0\ \forall\ \mathbf{x}, \mathbf{y} : g(\mathbf{x}) = g(\mathbf{y})$, hence, g achieves $p^\star = 0$.

Conversely, assume that $p^\star = 0$. To show that $\mathbf{x}_1, ..., \mathbf{x}_N$ are conditionally independent given $g(\mathbf{x}), C$, it suffices to show that $f(\mathbf{x}_i|g(\mathbf{x}), \mathbf{x_{-i}}) = f(\mathbf{x}_i|g(\mathbf{x}), \mathbf{y}_{-i}) \forall \mathbf{x}_i, \mathbf{x}_{-i}, \mathbf{y}_{-i} : g_{-i}(\mathbf{x}_{-i}) = g_{-i}(\mathbf{y}_{-i})$, see [31] for these steps. □

Observation. While according to Theorem 1, all choices of d that satisfy the required properties result in the same solution if $p^\star = 0$, they might behave differently for $p^\star > 0$. For example, the function $|\frac{f(\mathbf{x}_1|\mathbf{x}_2)}{f(\mathbf{x}_1|\mathbf{y}_2)} - \frac{f(\mathbf{y}_1|\mathbf{x}_2)}{f(\mathbf{y}_1|\mathbf{y}_2)}| + |\frac{f(\mathbf{x}_1|\mathbf{y}_2)}{f(\mathbf{x}_1|\mathbf{x}_2)} - \frac{f(\mathbf{y}_1|\mathbf{y}_2)}{f(\mathbf{y}_1|\mathbf{x}_2)}|$ cares about reducing the dependency between points with very low densities as it only depends on the ratio between densities, while the function $|f(\mathbf{x}_1, \mathbf{x}_2)f(\mathbf{y}_1, \mathbf{y}_2) - f(\mathbf{x}_1, \mathbf{y}_2)f(\mathbf{y}_1, \mathbf{x}_2)|$ has a small value if all the four density values are small.

Hardness of $\mathcal{P}_2$. The main difficulty in solving $\mathcal{P}_2$ is that the number of optimizing variables can be infinite, e.g., for continuous distributions. Even for discrete distributions the next Theorem 2 proves that $\mathcal{P}_2$ is NP-hard. We consider the case where $N = 2$ and a discrete distribution for $\mathbf{x}_1, \mathbf{x}_2|C$, hence, the size required to store the problem instance in memory is linear in the number of mass points of the distribution. Therefore, there is no polynomial time (as a function of the number of mass points of the distribution) algorithm that can solve $\mathcal{P}_2$ unless P=NP. The proof reduces the graph coloring problem to $\mathcal{P}_2$ and is provided in [31].

Theorem 2. *Problem $\mathcal{P}_2$ is NP-hard. Moreover, even given (optimal) g_{-i}, finding the optimal g_i is NP-hard, where $g_{-i} = [g_1, ..., g_{i-1}, g_{i+1}, ..., g_N]$.*

IV. Gaussian distributions for two sensors.

In this section we prove what we think is a surprising result: if the conditional distributions are Gaussian (a common enough assumption in the literature [2], [14], [27], [32]) it is (at least in the case we examined) possible to efficiently identify the optimal g that minimizes, or possibly eliminates (if such a g exists), the sensor observations dependency. In particular, we prove the next Theorem 3; our proof is constructive providing a method to solve $\mathcal{P}_2$ for Gaussian distributions. For simplicity we consider a Gaussian distribution with zero mean, and unit variance. For the distance $d(\mathbf{x}, \mathbf{y})$, we use a simplified version of the function $|f(\mathbf{x}_1, \mathbf{x}_2)f(\mathbf{y}_1, \mathbf{y}_2) - f(\mathbf{x}_1, \mathbf{y}_2)f(\mathbf{y}_1, \mathbf{x}_2)|$, namely

$$d(\mathbf{x}, \mathbf{y}) = e^{-x_1^2 - y_1^2 - x_2^2 - y_2^2 + 2\rho \log|(x_1 - y_1)(x_2 - y_2)|},$$

where ρ is the covariance of $x_1, x_2 | C_1$. It can be verified that $d(\mathbf{x}, \mathbf{y})$ satisfies the DS properties in Def. 1.

Theorem 3. *Consider binary hypothesis testing with two sensors observing scalar values x_1 and x_2. Assume the conditional distribution of x_1 and x_2 given C_i is a Gaussian distribution with zero mean, unit variance, and covariance ρ^2. Then there is an algorithm with computational complexity $O(2^{R_i} \log(\frac{1}{\epsilon}))$ that outputs a solution for $\mathcal{P}_2$ with objective value $p \leq p_2^\star + \epsilon$, where $p_2^\star$ is the optimal solution of $\mathcal{P}_2$.*

Proof. The detailed proof can be found in [31]; we next provide a proof outline. We focus on breaking the dependency for hypothesis 1, as breaking the dependency for 2 will be similar. By monotonicity of $\log$, optimizing over $\log(d(\mathbf{x}, \mathbf{y}))$ will give the same optimal point as optimizing over $d(\mathbf{x}, \mathbf{y})$. It is also easy to see that the optimization over the pair (x_1, y_1) and the pair (x_2, y_2) can be separated. Hence, we only consider the optimization over (x_1, y_1) as the optimization over (x_2, y_2) can be done similarly. In the following we focus on finding g_1, hence, we omit the subscript in g_1, x_1, y_1, R_1. Therefore, our problem reduces to

$$\mathcal{P}_3 : \min_{g:|ran(g)| \leq 2^R} \max_{x,y \in \mathbb{R}: g(x) = g(y)} h(x, y), \tag{6}$$

where $h(x, y) = -x^2 - y^2 + 2\rho \log|x - y|$.

In [31] we prove that the optimal $g_\star$ divides $\mathbb{R}^+ \cup \{0\}$ in 2^{R-1} quantization intervals (and because of symmetry the mirrored ones for $\mathbb{R}^- \cup \{0\}$); the value $g_\star(x)$ is simply the label for the interval x belongs in. To identify $g_\star$ we thus need to identify the $a_i \in R$ thresholds for these intervals, for $i = 1 \dots 2^{R-1}$. We prove in [31] that these can be found by solving the following system of equations

$$\log(p_2^\star) = h(0, a_2) = h(a_2, a_3) = ... = h(a_{2^{R-1}}, y_{a_{2^{R-1}}}),$$

where $p_2^\star$ is the optimal value of $\mathcal{P}_2$. These can be solved efficiently numerically (e.g., by using a bijection method). □

[2]The results in this section can be easily generalized for general mean and variances.

V. Approximation Algorithm for $\mathcal{P}_2$

For general distributions, we propose to solve the problem $\mathcal{P}_2$ by using alternating minimization over $g_1, ..., g_N$. Namely, we initialize g in some way (it can be a random initialization or a good guess; the initialization method we used in our simulations is provided in Section VI). We then proceed in iterations, where in each iteration, and for $i = 1 \dots N$, for the given choice of g_{-i} we find the associated optimal g_i and update it. We iterate until the solution does not improve by more than a threshold or we reach an upper bound of iterations. Note that alternating minimization may not find the global optimal g, but is guaranteed to reach a local optimal point. The complexity of the proposed algorithm is $O(2^{2R} m^2)$ for R bits and a discrete distribution with m mass points, and $O(2^{2R})$ for continuous distributions provided that the unconstrained maximization of $d(\mathbf{x}, \mathbf{y})$ over $\mathbf{x}$ for a given $\mathbf{y}$ can be performed in constant time.

With this approach, our problem is now reduced to:
Given $g_{-i} = [g_1, ..., g_{i-1}, g_{i+1}, ..., g_N]$, find the optimal g_i.
Unfortunately, as we proved in Theorem 2, this is still an NP-hard problem; our main contribution in this section is to construct an approximation algorithm that we describe next and to prove that it has constant multiplicative guarantees.

We start by defining

$$d_{g_i}(\mathbf{x}_i, \mathbf{y}_i) = \max_{\mathbf{x}_{-i}, \mathbf{y}_{-i}: g_{-i}(\mathbf{x}_{-i}) = g_{-i}(\mathbf{y}_{-i})} d(\mathbf{x}, \mathbf{y}).$$

Note that: $P(g) = \max_{\mathbf{x}_i, \mathbf{y}_i: g_i(\mathbf{x}_i) = g_i(\mathbf{y}_i)} d_{g_i}(\mathbf{x}_i, \mathbf{y}_i)$.
Our algorithm leverages the fact that since $d_{g_i}(\mathbf{x}_i, \mathbf{y}_i)$ considers the worst case $\mathbf{x}_{-i}, \mathbf{y}_{-i}$, then, if for any z_i, the values $d_{g_i}(\mathbf{x}_i, \mathbf{z}_i)$, $d_{g_i}(\mathbf{z}_i, \mathbf{y}_i)$ are small, $d_{g_i}(\mathbf{x}_i, \mathbf{y}_i)$ is also small[3]. More formally, we define a triangle-inequality like property:

Definition 2. *We say that $d : (\mathbf{R}^{n_1} \times ... \times \mathbf{R}^{n_N})^2 \to \mathbf{R}$ satisfies the **c-max property** if and only if for all $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{z}$, there is a universal constant c such that*

$$d(\mathbf{x}, \mathbf{y}) \leq c \max\{d(\mathbf{x}, (\mathbf{z}_i, \mathbf{y}_{-i})), d((\mathbf{z}_i, \mathbf{x}_{-i}), \mathbf{y})\}. \tag{7}$$

For example, for $N = 2$, it is easy to see that the function

$$d(\mathbf{x}, \mathbf{y}) = \frac{|f(\mathbf{x}_1, \mathbf{x}_2)f(\mathbf{y}_1, \mathbf{y}_2) - f(\mathbf{x}_1, \mathbf{y}_2)f(\mathbf{y}_1, \mathbf{x}_2)|}{\max_{\mathbf{x}} f(\mathbf{x})^2} \tag{8}$$

satisfies this property with $c = 2$.

Now, to design g_i, the algorithm selects a set of 2^{R_i} representative points $\mathbf{x}_i^{(k)}$ (and labels them as points $1, ..., 2^{R_i}$) such that every other $\mathbf{x}_i$ point in the space is close to one of the representatives in the d_{g_i} sense; in other words we find a set of $\{\mathbf{x}_i^{(k)}\}$ that "cover" the space of $\mathbf{x}_i$ in the sense of d_{g_i} (Fig. 2 visually shows this). The function $g_i(\mathbf{x}_i)$ simply conveys which of the 2^{R_i} representative points is closer to the observed $\mathbf{x}_i$. The pseudo-code is provided in Algorithm 1.

The following theorem proves that Algorithm 1 is within a factor c from the optimal.

[3]It can be shown that this property (and the **c-max property**) holds for any $d(\mathbf{x}, \mathbf{y})$ that is an increasing function of $|f(\mathbf{x})f(\mathbf{y}) - f(b_1\mathbf{x}_1 + (1 - b_1)\mathbf{y}_1, ...)f(b_1\mathbf{y}_1 + (1 - b_1)\mathbf{x}_1, ...)|, b_i \in \{0, 1\}$ with bounded derivative.

Algorithm 1 Finding g_i given g_{-i}

- Set $\mathbf{x}_i^{(1)} = \mathbb{E}[\mathbf{x}_i|C]$ and $g_i(\mathbf{x}_i^{(1)}) = 1$.
- For $k = 2, ..., 2^{R_i}$
  - Set $g(\mathbf{x}_i^{(k)}) = k$ with

$$\mathbf{x}_i^{(k)} = \arg \max_{\mathbf{x}_i \in \mathbb{R}^{n_i}} \min_{j=1,...,k-1} d_{g_i}(\mathbf{x}_i, \mathbf{x}_i^{(j)}). \quad (9)$$

- For any $\mathbf{x}_i \in \mathbb{R}^{n_i}$, set

$$g_i(\mathbf{x}_i) = \arg \min_{k=1,...,d} d_{g_i}(\mathbf{x}_i, \mathbf{x}_i^{(k)}). \quad (10)$$

Theorem 4. *Given a function d that is **DS** and c-max and functions g_{-i}, Algorithm 1 reaches a solution for g_i that is within a factor of c from the optimal solution.*

Proof. The proof, provided in [31], relies on the following ideas. As in each iteration, Alg. 1 finds the furthest point in d from its closest representative (and sets it as a new representative), once it terminates we have that for any non-representative point $\mathbf{x}_i$, $d(\mathbf{x}_i, \mathbf{x}_i^{(r)}) \leq d(\mathbf{x}_i^{(j)}, \mathbf{x}_i^{(k)})\ \forall j, k$, where $\mathbf{x}_i^{(r)}$ is the closest representative to $\mathbf{x}_i$. If the optimal solution differs from our solution, then the optimal g maps two different representative points to the same value (say $d(\mathbf{x}_i^{(j)}, \mathbf{x}_i^{(k)})$). Then

$$\begin{aligned} p = d(\mathbf{x}_i, \mathbf{y}_i) &\overset{(i)}{\leq} c \max\{d(\mathbf{x}_i, \mathbf{x}_i^{(r)}), d(\mathbf{y}_i, \mathbf{x}_i^{(r)})\} \\ &\leq c d(\mathbf{x}_i^{(j)}, \mathbf{x}_i^{(k)}) \leq c p_2^\star, \end{aligned} \quad (11)$$

where (i) can be shown to follow from the **c-max property**, p is the objective value achieved by Algorithm 1 and $\mathbf{x}_i, \mathbf{y}_i$ are the two points with the largest distance that are mapped to the same representative $\mathbf{x}_i^{(r)}$. □

VI. Numerical Evaluation

We numerically evaluate the performance of the approximation algorithm (in Section V) and compare it against baselines.

Setup. We assume binary hypothesis testing with $N = 2, n_1 = n_2 = 1$, and hypothesis (C_1, C_2), where conditioned on C_i the sensor observations follow a multivariate Gaussian distribution $N(\mu_j, \Sigma_j)$. As proved in [33], the performance gap between the optimal detector and one that assumes conditional independence appears when the covariance matrices Σ_1 and Σ_2 differ. With the intention to evaluate the improvement in bridging the gap, we set distribution parameters as $\mu_1 = (0, 0)$, $\mu_2 = (0.8, 0.8)$, $\Sigma_1 = \begin{bmatrix} 0.5 & 0 \\ 0 & 0.5 \end{bmatrix}$, and $\Sigma_2 = \begin{bmatrix} 1 & -0.8 \\ -0.8 & 1 \end{bmatrix}$. We compare our proposed scheme with the following baselines:

- **Baseline: Counting Rule [32]:** Each distributed node quantizes observations using B_i bits by comparing the log-likelihood ratio with $2^{B_i} - 1$ thresholds that are fine-tuned through simulations. Upon receiving the quantized observations, the fusion center forms a bit vector, and makes decision based on the *type* of this bit vector, see [32] for details.
- **Baseline: LQ Detector [7]:** Each distributed node quantizes observations as in **Counting Rule**. However, the fusion rule uses knowledge of the low order statistics of the quantized observations (in particular, the dependence), by optimizing a deflection metric in the class of linear-quadratic detectors, and comparing the metric with a threshold.
- **Our method:** We use the distance measure $d(\mathbf{x}, \mathbf{y})$ in (8). We compute the value of g_1 and g_2 following Algorithm 1, then quantize the ratio $\frac{f(\mathbf{x}_i|C_1)}{f(\mathbf{x}_i|g(\mathbf{x}),C_2)}$[4] using a set of thresholds that are fine tuned through simulations. For observation x_i, we use $R_i = 1$ bit to communicate the value of g_i for C_2 and $B_i - R_i$ bits to quantize the ratio described above. To design g, we apply the proposed algorithm in Section V, where g_2 is initialized by dividing the space of $\mathbf{x}_2$ into 2^{R_2} regions with equal probability. The optimization problem in (9) is solved by restricting the optimization space to be a set of 100 samples drawn from the distribution of the observations. At the central node, the same rule as in "Counting Rule" is applied.

Plots. Fig 3 plots the detection accuracy vs. communication cost B_i. We also plot the accuracy (89.5%) when using unquantized observations, which offers an upper limit on the performance we can achieve. Our scheme starts from 2 bits as we need one bit for g. However, it offers better performance for 2 bits and reaches 87.85% accuracy using 3 bits per observation, while the baselines do not improve beyond 84.58%.

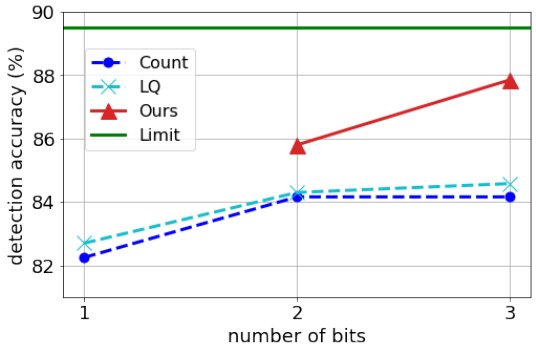

Fig. 3. Detection accuracy vs. number of bits per observation B_i

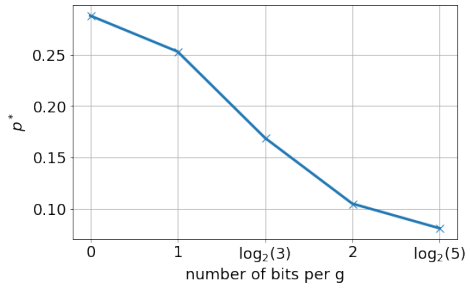

Fig. 4. Dependency decrease as we increase the number of bits for g.

Fig 4 plots p^*, the solution of $\mathcal{P}_2$ in (5) Section III, as a function of R_i, the number of bits that we use for g. We find that as we increase the range of g, our algorithm continuously decreases the dependency between x_1 and x_2, as measured by the distance $d(\mathbf{x}, \mathbf{y})$ in (8).

[4]As the sensor observations are conditionally independent given hypothesis 1, we only compute and communicate g for hypothesis 2.

REFERENCES

[1] J. N. Tsitsiklis *et al.*, "Decentralized detection," 1989.

[2] P. Willett, P. F. Swaszek, and R. S. Blum, "The good, bad and ugly: distributed detection of a known signal in dependent gaussian noise," *IEEE Transactions on signal processing*, vol. 48, no. 12, pp. 3266–3279, 2000.

[3] Q. Yan and R. S. Blum, "Distributed signal detection under the neyman-pearson criterion," *IEEE Transactions on Information Theory*, vol. 47, no. 4, pp. 1368–1377, 2001.

[4] J.-F. Chamberland and V. V. Veeravalli, "Decentralized detection in sensor networks," *IEEE transactions on signal processing*, vol. 51, no. 2, pp. 407–416, 2003.

[5] ——, "Asymptotic results for decentralized detection in power constrained wireless sensor networks," *IEEE Journal on Selected Areas in Communications*, vol. 22, pp. 1007–1015, 2004.

[6] X. Luo, M. Dong, and Y. Huang, "On distributed fault-tolerant detection in wireless sensor networks," *IEEE Transactions on computers*, vol. 55, no. 1, pp. 58–70, 2005.

[7] J. Unnikrishnan and V. V. Veeravalli, "Decentralized detection with correlated observations," in *2007 Conference Record of the Forty-First Asilomar Conference on Signals, Systems and Computers*. IEEE, 2007, pp. 381–385.

[8] V. V. Veeravalli and P. K. Varshney, "Distributed inference in wireless sensor networks," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 370, no. 1958, pp. 100–117, 2012.

[9] S. Barbarossa, S. Sardellitti, and P. Di Lorenzo, "Distributed detection and estimation in wireless sensor networks," in *Academic Press Library in Signal Processing*. Elsevier, 2014, vol. 2, pp. 329–408.

[10] S. Shahrampour, A. Rakhlin, and A. Jadbabaie, "Distributed detection: Finite-time analysis and impact of network topology," *IEEE Transactions on Automatic Control*, vol. 61, no. 11, pp. 3256–3268, 2015.

[11] O. A. Hanna, Y. H. Ezzeldin, T. Sadjadpour, C. Fragouli, and S. Diggavi, "On distributed quantization for classification," *IEEE Journal on Selected Areas in Information Theory*, vol. 1, no. 1, pp. 237–249, 2020.

[12] Z.-B. Tang, K. Pattipati, and D. Kleinman, "A distributed m-ary hypothesis testing problem with correlated observations," in *Proceedings of the 28th IEEE Conference on Decision and Control,*. IEEE, 1989, pp. 562–568.

[13] E. Drakopoulos and C.-C. Lee, "Optimum multisensor fusion of correlated local decisions," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 27, no. 4, pp. 593–606, 1991.

[14] P.-N. Chen and A. Papamarcou, "Likelihood ratio partitions for distributed signal detection in correlated gaussian noise," in *Proceedings of 1995 IEEE International Symposium on Information Theory*. IEEE, 1995, p. 118.

[15] T. Bréhard and V. Krishnamurthy, "Optimal data incest removal in bayesian decentralized estimation over a sensor network," in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, vol. 3. IEEE, 2007, pp. III–173.

[16] D. Franken and A. Hupper, "Improved fast covariance intersection for distributed data fusion," in *2005 7th International Conference on Information Fusion*, vol. 1. IEEE, 2005, pp. 7–pp.

[17] J.-F. Chamberland and V. V. Veeravalli, "How dense should a sensor network be for detection with correlated observations?" *IEEE Transactions on Information Theory*, vol. 52, no. 11, pp. 5099–5106, 2006.

[18] Z. Chair and P. Varshney, "Optimal data fusion in multiple sensor detection systems," *IEEE Transactions on Aerospace and Electronic Systems*, no. 1, pp. 98–101, 1986.

[19] J. N. Tsitsiklis, "Decentralized detection by a large number of sensors," *Mathematics of Control, Signals and Systems*, vol. 1, no. 2, pp. 167–182, 1988.

[20] M. Kam, Q. Zhu, and W. S. Gray, "Optimal data fusion of correlated local decisions in multiple sensor detection systems," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 28, no. 3, pp. 916–920, 1992.

[21] R. Viswanathan and P. K. Varshney, "Distributed detection with multiple sensors part i. fundamentals," *Proceedings of the IEEE*, vol. 85, no. 1, pp. 54–63, 1997.

[22] R. S. Blum, S. A. Kassam, and H. V. Poor, "Distributed detection with multiple sensors ii. advanced topics," *Proceedings of the IEEE*, vol. 85, no. 1, pp. 64–79, 1997.

[23] A. R. Benaskeur, "Consistent fusion of correlated data sources," in *IEEE 2002 28th Annual Conference of the Industrial Electronics Society. IECON 02*, vol. 4. IEEE, 2002, pp. 2652–2656.

[24] W. Niehsen, "Information fusion based on fast covariance intersection filtering," in *Proceedings of the Fifth International Conference on Information Fusion. FUSION 2002.(IEEE Cat. No. 02EX5997)*, vol. 2. IEEE, 2002, pp. 901–904.

[25] P. K. Varshney, *Distributed detection and data fusion*. Springer Science & Business Media, 2012.

[26] H. Durrant-Whyte and T. C. Henderson, "Multisensor data fusion," in *Springer handbook of robotics*. Springer, 2016, pp. 867–896.

[27] V. Aalo and R. Viswanathou, "On distributed detection with correlated sensors: Two examples," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 25, no. 3, pp. 414–421, 1989.

[28] J. Tsitsiklis and M. Athans, "On the complexity of decentralized decision making and detection problems," *IEEE Transactions on Automatic Control*, vol. 30, no. 5, pp. 440–446, 1985.

[29] H. Chen, B. Chen, and P. K. Varshney, "A new framework for distributed detection with conditionally dependent observations," *IEEE Transactions on Signal Processing*, vol. 60, no. 3, pp. 1409–1419, 2011.

[30] P. Yang, B. Chen, H. Chen, and P. K. Varshney, "Tandem distributed detection with conditionally dependent observations," in *2012 15th International Conference on Information Fusion*. IEEE, 2012, pp. 1808–1813.

[31] O. A. Hanna, X. Li, C. Fragouli, and S. Diggavi, "Can we break the dependency in distributed detection?" 2022. [Online]. Available: https://drive.google.com/file/d/18PhQeSUL_50dRy3yoBrW-Wfy7yMDiUlJ/view?usp=sharing

[32] V. Aalo and R. Viswanathan, "Asymptotic performance of a distributed detection system in correlated gaussian noise," 1992.

[33] H. Zhang, "The optimality of naive bayes," *AA*, vol. 1, no. 2, p. 3, 2004.