Submodular Sparse Sensing for Gaussian Detection with Correlated Observations

Mariohat Coutino, *Student Member, IEEE,* Sundeep Prabhakar Chepuri, *Member, IEEE,* and Geert Leus, *Fellow, IEEE*

***Abstract*—Detection of a signal under noise is a classical signal processing problem. When monitoring spatial phenomena under a fixed budget, i.e., either physical, economical or computational constraints, the selection of a subset of available sensors, referred to as sparse sensing, that meets both the budget and performance requirements is highly desirable. Unfortunately, the subset selection problem for detection under dependent observations is combinatorial in nature and suboptimal subset selection algorithms must be employed. In this work, different from the widely used convex relaxation of the problem, we leverage submodularity, the diminishing returns property, to provide practical algorithms suitable for large-scale subset selection. This is achieved by means of low-complexity greedy algorithms, which incur a reduced computational complexity compared to their convex counterparts.**

***Index Terms*—Greedy selection, sensor selection, sensor placement, sparse sensing, submodular optimization.**

I. Introduction

LARGE sensor networks are becoming pervasive in our daily life. They are found in monitoring activities, e.g, traffic flow and surveillance, as well as typical signal processing applications such as radar and seismic imaging. The data generated by these networks requires to undergo several processing steps before being used for inference tasks, such as estimation or detection. Due to the increase in the size of the network, managing the data throughput can become a challenging problem in itself. Hence, if a known inference task with fixed performance requirements is kept in mind during the design phase of a sampler, large data reduction benefits can be obtained by optimizing the number of deployed sensors. In realistic setups, the available budget for a particular measurement campaign is also constrained, e.g., limited processing power, reduced hardware costs, and physical space restrictions. Therefore, it is of great importance to only deploy the sensors that provide meaningful information to solve the problem at hand. However, there is always a trade-off between the performance and the sparsity of the deployed network when such constraints are enforced. This framework in which a reduced number of sensors is employed for data acquisition is here referred to as *sparse sensing*.

This work is part of the ASPIRE project (project 14926 within the STW OTP programme), which is financed by the Netherlands Organisation for Scientific Research (NWO). Mario Coutino is partially supported by CONACYT.

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

In this work, we are interested in the task of designing structured *sparse samplers* for detecting signals under correlated measurements. In particular, we focus on the detection problem for the case of Gaussian measurements with non-diagonal covariance matrices. Such problems are commonly found in practical applications such as sonar and radar systems [1], imaging [2], spectrum sensing for cognitive radio [3], and biometrics [4], to list a few. For this purpose, we consider a detection task in which a series of measurements, acquired in a distributed fashion, are gathered at a fusion center, e.g., the main processing unit, to perform a hypothesis test. We restrict ourselves to a binary decision problem, in which the fusion center has to decide between two available states $\{\mathcal{H}_0, \mathcal{H}_1\}$ given the observed data. Following the conventional detection theoretical approach, we provide sparse sampler design strategies for both the Neyman-Pearson and Bayesian setting. Furthermore, as our main goal is to provide a general and scalable framework capable of dealing with large-scale problems, we focus our attention on fast techniques leveraging submodularity and greedy methods. This approach differs from the current state-of-the-art that is fundamentally based on convex relaxations. Despite the fact that the typical convex relaxations provide approximate solutions to the sensor selection problem, they boil down to semidefinite programs. These problems, albeit being solvable efficiently, are computationally intensive and, for large datasets, do not scale very well.

A. Prior Art

The structured sparse sampler design problem consists of selecting the subset of measurements with the smallest cardinality possible such that some prescribed performance requirements are met. This problem is commonly referred to in the literature as sparse sensing or sensor selection [5]. Extensive research has been carried out in the area of sparse sensing for estimation [6]-[12] and detection [13]-[17] problems. However, much of this work depends on the convex optimization machinery for optimizing the performance metrics or their surrogates. Current efforts, spanning from the field of operational research and machine learning, have shown that greedy heuristics provide near-optimal solutions, given that the cost to optimize satisfies certain properties [18], [19]. For these setups, sparse sensing has mostly been studied for estimation purposes, using information theoretic measures such as entropy and mutual information as well as experiment design metrics [20]-[23], which exhibit the property of submodularity [32]. Similar to convex/concave functions,

submodular set functions have the property that they accept efficient algorithms for unconstrained exact minimization and near-optimal maximization [31]. Due to this property, some problems involving submodular set functions allow optimization algorithms that scale nicely, and in some instances even linearly, with the size of the input set. This fact has been fundamental for designing greedy sampling strategies in large scale problems [37]-[42], [50].

For the particular case of the detection task, the state-of-the-art structured sparse sampler design framework [13]-[17] aims to optimize surrogate functions of the probability of error for the case of binary hypothesis testing. For Gaussian processes with uncorrelated errors, the sampling problem can be solved *optimally* in linear time. These optimal solutions are possible as it can be shown that maximizing the divergence measures between the probability distributions [25], e.g., Kullback-Leibler (KL) divergence, J-divergence, or Bhattacharyya distance, is tantamount to optimizing the probability of error [17]. However, when correlated errors are considered, optimizing the divergence measures is not exactly equivalent to optimizing the probability of error. Therefore, only suboptimal solutions can be obtained by maximizing the divergences. Furthermore, even though such divergences are simpler to optimize than the actual error probabilities, the problem remains non-convex in the selection variables. As a result, convex approximations must be performed in order to solve the sensor selection problem, often leading to a semidefinite program. However, despite of being solvable efficiently, these semidefinite programs are not suitable for large-scale settings where our work takes the greatest interest.

B. Overview and Main Contributions

We concentrate on fast and near-optimal sparse sampler design for Gaussian detection problems with correlated errors. The typical surrogates for the *probability of miss detection* P_{m} in the Neyman-Pearson setting, and the *probability of error,* P_e, in the Bayesian setting, which are based on divergence measures between the two distributions, are in this work relaxed to provide submodular alternatives capable to tackle the sparse sampler design for large-scale problems.

The main idea behind this work is to show, that in certain situations, it might be possible to avoid the convex machinery [17] to solve the sensor selection problem for detection. This becomes important when large scales are considered and fast algorithms are highly desirable. Therefore, in this work, we mainly focus on cardinality constrained problems. In the following, our main contributions are highlighted.

- For Gaussian observations with common covariance and distinct means we derive a bound for the approximate submodularity of the signal-to-noise ratio (SNR) set function, which provides grounds for the direct application of a greedy heuristic to maximize this cost set function under certain conditions. For instances where the near-optimality guarantees are weak, we derive a submodular set function surrogate based on the Schur complement. While this surrogate establishes a link with traditional convex relaxations for sparse sensing, it accepts a near-optimal maximization using a greedy algorithm that despite its general polynomial complexity, scales linearly in the number of available sensors through its recursive description when the number of selected sensors is fixed. This method attains results comparable with the ones of convex relaxation, but at a significantly reduced computational complexity.
- For Gaussian observations with uncommon covariances and common means we show that the divergences between probability distributions are not submodular. Despite this, we present them as a difference of submodular functions, which can be approximately optimized. In cases where these decompositions are not readily available, we introduce surrogate decompositions based on the Schur complement. This approach provides local optimality guarantees without involving computationally intensive semidefinite programs as in the convex case.
- For the most general case of Gaussian observations with uncommon means and uncommon covariances, we show that the algorithms developed for the case of uncommon covariances and common means can be reused.

C. Outline and Notation

The rest of this paper is organized as follows. In Section II, the problem of sparse sampler design for detection is introduced, and the sensor selection metrics for both the Neyman-Pearson and Bayesian setting are discussed. The submodular optimization theory is introduced in Section III. In Section IV and Section V, submodular set function surrogates for the selection criteria are derived and a general framework to solve the sparse sampler design for Gaussian measurements is provided. Finally, conclusions are drawn in Section VI.

The notation used in this paper is the following. Upper (lower) bold faces letters are used to define matrices (column vectors). $\mathcal{N}(\mu, \sigma^2)$ is reserved to represent a Gaussian normal distribution with mean μ and variance σ^2. The notation $\sim$ is read as "is distributed according to". $(\cdot)^T$ and $(\cdot)^{-1}$ represent transposition and matrix inversion, respectively. $\mathrm{diag}(\cdot)$ refers to a diagonal matrix with its argument on the main diagonal. $\mathbf{I}$ and $\mathbf{1}$ denote the identity matrix and the all-one vector of appropriate size, respectively. $\det(\cdot)$ and $\log(\cdot)$ are the matrix determinant and natural logarithm, respectively. $\mathrm{tr}\{\cdot\}$ denotes the matrix trace operator. $[\mathbf{x}]_i$ and $[\mathbf{X}]_{i,j}$ denote the ith entry of the vector $\mathbf{x}$ and the (i, j) entry of the matrix $\mathbf{X}$, respectively. Calligraphic letters denote sets, e.g., $\mathcal{A}$, and the vector $\mathbf{1}_{\mathcal{A}}$, with $\mathcal{A} \subseteq \mathcal{V}$, denotes a vector with ones at the indices given by $\mathcal{A}$, and zeros in the complementary set, $\mathcal{V} \setminus \mathcal{A}$. $\lambda_{\max}\{\mathbf{A}\}$ and $\lambda_{\min}\{\mathbf{A}\}$ are the maximum eigenvalue and minimum eigenvalue of the matrix $\mathbf{A}$, respectively.

II. Problem Statement

Consider a set $\mathcal{X} = \{x_1, \ldots, x_M\}$ of M candidate measurements. These measurements can be temporal samples of temperature, spatial samples from wavefield measurements, etc. The samples are known to be related to the models

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

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

where $p_m(x|\mathcal{H}_i)$ for $i = 0, 1$ denotes the probability density function (pdf) of the mth measurement, x_m, conditioned on the state $\mathcal{H}_i$. By stacking the elements of $\mathcal{X}$ in a vector $\mathbf{x} = [x_1, x_2, \ldots, x_m]^T \in \mathbb{R}^M$, the pdf of the measurement set for the hypothesis $\mathcal{H}_i$ is denoted by $p(\mathbf{x}|\mathcal{H}_i)$.

We pose the acquisition of a reduced set $\mathcal{Y} \subseteq \mathcal{X}$ consisting of K measurements as a linear sensing problem where the rows of the sensing matrix are formed from a subset of rows of an identity matrix. The selected rows, indexed by $\mathcal{A}$, of the identity matrix are defined by a vector $\mathbf{w}$ whose entries belong to a binary alphabet set, i.e.,

$$\mathbf{w} = [w_1, w_2, \ldots, w_M]^T \in \{0, 1\}^M, \tag{6}$$

where $w_m = 1$ (0) indicates that the mth measurement is (not) selected. The subset of rows is then defined as

$$\mathcal{A} := \{m \mid w_m = 1, 1 \leq m \leq M\}. \tag{7}$$

The acquisition scheme can be formally expressed using the following linear model

$$\mathbf{y}_{\mathcal{A}} = \mathbf{\Phi}_{\mathcal{A}}\mathbf{x} \in \mathbb{R}^K, \tag{8}$$

where $\mathbf{y}_{\mathcal{A}} = [y_1, y_2, \ldots, y_K]^T$ is the reduced-size measurement vector whose entries belong to the set $\mathcal{Y} \subseteq \mathcal{X}$. The selection matrix $\mathbf{\Phi}_{\mathcal{A}}$ is a binary matrix composed of the rows of the identity matrix defined by the set $\mathcal{A}$ (non-zero entries of $\mathbf{w}$). Even though K is (possibly) unknown to us, we are interested only in cases where $K \ll M$, as it is desirable to perform inference on a reduced measurement set. As the notation based on either $\mathbf{w}$ or $\mathcal{A}$ is interchangeable, from this point on, we make no distinction between them.

The subset of measurements $\mathcal{Y}$ is finally used to solve the detection problem (2) given that the detection performance requirements, for a given application, are met. If the prior hypothesis probabilities are known, i.e., in a Bayesian setting, the optimal detector minimizes the probability of error, $P_e = P(\mathcal{H}_0|\mathcal{H}_1)P(\mathcal{H}_1) + P(\mathcal{H}_1|\mathcal{H}_0)P(\mathcal{H}_0)$, where $P(\mathcal{H}_i|\mathcal{H}_j)$ is the conditional probability of deciding $\mathcal{H}_i$ when $\mathcal{H}_j$ is true and $P(\mathcal{H}_i)$ is the prior probability of the ith hypothesis. When the prior hypothesis probabilities are unknown, i.e., in a Neyman-Pearson setting, the optimal detector aims to minimize the probability of miss detection (type II error), $P_m = P(\mathcal{H}_0|\mathcal{H}_1)$, for a fixed probability of false alarm (type I error), $P_{fa} = P(\mathcal{H}_1|\mathcal{H}_0)$.

In a more formal manner, the sensor selection problem for detection, in both settings, is given by

$$\textbf{Bayesian}: \quad \arg\min_{\mathcal{A}} \; P_e(\mathcal{A}) \quad \text{s. to} \;\; |\mathcal{A}| = K, \tag{9}$$

$$\textbf{Neyman-Pearson}: \quad \arg\min_{\mathcal{A}} \; P_m(\mathcal{A}) \quad \text{s. to} \; |\mathcal{A}| = K, \;\; P_{fa}(\mathcal{A}) \leq \alpha, \tag{10}$$

where $P_e(\mathcal{A})$, $P_m(\mathcal{A})$ and $P_{fa}(\mathcal{A})$ denote the error probabilities due to the measurement selection defined by the set $\mathcal{A}$, and α the prescribed false alarm rate.

As for the most general case, the performance metrics in (9) and (10) are not easy to optimize numerically, we present alternative measures that can be used as direct surrogates to solve the optimization problems (9) and (10). Here, we focus on metrics which provide a notion of distance between the hypotheses under test. That is, we are interested in *maximizing* the distance between two distinct probability distributions $p(\mathbf{y}_{\mathcal{A}}|\mathcal{H}_i)$ and $p(\mathbf{y}_{\mathcal{A}}|\mathcal{H}_j)$ using a divergence measure $\mathcal{D}(p(\mathbf{y}_{\mathcal{A}}|\mathcal{H}_i)\|p(\mathbf{y}_{\mathcal{A}}|\mathcal{H}_j)) \in \mathbb{R}_+$. They lead to tractable optimization methods and, in some particular cases such as for independent observations under uncorrelated Gaussian noise, they result in an optimal solution. A summary of the divergences, for Gaussian probability distributions, $\mathcal{N}(\boldsymbol{\theta}_i, \mathbf{\Sigma}_i)$, between the different hypotheses under test employed for different settings in this work is shown in Table I. Here, $\boldsymbol{\theta}_i$ and $\mathbf{\Sigma}_i$ denote the mean vector and the covariance matrix of the ith distribution, respectively. For a more detailed treatment of these divergence measures and their suitability for sensor selection, the reader is referred to [17], [24]-[27] and the references therein.

Using these divergence measures, the relaxed formulation of the sparse sensing problems (9) and (10) can be stated, respectively, as cardinality constraint (P-CC) and detection performance constraint (P-DC) problems:

$$\textbf{P-CC}: \quad \arg\max_{\mathcal{A}} \; f(\mathcal{A}) \quad \text{s. to} \;\; |\mathcal{A}| = K; \tag{11}$$

$$\textbf{P-DC}: \quad \arg\min_{\mathcal{A}} \; |\mathcal{A}| \quad \text{s. to} \;\; f(\mathcal{A}) \geq \lambda, \tag{12}$$

where $f(\mathcal{A})$ is one of the divergence measures, λ is the prescribed accuracy and K is the cardinality of the selected subset of measurements. For the sake of exposition, in this paper we mainly focus on cardinality constraints (i.e., a uniform matroid constraint). However, by allowing an increase in computational complexity and degradation of the near-optimality guarantees, the methods presented in this work can be extended to incorporate budget constraints representable by other kinds of matroids [28].

III. Preliminaries

In this section, some preliminaries about submodularity are provided. The main definitions and theorems related to submodular set functions used throughout the work are presented.

A. Submodularity

In many engineering applications we encounter the *diminishing returns* principle. That is, the gain of adding new information, e.g., a data measurement, to a large pool of measurements is smaller than the gain of adding the same piece of information to a smaller pool of measurements. This notion is mathematically captured by the next definition.

Definition 1. (Submodularity) *Let $\mathcal{V} = \{1, 2, \ldots, M\}$ refer to a ground set, then the set function $f : 2^{|\mathcal{V}|} \to \mathbb{R}$ is said to be submodular, if for every $\mathcal{A} \subseteq \mathcal{B} \subseteq \mathcal{V}$ and $v \in \mathcal{V} \setminus \mathcal{B}$ it holds that*

$$f(\mathcal{A} \cup \{v\}) - f(\mathcal{A}) \geq f(\mathcal{B} \cup \{v\}) - f(\mathcal{B}). \tag{13}$$

Similar to convex functions, submodular set functions have certain properties that make them convenient to optimize. For example, the unconstrained minimization of general submodular functions can be done in polynomial time [31] with respect to the size of the ground set $|\mathcal{V}|$.

TABLE I
SUMMARY OF DIVERGENCE MEASURES FOR GAUSSIAN PROBABILITY DISTRIBUTIONS.

Divergence	Expression	Setting
	Bhattacharyya	
$\mathcal{B}(\mathcal{H}_1 \| \mathcal{H}_0) :=$	$\frac{1}{8}(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_0)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_0) + \frac{1}{2}\log\left(\frac{\det(\boldsymbol{\Sigma})}{\sqrt{\det(\boldsymbol{\Sigma}_1)\det(\boldsymbol{\Sigma}_0)}}\right),\ \boldsymbol{\Sigma} = 0.5(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_0)$ (3)	Bayesian
	Kullback-Leibler	
$\mathcal{K}(\mathcal{H}_1 \| \mathcal{H}_0) :=$	$\frac{1}{2}\Big(\mathrm{tr}(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\Sigma}_1) + (\boldsymbol{\theta}_1 - \boldsymbol{\theta}_0)^T \boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_0) - N + \log(\det(\boldsymbol{\Sigma}_0)) - \log(\det(\boldsymbol{\Sigma}_1))\Big)$ (4)	Neyman-Pearson
	J-Divergence	
$\mathcal{D}_\mathrm{J}(\mathcal{H}_0 \| \mathcal{H}_1) :=$	$\mathcal{K}(\mathcal{H}_1 \| \mathcal{H}_0) + \mathcal{K}(\mathcal{H}_0 \| \mathcal{H}_1)$ (5)	Neyman-Pearson

Algorithm 1: GREEDY ALGORITHM.

Result: $\mathcal{A} : |\mathcal{A}| = K$
initialization $\mathcal{A} = \emptyset$, $k = 0$;
while $k < K$ **do**
  $a^* = \arg\max_{a \notin \mathcal{A}} f(\mathcal{A} \cup \{a\})$;
  $\mathcal{A} = \mathcal{A} \cup \{a^*\}$;
  $k = k + 1$;
end

Even though the maximization of general submodular set functions is an NP-hard problem, Nemhauser et al. [40] have shown that for the cardinality-constrained maximization [of the form (11)] of a non-decreasing submodular set function f, with $f(\emptyset) = 0$, the simple *greedy* procedure presented in Algorithm 1 finds a solution which provides at least a constant fraction $(1 - 1/e) \approx 63\%$ of the optimal value, where e is the base of the natural logarithm. In this context, a set function $f : 2^{|\mathcal{V}|} \to \mathbb{R}$ defined on the subsets of a ground set $\mathcal{V}$ is considered non-decreasing if and only if $f(\mathcal{B}) \geq f(\mathcal{A})$ holds for all sets $\mathcal{A} \subseteq \mathcal{B} \subseteq \mathcal{V}$.

Using similar arguments, Krause et al. [18] extended the near-optimality of the greedy heuristic for *approximately* submodular set functions or ϵ-submodular set functions:

Definition 2. (ϵ-Submodularity) [18] *A set function $f : 2^{|\mathcal{V}|} \to \mathbb{R}$ defined on the subsets of a ground set $\mathcal{V}$, is approximately submodular with constant $\epsilon > 0$ or ϵ-submodular, if for all sets $\mathcal{A} \subseteq \mathcal{B} \subseteq V$, and $v \in \mathcal{V} \backslash \mathcal{B}$ it holds that*

$$f(\mathcal{A} \cup \{v\}) - f(\mathcal{A}) \geq f(\mathcal{B} \cup \{v\}) - f(\mathcal{B}) - \epsilon. \tag{14}$$

For ϵ-submodular functions the greedy Algorithm 1 provides the following weaker guarantee.

Theorem 1. (ϵ-Near-Optimality) [18] Let $f : 2^{|\mathcal{V}|} \to \mathbb{R}$ be a normalized, i.e., $f(\emptyset) = 0$, non-decreasing, ϵ-submodular set function defined on the subsets of a finite ground set $\mathcal{V}$. Let $\mathcal{G}$ be the set of K elements obtained from Algorithm 1. Then,

$$f(\mathcal{G}) \geq \left(1 - \frac{1}{e}\right) f(\mathcal{A}_\mathrm{opt}) - K\epsilon, \tag{15}$$

where $\mathcal{A}_\mathrm{opt} := \arg\max_{\mathcal{A} \subseteq \mathcal{V}, |\mathcal{A}| = K} f(\mathcal{A})$ is the optimal set.

The result in Theorem 1 implies that for small $K\epsilon$, Algorithm 1 provides a good approximate solution for the maximization under cardinality constraints. As in practice it is observed that the lower bound from [40] is not tight, i.e., the greedy method performs much better than the lower bound [50], the expression provided for ϵ-submodular set functions in (15) is expected to be also a loose bound for the performance of Algorithm 1. In any case, Theorem 1 shows that the degradation on the approximation factor increases as K becomes larger.

Algorithm 2: SUPSUB PROCEDURE

Result: $\mathcal{A} : |\mathcal{A}| = K$
initialization $\mathcal{A}^0 = \emptyset$; $t = 0$;
while *not converged (i.e.,* $(\mathcal{A}^{t+1} \neq \mathcal{A}^t)$**)** **do**
  $\mathcal{A}^{t+1} := \underset{\mathcal{A},\ |\mathcal{A}|=k}{\mathrm{argmax}}\ g(\mathcal{A}) - m^h_{\mathcal{A}^t}(\mathcal{A})$;
  $t = t + 1$;
end

B. Difference of Submodular Functions

A notable result in combinatorial optimization arises from the fact that *any* set function can be expressed as a difference of two submodular set functions [41]. Therefore, the optimization problem

$$\max_{\mathcal{A} \subseteq \mathcal{V}} f(\mathcal{A}) \equiv \max_{\mathcal{A} \subseteq \mathcal{V}} [g(\mathcal{A}) - h(\mathcal{A})], \tag{16}$$

where the cost set function $f : 2^{|\mathcal{V}|} \to \mathbb{R}$ is expressed as the difference of two set functions $g : 2^{|\mathcal{V}|} \to \mathbb{R}$ and $h : 2^{|\mathcal{V}|} \to \mathbb{R}$, defined over a ground set $\mathcal{V}$ is, in general, NP-hard. Recent results from Iyer et al. [42] show that the general case of this problem is multiplicatively inapproximable. However, in this work we motivate the usage of practical methods, employing well-designed heuristics, to obtain good results when solving large-scale real-world problems.

Firstly, let us consider a heuristic from convex optimization for approximating the problem of minimizing the difference of convex functions. A typical heuristic is to linearize one of the convex functions with its Taylor series approximation. With such a linearization, the original nonconvex minimization problem can be transformed into a sequential minimization of a convex plus an affine function. In the literature this method is known as the convex-concave procedure (CCP) [45]. Similarly, for maximizing the difference of submodular set functions, it is possible to substitute one of the submodular set functions from (16) by its modular upper bound at every iteration as suggested in [42]. Algorithm 2 summarizes the supermodular-submodular (SupSub) procedure as described in [42] when the cardinality of the set is constrained for approximating the solution of (16).

In Algorithm 2, at every iteration, a submodular set function is maximized. This is due to the fact that the modular upper bound $m^h_{\mathcal{A}^t}$ of h, locally to $\mathcal{A}^t$ [40], preserves the submodularity of the cost. Using the characterization of submodular

set functions two tight modular upper bounds can be defined as follows

$$m^h_{\mathcal{A},1}(C) \triangleq h(\mathcal{A}) - \sum_{j\in\mathcal{A}\setminus C} h(\{j\}|\mathcal{A}\setminus\{j\}) + \sum_{j\in C\setminus\mathcal{A}} h(\{j\}|\emptyset), \quad (17)$$

$$m^h_{\mathcal{A},2}(C) \triangleq h(\mathcal{A}) - \sum_{j\in\mathcal{A}\setminus C} h(\{j\}|\mathcal{V}\setminus\{j\}) + \sum_{j\in C\setminus\mathcal{A}} h(\{j\}|C), \quad (18)$$

where $h(\mathcal{A}|C) \triangleq h(C\cup\mathcal{A}) - h(C)$ denotes the gain of adding $\mathcal{A}$ when C is already selected. In practice, either (17) or (18) can be employed in Algorithm 2 or both can be run in parallel choosing the one that is better. For a more in-depth treatment of these bounds, the reader is referred to [57]. These bounds follow similar arguments as the ones found in majorization-minimization algorithms [43] for general non-convex optimization.

Although the maximization of submodular functions is NP-hard, Algorithm 1 can be used to approximate at each step the maximum of the submodular set function in Algorithm 2. Furthermore, as the problem of submodular maximization with cardinality, matroid and knapsack constraints admits a constant factor approximation, the SupSub procedure can be extended to constrained minimization of a difference of two submodular functions. In addition, despite that the optimization of the difference of submodular set functions is inapproximable [42], similar to CCP, the SupSub procedure is guaranteed to reach a local optimum of the set function when the procedure converges [42], i.e., $\mathcal{A}^{t+1} = \mathcal{A}^t$.

The main reasons to prefer the SupSub procedure over, a possibly, submodular-supmodular (SubSup) procedure, where a modular lower bound of $g(\cdot)$ is used and the inner step consists of the minimization of a submodular function, are its computationally complexity and versatility. Even though unconstrained minimization of submodular set functions can be performed in polynomial time, the addition of constraints to the minimization of submodular set functions renders the problem NP-hard, for which there are no clear approximation guarantees. As a result, the SupSub is often preferred for optimizing differences of submodular functions.

IV. Observations with Uncommon Means

In this section, we illustrate how to design sparse samplers using the criteria presented in Section II for Gaussian observations with uncommon means. This kind of measurements arises often in communications as in the well-studied problem of detecting deterministic signals under Gaussian noise.

Consider the binary signal detection problem in (2). Furthermore, let us assume that the pdfs of the observations are multivariate Gaussians with uncommon means and equal covariance matrices. Then, the related conditional distributions, under each hypothesis, are given by

$$\begin{aligned}\mathcal{H}_0 &: \mathbf{y}_{\mathcal{A}} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_{\mathcal{A}})\\ \mathcal{H}_1 &: \mathbf{y}_{\mathcal{A}} \sim \mathcal{N}(\boldsymbol{\theta}_{\mathcal{A}}, \mathbf{\Sigma}_{\mathcal{A}}),\end{aligned} \quad (19)$$

where $\mathcal{A}\subseteq\mathcal{V}$ is the subset of selected sensors from the set of candidate sensors $\mathcal{V} = \{1, 2, \ldots, M\}$, and where $\boldsymbol{\theta}_{\mathcal{A}} = \mathbf{\Phi}_{\mathcal{A}}\boldsymbol{\theta} \in \mathbb{R}^K$ and $\mathbf{\Sigma}_{\mathcal{A}} = \mathbf{\Phi}_{\mathcal{A}}\mathbf{\Sigma}\mathbf{\Phi}^T_{\mathcal{A}} \in \mathbb{R}^{K\times K}$. The mean vector $\boldsymbol{\theta}$ and the covariance matrix $\mathbf{\Sigma}$ are assumed to be known a priori.

By observing the Bhattacharyya distance and the KL divergence in (3) and (4), respectively, it can be seen that for the probability distributions in (19) such metrics are reduced to the so-called signal-to-noise ratio function

$$s(\mathcal{A}) = \boldsymbol{\theta}^T_{\mathcal{A}}\mathbf{\Sigma}^{-1}_{\mathcal{A}}\boldsymbol{\theta}_{\mathcal{A}}. \quad (20)$$

Therefore, maximizing the signal-to-noise ratio, $s(\mathcal{A})$, directly maximizes the discussed divergence measures leading to an improvement in the detection performance. As a result, we are required to solve the following combinatorial problem

$$\underset{\mathcal{A}\subseteq\mathcal{V},|\mathcal{A}|=K}{\text{maximize}}\ s(\mathcal{A}) \quad (21)$$

Due to the hardness of the problem in (21), finding its exact solution requires an exhaustive search over $\binom{M}{K}$ possible combinations which for large M rapidly becomes intractable. Simplifications for the problem (21) can be derived using convex optimization [17]. Such approaches provide a sub-optimal solution in polynomial time when cast as a semidefinite program (SDP). Even though under the SDP framework, approximate solutions for (21) can be found efficiently, for large-scale problems near-optimal solutions obtained through Algorithm 1 are more attractive as for a fixed K the number of function evaluations required by the method scales linearly in the number of available sensors. Therefore, the complexity only depends on the efficient evaluation of the cost function.

A. ϵ-Submodularity of Signal-to-Noise Ratio

One may ask, can we apply the greedy heuristic in Algorithm 1 directly on the signal-to-noise ratio and still guarantee near optimality? The answer is no, in general. This is because, the signal-to-noise ratio is not a submodular function. However, although the signal-to-noise ratio set function is not submodular, we can try to quantify how far this set function is away from being submodular. For this purpose, we derive a bound for the ϵ-submodularity of the signal-to-noise ratio.

In the following, we present a key relationship between the parameter ϵ and the conditioning of the covariance matrix $\mathbf{\Sigma}$ to provide a bound on the approximate submodularity of the signal-to-noise ratio set function. This relation is summarized in the following theorem and corollary.

Theorem 2. *Let $\mathbf{\Sigma}$ be a non-diagonal covariance matrix, with minimum eigenvalue $\lambda_{\min}\{\mathbf{\Sigma}\} \neq 0$, maximum eigenvalue $\lambda_{\max}\{\mathbf{\Sigma}\}$, condition number $\kappa \coloneqq \lambda_{\max}\{\mathbf{\Sigma}\}/\lambda_{\min}\{\mathbf{\Sigma}\}$, that admits a decomposition $\mathbf{\Sigma} = a\boldsymbol{I} + \boldsymbol{S}$ where a is chosen as $a = \beta\lambda_{\min}$ with $\beta \in (0,1)$ to guarantee the positive definiteness of $\boldsymbol{S}$. Then the signal-to-noise ratio set function $s(\mathcal{A})$ is ϵ-approximately submodular with*

$$\epsilon \leq 4C_1\left(\frac{a}{(1-\beta)^2\lambda^2_{\min}\{\mathbf{\Sigma}\}} + \frac{\nu\kappa^2}{(1-\beta)^2}\right), \quad (22)$$

where $C_1 = \|\boldsymbol{\theta}\|^2_2$, with $\boldsymbol{\theta}$ being the mean vector, and $\nu = \lambda^{-1}_{\min}\{\mathbf{\Sigma}\}$.

Proof. See Appendix A. □

Corollary 1. *For the limiting case, $a \to 0$ or equivalently $\beta \to 0$, Theorem 2 reduces to*

$$\epsilon \leq 4C_1\kappa^2\lambda_{\min}^{-1}\{\mathbf{\Sigma}\}. \tag{23}$$

Proof. Follows from Theorem 2 (See Appendix A). □

From the results in (22) and (23), it can be seen that when the condition number of the covariance matrix $\mathbf{\Sigma}$ is low, e.g., weakly correlated matrices, the theoretical guarantee in Theorem 1 provides encouraging bounds for the greedy maximization of the signal-to-noise ratio. In this regime, several works have focused on sensor selection in the past. For example, in the limiting case $\mathbf{\Sigma} = \text{diag}\{\sigma_1^2, \sigma_2^2, \ldots, \sigma_M^2\}$, where $s(\mathcal{A})$ becomes a modular set function (the expression in (13) is met with equality), it has been shown that the optimization problem can be solved optimally by sorting [17]. Note that the characterization provided in Theorem 2, in terms of condition number of the matrix, excludes diagonal matrices. This is due to the fact that even if a diagonal matrix is ill-conditioned, the resulting set function is still a modular set function. Therefore, here we focus on matrices that model well correlated errors. Unfortunately, for arbitrary covariance matrices (especially badly conditioned matrices), the ϵ-submodular guarantee can be very loose. In that case, surrogate submodular set functions can be efficiently optimized using Algorithm 1 as a fast alternative for performing sensor selection in large-scale problems.

B. Signal-to-Noise Ratio Submodular Surrogate

Firstly, let us decompose the covariance matrix $\mathbf{\Sigma}$ as

$$\mathbf{\Sigma} = a\mathbf{I} + \mathbf{S}, \tag{24}$$

where $a \in \mathbb{R}$ and $\mathbf{S} \in \mathbb{R}^{M\times M}$ have been chosen as described in Theorem 2. Combining (20) and (24), it can be shown that the signal-to-noise ratio can be rewritten as [17]

$$s(\mathcal{A}) = \boldsymbol{\theta}_{\mathcal{A}}^T\mathbf{\Sigma}_{\mathcal{A}}^{-1}\boldsymbol{\theta}_{\mathcal{A}} \tag{25}$$

$$= \boldsymbol{\theta}^T\mathbf{S}^{-1}\boldsymbol{\theta} - \boldsymbol{\theta}^T\mathbf{S}^{-1}[\mathbf{S}^{-1} + a^{-1}\text{diag}(\mathbf{1}_{\mathcal{A}})]^{-1}\mathbf{S}^{-1}\boldsymbol{\theta}, \tag{26}$$

where the non-zero entries of the vector $\mathbf{1}_{\mathcal{A}}$ are given by the set $\mathcal{A}$. Then, considering that the signal-to-noise ratio is always non-negative we can use the Schur complement to express this condition as a linear matrix inequality (LMI) in $\mathbf{w}$,

$$\mathbf{M}_{\mathcal{A}} \coloneqq \begin{bmatrix} \mathbf{S}^{-1} + a^{-1}\text{diag}(\mathbf{1}_{\mathcal{A}}) & \mathbf{S}^{-1}\boldsymbol{\theta} \\ \boldsymbol{\theta}^T\mathbf{S}^{-1} & \boldsymbol{\theta}^T\mathbf{S}^{-1}\boldsymbol{\theta} \end{bmatrix} \geq 0, \tag{27}$$

which is similar to the LMI found in the convex program in [17]. Therefore, we can consider the following optimization problem as an approximation of (21)

$$\underset{\mathcal{A}\subseteq\mathcal{V};|\mathcal{A}|=K}{\arg\max} \quad f(\mathcal{A}) \tag{28}$$

where the cost set function has been defined as

$$f(\mathcal{A}) \triangleq \begin{cases} 0, & \text{if } \mathcal{A} = \emptyset \\ \log\det(\mathbf{M}_{\mathcal{A}}), & \text{if } \mathcal{A} \neq \emptyset \end{cases}. \tag{29}$$

The normalization of the cost is done to avoid the infinity negative cost due to the logarithm of zero.

In the following, we motivate why (28) is a good alternative for (21). First, notice that the determinant of $\mathbf{M}_{\mathcal{A}}$ consists of the product of two terms, where one of them is related to the signal-to-noise ratio $s(\mathcal{A})$. That is, using the generalization of the determinant for block matrices, we can decompose the determinant of the right-hand-side (RHS) of (27) as

$$\det(\mathbf{M}_{\mathcal{A}}) = \det\begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix} = \det(\mathbf{A})\det(\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B}) \tag{30}$$

$$= \gamma(\mathcal{A})s(\mathcal{A}) \tag{31}$$

where $\gamma(\mathcal{A}) = \det(\mathbf{S}^{-1} + a^{-1}\text{diag}(\mathbf{1}_{\mathcal{A}}))$ with $\gamma(\mathcal{A}) > 0$. This can always be achieved by appropriately choosing a.

From (31) we notice that the determinant of $\mathbf{M}_{\mathcal{A}}$ consists of the product of the signal-to-noise ratio $s(\mathcal{A})$, and $\gamma(\mathcal{A})$. The set function $\gamma(\mathcal{A})$ is a monotone nondecreasing set function of the selected set size. In addition, rewriting the SNR expression as

$$s(\mathcal{A}) = \boldsymbol{\theta}^T\mathbf{S}^{-1}\boldsymbol{\theta} - \frac{1}{\gamma(\mathcal{A})}\boldsymbol{\theta}^T\mathbf{S}^{-1}\text{adj}(\mathbf{S}^{-1} + a^{-1}\text{diag}(\mathbf{1}_{\mathcal{A}}))\mathbf{S}^{-1}\boldsymbol{\theta}, \tag{32}$$

where $\text{adj}(\mathbf{A})$ is the adjugate of $\mathbf{A}$ defined as the transpose of the cofactor matrix of $\mathbf{A}$, we observe that in order to keep the nondecreasing property of the SNR with respect to the set $\mathcal{A}$, the growth rate of $\gamma(\mathcal{A})$ should be larger than growth rate of the quadratic form in (32). Hence, it is reasonable to consider (31) as a surrogate function for $s(\mathcal{A})$. Finally, we remark that maximizing (31) effectively maximizes a modified version of (26) where the inverse of $\mathbf{S}^{-1} + a^{-1}\text{diag}(\mathbf{1}_{\mathcal{A}})$ has been substituted by its adjugate [cf. (32)].

In the following, we present a proposition that is required to provide guarantees for near optimality when the proposed submodular cost set function for the combinatorial problem (28) is maximized.

Proposition 1. (Monotonicity and Submodularity) *The cost set function in* (29) *is a monotone, nondecreasing, normalized submodular set function.*

Proof. See Appendix B. □

By the fact that the cost set function (29) is a normalized, nondecreasing submodular set function, (28) can be solved near-optimally for any cardinality size K using Algorithm 1.

C. Recursive Description of Cost Set Function

It is important to remark that most of the claims of scalability in submodular optimization works rely on the linear-time complexity with respect to the cardinality of the selected set. However, this claim might not translate in a fast optimization solver for all problem instances as the evaluation of the set function itself can be computationally expensive, and in certain situations, it might be a prohibitive endeavor.

Under this perspective, we demonstrate the suitability of a large-scale optimization of (28) by showing that it is possible to compute this set function recursively, alleviating the complexity of computing the determinant of an $(M+1)\times(M+1)$ matrix, which in general, has complexity $O((M+1)^3)$.

Let us consider the kth step of the greedy algorithm, with $\mathcal{A}_{k-1}$ denoting the subset of sensors selected upto this point.

First recall that the cost set function (29) can be expressed as [cf. (31)]

$$f(\mathcal{A}_k) = \ln(\det(\mathbf{S}^{-1} + a^{-1}\mathrm{diag}(\mathbf{1}_{\mathcal{A}_k}))s(\mathcal{A}_k)). \tag{33}$$

By applying the determinant lemma to (33) we obtain

$$f(\mathcal{A}_k) = \ln(\det(\mathbf{S}^{-1})\det(\mathbf{I} + a^{-1}\mathbf{S}_{\mathcal{A}_k})s(\mathcal{A}_k)), \tag{34}$$

where for $\mathcal{A}_k = \{m_1, \ldots, m_k\}$, we have defined $[\mathbf{S}_{\mathcal{A}_k}]_{i,j} = [\mathbf{S}]_{m_i,m_j}$. Here, m_i is the sensor index selected at the ith step. As the kth step of the greedy algorithm evaluates the cost set function for the set $\mathcal{A}_k = \mathcal{A}_{k-1} \cup \{i\}$, $\forall\, i \in \mathcal{V} \setminus \mathcal{A}$, in order to find the best sensor to add, the matrix in the second term of (34) can be written using the following block structure

$$\mathbf{I} + a^{-1}\mathbf{S}_{\mathcal{A}_k} = \left[\begin{array}{c|c} \mathbf{I} + a^{-1}\mathbf{S}_{\mathcal{A}_{k-1}} & \mathbf{s}_{\mathcal{A}_k} \\ \hline \mathbf{s}_{\mathcal{A}_k}^T & 1 + \alpha_{\mathcal{A}_k} \end{array}\right], \tag{35}$$

where for $\mathcal{A}_{k-1} = \{m_1, \ldots, m_{k-1}\}$, we have defined $[\mathbf{s}_{\mathcal{A}_k}]_j = [\mathbf{S}]_{m_j,i}$, and $\alpha_{\mathcal{A}_k} = [\mathbf{S}]_{i,i}$. Therefore, using the property of the determinant for block matrices, we can construct the following recursive evaluation for the cost set function

$$\begin{aligned} f(\mathcal{A}_k) = {} & \ln(\det(\mathbf{S}^{-1})\det(\mathbf{I} + a^{-1}\mathbf{S}_{\mathcal{A}_{k-1}}) \\ & \times (1 + \alpha_{\mathcal{A}_k} - \mathbf{s}_{\mathcal{A}_k}^T(\mathbf{I} + a^{-1}\mathbf{S}_{\mathcal{A}_{k-1}})^{-1}\mathbf{s}_{\mathcal{A}_k})s(\mathcal{A}_k)), \end{aligned} \tag{36}$$

where the matrix $\mathbf{I} + a^{-1}\mathbf{S}_{\mathcal{A}_{k-1}}$ is fixed for every $i \in \mathcal{V} \setminus \mathcal{A}_{k-1}$, and it only has to be updated when the sensor for the kth step has been chosen.

Remark 1. *We stress that for some choices of the parameter a, the matrix $\boldsymbol{M}_{\mathcal{A}}$ might be ill-conditioned. However, as the computation of the cost function is not performed directly on the matrix $\boldsymbol{M}_{\mathcal{A}}$ but through* (36)*, we only require that the recursive inversion of the matrix in expression* (36) *is numerically well-conditioned. This can be achieved in practice by selecting the value for a far from both* 0 *and* $\lambda_{\min}\{\boldsymbol{\Sigma}\}$*, e.g.,* $a = 0.5\lambda_{\min}\{\boldsymbol{\Sigma}\}$*. This approach avoids numerical problems that could arise due to the selection of the value of a.*

From (36), the computational advantages during function evaluations are clearly seen. First, computation of the inverse of the matrix $\mathbf{S}^{-1}$ is not needed as for any set the term $\det(\mathbf{S}^{-1})$ is constant. This contrasts with the convex method from [17] which requires the inversion of $\mathbf{S}$. Second, the rank-one update of the inverse in (36) as well as the computation of $s(\mathcal{A}_k)$ have worst-case complexity $O(K^2)$, which implies that the overall complexity of the proposed algorithm is about $O(MK^3)$. That is, different from its convex counterpart which has cubic complexity in the number of available sensors, the proposed method, for a fixed K, scales linearly with respect to the number of available sensors.

Furthermore, as seen in (36) it is possible to generate two solutions, without any extra computational expense, by the evaluation of the cost set function: (i) the solution for maximizing greedily the submodular surrogate $f(\cdot)$, and (ii) the solution of maximizing greedily the signal-to-noise ratio, $s(\cdot)$. Therefore, the proposed cost set function is perfectly suitable for large-scale problems, especially for instances with $M \gg K$, where computational complexity is of great importance. In addition, as two solutions can be built simultaneously, the one with the best performance can always be chosen as final solution. In addition, *lazy evaluations* [51], or *stochastic greedy selection* (SGS) [52] can be employed to further reduce the number of function evaluations required by the introduction of book keeping, i.e., lazy evaluations requires sorting, or by relaxing the guarantees to stochastic guarantees in the case of SGS.

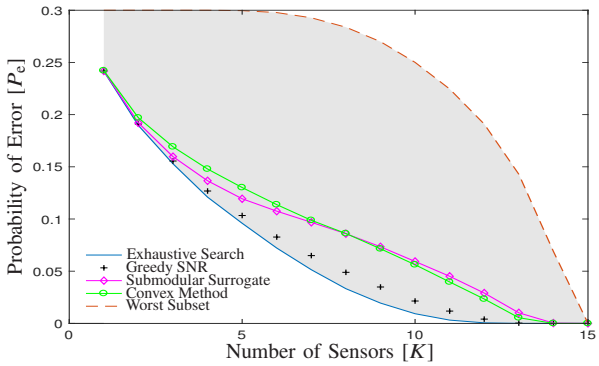


Fig. 1: Bayesian probability of error P_e for (19) with different subset sizes K when choosing from $M = 15$ available sensors. The probability of error for any random subset of K sensors will be in the shaded region of the plot.

D. Numerical Examples

To illustrate the performance of the submodular optimization machinery, we present two different examples for (19) under the Bayesian setting with $P(\mathcal{H}_0) = 0.3$. In both examples, the value for a has been fixed to $a = 0.5\lambda_{\min}\{\boldsymbol{\Sigma}\}$ to avoid heavily ill-conditioned matrices in our computations.

First, let us consider a small-scale sensor selection problem where the best K sensors have to be selected from a pool of $M = 15$ available sensors. This small scale example allows us to compare the developed algorithm with the optimal solution. In this example, 1000 Monte-Carlo runs are performed. The common covariance matrix $\boldsymbol{\Sigma}$, in each Monte-Carlo run, is generated using a superposition of M unit power Gaussian sources according to the standard far-field and narrowband array signal processing model [1], and the mean vectors $\boldsymbol{\theta}_i$ are considered i.i.d. Gaussian random unit vectors. We solve the problem by performing an exhaustive search over all possible $\binom{M}{K}$ combinations. The subset that maximizes and minimizes the P_e of the system is obtained and represents the worst and best possible selection, respectively. In addition, a comparison between the average performance of the greedy algorithm and the convex relaxation of the problem is shown in Fig. 1. In the plot, the P_e obtained by applying directly the greedy heuristic to the signal-to-noise ratio set function is denoted as *Greedy SNR*. From Fig. 1, it is seen that even though the submodular surrogate, given by expression (28), does not perform as good as optimizing the original signal-to-noise ratio set function, its performance is comparable to the one obtained by the convex relaxation approach. However, applying Algorithm 1 to the submodular surrogate incurs a significantly lower computational complexity due to its recursive implementation. In Fig. 1, the shaded area shows the region where all other sub-optimal samplers would lie for this problem.

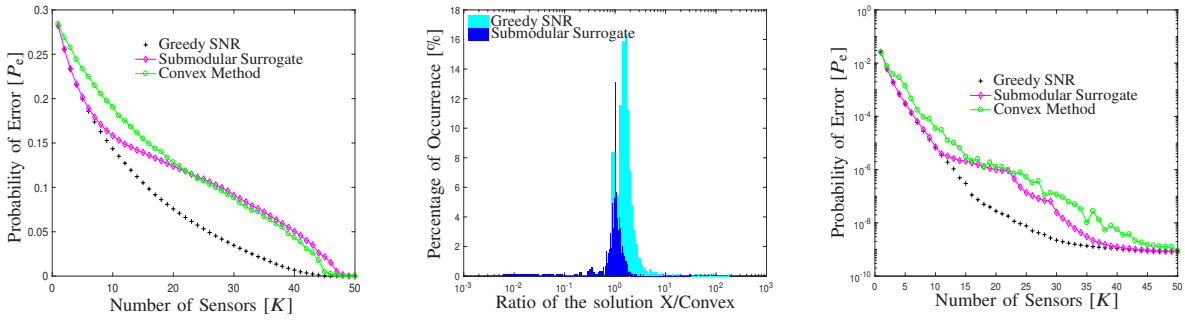


Fig. 2: Results for the random Toeplitz and uniform covariance matrices. (a) Bayesian probability of error P_e between the convex relaxation and the greedy heuristic for (19) with different subset sizes K when choosing from $M = 50$ available sensors and random Toeplitz covariance matrices. (b) Histogram of the distribution of the gain in SNR of different sensor selection strategies when the relaxed convex problem is considered as baseline. The sensor selection problem is solved for $M = 50$ available sensors over several realizations and different subset sizes. The height of the bar represents the relative frequency of the gain in the x-axis. (c) Bayesian probability of error P_e between the convex relaxation and the greedy heuristic for (19) with different subset sizes K when choosing from $M = 50$ available sensors and a uniform covariance matrix [c.f. (37)].

The previous example was intended to illustrate the performance of the discussed methods in comparison with the exhaustive search. However, for interesting problem sizes, exhaustive search solutions are not feasible even for small subset cardinalities. To illustrate the performance of the submodular surrogate for larger problem sizes, in the following example, instead of using the exhaustive search result as baseline, we compare the greedy heuristic with the convex relaxation for a problem of size $M = 50$. In Fig. 2a, the average performance over 1000 Monte-Carlo runs is shown, when the common covariance matrix $\mathbf{\Sigma}$, is considered to be a random Toeplitz symmetric matrix, and the mean vector i.i.d. Gaussian as before. Similar to the results from the previous example, the greedy rule from Algorithm 1 provides the lowest P_e when it is applied to the original signal-to-noise ratio function. As before, the submodular surrogate provides subsets with comparable system performance as the convex relaxation method with randomization, but with a reduced computational cost.

In Fig. 2b, we show the ratio between the SNR of the greedy and the submodular surrogate with respect to the solution of the relaxed convex problem for 100 Monte-Carlo realizations of problem (19) when random Toeplitz covariance matrices are considered for $\mathbf{\Sigma}$. The percentage of occurrence is shown in the vertical axis of the bar plot. In each Monte-Carlo run, the solution using the three approaches was computed, for the subset sizes $K = \{1, 6, 11, 16, 21, 26, 31, 36, 41, 46\}$. The histogram is computed over all subset sizes for each of the methods. It is evident from Fig. 2b that the greedy heuristic, when applied to the original signal-to-noise ratio, provides the best performance of all methods. As expected, the submodular surrogate set function provides similar results as the convex relaxation due to the fact that both are constructed from the Schur complement.

Finally, in Fig. 2c the comparison of the different methods is shown for the case when the covariance matrix $\mathbf{\Sigma}$ is considered to be a uniform correlation covariance matrix, i.e.,

$$\mathbf{\Sigma} = \begin{bmatrix} 1 & \rho & \rho & \dots & \rho \\ \rho & 1 & \rho & \dots & \rho \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \rho & \rho & \rho & \dots & 1 \end{bmatrix}, \tag{37}$$

with correlation factor $\rho = 0.43$. From Fig. 2c it is seen that the submodular surrogate outperforms the convex relaxation for all subset sizes. However, the best performance is achieved by the greedy heuristic applied directly to the signal-to-noise ratio set function.

E. When Does Greedy on the SNR Fail?

In the previous part, it has been numerically shown that the greedy heuristic applied directly to the signal-to-noise ratio set function might perform better than both the convex and submodular relaxations of the problem. However, we should be aware that the application of the greedy heuristic for a non-submodular maximization does not provide any optimality guarantees in general. Therefore, there might be problem instances in which the direct maximization of such a set function could lead to arbitrary bad results. In order to illustrate the importance of submodularity for the greedy heuristic, we show an example of the sensor selection problem in which the greedy method applied to the signal-to-noise performs worse than the submodular surrogate. Consider an example with $M = 3$ available sensors, from which we desire to obtain the best subset of $K = 2$ sensors which provides the highest signal-to-noise ratio. In addition, we consider the case where the difference of the mean vectors is the all-one vector, i.e., $\boldsymbol{\theta} = [1, 1, 1]^T$. The covariance matrix for the noise is given by the block matrix

$$\mathbf{\Sigma} = \begin{bmatrix} 1/(1-\rho^2) & -\rho/(1-\rho^2) & 0 \\ -\rho/(1-\rho^2) & 1/(1-\rho^2) & 0 \\ 0 & 0 & 1 \end{bmatrix},$$

where $\rho \in [0, 1)$. The signal-to-noise ratio set function is defined as $s(\mathcal{A}) = \mathbf{1}_{\mathcal{A}}^T \mathbf{\Sigma}_{\mathcal{A}}^{-1} \mathbf{1}_{\mathcal{A}}$. Since $s(\{1\}) = s(\{2\}) = (1-\rho^2)$ and $s(\{3\}) = 1$, Algorithm 1 will select $\{3\}$ first as $\rho \leq 1$, i.e., $\mathcal{A}_1 = \{3\}$. Then, either $\{1\}$ or $\{2\}$ are chosen next as both have the same set function value, i.e.,

$$s(\{3, 1\}) = s(\{3, 2\}) = s(\mathcal{A}_{\mathcal{G}}) = 2 - \rho^2,$$

where $\mathcal{A}_{\mathcal{G}}$ denotes the set obtained from the greedy SNR solution, i.e., obtained by greedily maximizing the SNR. However, the maximum of the set function is attained with

the set $\mathcal{A}^* = \{1, 2\}$ which provides the set function value $s(\mathbf{w}_{\mathcal{A}^*}) = 2 + 2\rho$. For the limiting case of $\rho \to 1$, we obtain

$$\lim_{\rho \to 1} \frac{s(\mathcal{A}_{\mathcal{G}})}{s(\mathcal{A}^*)} = 0.25.$$

Even though the greedy heuristic can provide good results in many cases, one should thus be aware that it could get stuck in solutions far from the optimal.

We will now show for the above example that, on average, applying the greedy heuristic to the submodular surrogate performs better than applying it to the original SNR cost set function. First, let us consider the following decomposition of the noise covariance matrix,

$$\mathbf{S} = \mathbf{\Sigma} - a\mathbf{I} = \begin{bmatrix} \frac{1}{1-\rho^2} - a & -\frac{\rho}{1-\rho^2} & 0 \\ -\frac{\rho}{1-\rho^2} & \frac{1}{1-\rho^2} - a & 0 \\ 0 & 0 & 1-a \end{bmatrix}, \tag{38}$$

for any a chosen as described in Theorem 2.

Then, the inverse of (38) can be expressed as

$$\mathbf{S}^{-1} = \begin{bmatrix} \frac{a\rho^2 - a + 1}{a^2\rho^2 - a^2 + 2a - 1} & -\frac{\rho}{a^2\rho^2 - a^2 + 2a - 1} & 0 \\ -\frac{\rho}{a^2\rho^2 - a^2 + 2a - 1} & -\frac{a\rho^2 - a + 1}{a^2\rho^2 - a^2 + 2a - 1} & 0 \\ 0 & 0 & -\frac{1}{a-1} \end{bmatrix}.$$

The submodular cost set function can be evaluated for each of the sensors by considering its factors as in (31), i.e.,

$$\begin{aligned} \gamma(\{i\}) &= \det\left(\mathbf{S}^{-1} + a^{-1}\mathbf{I}_{\{i\}}\right) = \frac{1}{a\,(a-1)\left(a^2\rho^2 - a^2 + 2a - 1\right)}, \\ s(\{i\}) &= 1 - \rho^2, \text{ for } i = 1, 2, \end{aligned}$$

and

$$\begin{aligned} \gamma(\{3\}) &= \det\left(\mathbf{S}^{-1} + a^{-1}\mathbf{I}_{\{3\}}\right) = \frac{1-\rho^2}{a\,(a-1)\left(a^2\rho^2 - a^2 + 2a - 1\right)} \\ s(\{3\}) &= 1. \end{aligned}$$

It is clear that the submodular cost set function provides the same value for any of the sensors, i.e.,

$$\gamma(\{1\})s(\{1\}) = \gamma(\{2\})s(\{2\}) = \gamma(\{3\})s(\{3\}).$$

Hence, if we break this tie arbitrarily, the possible values of the cost set function are

$$\begin{aligned} \gamma(\{1,2\})s(\{1,2\}) = \gamma(\{2,1\})s(\{2,1\}) &= \frac{2+2\rho}{a^2\,(a-1)\left(a^2\rho^2 - a^2 + 2a - 1\right)} \\ \gamma(\{3,1\})s(\{3,1\}) = \gamma(\{3,2\})s(\{3,2\}) &= \frac{2-\rho^2}{a^2\,(a-1)\left(a^2\rho^2 - a^2 + 2a - 1\right)}, \end{aligned}$$

where we consider the fact that the greedy heuristic does not select the 3rd sensor after the 1st or the 2nd sensor has been selected, i.e., the marginal gain is larger when the sensors $\{1, 2\}$ are selected. Therefore, the average value attained by the submodular method is

$$\begin{aligned} E\left[s(\mathcal{A}_{\mathcal{S}})\right] &= \tfrac{1}{3}(s(\{1,2\}) + s(\{2,1\})) + \tfrac{1}{6}(s(\{3,1\}) + s(\{3,2\})) \\ &= \tfrac{2}{3}s(\mathcal{A}^*) + \tfrac{2-\rho^2}{3}, \end{aligned}$$

where $\mathcal{A}_{\mathcal{S}}$ is the set returned by the maximization of the submodular surrogate. In the limiting case $\rho \to 1$, we have the following limit

$$\lim_{\rho \to 1} \frac{E\left[s(\mathcal{A}_{\mathcal{S}})\right]}{s(\mathcal{A}^*)} = 0.75,$$

which provides a higher approximation ratio compared with the previously seen greedy heuristic. However, it is clear that the proposed method also suffers from one of the drawbacks of greedy methods: when more than one possible solution obtains the same cost set function value, either ties should be broken arbitrarily or multiple branches have to be initialized.

Now, we show a larger instance of the previous example, where a set of $M = 200$ available sensors are considered. Furthermore, a block precision matrix $\mathbf{\Sigma}^{-1}$ with the following structure is considered for performing sensor selection

$$\mathbf{\Sigma}^{-1} = \begin{bmatrix} \mathbf{T} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \in \mathbb{R}^{M \times M} \tag{39}$$

where $\mathbf{T} = \text{Toeplitz}([1, \rho^1, \rho^2, \ldots, \rho^{\text{M}/2-1}]) \in \mathbb{R}^{\text{M}/2 \times \text{M}/2}$ is an exponential decaying Toeplitz matrix, and $\mathbf{I} \in \mathbb{R}^{\lfloor M/2 \rfloor \times \lfloor M/2 \rfloor}$ is the identity matrix. This kind of precision matrices could arise in systems where only a subset of sensors are calibrated, i.e., block of sensors whose precision matrix is the identity. The mean difference vector, i.e., $\boldsymbol{\theta} = \boldsymbol{\theta}_1 - \boldsymbol{\theta}_0$, is considered the all-ones vector, and ties in the selection are broken arbitrarily. In this example $\rho = 0.18$ has been fixed. From Fig. 3 it can be seen that even though for a small number of selected sensors both methods achieve similar SNR, the submodular surrogate outperforms the Greedy SNR method for most of the subset sizes. This result is expected due to the fact that the worst case bound given in Theorem 2 for ϵ-submodular set functions worsen as the size of the solution increases. More importantly, the submodular surrogate reaches the maximum SNR when half the sensors, i.e., for 50% compression, have been selected, whereas the Greedy SNR requires all the sensors to reach the maximum SNR. We want to emphasize that even though throughout the manuscript we focus on the cardinality constraint formulation (P-CC) [cf. (11)], the proposed methods are also suitable for performance constraint (P-DC) formulations [cf. (12)]. In many instances reducing the number of sensors too much might not lead to systems meeting minimum requirements, e.g., operational SNR. In this regard, observing Fig. 3, for a nominal system with a requirement of $SNR \geq 20dB$ (i.e., the performance constraint in (12)) only using Greedy SNR will lead to a solution containing a high number of sensors, while using the proposed surrogate leads to a solution that involves less sensors. Therefore, despite that for low values of K, there is no notable difference between both algorithms, in cases that a fixed SNR is required, using the proposed surrogate provides a clear advantage with respect to Greedy SNR.

V. Observations with Uncommon Covariances

In this section, we discuss sensor selection for detection when the data model for the hypotheses under test differ in their second-order statistics. For the case of Gaussian distributed measurements we can assume that the conditional distributions for the binary hypothesis test are given by

$$\begin{aligned} \mathcal{H}_0 &: \mathbf{x} \sim \mathcal{N}(\boldsymbol{\theta}, \mathbf{\Sigma}_0) \\ \mathcal{H}_1 &: \mathbf{x} \sim \mathcal{N}(\boldsymbol{\theta}, \mathbf{\Sigma}_1), \end{aligned} \tag{40}$$

where the mean vector $\boldsymbol{\theta} \in \mathbb{R}^M$ is shared by both hypotheses and the second-order statistics of the data are characterized by the $M \times M$ covariance matrices $\mathbf{\Sigma}_0$ and $\mathbf{\Sigma}_1$ for the hypothesis $\mathcal{H}_0$ and $\mathcal{H}_1$, respectively.

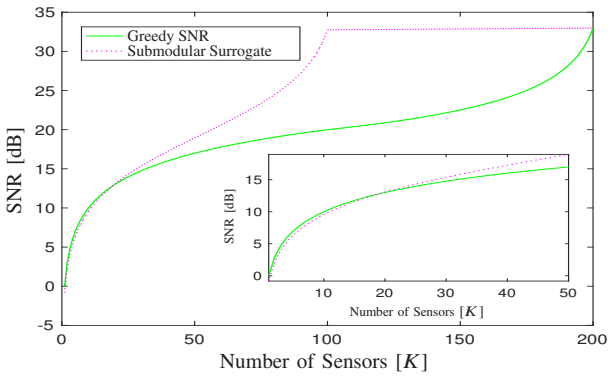


Fig. 3: Signal-to-noise ratio between the Greedy SNR and the submodular surrogate for different subset sizes K when choosing from $M = 200$ available sensors for an instance of the problem with precision matrix given by (39)

Using the metrics discussed in Section II, it has been suggested in [17] that the different distance measures between probability distributions can be applied to construct selection strategies using convex optimization. However, it turns out that some of the metrics to optimize can only be expressed as the difference of submodular functions, therefore the SupSub procedure described in Section III can be employed for its optimization. In the next section, we show how it is possible to decompose the divergence measures into the difference of submodular functions.

A. Submodular Decomposition of Divergence Measures

Unlike the case with commons means, the three distances discussed are not scaled versions of each other. For the linear model in (8) under Gaussian noise, the Bhattacharyya distance (3) is given as the following difference of submodular set functions

$$\begin{aligned} f(\mathcal{A}) &= \mathcal{B}(\mathcal{H}_1\|\mathcal{H}_0) \coloneqq g(\mathcal{A}) - h(\mathcal{A}); \\ g(\mathcal{A}) &= \frac{1}{2}\log\det(\mathbf{\Sigma}_{\mathcal{A}}); \\ h(\mathcal{A}) &= \frac{1}{4}(\log\det(\mathbf{\Sigma}_{0,\mathcal{A}}) + \log\det(\mathbf{\Sigma}_{1,\mathcal{A}})). \end{aligned} \tag{41}$$

The submodularity of $h(\mathcal{A})$ and $g(\mathcal{A})$ is clear as both functions are linear combinations of entropy functions. As a result, the Bhattacharyya distance can be approximately maximized using the SupSub procedure described in Algorithm 2.

Differently from the Bhattacharyya distance, the expressions for the KL divergence and the J-divergence in (4) and (5) for the distributions in (40) do not provide a direct decomposition in submodular set functions because in both divergences there are trace terms that cannot be expressed directly as a difference of submodular functions. Even though such decompositions exist [41], in general, finding them incurs exponential complexity [42]. However, similarly as in the case of the signal-to-noise ratio cost set function, a readily available submodular surrogate can be employed in order to optimize both distances using the SupSub procedure.

In order to obtain a submodular approximation of the trace term, let us consider the following set function

$$q(\mathbf{\Sigma}_{\mathcal{A}}, \mathbf{\Psi}_{\mathcal{A}}) = \operatorname{tr}\{\mathbf{\Sigma}_{\mathcal{A}}^{-1}\mathbf{\Psi}_{\mathcal{A}}\}, \tag{42}$$

where $\mathcal{A}$ is the index set of the selected sensors and $\mathbf{\Sigma}_{\mathcal{A}}$ and $\mathbf{\Psi}_{\mathcal{A}}$ are submatrices defined by the rows and columns of $\mathbf{\Sigma}$ and $\mathbf{\Psi}$, respectively, given by the elements of the set $\mathcal{A}$. Let us decompose one of the matrices as $\mathbf{\Sigma} = a\mathbf{I} + \mathbf{S}$, where a nonzero $a \in \mathbb{R}$ is chosen as described in Theorem 2 and therefore $\mathbf{S} \succ 0$. The set function in (42) is then equivalent to

$$\begin{aligned} q(\mathbf{\Sigma}_{\mathcal{A}}, \mathbf{\Psi}_{\mathcal{A}}) &= \operatorname{tr}\{\mathbf{S}^{-1}\mathbf{\Psi} - \mathbf{S}^{-1}[\mathbf{S}^{-1} + a^{-1}\operatorname{diag}(\mathbf{1}_{\mathcal{A}})]^{-1}\mathbf{S}^{-1}\mathbf{\Psi}\} \\ &= \operatorname{tr}\{\mathbf{\Psi}^{\frac{T}{2}}\mathbf{S}^{-\frac{1}{2}}(\mathbf{I} - \mathbf{S}^{-\frac{T}{2}}[\mathbf{S}^{-1} \\ &\qquad + a^{-1}\operatorname{diag}(\mathbf{1}_{\mathcal{A}})]^{-1}\mathbf{S}^{-\frac{1}{2}})\mathbf{S}^{-\frac{T}{2}}\mathbf{\Psi}^{\frac{1}{2}}\} \\ &= \sum_{i=0}^{M} \operatorname{tr}\{\mathbf{z}_i^T(\mathbf{I} - \mathbf{S}^{-\frac{T}{2}}[\mathbf{S}^{-1} \\ &\qquad + a^{-1}\operatorname{diag}(\mathbf{1}_{\mathcal{A}})]^{-1}\mathbf{S}^{-\frac{1}{2}})\mathbf{z}_i\}, \end{aligned}$$

where $\mathbf{z}_i$ has been defined as the i-th column of $\mathbf{S}^{-\frac{T}{2}}\mathbf{\Psi}^{\frac{1}{2}}$. Analogously to the uncommon means case, where the signal-to-noise ratio was replaced by its submodular surrogate, we can substitute $q(\mathbf{\Sigma}_{\mathcal{A}}, \mathbf{\Psi}_{\mathcal{A}})$ by the following submodular set function

$$q_{\text{sub}}(\mathbf{\Sigma}_{\mathcal{A}}, \mathbf{\Psi}_{\mathcal{A}}) \coloneqq \sum_{i=1}^{M} \log\det\begin{bmatrix} \mathbf{S}^{-1} + a^{-1}\operatorname{diag}(\mathbf{1}_{\mathcal{A}}) & \mathbf{S}^{-\frac{1}{2}}\mathbf{z}_i \\ \mathbf{z}_i^T\mathbf{S}^{-\frac{T}{2}} & \mathbf{z}_i^T\mathbf{z}_i \end{bmatrix}$$

which is submodular on the set of selected entries $\mathcal{A}$. It is clear that the set function $q_{\text{sub}}(\mathbf{\Sigma}_{\mathcal{A}}, \mathbf{\Psi}_{\mathcal{A}})$ is submodular as it is a non-negative combination of submodular set functions in $\mathcal{A}$. Furthermore, as this set function shares a similar structure with respect to the signal-to-noise ratio set function [cf. (20)], i.e.,

$$\begin{aligned} q_{\text{sub}}(\mathbf{\Sigma}_{\mathcal{A}}, \mathbf{\Psi}_{\mathcal{A}}) \quad &\coloneqq M\log\det(\mathbf{S}^{-1} + a^{-1}\operatorname{diag}(\mathbf{1}_{\mathcal{A}})) + \\ &\qquad \sum_{i=1}^{M} \log(\boldsymbol{\psi}_i^T\mathbf{\Phi}_{\mathcal{A}}^T\mathbf{\Sigma}_{\mathcal{A}}^{-1}\mathbf{\Phi}_{\mathcal{A}}\boldsymbol{\psi}_i), \end{aligned} \tag{43}$$

where $\boldsymbol{\psi}_i$ is the ith column of $\mathbf{\Psi}^{\frac{1}{2}}$, an efficient evaluation of (43) can be performed through a recursive definition similar to the one in (36). Unfortunately, as the summation is over M terms, this formulation leads to a worst-case complexity of $O(M^2K^3)$ for finding the solution through a greedy heuristic. However, for instances with $K \ll M$ this algorithm improves, in terms of speed, with respect to the cubic complexity of the convex relaxation.

After the introduction of the submodular set function q_{sub}, surrogates for the divergences $\mathcal{K}(\cdot)$ and $\mathcal{D}_{\text{J}}(\cdot)$ denoted as $\mathcal{K}_{\text{sub}}(\cdot)$ and $\mathcal{D}_{\text{J,sub}}(\cdot)$, respectively, can be obtained. The following is observed from these surrogates:

- $\mathcal{K}_{\text{sub}}(\cdot)$ can be expressed as a mixture of submodular and supermodular set functions as

$$\begin{aligned} \mathcal{K}_{\text{sub}}(\mathcal{H}_1\|\mathcal{H}_0) &= g(\mathcal{A}) - h(\mathcal{A}); \\ g(\mathcal{A}) &= \frac{1}{2}\log\det(\mathbf{\Sigma}_{0,\mathcal{A}}) + \frac{1}{2}q_{\text{sub}}(\mathbf{\Sigma}_{0,\mathcal{A}}, \mathbf{\Sigma}_{1,\mathcal{A}}); \\ h(\mathcal{A}) &= \frac{1}{2}\log\det(\mathbf{\Sigma}_{1,\mathcal{A}}). \end{aligned}$$

- $\mathcal{D}_{\text{J,sub}}(\cdot)$ is a submodular set function as it is a non-negative combination of two submodular functions, i.e.,

$$\mathcal{D}_{\text{J,sub}}(\mathcal{H}_1\|\mathcal{H}_0) = \frac{1}{2}(q_{\text{sub}}(\mathbf{\Sigma}_{0,\mathcal{A}}, \mathbf{\Sigma}_{1,\mathcal{A}}) + q_{\text{sub}}(\mathbf{\Sigma}_{1,\mathcal{A}}, \mathbf{\Sigma}_{0,\mathcal{A}})).$$

From these results, it is clear that $\mathcal{K}_{\text{sub}}(\cdot)$ can be optimized using the SupSub procedure in Algorithm 2 as in the case of the Bhattacharyya distance, while $\mathcal{D}_{\text{J,sub}}(\cdot)$ can be directly optimized using the greedy heuristic from Algorithm 1.

B. Uncommon Means and Uncommon Covariances

So far, only particular cases of the general hypothesis testing problem between two Gaussian distributions have been discussed. However, in the following, we show that the general case, i.e., distinct means and covariance matrices, can be solved using the same heuristics as discussed for the particular cases presented earlier in the previous sections.

To show that we can reuse the same machinery as in the uncommon covariances case, first let us consider the KL divergence:

$$\begin{aligned}\mathcal{K}(\mathcal{H}_1\|\mathcal{H}_0) \quad &= \tfrac{1}{2}\Big(\text{tr}(\boldsymbol{\Sigma}_{0,\mathcal{A}}^{-1}\tilde{\boldsymbol{\Sigma}}_{1,\mathcal{A}}) - M + \log\det(\boldsymbol{\Sigma}_{0,\mathcal{A}}) \\ &\qquad - \log\det(\boldsymbol{\Sigma}_{1,\mathcal{A}})\Big).\end{aligned} \tag{44}$$

In the above expression we have rewritten the quadratic form in terms of a trace [cf. Table I], and defined the matrix $\tilde{\boldsymbol{\Sigma}}_{1,\mathcal{A}} \coloneqq \boldsymbol{\Sigma}_{1,\mathcal{A}} + (\boldsymbol{\theta}_{1,\mathcal{A}} - \boldsymbol{\theta}_{0,\mathcal{A}})(\boldsymbol{\theta}_{1,\mathcal{A}} - \boldsymbol{\theta}_{0,\mathcal{A}})^T$ which remains symmetric. From (43) it can be seen that the decomposition proposed before can directly be used by just replacing $\boldsymbol{\Sigma}_{1,\mathcal{A}}$ by $\tilde{\boldsymbol{\Sigma}}_{1,\mathcal{A}}$ in the $q_{\text{sub}}(\cdot,\cdot)$ set function. Therefore, for this measure there is no distinction between these two cases. In addition, as the J-divergence is a sum of KL divergences, the decomposition of the J-divergence follows immediately.

For the Bhattacharyya distance, the quadratic form that appears in the expression for the general case, i.e., the term related to the distances between the means, needs to be added to the decomposition given in (40). As this new term is not submodular, the surrogate function proposed for the uncommon means case can be directly employed to provide a submodular set function for the decomposition. As a result, the surrogate distance measure for the Bhattacharyya distance can be given through the decomposition:

$$\begin{aligned}\mathcal{B}_{\text{sub}}(\mathcal{H}_1\|\mathcal{H}_0) &\coloneqq g(\mathcal{A}) - h(\mathcal{A}); \\ g(\mathcal{A}) &= \frac{1}{2}\log\det(\boldsymbol{\Sigma}_{\mathcal{A}}) + \frac{1}{8}\log\det(\mathbf{M}_{\mathcal{A}}); \\ h(\mathcal{A}) &= \frac{1}{4}(\log\det(\boldsymbol{\Sigma}_{0,\mathcal{A}}) + \log\det(\boldsymbol{\Sigma}_{1,\mathcal{A}})),\end{aligned} \tag{45}$$

which can be optimized using the approach discussed for the uncommon covariance case.

In summary, the greedy heuristic for the most general case of uncommon means and uncommon covariances can be developed using straightforward adaptations of the methods presented in Sec. V-A.

C. Numerical Examples

We demonstrate the applicability of the SupSub in Algorithm 2 for solving the maximization of the different divergences used for sensor selection, and its respective surrogates by comparing the results with the widely used CCP heuristic. To do so, first we perform an exhaustive search to solve the sensor selection problem for the test in (40) under the Neyman-Pearson setting. We find the subset of size K that maximizes the KL divergence, for random covariance matrices of size $M = 15$ and for random Toeplitz matrices of size $M = 50$. The results are shown in Fig. 4a and Fig. 4b, respectively. From these examples, it is seen that the greedy heuristic of Algorithm 1 applied to the KL divergence (*KL Greedy*), the SupSub procedure using both the original KL expression (*SupSub KL-Div*)[1] and the submodular surrogate (*SupSub Surrogate*) perform either better or as good as the CCP heuristic while incurring a much lower complexity. For random Toeplitz matrices, as seen in Fig. 4b, all the methods perform close to each other, however the CCP method achieves this performance with a higher computational load.

Binary Classification

Due to the non-monotonic behavior of the classification curves with respect to the number of features, i.e., the error of a classifier does not necessarily reduce when more features are used, a fast and reliable way to select the most relevant features for a given dataset is required. Therefore, in the following, we present two examples for binary classification where the KL divergence is used as a feature selection metric and it is optimized using the methods described in this work. In these examples, the PRTools Toolbox [54] is used for training classifiers. The built-in feature selection method, based on cross-validation, is used as baseline for comparison with the proposed methods based on the submodular machinery.

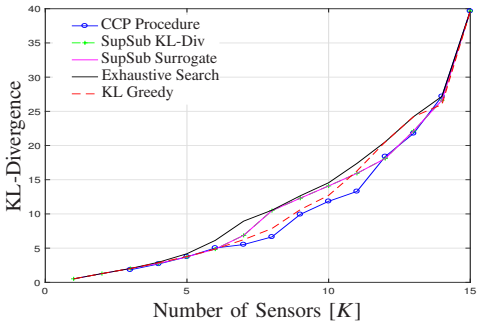
In the first example, we start by considering a simple case: two classes described by Gaussian distributions parametrized by their covariance matrices, $\{\boldsymbol{\Sigma}_0, \boldsymbol{\Sigma}_1\}$. In this scenario, the covariance matrices are a pair of Toeplitz matrices. The number of features considered for this example is 50. The trained classifier is the quadratic discriminant classifier (QDC) [53]. Furthermore, the 20/80 rule for training and testing has been used for the 500 objects contained in the dataset, i.e., 20% of the data set has been used for training the classifier and 80% for reporting its performance on unseen data. Additionally, random sampling of the objects for training has been performed. The selection of such a classifier is due to the nature of the dataset, i.e., as the assumption of Gaussianity of the features holds, QDC is the Bayes detector for equiprobable classes. The comparison of the classification soft error for the selected classifier is shown in Fig. 5a. The reported error in this figure is given by

$$e \coloneqq \frac{E_0}{|C_0|}P(C_0) + \frac{E_1}{|C_1|}P(C_1), \tag{46}$$

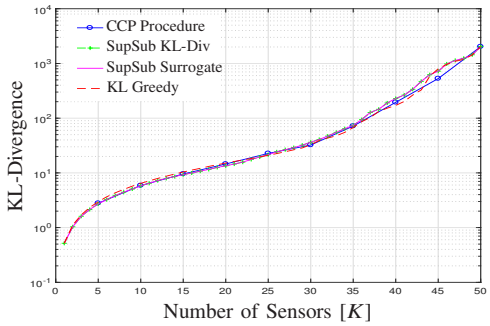
where E_i denotes the number of erroneously classified objects for the ith class, denoted by C_i, and $P(C_i)$ represents the prior probability for the ith class in the validation set.

As expected, the classification error decreases as the number of selected features increases as in this example QDC provides decision boundaries based on the log-likehood ratio test under

[1]This is done by computing the expressions of the modular upper bounds [cf. (17) and (18)] for the set function $q(\boldsymbol{\Sigma}_{\mathcal{A}}, \boldsymbol{\Psi}_{\mathcal{A}})$, despite that the function is not submodular.



(a)



(b)

Fig. 4: (a) KL divergence of the different sensor selection methods for different subset sizes K for random covariance matrices. (b) KL divergence of the different sensor selection methods for different subset sizes K for random Toeplitz matrices.

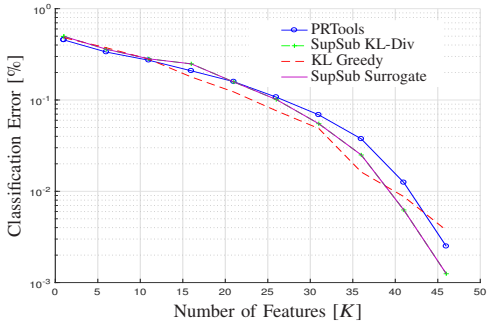
Gaussian assumptions. In this example, both methods based on the SupSub procedure provide a similar classification error, being mostly below the PRTools baseline result. In this scenario, for roughly half the number of sensors, the greedy heuristic over the KL divergence provides the lowest classification error.

Real Dataset Example

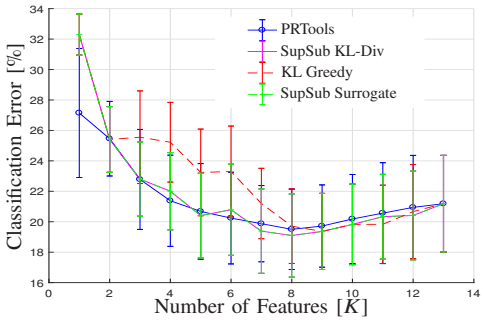
As a second example, we use the *Cleveland Heart Disease Data Set* [55] in which a set of 76 attributes from 303 patients are reported describing the presence of a heart disease. Due to the nature of the data, only 14 of the reported attributes are used as features, e.g., id number, social security number, and similar attributes are omitted. In the original dataset, the presence of heart disease is described by an integer number in the range $\{0, 1, \ldots, 4\}$, however in this scenario we consider a binary hypothesis test in which the label $l = 0$ represents a *healthy heart* and the labels $l \geq 1$ represent a patient with any kind of heart disease. For further information of the complete dataset the reader is referred to the related online repository [56]. Similarly to the previous case, only 20% of the data (randomly selected) is used to train the classifier selected for this problem. In this setup, the true covariance matrices for the features are considered for performing selection. That is, from the whole data set the second-order statistics for each feature, within a given class, are computed and the resulting covariance matrix is considered as the true covariance matrix for the data. The same criterion and baseline are used to perform the selection of the features from the dataset, and the results are reported over hundred random selections for the training subset. For this dataset, a support vector machine (SVM) was trained to discriminate between the healthy and unhealthy patients. In Fig. 5b, the average classification error, in percentage, is reported for each method with their respective 95% confidence interval. From this plot it can be observed that the methods based on the SupSub procedure produce subsets of features which attain a similar performance as the baseline, i.e., the PRTool built-in function optimizing over the QDC metric. However, the method that only uses the greedy heuristic to maximize the KL divergence obtains subsets with a worse performance for a small number of features. When the number of features is close to the maximum, the three methods based on the greedy rule perform slightly better, in both mean error and error deviation, than the baseline feature selection method. Notice the convex behavior of the classification error for the SVM classifier in Fig. 5b. Differently from the previous example, here the dataset structure is more complex and no Gaussian distribution properly describes it. Therefore, increasing the number of features could possibly overtrain the classifier hindering its generalization capabilities. However, it is important to notice than even when Gaussianity is not granted, the maximization of the KL divergence as a metric for feature selection leads to subsets with a smaller average classification error.

VI. Conclusions

In this paper, we have considered submodular optimization for model-based sparse sampler design for Gaussian signal detection with correlated data. Differently from traditional approaches based on convex optimization, in this work we have focused on efficient methods to solve the sensor selection problem using submodular set functions. We have shown how the discrete optimization of widely used performance metrics, for both Bayesian and Neyman-Pearson settings, can be approximated and solved using the submodular optimization machinery. For Gaussian observations with common covariance and uncommon means we bounded the ϵ-submodularity constant of the SNR set function, and derived a submodular surrogate based on the Schur complement for instances in which such a constant is large. We have shown that for series of practical classes of covariance matrices this surrogate leads to a performance comparable with the convex relaxation of the problem, but at a reduced computational complexity. For the case of common means and uncommon covariance, we propose to employ the SupSub procedure for maximizing the difference of submodular set functions. When the decomposition of the divergence measure into submodular functions is not straightforward, we introduce surrogate decompositions based on the Schur complement that can be evaluated efficiently. This approach can be easily adapted for the case of uncommon means uncommon covariances. Furthermore, a series of numerical examples with both synthetic and real data demonstrate the effectiveness of the proposed methods to perform both sensor and feature selection even when the data is not Gaussian distributed.



(a)



(b)

Fig. 5: (a) Classification soft error, weighted with class priors, when using QDC for a Gaussian binary classification problem. (b) Classification error for SVMs trained using different feature selection methods for the Heart-Cleveland data set.

APPENDIX A
PROOF OF THEOREM 2

Proof. First, consider the SNR set function that is defined as [cf. (20)]

$$s(\mathcal{A}) = \boldsymbol{\theta}_{\mathcal{A}}^T \boldsymbol{\Sigma}_{\mathcal{A}}^{-1} \boldsymbol{\theta}_{\mathcal{A}}. \tag{47}$$

Combining the above expression and the decomposition $\boldsymbol{\Sigma} = a\mathbf{I} + \mathbf{S}$, the signal-to-noise ratio can be rewritten as [17]

$$\begin{aligned} s(\mathcal{A}) &= \boldsymbol{\theta}_{\mathcal{A}}^T \boldsymbol{\Sigma}_{\mathcal{A}}^{-1} \boldsymbol{\theta}_{\mathcal{A}} && (48)\\ &= \boldsymbol{\theta}^T \mathbf{S}^{-1} \boldsymbol{\theta} - \boldsymbol{\theta}^T \mathbf{S}^{-1} [\mathbf{S}^{-1} + a^{-1}\mathrm{diag}(\mathbf{1}_{\mathcal{A}})]^{-1} \mathbf{S}^{-1} \boldsymbol{\theta}, && (49)\\ &= \boldsymbol{\theta}^T \mathbf{S}^{-1} \boldsymbol{\theta} + h(\mathcal{A}), && (50) \end{aligned}$$

where the non-zero entries of the vector $\mathbf{1}_{\mathcal{A}}$ are given by the set $\mathcal{A}$ and $h(\mathcal{A}) := -\boldsymbol{\theta}^T \mathbf{S}^{-1} \left[\mathbf{S}^{-1} + a^{-1}\mathrm{diag}(\mathbf{1}_{\mathcal{A}})\right]^{-1} \mathbf{S}^{-1} \boldsymbol{\theta}$.

For the sake of simplicity, we will rewrite the SNR as follows

$$s(\mathcal{A}) := C_1 \tilde{s}(\mathcal{A}), \tag{51}$$

where we have defined $C_1 = \|\boldsymbol{\theta}\|_2^2$, and $\tilde{s}(\mathcal{A})$ is the SNR set function with $\boldsymbol{\theta}$ being substituted by $\tilde{\boldsymbol{\theta}} = \boldsymbol{\theta}/\|\boldsymbol{\theta}\|_2$, i.e., the SNR set function is computed only considering the direction of the vector $\boldsymbol{\theta}$.

Now, let us assume that there exists a $\epsilon' \in \mathbb{R}_+$ such that

$$-\epsilon' \leq \tilde{s}(\mathcal{A}) - \hat{s}(\mathcal{A}) \leq \epsilon', \tag{52}$$

for any $\mathcal{A} \subseteq \mathcal{V}$ and some modular set function $\hat{s}(\mathcal{A})$. Using (52) and considering $\mathcal{A} \subseteq \mathcal{B} \subseteq \mathcal{V}$ and $i \notin \mathcal{B}$, we can obtain the following expression

$$\delta(\mathcal{A} \cup \{i\}) - \delta(\mathcal{A}) - \delta(\mathcal{B} \cup \{i\}) + \delta(\mathcal{B}) \geq -4\epsilon', \tag{53}$$

where we have defined $\delta(\mathcal{A}) = \tilde{s}(\mathcal{A}) - \hat{s}(\mathcal{A})$. Due to the modularity of $\hat{s}(\mathcal{A})$, i.e.,

$$\hat{s}(\mathcal{A} \cup \{i\}) - \hat{s}(\mathcal{A}) - \hat{s}(\mathcal{B} \cup \{i\}) + \hat{s}(\mathcal{B}) = 0, \tag{54}$$

and using expressions (53) and (54), we can show that

$$\tilde{s}(\mathcal{A} \cup \{i\}) - \tilde{s}(\mathcal{A}) - \tilde{s}(\mathcal{B} \cup \{i\}) + \tilde{s}(\mathcal{B}) \geq -4\epsilon'. \tag{55}$$

From this it is clear that the set function $\tilde{s}(\mathcal{A})$ is ϵ-submodular with an $\epsilon \leq 4\epsilon'$ (or equivalently, $s(\mathcal{A})$ is ϵ-submodular with an $\epsilon \leq 4\epsilon' C_1$). Therefore, for completing the proof we require to establish the bound in (52) for the specific ϵ' given in (22). In the following, we devote ourselves to this task.

For this proof, we select the following auxiliary set function:

$$\hat{s}(\mathcal{A}) = \tilde{\boldsymbol{\theta}}^T \mathbf{S}^{-1} \tilde{\boldsymbol{\theta}} + \hat{h}(\mathcal{A}), \tag{56}$$

where $\hat{h}(\mathcal{A})$ is chosen to be a *modular* [cf. (54)] set function and it is given by

$$\hat{h}(\mathcal{A}) = -\tilde{\boldsymbol{\theta}}^T \mathbf{S}^{-1} \left(a^{-1}\mathbf{I} + a^{-1}\mathrm{diag}(\mathbf{1}_{\mathcal{A}})\right)^{-1} \mathbf{S}^{-1} \tilde{\boldsymbol{\theta}}. \tag{57}$$

In (57), a scaled identity matrix has been introduced instead of the inverse of $\mathbf{S}$ [cf. (50)] to construct a modular set function. Here, it should be noticed that other set functions besides (57) could have been used for finding an upper bound on the ϵ constant. Depending on this choice, different bounds might be obtained.

To prove (52), we equivalently will establish the following inequalities

$$-\epsilon' \leq \tilde{h}(\mathcal{A}) - \hat{h}(\mathcal{A}) \leq \epsilon', \tag{58}$$

where we have defined $\tilde{h}(\mathcal{A}) = C_1^{-1} h(\mathcal{A})$. To obtain these inequalities, we can bound the difference of the positive definite (PD) matrices that are part of the quadratic forms in the set functions in (58). That is, we need to show that

$$\begin{aligned} -\epsilon'\mathbf{I} \quad &\leq \mathbf{S}^{-1} \left[a^{-1}\mathbf{I} + a^{-1}\mathrm{diag}(\mathbf{1}_{\mathcal{A}})\right]^{-1} \mathbf{S}^{-1} \\ &\quad -\mathbf{S}^{-1} \left[\mathbf{S}^{-1} + a^{-1}\mathrm{diag}(\mathbf{1}_{\mathcal{A}})\right]^{-1} \mathbf{S}^{-1} \leq \epsilon'\mathbf{I}. \end{aligned} \tag{59}$$

Considering $\mathrm{diag}(\mathbf{1}_{\mathcal{A}}) = \boldsymbol{\Phi}_{\mathcal{A}}^T \boldsymbol{\Phi}_{\mathcal{A}}$, we can apply the matrix inversion lemma to expand the difference of the matrices in between brackets above as

$$\boldsymbol{\Delta} := a\left(\mathbf{I} - \frac{1}{2}\boldsymbol{\Phi}_{\mathcal{A}}^T \boldsymbol{\Phi}_{\mathcal{A}}\right) - \mathbf{S} + \mathbf{S}\boldsymbol{\Phi}_{\mathcal{A}}^T \left(a\mathbf{I} + \boldsymbol{\Phi}_{\mathcal{A}} \mathbf{S} \boldsymbol{\Phi}_{\mathcal{A}}^T\right)^{-1} \boldsymbol{\Phi}_{\mathcal{A}} \mathbf{S}. \tag{60}$$

Hence (59) becomes,

$$-\epsilon'\mathbf{I} \leq \mathbf{S}^{-1} \boldsymbol{\Delta} \mathbf{S}^{-1} \leq \epsilon'\mathbf{I}. \tag{61}$$

As all terms in (60) are PD matrices, we upper bound the matrix in (60) by removing the negative terms in (60), and lower bound it by removing all the positive terms. That is,

$$-\left(\frac{a}{2}\mathbf{I} + \mathbf{S}\right) \leq \boldsymbol{\Delta} \leq a\mathbf{I} + \mathbf{S}\boldsymbol{\Phi}_{\mathcal{A}}^T \left(a\mathbf{I} + \boldsymbol{\Phi}_{\mathcal{A}} \mathbf{S} \boldsymbol{\Phi}_{\mathcal{A}}^T\right)^{-1} \boldsymbol{\Phi}_{\mathcal{A}} \mathbf{S}. \tag{62}$$

From the definition of $\mathbf{S}$ and a, we can notice that a possible lower bound for the expression above is given by

$$-\lambda_{\max}\{\boldsymbol{\Sigma}\}\mathbf{I} \leq \boldsymbol{\Delta}. \tag{63}$$

For the upper bound, we notice that by the maximum singular value of the second matrix, the following inequality holds

$$\begin{aligned}\boldsymbol{\Delta} &\preceq a\mathbf{I} + \sigma_{\max}\left\{\mathbf{S}\boldsymbol{\Phi}_{\mathcal{A}}^T\left(a\mathbf{I} + \boldsymbol{\Phi}_{\mathcal{A}}\mathbf{S}\boldsymbol{\Phi}_{\mathcal{A}}^T\right)^{-1}\boldsymbol{\Phi}_{\mathcal{A}}\mathbf{S}\right\}\mathbf{I} \\ &\preceq a\mathbf{I} + \lambda_{\min}^{-1}\left\{a\mathbf{I} + \boldsymbol{\Phi}_{\mathcal{A}}\mathbf{S}\boldsymbol{\Phi}_{\mathcal{A}}^T\right\}\sigma_{\max}^2\left\{\boldsymbol{\Phi}_{\mathcal{A}}\mathbf{S}\right\}\mathbf{I} \\ &\preceq a\mathbf{I} + \lambda_{\min}^{-1}\{a\mathbf{I} + \mathbf{S}\}\lambda_{\max}^2\{\mathbf{S}\}\mathbf{I}, \\ &\preceq (a + \nu\lambda_{\max}^2\{\mathbf{S}\})\mathbf{I},\end{aligned}$$

where the submultiplicativity and subadditivity of singular values, and the interlacing theorem for submatrices of PD matrices are used in the second and third inequality, respectively, and we have defined $\nu = \lambda_{\min}^{-1}\{\boldsymbol{\Sigma}\}$.

Considering that

$$\lambda_{\max}\{\boldsymbol{\Sigma}\} \leq a + \nu\lambda_{\max}^2\{\mathbf{S}\}, \tag{64}$$

we can bound the matrix in (60) by both sides as follows

$$-(a\mathbf{I} + \nu\lambda_{\max}^2\{\mathbf{S}\}\mathbf{I}) \preceq \boldsymbol{\Delta} \preceq a\mathbf{I} + \nu\lambda_{\max}^2\{\mathbf{S}\}\mathbf{I}. \tag{65}$$

Hence, we solely continue deriving the upper bound for the expression above as the obtained ϵ' will hold for both lower and upper bound.

Now, considering that the eigenvalues of $\boldsymbol{\Sigma}$ are larger than those of $\mathbf{S}$ by definition, $\mathbf{S}^{-1} \preceq \lambda_{\min}^{-1}\{\mathbf{S}\}\mathbf{I}$, and recalling that $a = \beta\lambda_{\min}\{\boldsymbol{\Sigma}\}$ we obtain

$$\mathbf{S}^{-1}\boldsymbol{\Delta}\mathbf{S}^{-1} \preceq \frac{a}{(1-\beta)^2\lambda_{\min}^2\{\boldsymbol{\Sigma}\}}\mathbf{I} + \frac{\nu\kappa^2}{(1-\beta)^2}\mathbf{I} = \epsilon'\mathbf{I}, \tag{66}$$

where κ is the condition number of the matrix $\boldsymbol{\Sigma}$, proving the result of the theorem.

For the limiting case, $a \to 0$ or equivalently $\beta \to 0$, we obtain:

$$\epsilon' = \frac{\kappa^2}{\lambda_{\min}\{\boldsymbol{\Sigma}\}}, \tag{67}$$

which shows a relation with the typical experiment design metrics, i.e., maximization of the minimum eigenvalue and log determinant (which promotes a good matrix condition). □

Appendix B
Proof of Proposition 1

Monotonicity:

Proof. Let us define the following:

$$\mathbf{T} = \begin{bmatrix}\mathbf{S}^{-1} & \mathbf{S}^{-1}\boldsymbol{\theta} \\ \boldsymbol{\theta}^T\mathbf{S}^{-1} & \boldsymbol{\theta}^T\mathbf{S}^{-1}\boldsymbol{\theta}\end{bmatrix},\ \mathbf{L}_{\mathcal{A}} = \begin{bmatrix}a^{-1}\mathrm{diag}(\mathbf{1}_{\mathcal{A}}) & \mathbf{0} \\ \mathbf{0} & \mathbf{0}\end{bmatrix}.$$

We can express the cost set function from (28) as $f(\mathcal{A}) = \log\det(\mathbf{T} + \mathbf{L}_{\mathcal{A}})$, where we have defined $\mathbf{M}_{\mathcal{A}} \coloneqq \mathbf{T} + \mathbf{L}_{\mathcal{A}}$. To prove the monotonicity of the set function we need to show

$$f(\mathcal{A}\cup\{i\}) - f(\mathcal{A}) = \log\frac{\det(\mathbf{M}_{\mathcal{A}} + \mathbf{L}_i)}{\det(\mathbf{M}_{\mathcal{A}})}.$$

Therefore, we should prove that $\det(\mathbf{M}_{\mathcal{A}} + \mathbf{L}_i) \geq \det(\mathbf{M}_{\mathcal{A}})$. This condition is implied by $\mathbf{M}_{\mathcal{A}} + \mathbf{L}_i \succeq \mathbf{M}_{\mathcal{A}}$, as $a \geq 0$. □

Submodularity :

Proof. Let us consider the previous definitions for $\mathbf{T}$ and $\mathbf{L}_{\mathcal{A}}$. We need to prove that the following expression is always positive

$$f(\mathcal{A}\cup i) - f(\mathcal{A}) - f(\mathcal{A}\cup\{i,j\}) + f(\mathcal{A}\cup j) = \log\frac{\det(\mathbf{M}_{\mathcal{A}} + \mathbf{L}_i)\det(\mathbf{M}_{\mathcal{A}} + \mathbf{L}_j)}{\det(\mathbf{M}_{\mathcal{A}})\det(\mathbf{M}_{\mathcal{A}} + \mathbf{L}_i + \mathbf{L}_j)} \geq 0$$

The above inequality is equivalent to

$$\frac{\det(\mathbf{M}_{\mathcal{A}} + \mathbf{L}_i)\det(\mathbf{M}_{\mathcal{A}} + \mathbf{L}_j)}{\det(\mathbf{M}_{\mathcal{A}})\det(\mathbf{M}_{\mathcal{A}} + \mathbf{L}_i + \mathbf{L}_j)} \geq 1$$

Noticing that $\mathbf{L}_i = a^{-1}\mathbf{e}_i\mathbf{e}_i^T$ is a dyadic product, and that $\mathbf{M}_{\mathcal{A}}$ and $\mathbf{M}_{\mathcal{A}} + \mathbf{L}_j$ are invertible by definition, we can apply the matrix determinant lemma and rewrite the previous expression as

$$\frac{\det(\mathbf{M}_{\mathcal{A}})\det(\mathbf{M}_{\mathcal{A}} + \mathbf{L}_j)(1 + a^{-1}\mathbf{e}_i^T\mathbf{M}_{\mathcal{A}}^{-1}\mathbf{e}_i)}{\det(\mathbf{M}_{\mathcal{A}})\mathbf{det}(\mathbf{M}_{\mathcal{A}} + \mathbf{L}_j)(1 + a^{-1}\mathbf{e}_i(\mathbf{M}_{\mathcal{A}} + \mathbf{L}_j)^{-1}\mathbf{e}_i)} \geq 1,$$

leading to

$$\frac{1 + a^{-1}\mathbf{e}_i^T\mathbf{M}_{\mathcal{A}}^{-1}\mathbf{e}_i}{1 + a^{-1}\mathbf{e}_i(\mathbf{M}_{\mathcal{A}} + \mathbf{L}_j)^{-1}\mathbf{e}_i} \geq 1.$$

Finally, the inequality for the last ratio can be proven using the following property of positive definite matrices. If $\mathbf{M} \succeq \mathbf{N}$, then $\mathbf{M}^{-1} \preceq \mathbf{N}^{-1}$. Hence,

$$a^{-1}\mathbf{e}_i^T(\mathbf{M}_{\mathcal{A}}^{-1} - (\mathbf{M}_{\mathcal{A}} + \mathbf{L}_i)^{-1})\mathbf{e}_i \geq 0,$$

which is always true for $a \geq 0$ and due to $\mathbf{M}_{\mathcal{A}} + \mathbf{L}_i \succeq \mathbf{M}_{\mathcal{A}}$. □

References

[1] H. L. Van Trees, *Array Processing*, Wiley, New York, 1971

[2] M. Withers, et al., "A comparison of select trigger algorithms for automated global seismic phase and event detection," *Bulletin of the Seismological Society of America* vol.88, no.1, pp.95-106, 1998.

[3] E. Axell, et al., "Spectrum sensing for cognitive radio: State-of-the-art and recent advances," *IEEE Sig. Proc. Mag.* vol.29,no.3,pp.101-116, 2012.

[4] I. Traore, et al., "Continuous Authentication Using Biometrics: Data, Models, and Metrics: Data, Models, and Metrics." *IGI Global*, 2011.

[5] S.P. Chepuri, and G. Leus, "Sparse Sensing for Statistical Inference," *Foundations and Trends in Sig. Proc.*, vol.9, no.34, pp.233-368, 2016.

[6] H. Jamali-Rad, et al., "Sparsity-aware sensor selection: Centralized and distributed algorithms," *IEEE Sig. Proc. Lett.*, vol.21, no.2, pp.217-220, 2014.

[7] Z. Quan, et al., "Innovations diffusion: A spatial sampling scheme for distributed estimation and detection," *IEEE Trans. Sig. Proc.*, vol.57, no.2, pp.738-751, 2009.

[8] J. Chaoyang, et al., "Sensor placement by maximal projection on minimum eigenspace for linear inverse problems," *IEEE Trans. Sig. Proc.*, vol.64, no.21, pp.5595-5610, 2015.

[9] A. Bertrand, and M. Moonen. "Efficient sensor subset selection and link failure response for linear MMSE signal estimation in wireless sensor networks,", *Proc. of the 18th European Sig. Proc. Conference, IEEE*, pp.1092-1096, Aug., 2010.

[10] S. Joshi, and S. Boyd. "Sensor selection via convex optimization." *IEEE Trans. Sig. Proc.*, vol.57, no.2, pp.451- 462, 2009.

[11] S. Liu, et al., "Sensor selection for estimation with correlated measurement noise," *IEEE Trans. Sig. Proc.*, vol.64, no.13, pp.3509-3522, 2016.

[12] M. Yilin, et al., "Sensor selection strategies for state estimation in energy constrained wireless sensor networks," *Automatica*, vol.47, no.7, pp.1330-1338, 2011.

[13] C. Yu, and P. K. Varshney, "Sampling design for Gaussian detection problems," *IEEE Trans. Sig. Proc.*, vol.45, no.9, pp.2328-2337, 1997.

[14] 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, 1983.

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

[16] D. Bajovic, et al., "Sensor selection for event detection in wireless sensor networks," *IEEE Trans. Sig. Proc.*, vol.59, no.10, pp.4938-4953, 2011.
[17] S.P. Chepuri, and G. Leus. "Sparse sensing for distributed detection," *IEEE Trans. Sig. Proc.*, vol.64, no.6, pp.1446-1460, 2015.
[18] A. Krause, A. Singh, and C. Guestrin, "Near-optimal sensor placements in Gaussian processes: Theory, efficient algorithms and empirical studies," *Journal of Machine Learning Research*, pp.235-284, Feb 2008 .
[19] B. Fang, D. Kempe, and R. Govindan, "Utility based sensor selection." *Proceedings of the 5th international conference on Inf. Proc. in sensor networks,* ACM, 2006.
[20] S. Manohar, S. Banerjee , and H. Vikalo, "Greedy sensor selection: Leveraging submodularity," *49th IEEE Conference on Decision and Control (CDC)*, 2010.
[21] J. Ranieri, A. Chebira, and M. Vetterli, "Near-optimal sensor placement for linear inverse problems," *IEEE Trans. Sig. Proc.*, vol.62, no.5, pp.1135-1146, 2014.
[22] D. E. Badawy, J. Ranieri, and M. Vetterli, "Near-optimal sensor placement for signals lying in a union of subspaces," *Proc. of the 22nd European Sig. Proc. Conference, IEEE*, 2014.
[23] S. Rao, S. P. Chepuri, and G. Leus, "Greedy sensor selection for non-linear models," *Proc. of the 6th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing, IEEE*, 2015.
[24] C. E. Shannon, R. G. Gallager, and E. R. Berlekamp, "Lower bounds to error probability for coding on discrete memoryless channels. I," *Information and Control*, vol.10, no.1, pp.65-103, 1967.
[25] T. Kailath, "The divergence and Bhattacharyya distance measures in signal selection," *IEEE Trans. Commun. Technol.*,vol.15,no.1,pp.52-60,1967.
[26] J. Harold, "An invariant form for the prior probability in estimation problems." *Proceedings of the Royal Society of London A: mathematical, physical and engineering sciences*. vol.186. no.1007, 1946.
[27] S. Kullback, *Information theory and statistics*, Courier Corp., 1997.
[28] C. Gruia, et al., "Maximizing a monotone submodular function subject to a matroid constraint," *SIAM Journal on Computing*, vol.40, no.6, pp.1740-1766, 2011.
[29] M. C. Shewry, and H. P. Wynn, "Maximum entropy sampling," *Journal of applied statistics*, vol.14, no.2, pp.165-170, 1987.
[30] T.M. Cover and J. A. Thomas, *Elements of information theory*, John Wiley & Sons, 2012.
[31] S. Fujishige, *Submodular functions and optimization*, Elsevier, 2005.
[32] L. Lovász, "Submodular functions and convexity," *The State of the Art Mathematical Programming,* Springer Berlin, pp.235-257, 1983.
[33] F. Bach, "Learning with submodular functions: A convex optimization perspective," arXiv preprint arXiv:1111.6453, 2011.
[34] S. Iwata, et al., "A combinatorial strongly polynomial algorithm for minimizing submodular functions," *Journal of the ACM (JACM)*, vol.48, no.4, pp.761-777, 2001.
[35] A. Schrijver, "A combinatorial algorithm minimizing submodular functions in strongly polynomial time," *Journal of Combinatorial Theory*, Series B, vol.80, no.2, pp.346-355, 2000.
[36] S. Jegelka, et al., "On fast approximate submodular minimization," *Advances in Neural Information Processing Systems*, 2011.
[37] R. Khanna, et al., "Scalable Greedy Feature Selection via Weak Submodularity," arXiv preprint arXiv:1703.02723, 2017.
[38] X. Yuan, et al., "Newton-Type Greedy Selection Methods for ℓ_0-Constrained Minimization," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2017.
[39] G. Shulkind, et al., "Sensor Array Design Through Submodular Optimization," arXiv preprint arXiv:1705.06616, 2017.
[40] G. L. Nemhauser, et al., "An analysis of approximations for maximizing submodular set functions," *Mathematical Programming*, vol.14, no.1, pp.265-294, 1978.
[41] M. Narasimhan, and J. A. Bilmes, "A submodular-supermodular procedure with applications to discriminative structure learning," arXiv preprint arXiv:1207.1404, 2012.
[42] R. Iyer, and J. Bilmes, "Algorithms for approximate minimization of the difference between submodular functions, with applications," arXiv preprint arXiv:1207.0560, 2012.
[43] Y. Sun, et al., "Majorization-minimization algorithms in Sig. Proc., communications, and machine learning," *IEEE Trans. Sig. Proc.*, vol.65, no.3, pp.794-816, 2017.
[44] S. M. Kay, *Fundamentals of statistical Sig. Proc., vol.II: Detection Theory,* Upper Saddle River, NJ: Prentice Hall, 1998.
[45] A. L. Yuille, and A. Rangarajan, "The concave-convex procedure," *Neural computation*, vol.15, no.4, pp.915-936, 2003.
[46] A. Krause, and D. Golovin, "Submodular function maximization," *Tractability: Practical Approaches to Hard Problems*, vol.3, no.19, 2012.
[47] S. Boyd, and L. Vandenberghe, *Convex optimization*. Cambridge University press, 2004.
[48] H. Jamali-Rad, et al., "Sparsity-aware sensor selection for correlated noise," *Proc. of the 17th Int. Conf. on Inf. Fusion, IEEE,* 2014.
[49] Z. Luo, et al, "Semidefinite relaxation of quadratic optimization problems," *IEEE Sig. Proc. Mag,* vol.27, no.3, 2010.
[50] L. Chamon, and A. Ribeiro, "Near-optimality of greedy set selection in the sampling of graph signals," *Proc. of the Global Conference on Signal and Information Processing, IEEE,* 2016.
[51] J. Leskovec, et al., "Cost-effective outbreak detection in networks," *Proc. of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM*, 2007.
[52] B . Mirzasoleiman, et al., "Lazier Than Lazy Greedy," *AAAI*, 2015.
[53] S. Theodoridis, et al., *Introduction to pattern recognition: a Matlab approach*, Academic Press, 2010.
[54] R.P.W. Duin, et al., *PRTools5, A Matlab Toolbox for Pattern Recognition*, Delft University of Technology, 2017.
[55] M.D. A. Janosi, *Heart-Cleveland dataset from the Hungarian Institute of Cardiology*. Budapest.
[56] https://archive.ics.uci.edu/ml/datasets/Heart+Disease
[57] S. Jegelka, et al., "Submodularity beyond submodular energies: coupling edges in graph cuts," *Proc. of the Conf. on Comp. Vision and Pattern Recognition, IEEE*, 2011