Sensor Placement for Distributed Sensing

Leus, Geert; Coutino, Mario; Chepuri, Sundeep Prabhakar

DOI
10.1002/9781394191048.ch8

Publication date
2023

Document Version
Accepted author manuscript

Published in
Sparse Arrays for Radar, Sonar, and Communications

Chapter 8

Sensor Placement for Distributed Sensing

Geert Leus,[1*] Mario Coutino,[2] and Sundeep Prabhakar Chepuri[3]

[1]*Faculty of Electrical Engineering, Mathematics and Computer Science (EEMCS), Delft University of Technology, 2628CD, Delft, The Netherlands*
[2] *Netherlands Organisation for Applied Scientific Research (TNO), 2597 AK, The Hague, The Netherlands*
[3]*Department of Electrical Communication Engineering (ECE), Indian Institute of Science, 560012, Bengaluru, India*

*Corresponding Author: Geert Leus; g.j.t.leus@tudelft.nl

Sensing the environment is of crucial importance in many different applications, such as autonomous driving, weather prediction, and hazard prevention to name a few. To improve the sensing outcome, sensor networks are often considered because they are capable of observing the environment from different viewpoints. Sensors are generally not cheap though; think for instance, about radar stations or satellites. As a result, we want to make optimal use of a limited amount of sensors, and thus sensor placement becomes a critical problem.

Sensor placement can be considered as an instance of sparse sensing, the more general field concerned with allocating a limited amount of resources to tackle a specific statistical inference task [Chepuri and Leus, 2016b]. Besides

sensor placement, it encompasses applications such as antenna selection and (non-uniform) sub-Nyquist sampling [Molisch and Win, 2004], [Blu et al., 2008], [Eldar and Michaeli, 2009], and [Vaidyanathan and Pal, 2011]. The basic idea of sparse sensing is that there is a large number of potential candidate measurements from which only a limited amount is selected due to cost considerations. Sparse sensing has been optimized for several different inference tasks, such as estimation [Krause et al., 2008, Joshi and Boyd, 2009, Ranieri et al., 2014, Chepuri and Leus, 2015, Liu et al., 2014] and detection [Bajovic et al., 2011], [Cambanis and Masry, 1983], [Yu and Varshney, 1997], [Coutino et al., 2018a], [Chepuri and Leus, 2016a], which are the two tasks we will focus on in this chapter. Different performance measures for these tasks have been considered ranging from the Cramér-Rao bound or minimum mean square estimation error, over frame potential, to the log likelihood ratio. Both linear as well as nonlinear observation models have been considered. Moreover, different solution approaches have been proposed. The two most popular ones are convex relaxation and greedy selection methods.

What is not often treated in sensor placement is the fact that the measurements can be conditionally dependent [Liu et al., 2014, Coutino et al., 2018a]. This can for instance occur when the candidate measurements are corrupted by interference from undesired sources which can be modeled as correlated noise. This complicates matters significantly and makes the sensor placement problem for estimation and detection tasks tough to handle. We show in this chapter how to tackle this scenario for both the convex relaxation and greedy selection methods. For simplicity of presentation, we only focus on the linear model in this chapter, but it can be easily extended to nonlinear models as well. Furthermore, we assume additive Gaussian noise, although that can also

be relaxed in some cases.

In Section 8.1, we introduce the sensor placement problem for a linear data model with additive Gaussian noise. As mentioned earlier, we will not only assume uncorrelated noise, but also handle the correlated case. Furthermore, the convex relaxation as well as the greedy selection methods will be presented. Finally, a field estimation and detection example will be introduced that will be used throughout the chapter for illustrative purposes. Sections 8.2 and 8.3 will then respectively discuss in detail the estimation and detection problem. We will introduce the related performance metrics and illustrate how the convex relaxation and greedy selection methods can be employed for these measures. The field estimation and detection example allows us to demonstrate the behavior of the solutions as well as the related trade-offs.

8.1. Data Model

In this section, we will detail the setup that will be considered throughout this chapter. We assume that K sensors can be distributed over M candidate positions, and the goal is to place these sensors in some optimal fashion. The candidate positions can represent the only available sensor positions, e.g., rooftops of buildings for a distributed radar application, or they are obtained by gridding some continuous area. The number of sensors we have available is generally much smaller than M, i.e., $K \ll M$

The potential measurement $x_m \in \mathbb{R}$ that is taken at position $m = 1, 2, \ldots, M$ is assumed to be linearly dependent on some parameter vector $\boldsymbol{\theta} \in \mathbb{R}^N$ cor-

rupted by addition noise $n_m \in \mathbb{R}$, i.e.,

$$x_m = \mathbf{h}_m^\top \boldsymbol{\theta} + n_m, \tag{8.1}$$

where $\mathbf{h}_m$ collects the linear coefficients related to the m-th measurement. Stacking the different candidate measurements into $\mathbf{x} = [x_1, \ldots, x_M]^\top$, we can write

$$\mathbf{x} = \mathbf{H}\boldsymbol{\theta} + \mathbf{n}, \tag{8.2}$$

where $\mathbf{H} = [\mathbf{h}_1, \ldots, \mathbf{h}_M]^\top$ and $\mathbf{n} = [n_1, \ldots, n_M]^\top$. Throughout, we will assume that the noise is Gaussian distributed with zero mean, i.e., $\mathbb{E}\{\mathbf{n}\} = \mathbf{0}$, and covariance matrix $\boldsymbol{\Sigma}$, i.e., $\mathbb{E}\{\mathbf{n}\mathbf{n}^\top\} = \boldsymbol{\Sigma}$. So in summary, the data model can be expressed as $\mathbf{x} \sim \mathcal{N}(\mathbf{H}\boldsymbol{\theta}, \boldsymbol{\Sigma})$.

Note that although such a linear Gaussian model is restrictive, some of the results in this chapter can be extended to non-linear non-Gaussian models as well. In the non-linear case, we can view (8.2) as a local linear model around a specific $\boldsymbol{\theta}$ where $\mathbf{H}$ represents the Jacobian matrix of the nonlinear model in $\boldsymbol{\theta}$. This will generally make $\mathbf{H}$ dependent on $\boldsymbol{\theta}$ which will require some adjustments to the proposed methods. The Gaussianity assumption, on the other hand, is required for the correlated noise case (non-diagonal $\boldsymbol{\Sigma}$), but can easily be dropped for the uncorrelated noise case (diagonal $\boldsymbol{\Sigma}$).

We cannot use all M measurements in $\mathbf{x}$ since we only have $K \ll M$ sensors available. To select the "best" K measurements, we will adopt the selection vector $\mathbf{w} \in \{0, 1\}^M$ whose m-th entry is one if the position m is selected or zero otherwise, and whose total number of non-zero entries is K, i.e., $\|\mathbf{w}\|_0 = K$.

The selected measurements can then be represented by

$$\mathbf{y} = \mathbf{\Phi}(\mathbf{w})\mathbf{x} = \mathbf{\Phi}(\mathbf{w})\mathbf{H}\boldsymbol{\theta} + \mathbf{\Phi}(\mathbf{w})\mathbf{n}, \tag{8.3}$$

where $\mathbf{\Phi}(\mathbf{w})$ is the submatrix of the $M \times M$ identity matrix $\mathbf{I}_M$ that is obtained by removing the zero rows from $\text{diag}(\mathbf{w})\mathbf{I}_M$. Note thereby that

$$\mathbf{\Phi}(\mathbf{w})\mathbf{\Phi}^\top(\mathbf{w}) = \mathbf{I}_K, \quad \mathbf{\Phi}^\top(\mathbf{w})\mathbf{\Phi}(\mathbf{w}) = \text{diag}(\mathbf{w}). \tag{8.4}$$

The problem we are tackling in this chapter is how to design the "best" $\mathbf{w}$ under the constraints $\mathbf{w} \in \{0,1\}^M$ and $\|\mathbf{w}\|_0 = K$. What we mean by "best" is determined by some inference performance metric $f : \{0,1\}^M \to \mathbb{R}$. Depending on the inference task, this metric can be the estimation error or miss detection probability for instance. Hence, the problem we need to solve is

$$\underset{\mathbf{w}\in\{0,1\}^M}{\arg\min}\ f(\mathbf{w}) \quad \text{s.t. } \|\mathbf{w}\|_0 = K. \tag{8.5}$$

This is known as the cardinality constrained (CC) problem.

In some cases, though, we might want to keep K as small as possible as long as we can guarantee a certain inference performance. In that case, we can formulate a performance constrained (PC) problem as

$$\underset{\mathbf{w}\in\{0,1\}^M}{\arg\min}\ \|\mathbf{w}\|_0 \quad \text{s.t. } f(\mathbf{w}) \leq \lambda, \tag{8.6}$$

where λ denotes the considered performance bound. Both problems are equivalent in that for every K in (8.5), there exists a related λ in (8.6).

Solving these problems would require a complex combinatorial search which quickly becomes intractable. As a result, in this chapter, we will focus on convex

relaxation and greedy selection approaches. The specific inference tasks we will aim at are

- Estimation of $\boldsymbol{\theta}$.
- Detection of $\boldsymbol{\theta}$, where $\boldsymbol{\theta}$ can be known or not.

In both cases, we assume the knowledge of $\mathbf{H}$ and the noise covariance matrix $\boldsymbol{\Sigma}$. Before we delve into the details of the estimation and detection inference tasks, we briefly overview the general idea of the convex relaxation and greedy selection methods.

8.1.1. Solution approaches

A common approach to tackle the CC and PC problems in (8.5) and (8.6) is adopting convex relaxation. More specifically, the l0-norm will be replaced by its convex hull, which is given by the l1-norm, whereas the Boolean constraint is replaced by a box constraint. Assuming further that $f(\cdot)$ is a convex function, potentially obtained after convexifying the desired performance measure, the CC and PC problems can respectively be relaxed to the following convex problems

$$\underset{\mathbf{w}\in[0,1]^M}{\arg\min}\ f(\mathbf{w}) \quad \text{s.t. } \|\mathbf{w}\|_1 = K. \tag{8.7}$$

and

$$\underset{\mathbf{w}\in[0,1]^M}{\arg\min}\ \|\mathbf{w}\|_1 \quad \text{s.t. } f(\mathbf{w}) \leq \lambda. \tag{8.8}$$

Note that the constraint $\|\mathbf{w}\|_1 = K$ in (8.7) can be considered convex, since due to $\mathbf{w} \in [0,1]^M$ it can be written as $\mathbf{1}^\top\mathbf{w} = 1$. For the same reason, we can also replace $\|\mathbf{w}\|_1$ in (8.8) by $\mathbf{1}^\top\mathbf{w}$.

Convex problems have been well-studied, although not all forms are nec-

essarily easy to solve. Furthermore, it is not always possible to quantify the suboptimality of the convex relaxation. Specific convex performance measures $f(\cdot)$ for the estimation and detection problem will be discussed in later sections. Note that an approximate Boolean solution for $\mathbf{w}$ can be obtained from the solution of the above problems by means of thresholding or randomization [Chepuri and Leus, 2016b].

Another method to handle the CC and PC problems is greedy selection. The idea behind this approach is rather intuitive and basically selects one sensor at a time (myopic method), either to add it to or to remove it from the previously selected sensors. Suppose that $\mathcal{X}$ represents the set of currently active or inactive sensors. Then a greedy approach would iteratively add to $\mathcal{X}$ the sensor that is obtained as

$$s^* = \arg\max_{s \notin \mathcal{X}} f(\mathcal{X} \cup s),$$

where $f(\mathcal{X})$ is the set function that correctly translates the performance measure $f(\mathbf{w})$. Note that this translation will depend on the fact whether $\mathcal{X}$ represents the active or inactive set of sensors. The CC problem is now handled by iterating till we have K active sensors, whereas for the PC problem the iterations will stop when the desired performance has been reached. Interestingly, if $f(\mathcal{X})$ is submodular, monotone nondecreasing, and normalized (i.e., $f(\emptyset) = 0$), then the greedy solution is near optimal and approaches the optimal cost with a factor $1 - 1/e$, where e is the Euler number [Nemhauser et al., 1978].

8.1.2. Running example

The running example that we will consider throughout this chapter[1] is that of spatial field estimation/detection. To simplify the presentation, we will consider a one-dimensional spatial field with candidate sensor positions that are uniformly spaced over the considered area. We further consider a spatially correlated field that is non-stationary (i.e., the correlation depends on the sensor location). More specifically, we assume the spatial correlation matrix of the field is given by

$$\mathbf{R} = \begin{bmatrix} 1 & \delta_1 & \delta_1^2 & \dots & \delta_1^{M-1} \\ \delta_1 & 1 & \delta_2 & \dots & \delta_2^{M-2} \\ \delta_1^2 & \delta_2 & 1 & \dots & \delta_3^{M-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \delta_1^{M-1} & \delta_2^{M-2} & \delta_3^{M-3} & \dots & 1 \end{bmatrix}.$$

To further simplify the setup and reduce the number of variables, we take $0 < \delta_1 = \delta < 1$ and $\delta_i = \delta_{i-1} - \kappa$, with $0 < \kappa < \delta/M$. This structure will allow us to analyze how the developed selection methods behave as a function of different correlation structures. Note however that we do not assume a stochastic signal in (8.3), but a deterministic one. That is why we further assume the observed deterministic field is somehow "smooth" within the class of fields characterized with correlation matrix $\mathbf{R}$. In other words, we define $\mathbf{H}$ as the matrix that stacks the $N \ll M$ unit-norm eigenvectors of $\mathbf{R}$ with the largest eigenvalues, i.e., $\mathbf{R} = \mathbf{H}\mathbf{\Lambda}\mathbf{H}^\top + \mathbf{H}_n\mathbf{\Lambda}_n\mathbf{H}_n^\top$, with $\mathbf{\Lambda} \in \mathbb{R}^{N\times N}$ the diagonal matrix of

[1]Software to reproduce results from the chapter are available at https://github.com/spchepuri/sensorplacement.git

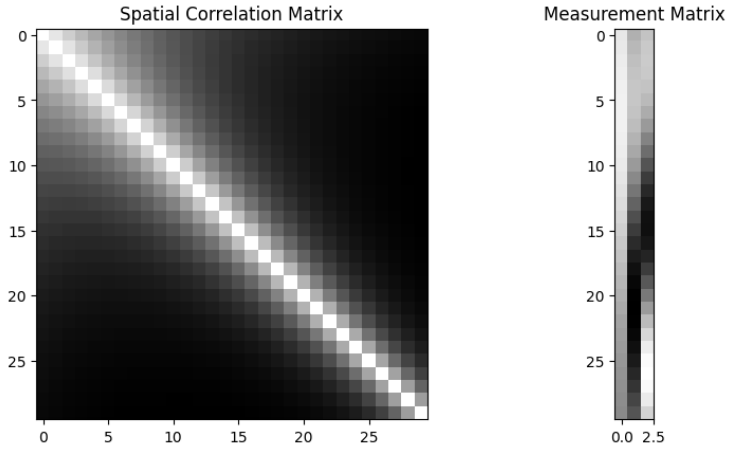

Figure 8.1: Illustration of the running example that will be used throughout this chapter. On the left is a representation of the spatial correlation matrix $\mathbf{R}$ whereas the related measurement matrix $\mathbf{H}$ is depicted on the right. The chosen parameters are $M = 30$, $N = 3$, $\delta = 0.9$ and $\kappa = 0.01$.

the N strongest eigenvalues of $\mathbf{R}$ and $\mathbf{\Lambda}_n \in \mathbb{R}^{(M-N)\times(M-N)}$ the diagonal matrix collecting the remaining eigenvalues. For the specific values of $M = 30$, $N = 3$, $\delta = 0.9$ and $\kappa = 0.01$, the spatial correlation matrix $\mathbf{R}$ and related measurement matrix $\mathbf{H}$ are depicted in Fig. 8.1.

As far as the additive Gaussian noise is concerned, two scenarios will be studied in this chapter: uncorrelated and correlated noise. In case of uncorrelated noise, we simply assume $\mathbf{\Sigma} = \sigma^2\mathbf{I}$. The correlated noise scenario is more complex. There we only focus on spatially stationary noise but we will assume a decaying correlation function as well as a constant correlation function. In

the first case, the noise covariance matrix is given by

$$\mathbf{\Sigma} = \sigma^2 \begin{bmatrix} 1 & \gamma & \gamma^2 & \dots & \gamma^{M-1} \\ \gamma & 1 & \gamma & \dots & \gamma^{M-2} \\ \gamma^2 & \gamma & 1 & \dots & \gamma^{M-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \gamma^{M-1} & \gamma^{M-2} & \gamma^{M-3} & \dots & 1 \end{bmatrix}, \tag{8.9}$$

whereas for the second case, we use

$$\mathbf{\Sigma} = \sigma^2[(1-\gamma)\mathbf{I} + \gamma \mathbf{1}\mathbf{1}^T]. \tag{8.10}$$

In both cases, we take $\gamma < 1$. Clearly, when $\gamma = 0$ we obtain the uncorrelated noise scenario. Note that in all scenarios the noise variance is given by σ^2.

8.2. Distributed Estimation

Consider the setting where the fusion center, which gathers a subset of sensor data, performs an estimate of the unknown parameter vector $\boldsymbol{\theta}$. Let us denote the estimator as $\hat{\boldsymbol{\theta}}(\mathbf{w})$, where we recall that $\mathbf{w}$ is the selection vector. The performance of such parameter estimation problems is characterized by the error covariance matrix denoted by $\mathbf{E}(\mathbf{w}) = \mathbb{E}\{[\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}(\mathbf{w})][\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}(\mathbf{w})]^\top\}$. Since $\hat{\boldsymbol{\theta}}(\mathbf{w})$ (and hence $\mathbf{E}(\mathbf{w})$) depends on the subset of measurements acquired via $\mathbf{w}$, various scalar measures of $\mathbf{E}(\mathbf{w})$ are optimized with respect to (w.r.t.) the selection variable $\mathbf{w}$. Next, we discuss some of the popular choices of performance measures.

8.2.1. Estimation Optimality Criteria

To measure the quality of estimation, the following scalar measures can be used.

- **A-optimality criterion or mean squared error.** The mean squared error in estimating $\boldsymbol{\theta}$ is

$$f(\mathbf{w}) = \mathbb{E}\{\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\|^2\} = \text{trace}(\mathbf{E}(\mathbf{w})), \tag{8.11}$$

  which is also the sum of the eigenvalues of $\mathbf{E}(\mathbf{w})$.
- **D-optimality criterion or volume of the confidence ellipsoid.** The η-confidence ellipsoid is the minimum volume ellipsoid that contains the estimation error $\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}$ with probability η. It can be shown to be proportional

to the determinant of $\mathbf{E}$:

$$f(\mathbf{w}) = \log\det(\mathbf{E}(\mathbf{w})). \tag{8.12}$$

- **E-optimality criterion or worst-case error variance.** Another criterion related to the volume of the confidence ellipsoid is the worst-case variance of the estimation error:

$$f(\mathbf{w}) = \max_{\|\mathbf{u}\|=1} \mathbf{u}^\top \mathbf{E}(\mathbf{w})\mathbf{u} = \lambda_{\max}(\mathbf{E}(\mathbf{w})), \tag{8.13}$$

  where $\lambda_{\max}(\cdot)$ denotes the maximum eigenvalue of its argument.

- **Worst-case coordinate error variance.** The worst-case coordinate error variance is the largest diagonal entry of the error covariance matrix $\mathbf{E}(\mathbf{w})$, i.e.,

$$f(\mathbf{w}) = \max_{1\le i\le M} [\mathbf{E}(\mathbf{w})]_{ii}, \tag{8.14}$$

  where $[\mathbf{E}(\mathbf{w})]_{ii}$ denotes the i-th diagonal entry of $\mathbf{E}(\mathbf{w})$.

Although there is no general answer to how one performance measure compares with another, all the above scalar measures are equally reasonable because they quantify the estimation quality.

8.2.2. Uncorrelated Observations

Suppose the noise covariance matrix is diagonal, i.e., $\mathbf{\Sigma} = \mathrm{diag}(\sigma_1^2, \cdots, \sigma_M^2)$ with $\sigma_1^2 = \cdots = \sigma_M^2 = \sigma^2$ being a special case. Since the parameter $\boldsymbol{\theta}$ is assumed deterministic, we can then focus on the maximum likelihood (ML)

estimate, which is given by

$$\hat{\boldsymbol{\theta}} = \left(\sum_{m=1}^{M} w_m \sigma_m^{-2} \mathbf{h}_m \mathbf{h}_m^\top \right)^{-1} \sum_{m=1}^{M} w_m x_m \mathbf{h}_m.$$

The estimation error $\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}$ then has zero mean, i.e., the estimate is unbiased, and covariance matrix

$$\mathbf{E}(\mathbf{w}) = \left(\sum_{m=1}^{M} w_m \sigma_m^{-2} \mathbf{h}_m \mathbf{h}_m^\top \right)^{-1}. \tag{8.15}$$

We can observe the explicit role of $\mathbf{w}$ in (8.15) for selecting the measurements that improve the conditioning of $\mathbf{E}(\mathbf{w})$.

Now we describe the sparse sampler design problem for the setting with uncorrelated observations using the above described scalar measures of $\mathbf{E}(\mathbf{w})$ and the CC problem formulation (8.7).

For the D-optimality criterion, (8.7) specializes to

$$\begin{aligned} &\underset{\mathbf{w} \in [0,1]^M}{\arg\max} \ \log\det \sum_{m=1}^{M} w_m \sigma_m^{-2} \mathbf{h}_m \mathbf{h}_m^\top \\ &\text{subject to } \mathbf{1}^\top \mathbf{w} = K, \end{aligned} \tag{8.16}$$

which is a convex problem as the objective is a concave function in $\mathbf{w}$, the sum constraint is linear, and the w_m variables are restricted to the interval $[0, 1]$. This problem can be solved using any one of the standard convex solvers. Similarly, we obtain convex problems for the A-optimality and E-optimality criteria as

$$\underset{\mathbf{w} \in [0,1]^M}{\arg\min} \ \operatorname{trace} \left(\sum_{m=1}^{M} w_m \sigma_m^{-2} \mathbf{h}_m \mathbf{h}_m^\top \right)^{-1} \tag{8.17}$$

$$\text{subject to}\, \mathbf{1}^\top \mathbf{w} = K,$$

and

$$\begin{aligned} &\underset{\mathbf{w}\in[0,1]^M}{\arg\max}\ \lambda_{\min}\left(\sum_{m=1}^{M} w_m \sigma_m^{-2} \mathbf{h}_m \mathbf{h}_m^\top\right) \\ &\text{subject to}\, \mathbf{1}^\top \mathbf{w} = K, \end{aligned} \tag{8.18}$$

respectively. Both these problem formulations can be transformed into their epigraph form, leading to a semi-definite program (SDP).

Also minimizing the objective function related to the worst-case coordinate error variance in (8.14) can be expressed in the epigraph form leading to the following SDP:

$$\begin{aligned} &\underset{t,\mathbf{w}\in[0,1]^M}{\text{minimize}} \quad t \\ &\text{subject to} \quad \mathbf{1}^\top \mathbf{w} = K, \\ &\qquad \begin{bmatrix} t & \mathbf{e}_j^\top \\ \mathbf{e}_j & \sum_{m=1}^{M} w_m \sigma_m^{-2} \mathbf{h}_m \mathbf{h}_m^\top \end{bmatrix} \succeq \mathbf{0}, \quad j = 0, 1, \ldots, M, \end{aligned} \tag{8.19}$$

where $\mathbf{e}_j$ is the j-th column of the identity matrix of size M.

Next, we discuss numerical experiments to illustrate the sensor placements obtained by solving the above convex programs. In Figure 8.2, we illustrate the sparse sampler obtained by optimizing the D-optimality criterion (i.e., maximizing the logdet function), the A-optimality criterion (i.e., minimizing the trace function) and the E-optimality criterion (i.e., maximizing the minimum eigenvalue function) while selecting $K = 6$ sensors in the CC problem setting. The system matrix is generated as explained in Section 8.1.2 with $M = 30$,

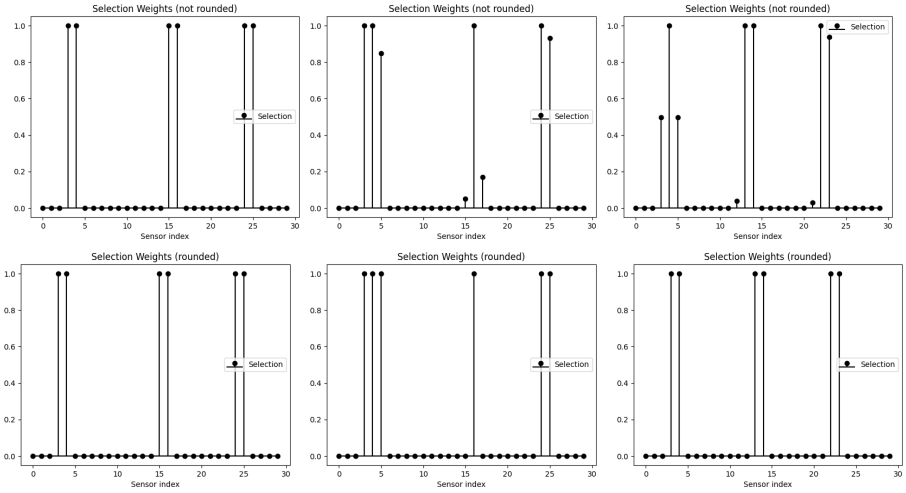

Figure 8.2: Comparison of the non-Boolean (top row) selection obtained by optimizing the D-optimality (left), A-optimality (middle) and E-optimality (right) criterion while selecting $K = 6$ measurements for the setting in our running example. The recovered Boolean solution after rounding (bottom row) from D-optimality (left), A-optimality (middle), and E-optimality (right) criteria.

$N = 3$, $\delta = 0.9$, and $\kappa = 0.01$. The noise variance is selected as $\sigma^2 = 1$. We observe that all three methods basically select three groups of sensors. There are three groups because we have $N = 3$ unknowns. If one measurement per group would be considered, a well-conditoned $\mathbf{H}$ matrix is obtained. The clustering into groups further boosts the energy of the measurements.

The D-optimality cost function can be shown to be a submodular function in the selection variables [Shamaiah et al., 2010]. This means that, as discussed earlier, we can greedily maximize the logdet function. In Figure 8.3, we can see that the selection pattern obtained by greedily maximizing the logdet function is similar as the rounded solutions shown in Figure 8.2.

Although it is difficult to characterize the optimality of the convex relaxation, it is optimal for the following special case. When the parameter to be

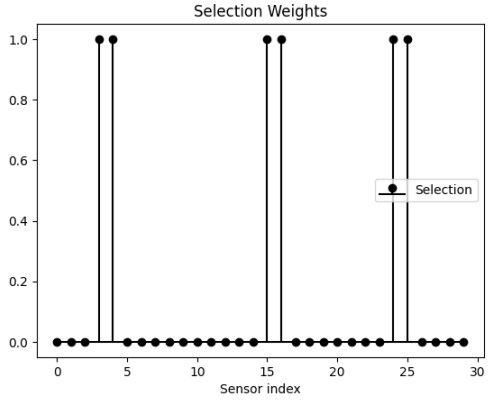

Figure 8.3: Selection obtained by greedily optimizing the D-optimality criterion while selecting $K = 6$ measurements for the setting in our running example.

estimated is a scalar θ with $N = 1$, the error covariance is also a scalar and readily given by $E(\mathbf{w}) = \sum_{n=1}^{N} w_m(h_m^2/\sigma_m^2)$. In this case, (8.5) can be optimally solved without requiring any convex relaxation. Minimizing the cost over the Boolean variable can then be optimally solved by rank ordering the signal-to-noise ratios h_m^2/σ_m^2 associated with each sensor. Specifically, the solution is obtained by setting the entries of w_m corresponding to the maximum entries of the set $\{h_m^2/\sigma_m^2, 1 \leq m \leq M\}$ to 1, and the remaining ones to zero, untill K sensors are selected (CC problem) or the performance bound is met (PC problem).

8.2.3. Correlated Observations

Now we discuss the general case with $\mathbf{\Sigma}$ being a non-diagonal matrix. The ML error covariance matrix can then be expressed as

$$\mathbf{E}(\mathbf{w}) = \left(\mathbf{H}^\top \mathbf{\Phi}^\top(\mathbf{w}) \left(\mathbf{\Phi}(\mathbf{w}) \mathbf{\Sigma} \mathbf{\Phi}^\top(\mathbf{w})\right)^{-1} \mathbf{\Phi}(\mathbf{w}) \mathbf{H}\right)^{-1}. \tag{8.20}$$

The error covariance matrix is now clearly a more intricate function of $\mathbf{w}$ compared to (8.15).

In what follows, we express the earlier discussed scalar functions of $\mathbf{E}(\mathbf{w})$ in a simpler form amenable to convex optimization. Firstly, we write the noise covariance matrix $\mathbf{\Sigma}$ as

$$\mathbf{\Sigma} = \alpha \boldsymbol{I}_M + \mathbf{S}, \tag{8.21}$$

where a nonzero $\alpha \in \mathbb{R}$ is chosen such that $\mathbf{S} \in \mathbb{R}^{M \times M}$ is invertible and well conditioned. Using (8.21) in (8.20), we obtain

$$\mathbf{E}(\mathbf{w}) = \mathbf{H}^\top \mathbf{\Phi}^\top(\mathbf{w}) \left(\alpha \boldsymbol{I}_K + \mathbf{\Phi}(\mathbf{w}) \mathbf{S} \mathbf{\Phi}^\top(\mathbf{w})\right)^{-1} \mathbf{\Phi}(\mathbf{w}) \mathbf{H}.$$

Since $\mathbf{\Phi}^\top(\mathbf{w})\mathbf{\Phi}(\mathbf{w}) = \mathrm{diag}(\mathbf{w})$, we have

$$\mathbf{\Phi}^\top(\mathbf{w}) \left(a \boldsymbol{I}_K + \mathbf{\Phi}(\mathbf{w}) \mathbf{S} \mathbf{\Phi}^\top(\mathbf{w})\right)^{-1} \mathbf{\Phi}(\mathbf{w}) = \mathbf{S}^{-1} - \mathbf{S}^{-1} \left[\mathbf{S}^{-1} + \alpha^{-1} \mathrm{diag}(\mathbf{w})\right]^{-1} \mathbf{S}^{-1}. \tag{8.22}$$

This is due to the matrix inversion lemma

$$\mathbf{C}(\mathbf{B}^{-1} + \mathbf{C}^\top \mathbf{A}^{-1} \mathbf{C})^{-1} \mathbf{C}^\top = \mathbf{A} - \mathbf{A}(\mathbf{A} + \mathbf{C}\mathbf{B}\mathbf{C}^\top)^{-1} \mathbf{A},$$

with $\mathbf{C} = \mathbf{\Phi}^\top(\mathbf{w})$, $\mathbf{B}^{-1} = \alpha \boldsymbol{I}_K$, and $\mathbf{A} = \mathbf{S}^{-1}$. Therefore, we have

$$\begin{aligned} \mathbf{E}(\mathbf{w}) &= \mathbf{H}^\top [\mathbf{S}^{-1} - \mathbf{S}^{-1} \left[\mathbf{S}^{-1} + \alpha^{-1} \mathrm{diag}(\mathbf{w})\right]^{-1} \mathbf{S}^{-1}] \mathbf{H}. \\ &= \mathbf{H}^\top \mathbf{S}^{-1} \mathbf{H} - \mathbf{H}^\top \mathbf{S}^{-1} \left[\mathbf{S}^{-1} + \alpha^{-1} \mathrm{diag}(\mathbf{w})\right]^{-1} \mathbf{S}^{-1} \mathbf{H}. \end{aligned} \tag{8.23}$$

Using this simpler form, for instance, the E-optimality constraint can be expressed as a linear matrix inequality in $\mathbf{w}$, i.e., $\lambda_{\min}(\mathbf{E}(\mathbf{w})) \geq \lambda$ can be ex-

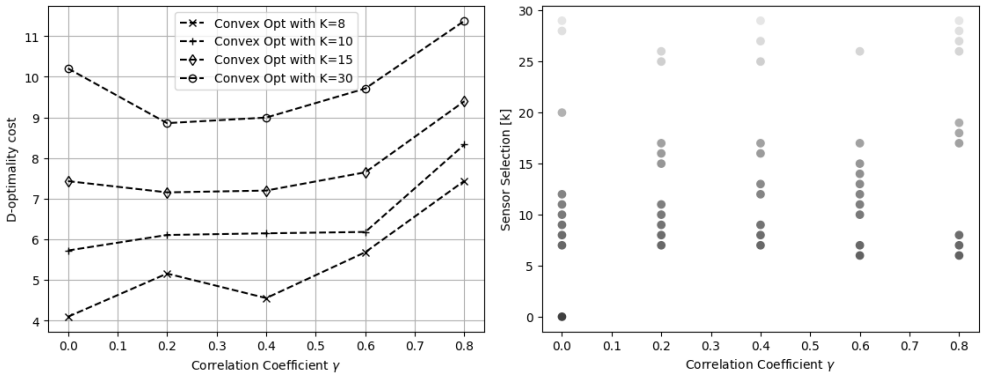

Figure 8.4: (Left) Sensor placements obtained from (8.25) for different K and correlation coefficients γ when the noise has a spatially constant correlation function as in (8.10). (Right) Sensor placements for different γ for $K = 10$.

pressed as

$$\begin{bmatrix} \mathbf{S}^{-1} + \alpha^{-1}\mathrm{diag}(\mathbf{w}) & \mathbf{S}^{-1}\mathbf{H} \\ \mathbf{H}^\top\mathbf{S}^{-1} & \mathbf{H}^\top\mathbf{S}^{-1}\mathbf{H} - \lambda \boldsymbol{I}_N \end{bmatrix} \succeq \mathbf{0}_{M+N} \tag{8.24}$$

with $\mathbf{S}^{-1}+\alpha^{-1}\mathrm{diag}(\mathbf{w})$ being positive definite, which determines the choices of α. Thus, in the case of correlated observations and focusing on the PC problem, the sparse sampler can be designed by solving the following convex problem:

$$\begin{aligned} &\underset{\mathbf{w}\in[0,1]^M}{\arg\min} \quad \mathbf{1}^\top\mathbf{w} \\ &\text{subject to} \begin{bmatrix} \mathbf{S}^{-1} + \alpha^{-1}\mathrm{diag}(\mathbf{w}) & \mathbf{S}^{-1}\mathbf{H} \\ \mathbf{H}^\top\mathbf{S}^{-1} & \mathbf{H}^\top\mathbf{S}^{-1}\mathbf{H} - \lambda \boldsymbol{I}_N \end{bmatrix} \succeq \mathbf{0}_{M+N}. \end{aligned} \tag{8.25}$$

This convex program can be solved using any off-the-shelf convex solvers.

We next illustrate sensor placements obtained for the two noise correlation matrices that were introduced in Section 8.1.2 [cf. (8.9) and (8.10)]. We use the same system matrix as in the uncorrelated case with $M = 30$, $N = 3$, $\delta = 0.9$,

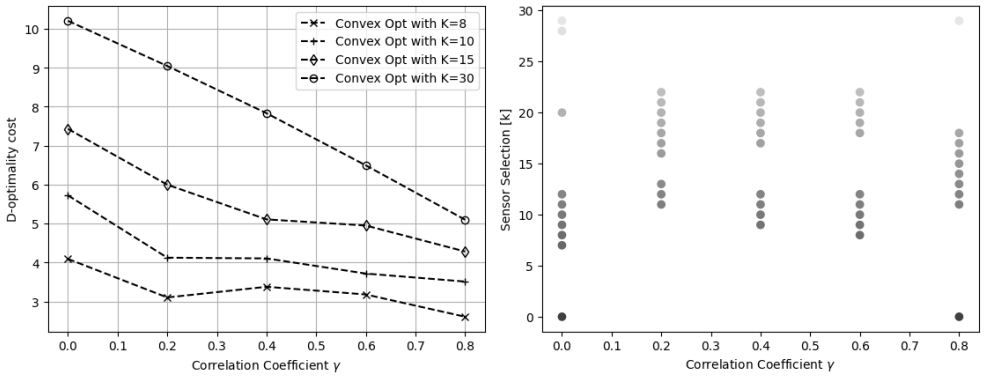

Figure 8.5: (Left) Sensor placements obtained from (8.25) for different K and correlation coefficients γ when the noise has a spatially decaying correlation function as in (8.9). (Right) Sensor placements for different γ for $K = 10$.

$\kappa = 0.01$, and $\sigma^2 = 1$. We set $\alpha = 10^{-4}$. While the sensor placements obtained in the uncorrelated case led to a well-conditioned system matrix, the sensor placement design with correlated observations aims to improve the conditioning taking into account the correlation across observations as measured by the error covariance matrix (8.20).

For the spatially constant correlation function, where all observations are equally correlated with the correlation coefficient γ, we can see in Figure 8.4 that as γ increases, the D-optimality cost increases for a given number of sensors. This, in other words, means that if there is more correlation across sensors, fewer observations lead to better estimation quality. On the other hand, with the decaying correlation function (8.9), the effect is the opposite as seen in Figure 8.5 with a more clustered placement.

8.3. Distributed Detection

In this section, we focus on the binary detection problem related to the model (8.2). More specifically, the two hypotheses (states) we consider are:

$$\mathcal{H}_0 : \mathbf{x} = \mathbf{n}, \tag{8.26}$$

$$\mathcal{H}_1 : \mathbf{x} = \mathbf{H}\boldsymbol{\theta} + \mathbf{n}, \tag{8.27}$$

where we will distinguish between the cases where $\boldsymbol{\theta}$ is known or not.

Irrespective of the scenario, we can relate the measurements $\mathbf{x}$ to the model

$$\mathcal{H}_0 : \mathbf{x} \sim p(\mathbf{x}|\mathcal{H}_0), \tag{8.28}$$

$$\mathcal{H}_1 : \mathbf{x} \sim p(\mathbf{x}|\mathcal{H}_1), \tag{8.29}$$

where $p(\mathbf{x}|\mathcal{H}_i)$ for $i = 0, 1$ denotes the probability density function (pdf) of $\mathbf{x}$, conditioned on the state $\mathcal{H}_i$. As we discussed in Section 8.1, we assume that $p(\mathbf{x}|\mathcal{H}_0) = \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ while $p(\mathbf{x}|\mathcal{H}_1) = \mathcal{N}(\mathbf{H}\boldsymbol{\theta}, \boldsymbol{\Sigma})$, although we will often use the more general notation throughout this section.

The goal now is to solve the CC and/or PC problem in (8.5) and (8.6), respectively, for some specific detection performance measure $f(\mathbf{w})$. We next discuss some typical detection performance measures, which we approximate by more manageable functions in the next section. If the prior hypothesis probabilities are known, i.e., in a Bayesian setting, the optimal detector minimizes the probability of error, $P_{\mathrm{e}} = P(\mathcal{H}_0|\mathcal{H}_1)P(\mathcal{H}_1) + P(\mathcal{H}_1|\mathcal{H}_0)P(\mathcal{H}_0)$, 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), $P_{\mathrm{m}} = P(\mathcal{H}_0|\mathcal{H}_1)$, for a fixed probability of false alarm (type I error), $P_{\mathrm{fa}} = P(\mathcal{H}_1|\mathcal{H}_0)$. The sensor selection problems for detection can then be expressed as in (8.5) and (8.6) with $f(\mathbf{w})$ replaced by either $P_{\mathrm{e}}(\mathbf{w})$ (Bayesian setting) or $P_{\mathrm{m}}(\mathbf{w})$ (Neyman-Pearson setting), where in the latter case we implicitly assume that $P_{\mathrm{fa}}(\mathbf{w})$ is fixed. Note that $P_{\mathrm{e}}(\mathbf{w})$, $P_{\mathrm{m}}(\mathbf{w})$ and $P_{\mathrm{fa}}(\mathbf{w})$ denote the error probabilities due to the measurement selection defined by $\mathbf{w}$.

In general, the above performance measures are not easy to optimize numerically. As a result, in the following, we present alternative measures for both known and unknown $\boldsymbol{\theta}$ cases that can be used as direct surrogates to solve the optimization problems.

8.3.1. Known θ Parameter

Even when $\boldsymbol{\theta}$ is known, the error probabilities P_{e} and P_{m} (for a fixed P_{fa}) might not admit a known closed-form expression or their expressions might not be favorable for numerical optimization. We therefore present several weaker and simpler substitutes, which can be optimized instead of the error probabilities. These substitutes are based on the notion of distance (closeness or divergence) between the two distributions of the observations under test. They lead to tractable, if not always optimal (in terms of the error probabilities) design procedures for sparse sampling. Nevertheless, optimizing these distance measures improves the performance of any practical system.

8.3.1.1. Optimality Criteria

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

$$l(\mathbf{y}) = \frac{p(\mathbf{y}|\mathcal{H}_1)}{p(\mathbf{y}|\mathcal{H}_0)}. \tag{8.30}$$

In what follows, we consider a number of distance measures that belong to the general class of Ali-Silvey distances [Ali and Silvey, 1966], which are of the form

$$\varphi\big(\mathbb{E}_{|\mathcal{H}_i}\{\phi[l(\mathbf{y})]\}\big), \tag{8.31}$$

where $\varphi(\cdot)$ is an increasing real-valued function, $\phi[\cdot]$ is a continuous convex function on $(0, \infty)$, and $\mathbb{E}_{|\mathcal{H}_i}\{\phi[l(\mathbf{y})]\}$ indicates that $\phi[l(\mathbf{y})]$ is averaged under the pdf $p(\mathbf{y}|\mathcal{H}_i)$.

Known Prior Probabilities. Under the assumption of known prior distributions for both hypotheses, a Bayesian detector can be devised. This kind of detector minimizes, and makes a decision based on comparing the optimal statistic to a threshold, i.e.,

$$\log l(\mathbf{y}) = \log \frac{p(\mathbf{y}|\mathcal{H}_1)}{p(\mathbf{y}|\mathcal{H}_0)} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \log \frac{P(\mathcal{H}_0)}{P(\mathcal{H}_1)}. \tag{8.32}$$

In the Bayesian setting, our goal is to choose the best subset of measurements that results in a prescribed Bayesian probability of error. The best achievable exponent in the Bayesian probability of error is parameterized by the Chernoff information (sometimes also referred to as the Chernoff distance) [Cover and Thomas, 1991], and it is given by

$$C(\mathcal{H}_1||\mathcal{H}_0) = -\log \min_{0 \le n \le 1} \mathbb{E}_{|\mathcal{H}_0}\{[l(\mathbf{y})]^n\}. \tag{8.33}$$

Due to the involved minimization over n, the Chernoff information in (8.33) is difficult to optimize over. Therefore, we use a special case of the Chernoff information called the Bhattacharyya distance as the optimization criterion, where the Bhattacharyya distance is obtained by fixing $n = 0.5$ in (8.33). The Bhattacharyya distance is given by

$$\mathcal{B}(\mathcal{H}_1||\mathcal{H}_0) = -\log \rho, \tag{8.34}$$

where the Bhattacharyya coefficient ρ is

$$\rho = \mathbb{E}_{|\mathcal{H}_0}\{\sqrt{l(\mathbf{y})}\}. \tag{8.35}$$

It is easy to verify from (8.35) that the Bhattacharya distance is symmetric, which means $\mathcal{B}(\mathcal{H}_1||\mathcal{H}_0) = \mathcal{B}(\mathcal{H}_0||\mathcal{H}_1)$. More importantly, the upper and lower bounds for the Bayesian probability of error can be obtained using the Bhattacharyya coefficient. The bounds are given as follows [Kadota and Shepp, 1967]

$$\frac{1}{2}\min\big(P(\mathcal{H}_0), P(\mathcal{H}_1)\big)\rho^2 \leq P_e \leq \sqrt{P(\mathcal{H}_0)P(\mathcal{H}_1)}\rho. \tag{8.36}$$

Therefore, in place of the Bayesian error probability, we could maximize the Bhattacharyya distance. Furthermore, when $\int [p(\mathbf{y}|\mathcal{H}_1)]^n [p(\mathbf{y}|\mathcal{H}_0)]^{1-n} d\mathbf{y}$ is symmetric in n and the observations are independent and identically distributed, the Bhattacharyya distance is exponentially the best [Kailath, 1967], i.e.,

$$P_e \overset{\text{as.}}{=} \exp\big(-\mathcal{B}(\mathcal{H}_1||\mathcal{H}_0)\big) \text{ for } P_e \to 0. \tag{8.37}$$

A similar bound as that in (8.36) can be obtained linking the Bhattacharyya distance and the more general Bayes risk

$$\mathcal{R} = \sum_{i=0}^{1} \sum_{j=0}^{1} C_{ij} \Pr(\mathcal{H}_i|\mathcal{H}_j) \Pr(\mathcal{H}_j). \tag{8.38}$$

The bound is given by [Kobayashi and Thomas, 1967]

$$\mathcal{R}_0 + \mathcal{R}_2 \rho^2 \leq \mathcal{R} \leq \mathcal{R}_0 + \sqrt{\mathcal{R}_1} \rho, \tag{8.39}$$

for constants $\mathcal{R}_0, \mathcal{R}_1, \mathcal{R}_2$ dependent on the costs C_{ij} and the prior probabilities $P(\mathcal{H}_i)$, which further motivates the choice of the Bhachattaryya distance for selecting measurements in the Bayesian setting.

Note that under the independence assumption of the measurements, the likelihood $l(\mathbf{y})$ can be factorized as the product of the *local* measurement likelihoods $l_m(x) := p_m(x|\mathcal{H}_1)/p_m(x|\mathcal{H}_0)$, i.e.,

$$l(\mathbf{y}) = \frac{p(\mathbf{y}|\mathcal{H}_1)}{p(\mathbf{y}|\mathcal{H}_0)} = \prod_{m=1}^{M} [l_m(x)]^{w_m}, \tag{8.40}$$

which implies linearity in the selection variables of the Bhattacharyya distance [Chepuri and Leus, 2016b], i.e.,

$$\mathcal{B}(\mathcal{H}_1||\mathcal{H}_0) = \sum_{m=1}^{M} w_m \mathcal{B}_m(\mathcal{H}_1||\mathcal{H}_0), \tag{8.41}$$

where $\mathcal{B}_m(\mathcal{H}_1||\mathcal{H}_0) := -\log \mathbb{E}_{|\mathcal{H}_0}\{\sqrt{l_m(x)}\}$ is the mth local Bhattacharya distance.

While the linearity of (8.41) eases the optimization over $\mathbf{w}$, in general, the independence on measurements of the Bhattacharya distance allows for an

offline design of the selection variable.

Unknown Prior Probabilities. In the case that prior probabilities are unknown, we typically design detectors based on the Neyman-Pearson approach. In this setting, one of the error probabilities is considered fixed, e.g., P_{fa}, and the other error probability P_{m} is minimized. When this procedure is followed, the decision is again based upon the log-likelihood ratio test

$$\log l(\mathbf{y}) = \log \frac{p(\mathbf{y}|\mathcal{H}_1)}{p(\mathbf{y}|\mathcal{H}_0)} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \gamma, \tag{8.42}$$

where now the threshold γ is obtained by fixing P_{fa}.

For a Neyman-Pearson problem, it is known that the best achievable error exponent in the probability of error, e.g., in P_{m}, is given by the relative entropy or Kullback-Leibler (KL) divergence

$$\log P_{\text{m}} \overset{\text{as.}}{=} -\mathcal{K}(\mathcal{H}_1|\mathcal{H}_0) \text{ for } P_{\text{m}} \to 0. \tag{8.43}$$

where the Kullback-Leibler divergence is defined as

$$\mathcal{K}(\mathcal{H}_1|\mathcal{H}_0) := \mathbb{E}_{|\mathcal{H}_1}\{\log l(\mathbf{y})\}. \tag{8.44}$$

These properties allow to construct an upper and lower bound for P_{m}, for P_{fa} values of practical interest, i.e., $P_{\text{fa}} \approx 0$ ([Chepuri and Leus, 2016b]). That is,

$$\exp(-\mathcal{K}(\mathcal{H}||\mathcal{H}_0)) \leq P_{\text{m}} \leq \frac{1}{1 + \frac{(\mathcal{K}(\mathcal{H}_1||\mathcal{H}_0) - \log\gamma)^2}{v^2}}, \tag{8.45}$$

where v^2 is the variance of the log-likehood ratio and γ is the threshold for a target P_{fa}.

Similarly to the Bhattacharyya distance, when the measurements are conditional independent on the hyphotesis $\mathcal{H}$, the KL divergence is linear in $\mathbf{w}$

$$\mathcal{K}(\mathcal{H}_1||\mathcal{H}_0) = \sum_{m=1}^{M} w_m \mathcal{K}_m(\mathcal{H}_1||\mathcal{H}_0), \tag{8.46}$$

where $\mathcal{K}_m(\mathcal{H}_1||\mathcal{H}_0) := \mathbb{E}_{|\mathcal{H}_1}\{\log l_m(x)\}$ is the mth local KL divergence.

Note that due to the asymmetry of the KL divergence, it cannot be considered a proper distance. However, this property can be used if instead of minimizing the P_{m} for a given P_{fa} we want to minimize a given P_{fa} for a given P_{m}. To do so, we only need to interchange $\mathcal{H}_1$ and $\mathcal{H}_0$ in the previous expressions. When the symmetry of the metric is of importance, instead of the KL divergence, its symmetric version, the J-divergence can be employed:

$$\mathcal{D}(\mathcal{H}_1||\mathcal{H}_0) := \mathcal{K}(\mathcal{H}_1||\mathcal{H}_0) + \mathcal{K}(\mathcal{H}_0||\mathcal{H}_1). \tag{8.47}$$

As it is composed by two KL-divergence terms, this distance is also linear in $\mathbf{w}$ and can be used to find an upper and lower bound for $P_{\mathrm{e}} = 0.5(P_{\mathrm{fa}} + P_{\mathrm{m}})$ for general distributions. For arbitrary P_{e}, the upper bound is only possible to obtain when the observations are Gaussian [Kadota and Shepp, 1967].

Under the assumption of Gaussian observations, the discussed metrics between the different hypotheses under test are shown in Table 8.1. Here, $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$ denote the mean vector and the covariance matrix of the ith distribution, respectively.

8.3.1.2. Sparse Sampler Design

Using the discussed surrogate measures as $f(\mathbf{w})$, a new formulation of the sparse sensing problems (8.5) and (8.6) can be stated. In what follows, we

Table 8.1: Summary of metrics for Gaussian probability distributions.

Metric	Expression		Setting
	Bhattacharyya		
$\mathcal{B}(\mathcal{H}_1\|\mathcal{H}_0) \coloneqq$	$\frac{1}{8}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_0)+\frac{1}{2}\log\left(\frac{\det(\boldsymbol{\Sigma})}{\sqrt{\det(\boldsymbol{\Sigma}_1)\det(\boldsymbol{\Sigma}_0)}}\right),\ \boldsymbol{\Sigma}=0.5(\boldsymbol{\Sigma}_1+\boldsymbol{\Sigma}_0)$	(8.51)	Bayesian
	Kullback-Leibler		
$\mathcal{K}(\mathcal{H}_1\|\mathcal{H}_0) \coloneqq$	$\frac{1}{2}\Big(\mathrm{tr}(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\Sigma}_1)+(\boldsymbol{\mu}_1-\boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_0)-N+\log\big(\det(\boldsymbol{\Sigma}_0)\big)-\log\big(\det(\boldsymbol{\Sigma}_1)\big)\Big)$	(8.52)	Neyman-Pearson
	J-Divergence		
$\mathcal{D}(\mathcal{H}_0\|\mathcal{H}_1) \coloneqq$	$\mathcal{K}(\mathcal{H}_1\|\mathcal{H}_0)+\mathcal{K}(\mathcal{H}_0\|\mathcal{H}_1)$	(8.53)	Neyman-Pearson

present different approaches for designing sparse samplers for detection, giving special attention to the Gaussian case for correlated observations.

Conditionally Independent Observations. As discussed before, when the observations are considered to be conditionally independent, the three discussed metrics are linear in $\mathbf{w}$. That is, if we collected the local distances/divergences in a vector $\mathbf{d} \in \mathbb{R}^M$, the CC and PC problems respectively simplify to

$$\underset{\mathbf{w}\in\{0,1\}^M}{\arg\max}\ \mathbf{d}^\top\mathbf{w} \quad \text{s.t.} \quad \|\mathbf{w}\|_0 = K, \tag{8.54}$$

$$\underset{\mathbf{w}\in\{0,1\}^M}{\arg\min}\ \|\mathbf{w}\|_0 \quad \text{s.t.} \quad \mathbf{d}^\top\mathbf{w} \geq \lambda. \tag{8.55}$$

Due to the linearity of the metric, it can be shown that for both problems a greedy strategy finds the optimal solution. More specifically, the largest entries in $\mathbf{d}$ are selected till we have K sensors (CC problem) or the desired performance has been reached (PC problem). As a result, the proposed selection mechanism is very attractive as its complexity is mainly that of sorting a list. An example of the performance, in terms of probability of error, of the selection for a case where $\boldsymbol{\Sigma} = \mathbf{I}$ and $\boldsymbol{\theta} = \mathbf{1}$ with prior probabilities $P(\mathcal{H}_0) = 0.3$ and $P(\mathcal{H}_1) = 0.7$ is shown in Figure 8.6.

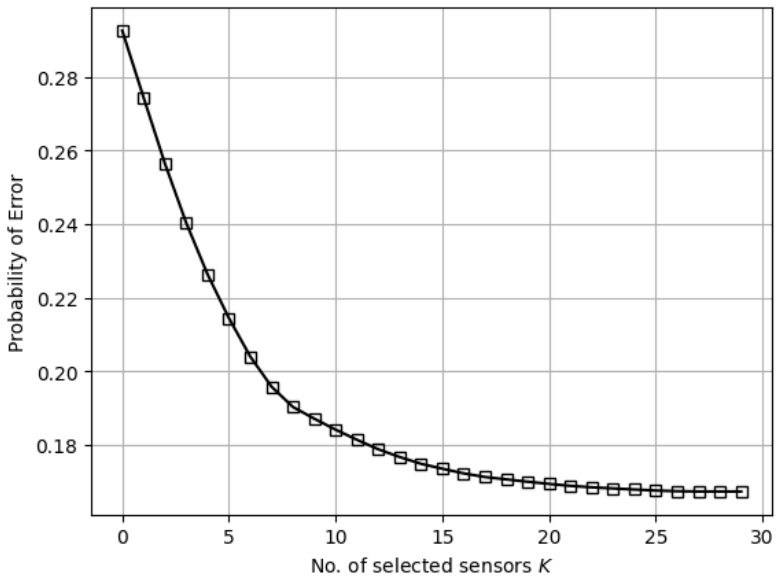

Figure 8.6: Performance of the optimal selection with conditionally independent observations in the Bayesian setting. Here, $\mathbf{\Sigma} = \mathbf{I}$, $\boldsymbol{\theta} = \mathbf{1}$ and $P(\mathcal{H}_0) = 0.3$ and $P(\mathcal{H}_1) = 0.7$.

As we will see next, for dependent observations, in general no simple mechanism exists to obtain optimal selections. However, similar to the methods developed for estimation, we discuss alternatives leveraging the convex and submodular machinery for optimizing the introduced metrics.

Dependent Observations. Although the conditional independence assumption is valid in circumstances where the noise is caused solely by the measurement acquisition, when sensors experience external noise sources or if the signal itself is stochastic in nature the independence assumption might not hold anymore. As a result, the additive property of the metrics is not valid anymore. Thus, computing the optimal sparse sampler under dependent observations requires considering the non-linear (and possibly non-convex) metrics. So, in general, we will have to solve the sparse sampler design problem [cf. (8.5)-(8.6)]

using a global optimization technique [Horst and Pardalos, 2013].

Despite that there might not be a general recipe to solve problems (8.5) and (8.6) for arbitrary measurement distributions, in the particular case of correlated Gaussian observations, the presented metrics allow for a selection mechanism based on convex and submodular optimization that enables an efficient offline sampler design.

When the two distributions only differ by their means as in our binary hypothesis problem formulation [cf. (8.26)-(8.27)], from Table 8.1 it is seen that all metrics (upto a scale factor) coincide; that is

$$\mathcal{B}(\mathcal{H}_1||\mathcal{H}_0) \propto \mathcal{K}(\mathcal{H}_1||\mathcal{H}_0) \propto \mathcal{D}(\mathcal{H}_1||\mathcal{H}_0) \propto (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0), \quad (8.56)$$

which can be interpreted as a measure of the signal-to-noise ratio, effectively defining the detection performance. By considering the selection matrix and defining $\mathbf{m} := \boldsymbol{\mu}_1 - \boldsymbol{\mu}_0$, we can consider the quantity

$$s(\mathbf{w}) := \mathbf{m}^\top \boldsymbol{\Phi}^\top(\mathbf{w}) \left(\boldsymbol{\Phi}(\mathbf{w}) \boldsymbol{\Sigma} \boldsymbol{\Phi}^\top(\mathbf{w})\right)^{-1} \boldsymbol{\Phi}(\mathbf{w}) \mathbf{m} \quad (8.57)$$

as our performance metric and solve, for instance, the PC problem

$$\underset{w \in \{0,1\}^M}{\arg\min} \quad \|\mathbf{w}\|_0 \quad \text{s.t} \quad s(\mathbf{w}) \geq \lambda. \quad (8.58)$$

Although the above problem is still hard to solve, we can consider the following convex relaxation

$$\underset{\mathbf{w} \in [0,1]^M}{\arg\min} \ \mathbf{1}^\top \mathbf{w} \quad (8.59)$$

$$\text{s.t} \begin{bmatrix} \mathbf{S}^{-1} + \alpha^{-1}\text{diag}(\mathbf{w}) & \mathbf{S}^{-1}\mathbf{m} \\ \mathbf{m}^\top \mathbf{S}^{-1} & \lambda' \end{bmatrix} \geq 0,$$

where $\lambda' := \lambda - \mathbf{m}^\top \mathbf{S}^{-1}\mathbf{m}$. The linear matrix inequality (LMI) is derived in a similar way as was done for estimation in correlated noise, i.e., by decomposing the covariance matrix $\mathbf{\Sigma}$ as $\mathbf{\Sigma} = \alpha \mathbf{I}_M + \mathbf{S}$ with $\alpha \in \mathbb{R}$ chosen such that $\mathbf{S}$ is invertible and well conditioned, and by using the equivalence of the inequality

$$\mathbf{m}^\top \mathbf{S}^{-1}[\mathbf{S}^{-1} + \alpha^{-1}\text{diag}(\mathbf{w})]^{-1}\mathbf{S}^{-1}\mathbf{m} \leq \lambda' \tag{8.60}$$

with the LMI in (8.59). Note that (8.60) is derived from the constraint in (8.58) using the matrix inversion lemma and the properties of the selection matrix.

Analogously, the CC problem can be formulated using the same LMI as

$$\begin{aligned} &\underset{\mathbf{w}\in[0,1]^M, t}{\arg\min} \quad t \\ &\text{s.t } \mathbf{1}^\top \mathbf{w} = K, \\ &\begin{bmatrix} \mathbf{S}^{-1} + \alpha^{-1}\text{diag}(\mathbf{w}) & \mathbf{S}^{-1}\mathbf{m} \\ \mathbf{m}^\top \mathbf{S}^{-1} & t \end{bmatrix} \geq 0, \end{aligned} \tag{8.61}$$

with an auxiliary variable t that has been used to formulate the epigraph form of a problem using the left hand side of (8.60) as cost function.

As the problems (8.59) and (8.61) are convex, they can be solved efficiently using any off-the-shelf convex optimization method. Although the resulting solution is not necessarily binary, procedures such as randomized rounding can be employed to recover a feasible selection vector with good performance from the continuous solution. Although the above convex relaxations provide a way to tackle the sparse sampler design for unequal means, greedy methods can still

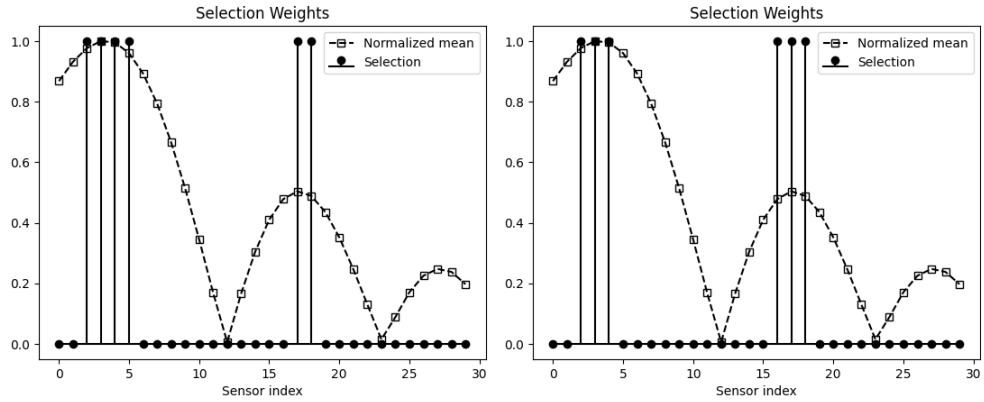

Figure 8.7: Comparison of the selection of the convex (left) and submodular (right) optimization approaches when selecting $K = 6$ measurements considering the setting of our running example under equi-correlated noise [cf. (8.10)] with $\gamma = 0.5$. In dashed lines the normalized mean is shown for illustration. Here, $\boldsymbol{\theta} = \mathbf{1}$.

be employed in several instances while obtaining a comparable or even better detection performance. For example, as suggested in [Coutino et al., 2018b], a greedy procedure that constructs a solution by greedily selecting the best solution of the original cost function $s(\mathbf{w})$ and the surrogate

$$s'(\mathbf{w}) := \begin{cases} 0 & \text{if} \quad \mathbf{w} = \mathbf{0} \\ \log\det \boldsymbol{M}(\mathbf{w}) & \text{otherwise} \end{cases} \tag{8.62}$$

where $\boldsymbol{M}(\mathbf{w})$ denotes the matrix in the left hand side of the LMI in (8.59) and (8.61) can be considered. As the cost function (8.62) is submodular, its greedy optimization achieves provable near-optimality guarantees. Further, because at every step it can be efficiently computed, i.e., a determinant update, this mechanism for sparse sampler design can be employed when solving the semidefinite program in (8.59) or (8.61) is not practically feasible.

An illustration of the different selections made by the convex and greedy

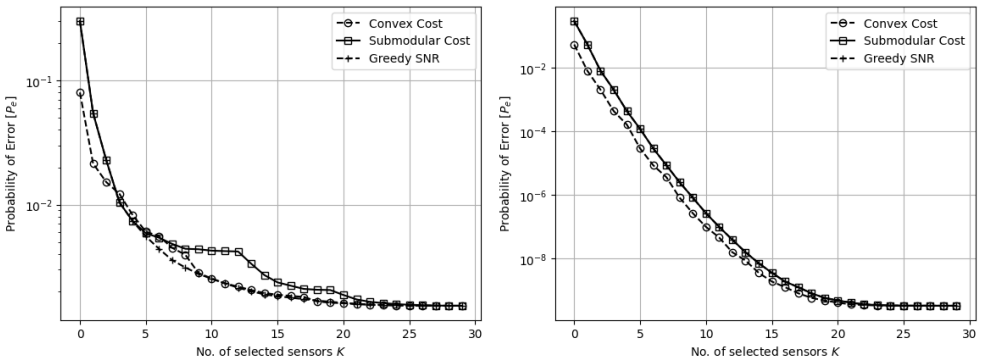

Figure 8.8: Comparison of the selection of the convex and submodular optimization approaches when selecting different number of measurements considering spatially decaying [left; cf. (8.9)] and spatially constant [right; cf. (8.10)] correlated noise. In dashed-crossed lines the performance of greedily optimizing the SNR function is shown. Here, $\boldsymbol{\theta} = \mathbf{1}$ and $\gamma = 0.5$ for both types of noise.

selection approaches is shown in Fig. 8.7. In this example, we consider again $\boldsymbol{\theta} = \mathbf{1}$ and ask the methods to select a set of cardinality $K = 6$ when the noise is equi-correlated with parameter $\gamma = 0.5$ [cf. (8.10)]. We further take $\sigma^2 = 1$. As the convex solution not necessarily provides an integer solution, here, we have selected the K entries of $\mathbf{w}$ with the highest value. To illustrate the possible impact of the mean of different measurements, the means normalized with respect to the maximum are also shown. Finally, a complete comparison for our running example, under spatially decaying or spatially constant correlated noise, is shown in Fig. 8.8. Here, we compare the methods under the same setting across different selection cardinalities K and set $\gamma = 0.5$ and $\sigma^2 = 1$ for both cases. In addition, for reference, we also include the performance of performing greedy optimization directly on the SNR function. Note that while the three methods achieve similar performance, the greedy methods are cheap alternatives to convex optimization which in several instances outperforms the latter even if the function is not submodular, i.e., for the Greedy

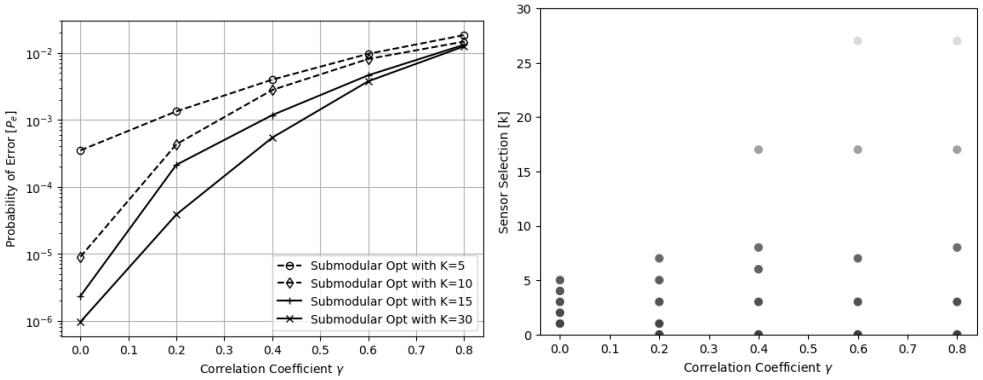

Figure 8.9: (Left) Sensor selection of the submodular strategy for different numbers of sensors K and correlation coefficients γ when the noise correlation is giving by the spatially decaying correlation in (8.9). (Right) Selected sensors for different γ for $K = 5$.

SNR method. However, it is known that in certain instances this might not be the case as greedy optimization on non-submodular functions can fail, see, e.g., [Coutino et al., 2018b].

To illustrate the effect of the correlation pattern and strength, we consider the sensor selection problem when the noise correlation is given by respectively (8.9) and (8.10) when using the submodular selection strategy.

The result of the selection for a spatially decaying correlation function is shown in Fig. 8.9. While typically for estimation tasks the strategy selects sensors that lead to a well-conditioned measurement matrix, for detection the focus lays mostly on selecting measurements with high signal-to-noise ratio. This introduces a natural trade off between correlation and energy on the measurements. For instance, in the case of a spatially decaying noise correlation matrix, sensors seem to spread more as the correlation coefficient increases. As a result, more sensors are needed to reach the same performance as the correlation grows.

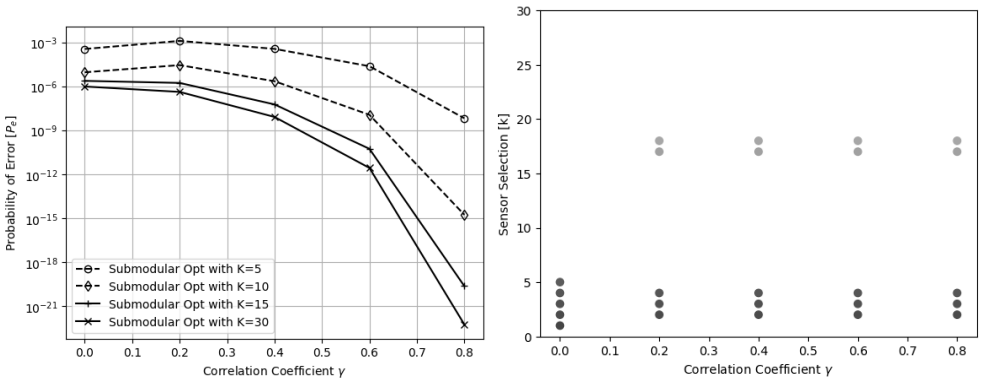

Figure 8.10: (Left) Sensor selection of the submodular strategy for different numbers of sensors K and correlation coefficients γ when the noise correlation is giving by the spatially constant correlation in (8.10). (Right) Selected sensors for different γ for $K = 5$.

This trade off becomes less apparent when an equi-correlated noise matrix is considered. In this setting, there is no need for the sensors to have a particular arrangement to leverage correlation. This naturally leads to clear clusters based on the signal-to-noise ratio. Also, it allows for a reduced number of required sensors when the correlation coefficient increases. These effects are illustrated in Fig. 8.10.

8.3.2. Unknown θ Parameters

In many practical detection problems, complete knowledge of the signal to detect is not available. For example, only knowledge of the subspace described by our matrix $\mathbf{H}$ might be known but the precise values of $\boldsymbol{\theta}$ not. This situation can arise in direction of arrival estimation with prior knowledge of the angular sector where the signal might arrive or in the communications domain where user codes are known but not whether they are active or not.

Optimality Criteria. To tackle the design of sparse samplers for this setting, let us recall our binary hypothesis test problem

$$\mathcal{H}_0 : \mathbf{x} = \mathbf{n}, \tag{8.63}$$

$$\mathcal{H}_1 : \mathbf{x} = \mathbf{H}\boldsymbol{\theta} + \mathbf{n}. \tag{8.64}$$

while keeping in mind that in this setting $\boldsymbol{\theta}$ is an unknown parameter.

As $\boldsymbol{\theta}$ is unknown, for addressing the detection problem, we need to make use of the generalized likelihood ratio test (GRLT) which decides $\mathcal{H}_1$ if

$$L_G(\mathbf{x}) = \frac{p(\mathbf{x}; \hat{\boldsymbol{\theta}}_1)}{p(\mathbf{x}; \hat{\boldsymbol{\theta}}_0)} > \gamma \tag{8.65}$$

where $\hat{\boldsymbol{\theta}}_i$ is the ML estimate of $\boldsymbol{\theta}$ under $\mathcal{H}_i$. In our particular case, these estimates are given, respectively, by

$$\hat{\boldsymbol{\theta}}_0 := \mathbf{0}, \tag{8.66}$$

$$\begin{aligned}\hat{\boldsymbol{\theta}}_1 = &\left(\mathbf{H}^\top \boldsymbol{\Phi}^\top(\mathbf{w}) \left(\boldsymbol{\Phi}(\mathbf{w})\boldsymbol{\Sigma}\boldsymbol{\Phi}^\top(\mathbf{w})\right)^{-1} \boldsymbol{\Phi}(\mathbf{w})\mathbf{H}\right)^{-1} \\ &\times \mathbf{H}^\top \boldsymbol{\Phi}^\top(\mathbf{w}) \left(\boldsymbol{\Phi}(\mathbf{w})\boldsymbol{\Sigma}\boldsymbol{\Phi}^\top(\mathbf{w})\right)^{-1} \boldsymbol{\Phi}(\mathbf{w})\mathbf{x},\end{aligned} \tag{8.67}$$

where to compute the ML estimate under $\mathcal{H}_1$ prewhitening after sampling has been performed. In this case, it can be shown that the logarithm of the GRLT is distributed as [Kay, 2009]

$$2 \ln L_G(\mathbf{x}) \sim \begin{cases} \chi^2_N & \text{under } \mathcal{H}_0, \\ \chi^2_N(\lambda(\mathbf{w}; \boldsymbol{\theta})) & \text{under } \mathcal{H}_1, \end{cases} \tag{8.68}$$

where N is the dimensionality of $\boldsymbol{\theta}$ and

$$\lambda(\mathbf{w};\boldsymbol{\theta}) = \boldsymbol{\theta}^{\top}\mathbf{H}^{\top}\boldsymbol{\Phi}^{\top}(\mathbf{w})\left(\boldsymbol{\Phi}(\mathbf{w})\boldsymbol{\Sigma}\boldsymbol{\Phi}^{\top}(\mathbf{w})\right)^{-1}\boldsymbol{\Phi}(\mathbf{w})\mathbf{H}\boldsymbol{\theta} \tag{8.69}$$

is the non-centrality parameter of the chi-squared distribution under $\mathcal{H}_1$. This leads to a detector performance defined by

$$P_{\text{fa}} = Q_{\chi^2_N}(\gamma'), \tag{8.70}$$

$$P_{\text{d}} = Q_{\chi^2_N(\lambda(\mathbf{w};\boldsymbol{\theta}))}(\gamma'), \tag{8.71}$$

where γ' is the modified threshold for the test in (8.68).

Based on this performance metric, we can design a sparse sampler which maximizes P_{d} for a fixed P_{fa}. This is achieved by noticing that the P_{d} is a monotone function of $\lambda(\mathbf{w};\boldsymbol{\theta})$. Therefore, by maximizing the non-centrality parameter (8.69), the power of the test can be maximized. However, as $\boldsymbol{\theta}$ is unknown a priori, direct maximization of $\lambda(\mathbf{w};\boldsymbol{\theta})$ is not possible. In the following, we briefly discuss three alternatives to design sparse samplers using the noncentrality parameter as selection criterion for the designs. For more details and an extended formulation of the unknown parameter case, see, e.g., [Coutino et al., 2017]

Average Design. If there is some prior belief that the parameter lies within a known set $\boldsymbol{\Omega}$, a straightforward design approach can be based on maximizing the average noncentrality parameter, i.e.,

$$\max_{\mathbf{w}\in[0,1]^M} \int_{\boldsymbol{\theta}\in\boldsymbol{\Omega}} \lambda(\mathbf{w};\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}. \tag{8.72}$$

This design is equivalent to the one obtained through a Bayesian approach where a prior $p(\boldsymbol{\theta})$ is given to all the possible vectors $\boldsymbol{\theta} \in \boldsymbol{\Omega}$. This approach

requires an integration over the domain $\boldsymbol{\Omega}$. Note that when $\boldsymbol{\Omega}$ is discretized and the integral is approximated by a summation, the previously discussed approach based on LMIs, see, e.g., (8.61), can be directly applied as $\lambda(\mathbf{w}, \boldsymbol{\theta})$ exhibits the same structure as the SNR expression in (8.57).

Worst Case (Max-Min) Design. Alternatively, we could consider a worst-case design which aims to select samples that maximize the noncentrality parameter in the worst case, i.e.,

$$\max_{\mathbf{w}\in[0,1]^M} \min_{\boldsymbol{\theta}\neq\mathbf{0}} \lambda(\mathbf{w};\boldsymbol{\theta}). \tag{8.73}$$

Given the structure of the noncentrality parameter, this design problem becomes an eigenvalue maximization problem; that is

$$\max_{\mathbf{w}\in[0,1]^M} \lambda_{\min}\left(\mathbf{H}^\top\boldsymbol{\Phi}^\top(\mathbf{w})\left(\boldsymbol{\Phi}(\mathbf{w})\boldsymbol{\Sigma}\boldsymbol{\Phi}^\top(\mathbf{w})\right)^{-1}\boldsymbol{\Phi}(\mathbf{w})\mathbf{H}\right), \tag{8.74}$$

which is analogous to the sparse sampler design problem using the E-optimality criterion when the ML error covariance matrix is used [cf. (8.20)].

Log-Det Design. In many instances, the max-min criterion for a composite hypothesis test has an inherently pessimistic nature [Feder and Merhav, 2002]. Therefore, a less stringent design, based on the D-optimality criterion could be to design samplers based on the problem

$$\max_{\mathbf{w}\in[0,1]^M} \log\det\left(\mathbf{H}^\top\boldsymbol{\Phi}^\top(\mathbf{w})\left(\boldsymbol{\Phi}(\mathbf{w})\boldsymbol{\Sigma}\boldsymbol{\Phi}^\top(\mathbf{w})\right)^{-1}\boldsymbol{\Phi}(\mathbf{w})\mathbf{H}\right), \tag{8.75}$$

which instead of only focusing on the minimum eigenvalue, considers the energy distribution across the different directions as it aims to maximize the product of

the eigenvalues. Note that, again, this problem is analogous to that of sparse samplers for estimation when the D-optimality criterion and the ML error covariance are used.

8.4. Conclusions

Distributed sensing has been investigated in this chapter. More specifically, we looked at the optimal placement of sensors to carry out a specific inference task such as estimation or detection. To simplify the presentation, a linear Gaussian model was considered and the task was to estimate the unknown variables or to detect whether there is a signal or not. The latter detection problem was studied for known and unknown signals. Even for the considered linear model, the problem statement is not easy to tackle, especially in case of correlated noise (i.e., conditionally dependent observations). Simulations on a field estimation problem show that sensors are selected that result in a well-conditioned and highly energetic measurement matrix. Furthermore, in case of correlated noise, this behavior is traded off with the correlation function. For a spatially decaying noise correlation function, this can potentially result in requiring more sensors to reach the same estimation performance as for uncorrelated noise. In case of a constant noise correlation function, sensors do not have to be close to each other to exploit the correlation, and we see that less sensors are needed to reach the same performance as in the uncorrelated noise case. For field detection a similar behavior is observed except that in this case the conditioning of the selected measurement matrix is not that important as the focus is more on selecting measurements with a high signal-to-noise ratio.

Bibliography

[Ali and Silvey, 1966] Ali, S. and Silvey, S. D. (1966). A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Stat. Society. Series B (Methodological)*, 28(1):131–142.

[Bajovic et al., 2011] Bajovic, D., Sinopoli, B., and Xavier, J. (2011). Sensor selection for event detection in wireless sensor networks. *IEEE Trans. Signal Process.*, 59(10):4938–4953.

[Blu et al., 2008] Blu, T., Dragotti, P.-L., Vetterli, M., Marziliano, P., and Coulot, L. (2008). Sparse sampling of signal innovations. *IEEE Signal Process. Mag.*, 25(2):31–40.

[Cambanis and Masry, 1983] Cambanis, S. and Masry, E. (1983). Sampling designs for the detection of signals in noise. *IEEE Trans. Info. Theory*, 29(1):83–104.

[Chepuri and Leus, 2015] Chepuri, S. P. and Leus, G. (2015). Sparsity-promoting sensor selection for non-linear measurement models. *IEEE Trans. Sig. Proc.*, 63(3):684–698.

[Chepuri and Leus, 2016a] Chepuri, S. P. and Leus, G. (2016a). Sparse sensing for distributed detection. *IEEE Trans. Signal Process.*, 64(6):1446–1460.

[Chepuri and Leus, 2016b] Chepuri, S. P. and Leus, G. (2016b). Sparse sensing for statistical inference. *Foundations and Trends in Signal Processing*, 9(3–4):233–368.

[Coutino et al., 2018a] Coutino, M., Chepuri, S., and Leus, G. (2018a). Near-optimal sparse sensing for Gaussian detection with correlated observations. *IEEE Tran. on Signal Proc.*, 66(15):4025–4039.

[Coutino et al., 2017] Coutino, M., Chepuri, S. P., and Leus, G. (2017). Sparse sensing for composite matched subspace detection. In *2017 IEEE 7th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, pages 1–5. IEEE.

[Coutino et al., 2018b] Coutino, M., Chepuri, S. P., and Leus, G. (2018b). Submodular sparse sensing for gaussian detection with correlated observations. *IEEE Transactions on Signal Processing*, 66(15):4025–4039.

[Cover and Thomas, 1991] Cover, T. M. and Thomas, J. A. (1991). Information theory and statistics. *Elements of information theory*, 1(1):279–335.

[Eldar and Michaeli, 2009] Eldar, Y. C. and Michaeli, T. (2009). Beyond bandlimited sampling. *IEEE Signal Processing Magazine*, 26(3):48–68.

[Feder and Merhav, 2002] Feder, M. and Merhav, N. (2002). Universal composite hypothesis testing: A competitive minimax approach. *IEEE Transactions on information theory*, 48(6):1504–1517.

[Horst and Pardalos, 2013] Horst, R. and Pardalos, P. M. (2013). *Handbook of global optimization*, volume 2. Springer Science & Business Media.

[Joshi and Boyd, 2009] Joshi, S. and Boyd, S. (2009). Sensor selection via convex optimization. *IEEE Trans. Signal Process.*, 57(2):451–462.

[Kadota and Shepp, 1967] Kadota, T. and Shepp, L. (1967). On the best finite set of linear observables for discriminating two gaussian signals. *IEEE Transactions on Information Theory*, 13(2):278–284.

[Kailath, 1967] Kailath, T. (1967). The divergence and bhattacharyya distance measures in signal selection. *IEEE transactions on communication technology*, 15(1):52–60.

[Kay, 2009] Kay, S. M. (2009). *Fundamentals of statistical processing, Volume 2: Detection theory*. Pearson Education India.

[Kobayashi and Thomas, 1967] Kobayashi, H. and Thomas, J. B. (1967). Distance measures and related criteria. In *5th Annu. Allerton Conf. Circuit System Theory*, pages 491–500.

[Krause et al., 2008] Krause, A., Singh, A., and Guestrin, C. (2008). Near-optimal sensor placements in Gaussian processes: Theory, efficient algorithms and empirical studies. *The Journal of Machine Learning Research*, 9(2):235–284.

[Liu et al., 2014] Liu, S., Fardad, M., Varshney, P. K., and Masazade, E. (2014). Optimal periodic sensor scheduling in networks of dynamical systems. *IEEE Trans. Signal Process.*, 62(12):3055–3068.

[Molisch and Win, 2004] Molisch, A. and Win, M. (2004). Mimo systems with antenna selection. *IEEE Microwave Magazine*, 5(1):46–56.

[Nemhauser et al., 1978] Nemhauser, G. L., Wolsey, L. A., and Fisher, M. L. (1978). An analysis of approximations for maximizing submodular set functions— I. *Mathematical Programming*, 14(1):265–294.

[Ranieri et al., 2014] Ranieri, J., Chebira, A., and Vetterli, M. (2014). Near-optimal sensor placement for linear inverse problems. *IEEE Trans. Signal Process.*, 62(5):1135–1146.

[Shamaiah et al., 2010] Shamaiah, M., Banerjee, S., and Vikalo, H. (2010). Greedy sensor selection: Leveraging submodularity. In *49th IEEE conference on decision and control (CDC)*, pages 2572–2577. IEEE.

[Vaidyanathan and Pal, 2011] Vaidyanathan, P. P. and Pal, P. (2011). Sparse sensing with co-prime samplers and arrays. *IEEE Trans. Signal Process.*, 59(2):573–586.

[Yu and Varshney, 1997] Yu, C.-T. and Varshney, P. K. (1997). Sampling design for Gaussian detection problems. *IEEE Trans. Signal Process.*, 45(9):2328–2337.