CHAPTER

9 SPARSITY-AWARE LEARNING: CONCEPTS AND THEORETICAL FOUNDATIONS

CHAPTER OUTLINE

9.1 INTRODUCTION

In Chapter 3, the notion of regularization was introduced as a tool to address a number of problems that are usually encountered in machine learning. Improving the performance of an estimator by shrinking the norm of the minimum variance unbiased (MVU) estimator, guarding against overfitting,

Machine Learning. http://dx.doi.org/10.1016/B978-0-12-801522-3.00009-4

coping with ill-conditioning, and providing a solution to an underdetermined set of equations are some notable examples where regularization has provided successful answers. Some of the advantages were demonstrated via the ridge regression concept, where the least-squares (LS) cost function was combined, in a tradeoff rationale, with the squared Euclidean norm of the desired solution.

In this and the next chapter, our interest will be on alternatives to the Euclidean norm, and in particular the focus will revolve around the ℓ_1 norm; this is the sum of the absolute values of the components comprising a vector. Although seeking a solution to a problem via the ℓ_1 norm regularization of a cost function has been known and used since the 1970s, it is only recently that it has become the focus of attention of a massive volume of research in the context of compressed sensing. At the heart of this problem lies an underdetermined set of linear equations, which, in general, accepts an infinite number of solutions. However, in a number of cases, an extra piece of information is available: the true model, whose estimate we want to obtain, is sparse; that is, only a few of its coordinates are nonzero. It turns out that a large number of commonly used applications can be cast under such a scenario and can benefit by sparse modeling.

Besides its practical significance, sparsity-aware learning has offered to the scientific community novel theoretical tools and solutions to problems that only a few years ago seemed intractable. This is also a reason that this is an interdisciplinary field of research encompassing scientists from, for example, mathematics, statistics, machine learning, and signal processing. Moreover, it has already been applied in many areas, ranging from biomedicine to communications and astronomy. At the time this book was compiled, there was a "research happening" in the field, which posed some difficulties in assembling related material. We made an effort to present in a unifying way the basic notions and ideas that run across this field. Our goal is to provide the reader with an overview of the major contributions that have taken place in the theoretical and algorithmic fronts and have been consolidated as a distinct scientific area.

In the current chapter, the focus is on presenting the main concepts and theoretical foundations related to sparsity-aware learning techniques. We start by reviewing various norms, then we move on to establish conditions on the recovery of sparse vectors, or vectors that are sparse in a transform domain, using less observations than the dimension of the corresponding space. Geometry plays an important part in our approach. Finally, some theoretical advances that tie sparsity and sampling theory are presented. At the end of the chapter, a case study concerning image de-noising is discussed.

9.2 SEARCHING FOR A NORM

Mathematicians have been very imaginative in proposing various norms in order to equip linear spaces. Among the most popular norms used in functional analysis are the ℓ_p norms. To tailor things to our needs, given a vector $\boldsymbol{\theta} \in \mathbb{R}^l$, its ℓ_p norm is defined as

$$\|\boldsymbol{\theta}\|_p := \left(\sum_{i=1}^{l} |\theta_i|^p \right)^{1/p}. \tag{9.1}$$

For $p = 2$, the Euclidean or ℓ_2 norm is obtained, and for $p = 1$, (9.1) results in the ℓ_1 norm, that is,

$$\|\boldsymbol{\theta}\|_1 = \sum_{i=1}^{l} |\theta_i|. \tag{9.2}$$

If we let $p \longrightarrow \infty$, then we get the ℓ_∞ norm; let $|\theta_{i_{\max}}| \coloneqq \max\{|\theta_1|, |\theta_2|, \ldots, |\theta_l|\}$, and notice that

$$\|\boldsymbol{\theta}\|_\infty \coloneqq \lim_{p\to\infty} \left(|\theta_{i_{\max}}|^p \sum_{i=1}^{l} \left(\frac{|\theta_i|}{|\theta_{i_{\max}}|} \right)^p \right)^{1/p} = |\theta_{i_{\max}}|, \tag{9.3}$$

that is, $\|\boldsymbol{\theta}\|_\infty$ is equal to the maximum of the absolute values of the coordinates of $\boldsymbol{\theta}$. One can show that all the ℓ_p norms are true norms for $p \geq 1$; that is, satisfy all four requirements that a function $\mathbb{R}^l \longmapsto [0, \infty)$ must respect in order to be called a norm, that is,

1. $\|\boldsymbol{\theta}\|_p \geq 0$,
2. $\|\boldsymbol{\theta}\|_p = 0 \Leftrightarrow \boldsymbol{\theta} = \mathbf{0}$,
3. $\|\alpha\boldsymbol{\theta}\|_p = |\alpha| \, \|\boldsymbol{\theta}\|_p, \forall \alpha \in \mathbb{R}$,
4. $\|\boldsymbol{\theta}_1 + \boldsymbol{\theta}_2\|_p \leq \|\boldsymbol{\theta}_1\|_p + \|\boldsymbol{\theta}_2\|_p$.

The third condition enforces the norm function to be (*positively*) *homogeneous* and the fourth one is the *triangle inequality*. These properties also guarantee that any function that is a norm is also a convex one (Problem 9.3). Though strictly speaking, if we allow $p > 0$ to take values less than one in (9.1), the resulting function is not a true norm (Problem 9.8), we can still call them norms, albeit knowing that this is an abuse of the definition of a norm. An interesting case, which will be used extensively in this chapter, is the ℓ_0 norm, which can be obtained as the limit, for $p \longrightarrow 0$, of

$$\|\boldsymbol{\theta}\|_0 \coloneqq \lim_{p\to 0} \|\boldsymbol{\theta}\|_p^p = \lim_{p\to 0} \sum_{i=1}^{l} |\theta_i|^p = \sum_{i=1}^{l} \chi_{(0,\infty)}(|\theta_i|), \tag{9.4}$$

where $\chi_{\mathcal{A}}(\cdot)$ is the characteristic function with respect to a set $\mathcal{A}$, defined as

$$\chi_{\mathcal{A}}(\tau) \coloneqq \begin{cases} 1, & \text{if } \tau \in \mathcal{A}, \\ 0, & \text{if } \tau \notin \mathcal{A}. \end{cases}$$

That is, the ℓ_0 norm is equal to the number of nonzero components of the respective vector. It is very easy to check that this function is not a true norm. Indeed, this is not homogeneous, that is, $\|\alpha\boldsymbol{\theta}\|_0 \neq |\alpha| \, \|\boldsymbol{\theta}\|_0, \forall \alpha \neq 1$. Figure 9.1 shows the isovalue curves, in the two-dimensional space, that correspond to $\|\boldsymbol{\theta}\|_p = \rho \equiv 1$, for $p = 0, 0.5, 1, 2$, and ∞. Observe that for the Euclidean norm the isovalue curve has the shape of a "ball" and for the ℓ_1 norm the shape of a rhombus. We refer to them as the ℓ_2 and the ℓ_1 balls, respectively, by slightly "abusing" the meaning of a ball.[1] Observe that in the case of the ℓ_0 norm, the isovalue curve comprises both the horizontal and the vertical axes, excluding the $(0, 0)$ element. If we restrict the size of the ℓ_0 norm to be less than one, then the corresponding set of points becomes a singleton, that is, $(0, 0)$. Also, the set of all the two-dimensional points that have ℓ_0 norm less than or equal to two is the $\mathbb{R}^2$ space. This, slightly "strange" behavior, is a consequence of the discrete nature of this "norm."

Figure 9.2 shows the graph of $|\cdot|^p$, which is the individual contribution of each component of a vector to the ℓ_p norm, for different values of p. Observe that (a) for $p < 1$, the region that is formed above the graph (epigraph, see Chapter 8) is not a convex one, which verifies what we have already said, that is, the respective function is not a true norm; and (b) for values of the argument $|\theta| > 1$, the

[1] Strictly speaking, a ball must also contain all the points in the interior, that is, all concentric spheres of smaller radius, Chapter 8.

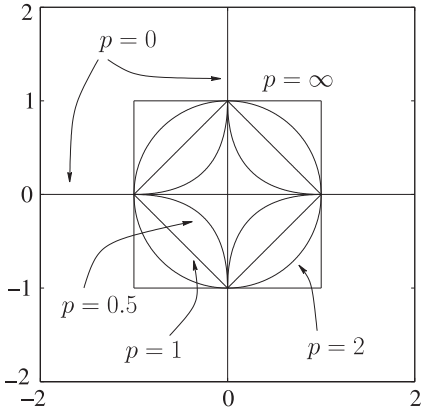

FIGURE 9.1

The isovalue curves for $\|\boldsymbol{\theta}\|_p = 1$ and for various values of p, in the two-dimensional space. Observe that for the ℓ_0 norm, the respective values cover the two axes with the exception of the point $(0, 0)$. For the ℓ_1 norm, the isovalue curve is a rhombus, and for the ℓ_2 (Euclidean) norm, it is a circle.

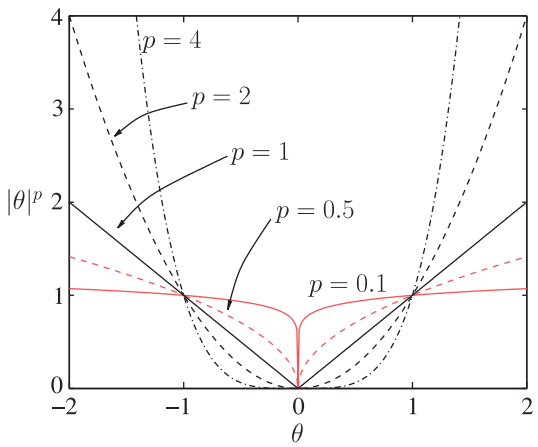

FIGURE 9.2

Observe that the epigraph, that is, the region above the graph, is nonconvex for values $p < 1$, indicating the nonconvexity of the respective $|\cdot|^p$ function. The value $p = 1$ is the smallest one for which convexity is retained. Also note that, for large values of $p > 1$, the contribution of small values of $|\theta| < 1$ to the respective norm becomes insignificant.

larger the value of $p \geq 1$ and the larger the value of $|\theta|$, the higher the contribution of the respective component to the norm. Hence, if ℓ_p norms, $p \geq 1$, are used in the regularization method, components with large values become the dominant ones and the optimization algorithm will concentrate on these by penalizing them to get smaller so that the overall cost can be reduced. The opposite is true for values $|\theta| < 1$; ℓ_p, $p > 1$ norms tend to push the contribution of such components to zero. The ℓ_1 norm is the only one (among $p \geq 1$) that retains relatively large values even for small values of $|\theta| < 1$ and, hence, components with small values can still have a say in the optimization process and can be penalized by being pushed to smaller values. Hence, if the ℓ_1 norm is used to replace the ℓ_2 one in (3.39), *only those components of the vector that are really significant in reducing the model misfit measuring term in the regularized cost function will be kept, and the rest will be forced to zero*. The same tendency, yet more aggressive, is true for $0 \leq p < 1$. The extreme case is when one considers the ℓ_0 norm. Even a small increase of a component from zero makes its contribution to the norm large, so the optimizing algorithm has to be very "cautious" in making an element nonzero.

In a nutshell, from all the true norms ($p \geq 1$), the ℓ_1 is the only one that shows respect to small values. The rest of the ℓ_p norms, $p > 1$, just squeeze them to make their values even smaller, and care mainly for the large values. We will return to this point very soon.

9.3 THE LEAST ABSOLUTE SHRINKAGE AND SELECTION OPERATOR (LASSO)

In Chapter 3, we discussed some of the benefits in adopting the regularization method for enhancing the performance of an estimator. In this chapter, we will see and study more reasons that justify the use of regularization. The first one refers to what is known as the *interpretation* power of an estimator. For example, in the regression task, we want to select those components, θ_i, of $\boldsymbol{\theta}$ that have the most important say in the formation of the output variable. This is very important if the number of parameters, l, is large and we want to concentrate on the most important of them. In a classification task, not all features are informative, hence one would like to keep the most informative of them and make the less informative ones equal to zero. Another related problem refers to those cases where we know, a priori, that a number of the components of a parameter vector are zero, but we do not know which ones. Now, the discussion at the end of the previous section becomes more meaningful. Can we use, while regularizing, an appropriate norm that can assist the optimization process (a) in unveiling such zeros or (b) to put more emphasis on the most significant of its components, those that play a decisive role in reducing the misfit measuring term in the regularized cost function, and set the rest of them equal to zero? Although the ℓ_p norms, with $p < 1$, seem to be the natural choice for such a regularization, the fact that they are not convex makes the optimization process hard. The ℓ_1 norm is the one that is "closest" to them, yet it retains the computationally attractive property of convexity.

The ℓ_1 norm has been used for such problems for a long time. In the 1970s, it was used in seismology [27, 85], where the reflected signal that indicates changes in the various earth substrates is a sparse one, that is, very few values are relatively large and the rest are small and insignificant. Since then, it has been used to tackle similar problems in different applications (e.g., [40, 80]). However, one can trace two papers that were catalytic in providing the spark for the current strong interest around the ℓ_1 norm. One came from statistics, [88], which addressed the LASSO task (first formulated, to our knowledge, in [80]), to be discussed next, and the other from the signal analysis community, [26], which formulated the *Basis Pursuit*, to be discussed in a later section.

We first address our familiar regression task

$$\boldsymbol{y} = X\boldsymbol{\theta} + \boldsymbol{\eta}, \quad \boldsymbol{y}, \boldsymbol{\eta} \in \mathbb{R}^N, \ \boldsymbol{\theta} \in \mathbb{R}^l, \ N \geq l,$$

and obtain the estimate of the unknown parameter $\boldsymbol{\theta}$ via the LS loss, regularized by the ℓ_1 norm, that is, for $\lambda \geq 0$,

$$\begin{aligned} \hat{\boldsymbol{\theta}} &:= \arg\min_{\boldsymbol{\theta} \in \mathbb{R}^l} L(\boldsymbol{\theta}, \lambda) \qquad (9.5) \\ &:= \arg\min_{\boldsymbol{\theta} \in \mathbb{R}^l} \left(\sum_{n=1}^{N} (y_n - \boldsymbol{x}_n^{\mathrm{T}} \boldsymbol{\theta})^2 + \lambda \, \|\boldsymbol{\theta}\|_1 \right) \\ &= \arg\min_{\boldsymbol{\theta} \in \mathbb{R}^l} \left((\boldsymbol{y} - X\boldsymbol{\theta})^{\mathrm{T}} (\boldsymbol{y} - X\boldsymbol{\theta}) + \lambda \, \|\boldsymbol{\theta}\|_1 \right). \qquad (9.6) \end{aligned}$$

Following the discussion with respect to the bias term given in Section 3.8 and in order to simplify the analysis, we will assume hereafter, without harming generality, that the data are of zero mean values. If this is not the case, the data can be centered by subtracting their respective sample means.

It turns out that the task in (9.6) can be equivalently written in the following two formulations:

$$\begin{aligned} \hat{\boldsymbol{\theta}} : \quad & \min_{\boldsymbol{\theta} \in \mathbb{R}^l} (\boldsymbol{y} - X\boldsymbol{\theta})^{\mathrm{T}} (\boldsymbol{y} - X\boldsymbol{\theta}), \\ \text{s.t.} \quad & \|\boldsymbol{\theta}\|_1 \leq \rho, \end{aligned} \qquad (9.7)$$

or

$$\begin{aligned} \hat{\boldsymbol{\theta}} : \quad & \min_{\boldsymbol{\theta} \in \mathbb{R}^l} \|\boldsymbol{\theta}\|_1, \\ \text{s.t.} \quad & (\boldsymbol{y} - X\boldsymbol{\theta})^{\mathrm{T}} (\boldsymbol{y} - X\boldsymbol{\theta}) \leq \epsilon, \end{aligned} \qquad (9.8)$$

given the user-defined parameters $\rho, \epsilon \geq 0$. The formulation in (9.7) is known as the LASSO and the one in (9.8) as the *basis pursuit de-noising* (BPDN) (e.g., [15]). All three formulations are equivalent for specific choices of λ, ϵ, and ρ (see, e.g., [14]). Observe that the minimized cost function in (9.6) corresponds to the Lagrangian of the formulation in (9.7). However, this functional dependence among λ, ϵ, and ρ is hard to compute, unless the columns of X are mutually orthogonal. Moreover, this equivalence does not necessarily imply that all three formulations are equally easy or difficult to solve. As we will see later in this chapter, algorithms have been developed along each one of the previous formulations. From now on, we will refer to all three formulations as the LASSO task, in a slight abuse of the standard terminology, and the specific formulation will be apparent from the context, if not stated explicitly.

We know that ridge regression admits a closed form solution, that is,

$$\hat{\boldsymbol{\theta}}_R = \left(X^{\mathrm{T}} X + \lambda I \right)^{-1} X^{\mathrm{T}} \boldsymbol{y}.$$

In contrast, this is not the case for LASSO, and its solution requires iterative techniques. It is straightforward to see that LASSO can be formulated as a standard convex quadratic problem with linear inequalities. Indeed, we can rewrite (9.6) as

$$\begin{aligned} \min_{\{\theta_i, u_i\}_{i=1}^{l}} \quad & (\boldsymbol{y} - X\boldsymbol{\theta})^{\mathrm{T}} (\boldsymbol{y} - X\boldsymbol{\theta}) + \lambda \sum_{i=1}^{l} u_i \\ \text{s.t.} \quad & \begin{cases} -u_i \leq \theta_i \leq u_i, \\ u_i \geq 0, \end{cases} \quad i = 1, 2, \ldots, l, \end{aligned}$$

which can be solved by any standard convex optimization method (e.g., [14, 100]). The reason that developing algorithms for the LASSO has been a hot research topic is due to the emphasis on obtaining *efficient* algorithms by exploiting the specific nature of this task, especially for cases where l is very large, as is often the case in practice.

In order to get better insight into the nature of the solution that is obtained by LASSO, let us assume that the regressors are mutually orthogonal and of unit norm, hence $X^{\mathrm{T}}X = I$. Orthogonality of the input matrix helps to decouple the coordinates and results to l one-dimensional problems that can be solved analytically. For this case, the LS estimate becomes

$$\hat{\boldsymbol{\theta}}_{\mathrm{LS}} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}\boldsymbol{y} = X^{\mathrm{T}}\boldsymbol{y},$$

and the ridge regression gives

$$\hat{\boldsymbol{\theta}}_R = \frac{1}{1+\lambda}\hat{\boldsymbol{\theta}}_{\mathrm{LS}}, \tag{9.9}$$

that is, every component of the LS estimate is simply shrunk by the *same* factor, $\frac{1}{1+\lambda}$.

In the case of the ℓ_1 regularization, the minimized Lagrangian function is no more differentiable, due to the presence of the absolute values in the ℓ_1 norm. So, in this case, we have to consider the notion of the subdifferential. It is known (Chapter 8) that if the zero vector belongs to the subdifferential set of a convex function at a point, this means that this point corresponds to a minimum of the function. Taking the subdifferential of the Lagrangian defined in (9.6) and recalling that the subdifferential set of a differentiable function includes as its *single* element the respective gradient, the resulting from the ℓ_1 regularized task estimate, $\hat{\boldsymbol{\theta}}_1$, must satisfy

$$\mathbf{0} \in -2X^{\mathrm{T}}\boldsymbol{y} + 2X^{\mathrm{T}}X\boldsymbol{\theta} + \lambda\partial\left\|\boldsymbol{\theta}\right\|_1,$$

where ∂ stands for the subdifferential set (Chapter 8). If X has orthonormal columns, the previous equation can be written component-wise as follows:

$$0 \in -\hat{\theta}_{\mathrm{LS},i} + \hat{\theta}_{1,i} + \frac{\lambda}{2}\partial\left|\hat{\theta}_{1,i}\right|, \quad \forall i, \tag{9.10}$$

where the subdifferential of the function $|\cdot|$, derived in Example 8.4 (Chapter 8), is given as

$$\partial|\theta| = \begin{cases} \{1\}, & \text{if } \theta > 0, \\ \{-1\}, & \text{if } \theta < 0, \\ [-1, 1], & \text{if } \theta = 0. \end{cases}$$

Thus, we can now write for each component of the LASSO optimal estimate

$$\hat{\theta}_{1,i} = \begin{cases} \hat{\theta}_{\mathrm{LS},i} - \dfrac{\lambda}{2}, & \text{if } \hat{\theta}_{1,i} > 0, \quad (9.11) \\ \hat{\theta}_{\mathrm{LS},i} + \dfrac{\lambda}{2}, & \text{if } \hat{\theta}_{1,i} < 0. \quad (9.12) \end{cases}$$

Notice that (9.11) can only be true if $\hat{\theta}_{\mathrm{LS},i} > \frac{\lambda}{2}$, and (9.12) only if $\hat{\theta}_{\mathrm{LS},i} < -\frac{\lambda}{2}$. Moreover, in the case where $\hat{\theta}_{1,i} = 0$, then (9.10) and the subdifferential of $|\cdot|$ suggest that necessarily $\left|\hat{\theta}_{\mathrm{LS},i}\right| \leq \frac{\lambda}{2}$. Concluding, we can write in a more compact way that

$$\hat{\theta}_{1,i} = \text{sgn}(\hat{\theta}_{\text{LS},i}) \left(\left| \hat{\theta}_{\text{LS},i} \right| - \frac{\lambda}{2} \right)_+ : \quad \text{Soft Thresholding Operation,} \tag{9.13}$$

where $(\cdot)_+$ denotes the "positive part" of the respective argument; it is equal to the argument if this is nonnegative, and zero otherwise. This is very interesting indeed. In contrast to the ridge regression that shrinks all coordinates of the unregularized LS solution by the same factor, LASSO forces all coordinates, whose absolute value is less than or equal to $\lambda/2$, to zero, and the rest of the coordinates are reduced, in absolute value, by the same amount $\lambda/2$. This is known as *soft thresholding*, to distinguish it from the *hard thresholding* operation; the latter is defined as $\theta \cdot \chi_{(0,\infty)} \left(|\theta| - \frac{\lambda}{2} \right)$, $\theta \in \mathbb{R}$, where $\chi_{(0,\infty)}(\cdot)$ stands for the characteristic function with respect to the set $(0, \infty)$. Figure 9.3 shows the graphs illustrating the effect that the ridge regression, LASSO, and hard thresholding have on the unregularized LS solution, as a function of its value (horizontal axis). Note that our discussion here, simplified via the orthonormal input matrix case, has quantified what we said before about the tendency of the ℓ_1 norm to push small values to become *exactly zero*. This will be further strengthened, via a more rigorous mathematical formulation, in Section 9.5.

Example 9.1. Assume that the unregularized LS solution, for a given regression task, $\boldsymbol{y} = X\boldsymbol{\theta} + \boldsymbol{\eta}$, is given by

$$\hat{\boldsymbol{\theta}}_{\text{LS}} = [0.2, -0.7, 0.8, -0.1, 1.0]^{\text{T}}.$$

Derive the solutions for the corresponding ridge regression and ℓ_1 norm regularization tasks. Assume that the input matrix X has orthonormal columns and that the regularization parameter is $\lambda = 1$. Also, what is the result of hard thresholding the vector $\hat{\boldsymbol{\theta}}_{\text{LS}}$ with threshold equal to 0.5?

We know that the corresponding solution for the ridge regression is

$$\hat{\boldsymbol{\theta}}_R = \frac{1}{1+\lambda} \hat{\boldsymbol{\theta}}_{\text{LS}} = [0.1, -0.35, 0.4, -0.05, 0.5]^{\text{T}}.$$

The solution for the ℓ_1 norm regularization is given by soft thresholding, with threshold equal to $\lambda/2 = 0.5$, hence the corresponding vector is

$$\hat{\boldsymbol{\theta}}_1 = [0, -0.2, 0.3, 0, 0.5]^{\text{T}}.$$

The result of the hard thresholding operation is the vector $[0, -0.7, 0.8, 0, 1.0]^{\text{T}}$.

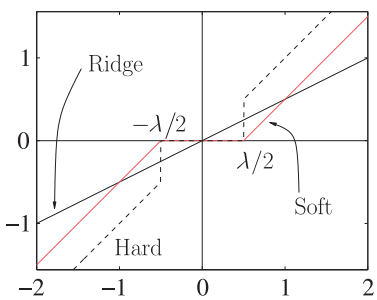

FIGURE 9.3

Output-input curves for the hard thresholding, soft thresholding operators together with the linear operator associated with the ridge regression, for the same value of $\lambda = 1$.

Remarks 9.1.

- The hard and soft thresholding rules are only two possibilities out of a larger number of alternatives. Note that the hard thresholding operation is defined via a discontinuous function, and this makes this rule unstable in the sense of being very sensitive to small changes of the input. Moreover, this shrinking rule tends to exhibit large variance in the resulting estimates. The soft thresholding rule is a continuous function, but, as readily seen from the graph in Figure 9.3, it introduces bias even for the large values of the input argument. In order to ameliorate such shortcomings, a number of alternative thresholding operators have been introduced and studied both theoretically and experimentally. Although these are not within the mainstream of our interest, we provide two popular examples for the sake of completeness—the *smoothly clipped absolute deviation* (SCAD) thresholding rule:

$$\hat{\theta}_{\text{SCAD}} = \begin{cases} \text{sgn}(\theta)\,(|\theta| - \lambda_{\text{SCAD}})_+, & |\theta| \leq 2\lambda_{\text{SCAD}}, \\ \dfrac{(\alpha-1)\theta - \alpha\lambda_{\text{SCAD}}\,\text{sgn}(\theta)}{\alpha - 2}, & 2\lambda_{\text{SCAD}} < |\theta| \leq \alpha\lambda_{\text{SCAD}}, \\ \theta, & |\theta| > \alpha\lambda_{\text{SCAD}}, \end{cases}$$

and the *nonnegative garrote* thresholding rule:

$$\hat{\theta}_{\text{garr}} = \begin{cases} 0, & |\theta| \leq \lambda_{\text{garr}}, \\ \theta - \dfrac{\lambda_{\text{garr}}^2}{\theta}, & |\theta| > \lambda_{\text{garr}}. \end{cases}$$

Figure 9.4 shows the respective graphs. Observe that, in both cases, an effort has been made to remove the discontinuity (associated with the hard thresholding) and to remove/reduce the bias for large values of the input argument. The parameter $\alpha > 2$ is a user-defined one. For a more detailed discussion on this topic, the interested reader can refer, for example, to [2].

9.4 SPARSE SIGNAL REPRESENTATION

In the previous section, we brought into our discussion the need to take special care for zeros. Sparsity is an attribute that is met in a plethora of natural signals, because nature tends to be parsimonious. The notion of and need for parsimonious models was also discussed in Chapter 3, in the context of inverse problems in machine learning tasks. In this section, we will briefly present a number of application cases where the existence of zeros in a mathematical expansion is of paramount importance, hence, it justifies our search for and development of related analysis tools.

In Chapter 4, we discussed the task of echo cancellation. In a number of cases, the echo path, represented by a vector comprising the values of the impulse response samples, is a sparse one. This is the case, for example, in internet telephony and in acoustic and network environments (e.g., [3, 10, 73]). Figure 9.5 shows the impulse response of such an echo path. The impulse response of the echo path is of short duration; however, the delay with which it appears is not known. So, in order to model it, one has to use a long impulse response, yet only a relatively small number of the coefficients will be significant and the rest will be close to zero. Of course, one could ask, why not use an LMS or an RLS,

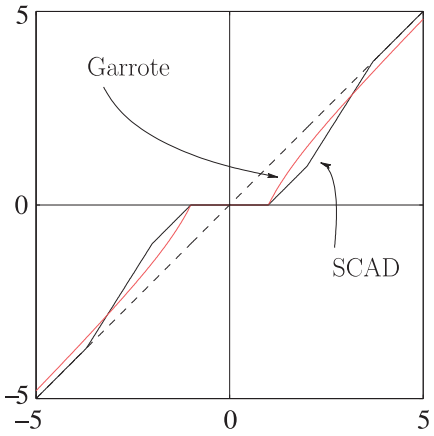

FIGURE 9.4

Output-input graph for the SCAD and nonnegative garrote rules with parameters $\alpha = 3.7$, and $\lambda_{\text{SCAD}} = \lambda_{\text{garr}} = 1$. Observe that both rules smooth out the discontinuity associated with the hard thresholding rule. Notice, also, that the SCAD rule removes the bias associated with the soft thresholding rule for large values of the input variable. On the contrary, the garrote thresholding rule allows some bias for large input values, which diminishes as λ_{garr} gets smaller and smaller.

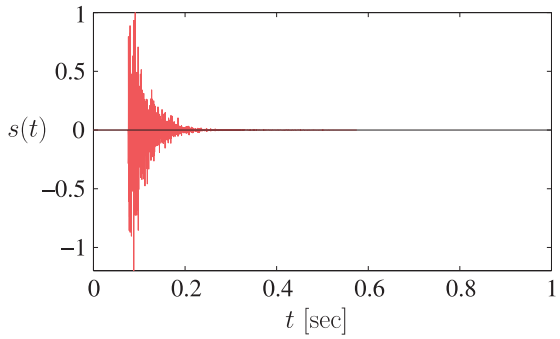

FIGURE 9.5

The impulse response function of an echo-path in a telephone network. Observe that although it is of relatively short duration, it is not a priori known where exactly in time it will occur.

and eventually the significant coefficients will be identified? The answer is that this turns out not to be the most efficient way to tackle such problems, because the convergence of the algorithm can be very slow. In contrast, if one embeds, somehow, into the problem the a priori information concerning the existence of (almost) zero coefficients, then the convergence speed can be significantly increased and also better error floors can be attained.

A similar situation occurs in wireless communication systems, which involve multipath channels. A typical application is in high-definition television (HDTV) systems that the involved communications

channels consist of a *few* nonnegligible coefficients, some of which may have quite large time delays with respect to the main signal (see, e.g., [4, 32, 52, 77]). If the information signal is transmitted at high symbol rates through such a dispersive channel, then the introduced intersymbol interference (ISI) has a span of several tens up to hundreds of symbol intervals. This in turn implies that quite long channel estimators are required at the receiver's end in order to reduce effectively the ISI component of the received signal, although only a small part of it has values substantially different to zero. The situation is even more demanding whenever the channel frequency response exhibits deep nulls. More recently, sparsity has been exploited in channel estimation for multicarrier systems, both for single antenna as well as for multiple-input-multiple-output (MIMO) systems [46, 47]. A thorough, in-depth treatment related to sparsity in multipath communication systems is provided in [5].

Another example, which might be more widely known, is that of signal compression. It turns out that if the signal modalities with which we communicate (e.g., speech) and also sense the world (e.g., images, audio) are transformed into a suitably chosen domain then they are sparsely represented; only a relatively small number of the signal components in this domain are large, and the rest are close to zero. As an example, Figure 9.6a shows an image and Figure 9.6b the plot of the magnitude of the obtained discrete cosine transform (DCT) components, which are computed by writing the corresponding image array as a vector in lexicographic order. Note that more than 95% of the total energy is contributed by only 5% of the largest components. This is at the heart of any compression technique. Only the large coefficients are chosen to be coded and the rest are considered to be zero. Hence, significant gains are obtained in memory/bandwidth requirements while storing/transmitting such signals, without much perceptual loss. Depending on the modality, different transforms are used. For example, in JPEG-2000, an image array, represented in terms of a vector that contains the intensity of the gray levels of the image

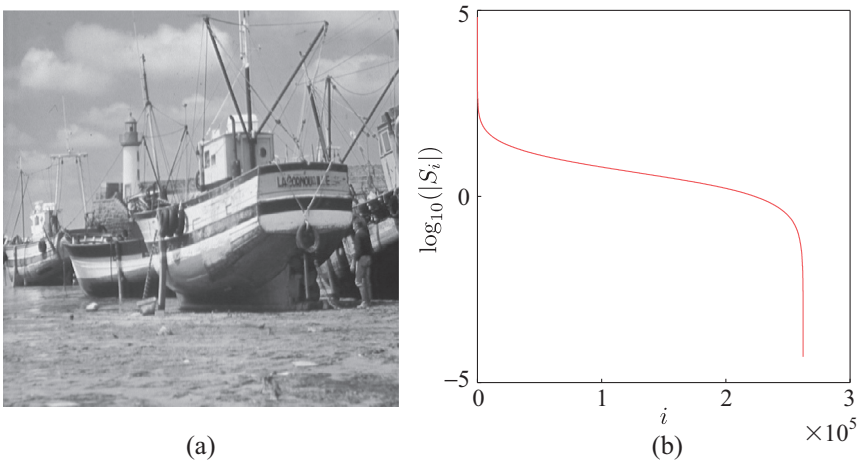

FIGURE 9.6

(a) A 512 × 512 pixel image and (b) the magnitude of its DCT components in descending order and logarithmic scale. Note that more than 95% of the total energy is contributed by only 5% of the largest components.

pixels, is transformed via the discrete wavelet transform (DWT) and results in a transformed vector that comprises only a few large components.

Let

$$\boldsymbol{S} = \Phi^H \boldsymbol{s},\ \boldsymbol{s}, \boldsymbol{S} \in \mathbb{C}^l, \tag{9.14}$$

where $\boldsymbol{s}$ is the vector of the "raw" signal samples, $\boldsymbol{S}$ is the (complex-valued) vector of the transformed ones, and Φ is the $l \times l$ transformation matrix. Often, this is an orthonormal/unitary matrix, $\Phi^H \Phi = \boldsymbol{I}$. Basically, a transform is nothing more than a projection of a vector on a new set of coordinate axes, which comprise the columns of the transformation matrix Φ. Celebrated examples of such transforms are the wavelet, the discrete Fourier (DFT), and the discrete cosine (DCT) transforms (e.g., [86]). In such cases, where the transformation matrix is orthonormal, one can write that

$$\boldsymbol{s} = \Psi \boldsymbol{S}, \tag{9.15}$$

where $\Psi = \Phi$. Equation (9.14) is known as the *analysis* and (9.15) as the *synthesis* equation.

Compression via such transforms exploits the fact that many signals in nature, which are rich in context, can be *compactly* represented in an appropriately chosen basis, depending on the modality of the signal. Very often, the construction of such bases tries to "imitate" the sensory systems that the human brain has developed in order to sense these signals; and we know that nature (in contrast to modern humans) does not like to waste resources. A standard compression task comprises the following stages: (a) Obtain the l components of $\boldsymbol{S}$ via the analysis step (9.14); (b) keep the k most significant of them; (c) code these values, as well as their respective locations in the transformed vector $\boldsymbol{S}$; and (d) obtain the (approximate) original signal $\boldsymbol{s}$ when needed (after storage or transmission), via the synthesis Eq. (9.15), where in place of $\boldsymbol{S}$ only its k most significant components are used, which are the ones that were coded, while the rest are set equal to zero. However, there is something unorthodox in this process of compression as it has been practiced until very recently. One processes (transforms) large signal vectors of l coordinates, where l in practice can be quite large, and then uses only a small percentage of the transformed coefficients, while the rest are simply ignored. Moreover, one has to store/transmit the location of the respective large coefficients that are finally coded. A natural question that is raised is the following: Because $\boldsymbol{S}$ in the synthesis equation is (approximately) sparse, can one compute it via an alternative path than the analysis equation in (9.14)? The issue here is to investigate whether one could use a more informative way of sampling the available raw data so that less than l samples/observations are sufficient to recover all the necessary information. The ideal case would be to recover it via a set of k such samples, because this is the number of the significant free parameters. On the other hand, if this sounds a bit extreme, can one obtain N ($k < N \ll l$) such signal-related measurements, from which s can eventually be retrieved? It turns out that such an approach is possible and it leads to the solution of an *underdetermined* system of linear equations, under the constraint that the unknown target vector is a sparse one.

The importance of such techniques becomes even more apparent when, instead of an orthonormal basis, as discussed before, a more general type of expansion is adopted, in terms of what is known as *overcomplete dictionaries*. A dictionary [65] is a collection of parameterized waveforms, which are discrete-time signal samples, represented as vectors $\boldsymbol{\psi}_i \in \mathbb{C}^l$, $i \in \mathcal{I}$, where $\mathcal{I}$ is an integer index set. For example, the columns of a DFT or a DWT matrix comprise a dictionary. These are two examples of what are known as *complete* dictionaries, which consist of l (orthonormal) vectors, that is, a number equal to the length of the signal vector. However, in many cases in practice, using such dictionaries is

very restrictive. Let us take, for example, a segment of audio signal, from a news media or a video, that needs to be processed. This consists, in general, of different types of signals, namely speech, music, and environmental sounds. For each type of these signals, different signal vectors may be more appropriate in the expansion for the analysis. For example, music signals are characterized by a strong harmonic content and the use of sinusoids seems to be best for compression, while for speech signals a Gabor type signal expansion (sinusoids of various frequencies weighted by sufficiently narrow pulses at different locations in time [31, 86]), may be a better choice. The same applies when one deals with an image. Different parts of an image, such as parts that are smooth or contain sharp edges, may demand a different expansion vector set for obtaining the best overall performance. The more recent tendency, in order to satisfy such needs, is to use *overcomplete* dictionaries. Such dictionaries can be obtained, for example, by concatenating different dictionaries together, for example, a DFT and a DWT matrix to result in a combined $l \times 2l$ transformation matrix. Alternatively, a dictionary can be "trained" in order to effectively represent a set of available signal exemplars, a task that is often referred to as dictionary learning [75, 78, 89, 99]. While using such overcomplete dictionaries, the synthesis equation takes the form

$$\boldsymbol{s} = \sum_{i \in \mathcal{I}} \theta_i \boldsymbol{\psi}_i. \tag{9.16}$$

Note that, now, the analysis is an ill-posed problem, because the elements $\{\boldsymbol{\psi}_i\}_{i \in \mathcal{I}}$ (usually called *atoms*) of the dictionary are not linearly independent, and there is not a unique set of coefficients $\{\theta_i\}_{i \in \mathcal{I}}$ that generates $\boldsymbol{s}$. Moreover, we expect most of these coefficients to be (nearly) zero. Note that, in such cases, the cardinality of $\mathcal{I}$ is larger than l. This necessarily leads to underdetermined systems of equations with infinite many solutions. The question that is now raised is whether we can exploit the fact that most of these coefficients are known to be zero, in order to come up with a unique solution. If yes, under which conditions is such a solution possible? We will return to the task of learning dictionaries in Chapter 19.

Besides the previous examples, there are a number of cases where an underdetermined system of equations is the result of our inability to obtain a sufficiently large number of measurements, due to physical and technical constraints. This is the case in MRI imaging, which will be presented in more detail later in the chapter.

9.5 IN SEARCH OF THE SPARSEST SOLUTION

Inspired by the discussion in the previous section, we now turn our attention to the task of solving underdetermined systems of equations by imposing the sparsity constraint on the solution. We will develop the theoretical setup in the context of regression and we will adhere to the notation that has been adopted for this task. Moreover, we will focus on the real-valued data case in order to simplify the presentation. The theory can be readily extended to the more general complex-valued data case (see, e.g., [64, 98]). We assume that we are given a set of observations/measurments, $\boldsymbol{y} := [y_1, y_2, \ldots, y_N]^{\mathrm{T}} \in \mathbb{R}^N$, according to the linear model

$$\boldsymbol{y} = X\boldsymbol{\theta}, \quad \boldsymbol{y} \in \mathbb{R}^N, \ \boldsymbol{\theta} \in \mathbb{R}^l, \ l > N, \tag{9.17}$$

where X is the $N \times l$ input matrix, which is assumed to be of full row rank, that is, $\mathrm{rank}(X) = N$. Our starting point is the noiseless case. The linear system of equations in (9.17) is an underdetermined one

and accepts an infinite number of solutions. The set of possible solutions lies in the intersection of the N hyperplanes[2] in the l-dimensional space,

$$\left\{\boldsymbol{\theta} \in \mathbb{R}^l : y_n = \boldsymbol{x}_n^{\mathrm{T}}\boldsymbol{\theta}\right\}, \quad n = 1, 2, \ldots, N.$$

We know from geometry that the intersection of N nonparallel hyperplanes (which in our case is guaranteed by the fact that X has been assumed to be full row rank, hence $\boldsymbol{x}_n$, $n = 1, 2, \ldots, N$, are linearly independent) is a plane of dimensionality $l - N$ (e.g., the intersection of two (nonparallel) (hyper)planes in the three-dimensional space is a straight line, that is, a plane of dimensionality equal to one). In a more formal way, the set of all possible solutions, to be denoted as Θ, is an *affine* set. An affine set is the translation of a linear subspace by a constant vector. Let us pursue this a bit further, because we will need it later on.

Let the null space of X be the set null(X) (sometimes, denoted as $\mathcal{N}(X)$), defined as the linear subspace

$$\text{null}(X) = \left\{\boldsymbol{z} \in \mathbb{R}^l : X\boldsymbol{z} = \boldsymbol{0}\right\}.$$

Obviously, if $\boldsymbol{\theta}_0$ is a solution to (9.17), that is, $\boldsymbol{\theta}_0 \in \Theta$, then it is easy to verify that $\forall \boldsymbol{\theta} \in \Theta, X(\boldsymbol{\theta} - \boldsymbol{\theta}_0) = \boldsymbol{0}$, or $\boldsymbol{\theta} - \boldsymbol{\theta}_0 \in \text{null}(X)$. As a result,

$$\Theta = \boldsymbol{\theta}_0 + \text{null}(X),$$

and Θ is an affine set. We also know from linear algebra basics (and it is easy to show it; Problem 9.9), that the null space of a full row rank matrix, $N \times l$, $l > N$, is a subspace of dimensionality $l - N$. Figure 9.7 illustrates the case for one measurement sample in the two-dimensional space, $l = 2$ and

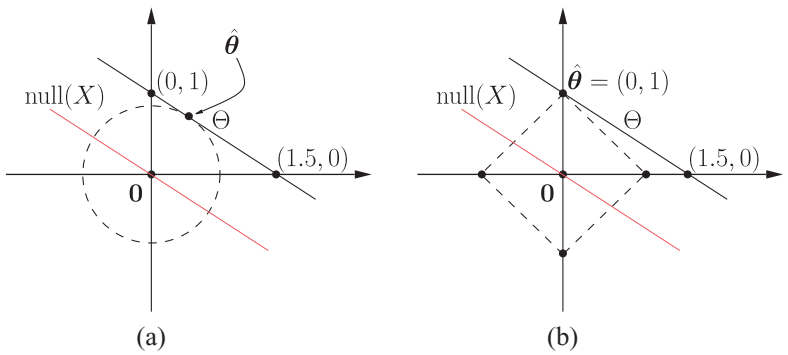

FIGURE 9.7

The set of solutions Θ is an affine set (gray line), which is a translation of the null(X) subspace (red line). (a) The ℓ_2 norm minimizer. The dotted circle corresponds to the smallest ℓ_2 ball that intersects the set Θ. As such, the intersection point, $\hat{\boldsymbol{\theta}}$, is the ℓ_2 norm minimizer of the task in (9.18). Notice that the vector $\hat{\boldsymbol{\theta}}$ contains no zero component. (b) The ℓ_1 norm minimizer. The dotted rhombus corresponds to the smallest ℓ_1 ball that intersects Θ. Hence, the intersection point, $\hat{\boldsymbol{\theta}}$, is the solution of the constrained ℓ_1 minimization task of (9.21). Notice that the obtained estimate $\hat{\boldsymbol{\theta}} = (0, 1)$ contains a zero.

[2] In $\mathbb{R}^l$, a hyperplane is of dimension $l - 1$. A plane has dimension lower than $l - 1$.

$N = 1$. The set of solutions Θ is a straight line, which is the translation of the linear subspace crossing the origin (the null(X)). Therefore, if one wants to select a *single* point among all the points that lie in the affine set of solutions, Θ, then an extra constraint/a priori knowledge has to be imposed.

In the sequel, three such possibilities are examined.

The ℓ_2 norm minimizer

Our goal now becomes to pick a point in (the affine set) Θ that corresponds to the minimum ℓ_2 norm. This is equivalent to solving the following constrained task:

$$\begin{aligned} \min_{\boldsymbol{\theta}\in\mathbb{R}^l} \quad & \|\boldsymbol{\theta}\|_2^2 \\ \text{s.t.} \quad & \boldsymbol{x}_n^{\mathrm{T}}\boldsymbol{\theta} = y_n, \quad n = 1, 2, \ldots, N. \end{aligned} \tag{9.18}$$

We already know from Section 6.4 (and one can rederive it by employing Lagrange multipliers; Problem 9.10) that the previous optimization task accepts a *unique* solution given in closed form as

$$\hat{\boldsymbol{\theta}} = X^{\mathrm{T}}\left(XX^{\mathrm{T}}\right)^{-1}\boldsymbol{y}. \tag{9.19}$$

The geometric interpretation of this solution is provided in Figure 9.7a, for the case of $l = 2$ and $N = 1$. The radius of the Euclidean norm ball keeps increasing, until it touches the plane that contains the solutions. This point is the one with the minimum ℓ_2 norm or, equivalently, the point that lies closest to the origin. Equivalently, the point $\hat{\boldsymbol{\theta}}$ can be seen as the (metric) projection of $\mathbf{0}$ onto Θ.

Minimizing the ℓ_2 norm in order to solve a linear set of underdetermined equations has been used in various applications. The closest to us is in the context of determining the unknown coefficients in an expansion using an overcomplete dictionary of functions (vectors) [35]. A main drawback of this method is that it is not sparsity preserving. There is no guarantee that the solution in (9.19) will give zeros even if the true model vector $\boldsymbol{\theta}$ has zeros. Moreover, the method is *resolution limited* [26]. This means that even if there may be a sharp contribution of specific atoms in the dictionary, this is not portrayed in the obtained solution. This is a consequence of the fact that the information provided by XX^{T} is a global one, containing all atoms of the dictionary in an "averaging" fashion, and the final result tends to smooth out the individual contributions, especially when the dictionary is overcomplete.

The ℓ_0 norm minimizer

Now we turn our attention to the ℓ_0 norm (once more, it is pointed out that this is an abuse of the definition of the norm, as stated before), and we make sparsity our new flag under which a solution will be obtained. The task now becomes

$$\begin{aligned} \min_{\boldsymbol{\theta}\in\mathbb{R}^l} \quad & \|\boldsymbol{\theta}\|_0 \\ \text{s.t.} \quad & \boldsymbol{x}_n^{\mathrm{T}}\boldsymbol{\theta} = y_n, \; n = 1, 2, \ldots, N, \end{aligned} \tag{9.20}$$

that is, from all the points that lie on the plane of all possible solutions find the *sparsest* one, that is, the one with the least number of nonzero elements. As a matter of fact, such an approach is within the spirit of *Occam's razor* rule—it corresponds to the smallest number of parameters that can explain the obtained observations. The points that are now raised are:

- Is a solution to this problem unique, and under which conditions?
- Can a solution be obtained with low enough complexity in realistic time?

We postpone the answer to the first question until later. As for the second one, the news is not good. Minimizing the ℓ_0 norm under a set of linear constraints is a task of combinatorial nature, and as a matter of fact, the problem is, in general, NP-hard [72]. The way to approach the problem is to consider all possible combinations of zeros in $\boldsymbol{\theta}$, removing the respective columns of X in (9.17), and check whether the system of equations is satisfied; keep as solutions the ones with the smallest number of nonzero elements. Such a searching technique exhibits complexity of an exponential dependence on l. Figure 9.7a illustrates the two points ((1.5, 0) and (0, 1)) that comprise the solution set of minimizing the ℓ_0 norm for the single measurement (constraint) case.

The ℓ_1 norm minimizer

The current task is now given by

$$\begin{aligned} \min_{\boldsymbol{\theta}\in\mathbb{R}^l} \quad & \|\boldsymbol{\theta}\|_1 \\ \text{s.t.} \quad & \boldsymbol{x}_n^{\mathrm{T}}\boldsymbol{\theta} = y_n,\ n = 1, 2, \ldots, N. \end{aligned} \tag{9.21}$$

Figure 9.7b illustrates the geometry. The ℓ_1 ball is increased until it touches the affine set of the possible solutions. For this specific geometry, the solution is the point (0, 1), which is a sparse solution. In our discussion in Section 9.2, we saw that the ℓ_1 norm is the one, out of all ℓ_p, $p \geq 1$ norms, that bears some similarity with the sparsity-promoting (nonconvex) ℓ_p, $p < 1$ "norms." Also, we have commented that the ℓ_1 norm encourages zeros when the respective values are small. In the sequel, we will state one lemma that establishes this zero-favoring property in a more formal way. The ℓ_1 norm minimizer is also known as *Basis Pursuit* and it was suggested for decomposing a vector signal in terms of the atoms of an overcomplete dictionary [26].

The ℓ_1 minimizer can be brought into the standard linear programming (LP) form and then can be solved by recalling any related method; the simplex method and the more recent interior point methods are two possibilities (see, e.g., [14, 33]). Indeed, consider the LP task

$$\begin{aligned} \min_{\boldsymbol{x}} \quad & \boldsymbol{c}^{\mathrm{T}}\boldsymbol{x}, \\ \text{s.t.} \quad & A\boldsymbol{x} = \boldsymbol{b}, \\ & \boldsymbol{x} \geq \boldsymbol{0}. \end{aligned}$$

To verify that our ℓ_1 minimizer can be cast in the previous form, notice first that any l-dimensional vector $\boldsymbol{\theta}$ can be decomposed as

$$\boldsymbol{\theta} = \boldsymbol{u} - \boldsymbol{v}, \quad \boldsymbol{u} \geq \boldsymbol{0}, \boldsymbol{v} \geq \boldsymbol{0}.$$

Indeed, this holds true if, for example,

$$\boldsymbol{u} := \boldsymbol{\theta}_+, \quad \boldsymbol{v} := (-\boldsymbol{\theta})_+,$$

where $\boldsymbol{x}_+$ stands for the vector obtained after keeping the positive components of $\boldsymbol{x}$ and setting the rest equal to zero. Moreover, notice that

$$\|\boldsymbol{\theta}\|_1 = [1, 1, \ldots, 1]\begin{bmatrix} \boldsymbol{\theta}_+ \\ (-\boldsymbol{\theta})_+ \end{bmatrix} = [1, 1, \ldots, 1]\begin{bmatrix} \boldsymbol{u} \\ \boldsymbol{v} \end{bmatrix}.$$

Hence, our ℓ_1 minimization task can be recast in the LP form, if

$$\begin{aligned} \boldsymbol{c} &:= [1, 1, \ldots, 1]^{\mathrm{T}}, \quad \boldsymbol{x} := [\boldsymbol{u}^{\mathrm{T}}, \boldsymbol{v}^{\mathrm{T}}]^{\mathrm{T}}, \\ A &:= [X, -X], \qquad\quad \boldsymbol{b} := \boldsymbol{y}. \end{aligned}$$

Characterization of the ℓ_1 norm minimizer

Lemma 9.1. *An element $\boldsymbol{\theta}$ in the affine set, Θ, of the solutions of the underdetermined linear system (9.17), has minimal ℓ_1 norm if and only if the following condition is satisfied:*

$$\left| \sum_{i:\ \theta_i \neq 0} \text{sgn}(\theta_i) z_i \right| \leq \sum_{i:\ \theta_i = 0} |z_i|, \quad \forall \boldsymbol{z} \in \text{null}(X). \tag{9.22}$$

Moreover, the ℓ_1 minimizer is unique if and only if the inequality in (9.22) is a strict one for all $\boldsymbol{z} \neq \mathbf{0}$ (see, e.g., [74] and Problem 9.11).

Remarks 9.2.

- The previous lemma has a very interesting and important consequence. If $\hat{\boldsymbol{\theta}}$ is the *unique* minimizer of (9.21), then

$$\text{card}\{i : \hat{\theta}_i = 0\} \geq \dim(\text{null}(X)), \tag{9.23}$$

where card$\{\cdot\}$ denotes the cardinality of a set. In words, the number of zero coordinates of the unique minimizer cannot be smaller than the dimension of the null space of X. Indeed, if this is not the case, then the unique minimizer could have less zeros than the dimensionality of null(X). This means that we can always find a $\boldsymbol{z} \in \text{null}(X)$, which has zeros in the same locations where the coordinates of the unique minimizer are zero, and at the same time it is not identically zero, that is, $\boldsymbol{z} \neq \mathbf{0}$ (Problem 9.12). However, this would violate (9.22), which in the case of uniqueness holds as a strict inequality.

Definition 9.1. A vector $\boldsymbol{\theta}$ is called k-sparse if it has *at most* k nonzero components.

Remarks 9.3.

- If the minimizer of (9.21) is *unique*, then it is a k-sparse vector with

$$k \leq N.$$

This is a direct consequence of the Remark 9.2, and the fact that for the matrix X,

$$\dim(\text{null}(X)) = l - \text{rank}(X) = l - N.$$

Hence, the number of the nonzero elements of the unique minimizer must be at most equal to N.

If one resorts to geometry, all the previously stated results become crystal clear.

Geometric interpretation

Assume that our target solution resides in the three-dimensional space and that we are given one measurement

$$y_1 = \boldsymbol{x}_1^{\mathrm{T}} \boldsymbol{\theta} = x_{11}\theta_1 + x_{12}\theta_2 + x_{13}\theta_3.$$

Then the solution lies in the two-dimensional (hyper)plane, which is described by the previous equation. To get the minimal ℓ_1 solution we keep increasing the size of the ℓ_1 ball[3] (the set of all points that have

[3] Observe that in the three-dimensional space the ℓ_1 ball looks like a diamond.

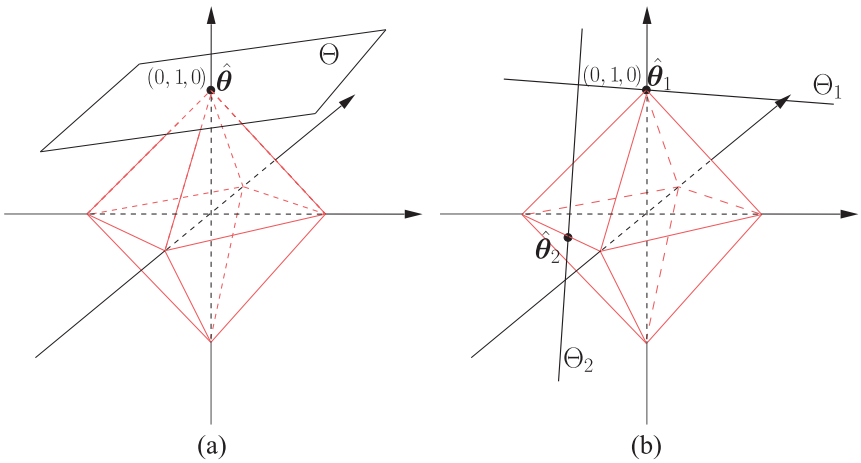

FIGURE 9.8

(a) The ℓ_1 ball intersecting with a plane. The only possible scenario, for the existence of a unique common intersecting point of the ℓ_1 ball with a plane in the Euclidean $\mathbb{R}^3$ space, is for the point to be located at one of the vertices of the ℓ_1 ball, that is, to be a 1-sparse vector. (b) The ℓ_1 ball intersecting with lines. In this case, the sparsity level of the unique intersecting point is relaxed; it could be a 1- or a 2-sparse vector.

equal ℓ_1 norm) until it touches this plane. The only way that these two geometric objects have a single point in common (unique solution) is when they meet at a vertex of the diamond. This is shown in Figure 9.8a. In other words, the resulting solution is 1-sparse, having two of its components equal to zero. This complies with the finding stated in Remark 9.3, because now $N = 1$. For any other orientation of the plane, this will either cut across the ℓ_1 ball or will share with the diamond an edge or a side. In both cases, there will be infinite solutions.

Let us now assume that we are given an extra measurement,

$$y_2 = x_{21}\theta_1 + x_{22}\theta_2 + x_{23}\theta_3.$$

The solution now lies in the intersection of the two previous planes, which is a straight line. However, now, we have more alternatives for a unique solution. A line, for example, Θ_1, can either touch the ℓ_1 ball at a vertex (1-sparse solution) or, as shown in Figure 9.8b, it can touch the ℓ_1 ball at one of its edges, for example, Θ_2. The latter case corresponds to a solution that lies on a two-dimensional subspace, hence it will be a 2-sparse vector. This also complies with the findings stated in Remark 9.3, because in this case we have $N = 2$, $l = 3$, and the sparsity level for a unique solution can be either 1 or 2.

Note that uniqueness is associated with the particular geometry and orientation of the affine set, which is the set of all possible solutions of the underdetermined system of equations. For the case of the squared ℓ_2 norm, the solution is always unique. This is a consequence of the (hyper)spherical shape formed by the Euclidean norm. From a mathematical point of view, the squared ℓ_2 norm is a strictly convex function. This is not the case for the ℓ_1 norm, which is convex, albeit not a strictly convex function (Problem 9.13).

Example 9.2. Consider a sparse vector parameter $[0, 1]^{\mathrm{T}}$, which we assume to be unknown. We will use one measurement to *sense* it. Based on this single measurement, we will use the ℓ_1 minimizer of (9.21) to recover its true value. Let us see what happens.

We will consider three different values of the "sensing" (input) vector $\boldsymbol{x}$ in order to obtain the measurement $y = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\theta}$: a) $\boldsymbol{x} = [\frac{1}{2}, 1]^{\mathrm{T}}$, b) $\boldsymbol{x} = [1, 1]^{\mathrm{T}}$, and c) $\boldsymbol{x} = [2, 1]^{\mathrm{T}}$. The resulting measurement, after sensing $\boldsymbol{\theta}$ by $\boldsymbol{x}$, is $y = 1$ for all three previous cases.

Case a: The solution will lie on the straight line

$$\Theta = \left\{ [\theta_1, \theta_2]^{\mathrm{T}} \in \mathbb{R}^2 : \frac{1}{2}\theta_1 + \theta_2 = 1 \right\},$$

which is shown in Figure 9.9a. For this setting, expanding the ℓ_1 ball, this will touch the straight line (our solution's affine set) at the vertex $[0, 1]^{\mathrm{T}}$. This is a unique solution, hence it is sparse, and it coincides with the true value.

Case b: The solution lies on the straight line

$$\Theta = \left\{ [\theta_1, \theta_2]^{\mathrm{T}} \in \mathbb{R}^2 : \theta_1 + \theta_2 = 1 \right\},$$

which is shown in Figure 9.9b. For this setup, there is an infinite number of solutions, including two sparse ones.

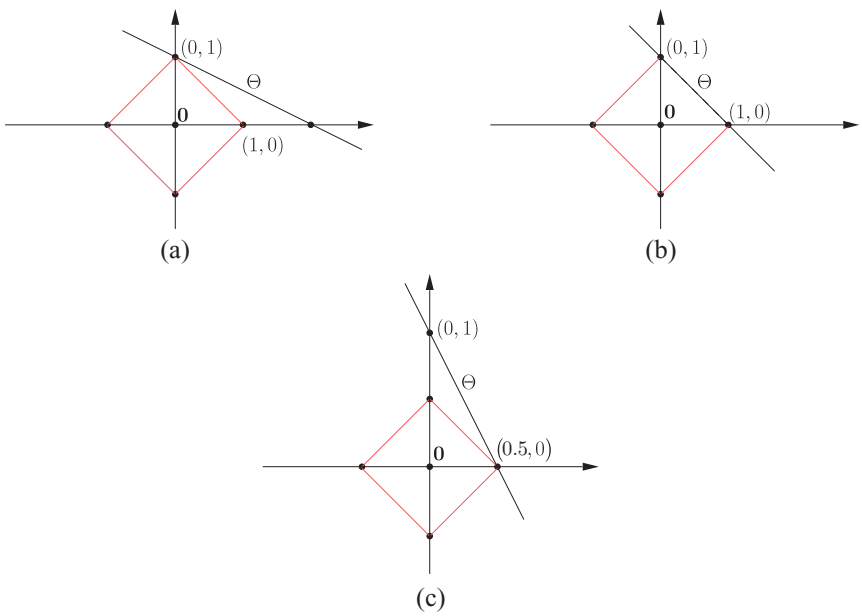

FIGURE 9.9

(a) Sensing with $\boldsymbol{x} = [\frac{1}{2}, 1]^{\mathrm{T}}$, (b) sensing with $\boldsymbol{x} = [1, 1]^{\mathrm{T}}$, (c) sensing with $\boldsymbol{x} = [2, 1]^{\mathrm{T}}$. The choice of the sensing vector $\boldsymbol{x}$ is crucial to unveiling the true sparse solution (0,1). Only the sensing vector $\boldsymbol{x} = [\frac{1}{2}, 1]^{\mathrm{T}}$ identifies uniquely the desired (0,1).

Case c: The affine set of solutions is described by

$$\Theta = \left\{ [\theta_1, \theta_2]^{\mathrm{T}} \in \mathbb{R}^2 : 2\theta_1 + \theta_2 = 1 \right\},$$

which is sketched in Figure 9.9c. The solution in this case is sparse, but it is not the correct one.

This example is quite informative. *If we sense (measure) our unknown parameter vector with appropriate sensing (input) data, the use of the ℓ_1 norm can unveil the true value of the parameter vector, even if the system of equations is underdetermined, provided that the true parameter is sparse*. This becomes our new goal; to investigate whether what we have just said can be generalized, and under which conditions it holds true. In such a case, the choice of the regressors (which we called sensing vectors) and hence the input matrix (which we will refer to more and more frequently as the sensing matrix) acquire extra significance. It is not enough for the designer to care only for the rank of the matrix, that is, the linear independence of the sensing vectors. One has to make sure that the corresponding affine set of the solutions has such an orientation so that the touch with the ℓ_1 ball (as this increases from zero to meet this plane) is a "gentle" one; that is, they meet at a single point, and more important at the correct one, which is the point that represents the true value of the sparse parameter vector.

Remarks 9.4.

- Often in practice, the columns of the input matrix, X, are normalized to unit ℓ_2 norm. Although ℓ_0 norm is insensitive to the values of the nonzero components of $\boldsymbol{\theta}$, this is not the case with the ℓ_1 and ℓ_2 norms. Hence, while trying to minimize the respective norms and at the same time fulfill the constraints, components that correspond to columns of X with high energy (norm) are favored over the rest. Hence, the latter become more popular candidates to be pushed to zero. In order to avoid such situations, the columns of X are normalized to unity by dividing each element of the column vector by the respective (Euclidean) norm.

9.6 UNIQUENESS OF THE ℓ_0 MINIMIZER

Our first goal is to derive *sufficient* conditions that guarantee uniqueness of the ℓ_0 minimizer, which has been defined in Section 9.5.

Definition 9.2. The *spark* of a full row rank $N \times l$ ($l \geq N$) matrix, X, denoted as spark(X), is the *smallest* number of its linearly dependent columns.

According to the previous definition, *any* $m <$ spark(X) columns of X is, necessarily, *linearly independent*. The spark of a square, $N \times N$, full rank matrix is equal to $N + 1$.

Remarks 9.5.

- In contrast to the rank of a matrix, which can be easily determined, its spark can only be obtained by resorting to a combinatorial search over all possible combinations of the columns of the respective matrix (see, e.g., [15, 37]). The notion of the spark was used in the context of sparse representation, under the name *Uniqueness Representation Property*, in [53]. The name "spark" was coined in [37]. An interesting discussion relating this matrix index with indices used in other disciplines, is given in [15].
- Note that the notion of "spark" is related to the notion of the minimum Hamming weight of a linear code in coding theory (e.g., [60]).

Example 9.3. Consider the following matrix

$$X = \begin{bmatrix} 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix}.$$

The matrix has rank equal to 4 and spark equal to 3. Indeed, any pair of columns is linearly independent. On the other hand, the first, the second, and the fifth columns are linearly dependent. The same is also true for the combination of the second, third, and sixth columns. Also, the maximum number of linearly independent columns is four.

Lemma 9.2. *If* null(X) *is the null space of* X*, then*

$$\|\boldsymbol{\theta}\|_0 \geq \text{spark}(X), \quad \forall \boldsymbol{\theta} \in \text{null}(X), \boldsymbol{\theta} \neq \mathbf{0}.$$

Proof: To derive a contradiction, assume that there exists a $\boldsymbol{\theta} \in \text{null}(X)$, $\boldsymbol{\theta} \neq \mathbf{0}$, such that $\|\boldsymbol{\theta}\|_0 < \text{spark}(X)$. Because by definition $X\boldsymbol{\theta} = \mathbf{0}$, there exists a number of $\|\boldsymbol{\theta}\|_0$ columns of X that are linearly dependent. However, this contradicts the minimality of spark(X), and the claim of Lemma 9.2 is established.

Lemma 9.3. *If a linear system of equations,* $X\boldsymbol{\theta} = \mathbf{y}$*, has a solution that satisfies*

$$\boxed{\|\boldsymbol{\theta}\|_0 < \frac{1}{2}\text{spark}(X),}$$

then this is the sparsest possible solution. In other words, this is, necessarily, the unique solution of the ℓ_0 *minimizer.*

Proof: Consider any other solution $\boldsymbol{h} \neq \boldsymbol{\theta}$. Then, $\boldsymbol{\theta} - \boldsymbol{h} \in \text{null}(X)$, that is,

$$X(\boldsymbol{\theta} - \boldsymbol{h}) = \mathbf{0}.$$

Thus, according to Lemma 9.2,

$$\text{spark}(X) \leq \|\boldsymbol{\theta} - \boldsymbol{h}\|_0 \leq \|\boldsymbol{\theta}\|_0 + \|\boldsymbol{h}\|_0 . \tag{9.24}$$

Observe that although the ℓ_0 "norm" is not a true norm, it can be readily verified by simple inspection and reasoning that the triangular property is satisfied. Indeed, by adding two vectors together, the resulting number of nonzero elements will always be at most equal to the total number of nonzero elements of the two vectors. Therefore, if $\|\boldsymbol{\theta}\|_0 < \frac{1}{2}\text{spark}(X)$, then (9.24) suggests that

$$\|\boldsymbol{h}\|_0 > \frac{1}{2}\text{spark}(X) > \|\boldsymbol{\theta}\|_0 .$$

Remarks 9.6.

- Lemma 9.3 is a very interesting result. We have a sufficient condition to check whether a solution is the unique optimal in a generally NP-hard problem. Of course, although this is nice from a theoretical point of view, it is not of much use by itself, because the related bound (the spark) can only be obtained after a combinatorial search. In the next section, we will see that we can relax the bound by involving another index in place of the spark, which can be easily computed.
- An obvious consequence of the previous lemma is that if the unknown parameter vector is a sparse one with k nonzero elements, then if matrix X is chosen in order to have spark$(X) > 2k$, the true

parameter vector is necessarily the sparsest one that satisfies the set of equations, and the (unique) solution to the ℓ_0 minimizer.

- In practice, the goal is to sense the unknown parameter vector by a matrix that has a spark as high as possible, so that the previously stated sufficiency condition covers a wide range of cases. For example, if the spark of the input matrix is equal to three, then one can check for optimal sparse solutions up to a sparsity level of $k = 1$. From the respective definition, it is easily seen that the values of the spark are in the range $1 < \text{spark}(X) \leq N + 1$.
- Constructing an $N \times l$ matrix X in a random manner, by generating i.i.d. entries, guarantees with high probability that $\text{spark}(X) = N + 1$; that is, any N columns of the matrix are linearly independent.

9.6.1 MUTUAL COHERENCE

Because the spark of a matrix is a number that is difficult to compute, our interest shifts to another index, which can be derived more easily and at the same time offers a useful bound on the spark. The *mutual coherence* of an $N \times l$ matrix X [65], denoted as $\mu(X)$, is defined as

$$\boxed{\mu(X) := \max_{1 \leq i < j \leq l} \frac{|\mathbf{x}_i^{\mathrm{T}} \mathbf{x}_j|}{\|\mathbf{x}_i\| \, \|\mathbf{x}_j\|} : \quad \text{Mutual Coherence,}} \tag{9.25}$$

where $\mathbf{x}_i$, $i = 1, 2, \ldots, l$, denote the columns of X (notice the difference in notation between a row $\boldsymbol{x}_i^{\mathrm{T}}$ and a column[4] $\mathbf{x}_i$ of the matrix X). This number reminds us of the correlation coefficient between two random variables. Mutual coherence is bounded as $0 \leq \mu(X) \leq 1$. For a square orthogonal matrix, X, $\mu(X) = 0$. For general matrices, with $l > N$, $\mu(X)$ satisfies

$$\sqrt{\frac{l - N}{N(l - 1)}} \leq \mu(X) \leq 1,$$

which is known as the *Welch bound* [97] (Problem 9.15). For large values of l, the lower bound becomes, approximately, $\mu(X) \geq \frac{1}{\sqrt{N}}$. Common sense reasoning guides us to construct input (sensing) matrices of mutual coherence as small as possible. Indeed, the purpose of the sensing matrix is to "measure" the components of the unknown vector and "store" this information in the measurement vector $\boldsymbol{y}$. Thus, this should be done in such a way that $\boldsymbol{y}$ retains as much information about the components of $\boldsymbol{\theta}$ as possible. This can be achieved if the columns of the sensing matrix, X, are as "independent" as possible. Indeed, $\boldsymbol{y}$ is the result of a combination of the columns of X, each one weighted by a different component of $\boldsymbol{\theta}$. Thus, if the columns are as "independent" as possible, then the information regarding each component of $\boldsymbol{\theta}$ is contributed by a different direction, making its recovery easier. This is easier understood if X is a square orthogonal matrix. In the more general case of a nonsquare matrix, the columns should be made as "orthogonal" as possible.

Example 9.4. Assume that X is an $N \times 2N$ matrix, formed by concatenating two orthonormal bases together,

$$X = [I, W],$$

[4] Not to be confused with the roman font used for random variables in previous chapters.

where I is the identity matrix, having as columns the vectors $\boldsymbol{e}_i$, $i = 1, 2, \ldots, N$, with elements equal to

$$\delta_{ir} = \begin{cases} 1, & \text{if } i = r, \\ 0, & \text{if } i \neq r, \end{cases}$$

for $r = 1, 2, \ldots, N$. The matrix W is the orthonormal DFT matrix, defined as

$$W = \frac{1}{\sqrt{N}} \begin{bmatrix} 1 & 1 & \ldots & 1 \\ 1 & W_N & \ldots & W_N^{N-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & W_N^{N-1} & \ldots & W_N^{(N-1)(N-1)} \end{bmatrix},$$

where

$$W_N := \exp\left(-j\frac{2\pi}{N}\right).$$

Such an overcomplete dictionary could be used to represent signal vectors in terms of the expansion in (9.16), which comprise the sum of sinusoids with very narrow, spiky-like pulses. The inner products between any two columns of I and between any two columns of W are zero, due to orthogonality. On the other hand, it is easy to see that the inner product between any column of I and any column of W has absolute value equal to $\frac{1}{\sqrt{N}}$. Hence, the mutual coherence of this matrix is $\mu(X) = \frac{1}{\sqrt{N}}$. Moreover, observe that the spark of this matrix is $\text{spark}(X) = N + 1$.

Lemma 9.4. *For any $N \times l$ matrix X, the following inequality holds:*

$$\text{spark}(X) \geq 1 + \frac{1}{\mu(X)}. \tag{9.26}$$

The proof is given in [37] and it is based on arguments that stem from matrix theory applied on the Gram matrix, $X^{\mathrm{T}}X$, of X (Problem 9.16). A "superficial" look at the previous bound is that for very small values of $\mu(X)$ the spark can be larger than $N + 1$! Looking at the proof, it is seen that in such cases the spark of the matrix attains its maximum value $N + 1$.

The result complies with common sense reasoning. The smaller the value of $\mu(X)$, the more independent the columns of X, hence the higher the value its spark is expected to be. Based on this lemma, we can now state the following theorem, first given in [37]. Combining Lemma 9.3 and (9.26), we come to the following important theorem.

Theorem 9.1. *If the linear system of equations in (9.17) has a solution that satisfies the condition*

$$\boxed{\|\boldsymbol{\theta}\|_0 < \frac{1}{2}\left(1 + \frac{1}{\mu(X)}\right),} \tag{9.27}$$

then this solution is the sparsest one.

Remarks 9.7.

- The bound in (9.27) is "psychologically" important. It relates an easily computed bound to check whether the solution to an NP-hard task is the optimal one. However, it is not a particularly good bound and it restricts the range of values in which it can be applied. As we saw in Example 9.4, while the maximum possible value of the spark of a matrix was equal to $N + 1$, the minimum

possible value of the mutual coherence was $\frac{1}{\sqrt{N}}$. Therefore, the bound based on the mutual coherence restricts the range of sparsity, that is, $\|\boldsymbol{\theta}\|_0$, where one can check optimality, to around $\frac{1}{2}\sqrt{N}$. Moreover, as the previously stated Welch bound suggests, this $\mathcal{O}(\frac{1}{\sqrt{N}})$ dependence of the mutual coherence seems to be a more general trend and not only the case for Example 9.4 (see, e.g., [36]). On the other hand, as we have already stated in the Remarks 9.6, one can construct random matrices with spark equal to $N+1$; hence, using the bound based on the spark, one could expand the range of sparse vectors up to $\frac{1}{2}N$.

9.7 EQUIVALENCE OF ℓ_0 AND ℓ_1 MINIMIZERS: SUFFICIENCY CONDITIONS

We have now come to the crucial point where we will establish the conditions that guarantee the equivalence between the ℓ_1 and the ℓ_0 minimizers. Hence, under such conditions, a problem that is in general an NP-hard one *can be solved via a tractable convex optimization task*. Under these conditions, the zero value encouraging nature of the ℓ_1 norm, which has already been discussed, obtains a much higher stature; it provides the sparsest solution.

9.7.1 CONDITION IMPLIED BY THE MUTUAL COHERENCE NUMBER

Theorem 9.2. *Let the underdetermined system of equations*

$$\boldsymbol{y} = X\boldsymbol{\theta},$$

where X is an $N \times l$ ($N < l$) full row rank matrix. If a solution exists and satisfies the condition

$$\|\boldsymbol{\theta}\|_0 < \frac{1}{2}\left(1 + \frac{1}{\mu(X)}\right), \tag{9.28}$$

then this is the unique solution of both, the ℓ_0 as well as the ℓ_1 minimizers.

This is a very important theorem, and it was shown independently in [37, 54]. Earlier versions of the theorem addressed the special case of a dictionary comprising two orthonormal bases, [36, 48]. A proof is also summarized in [15] (Problem 9.17). This theorem established, for the first time, what it was until then empirically known: often, the ℓ_1 and ℓ_0 minimizers result in the same solution.

Remarks 9.8.

- The theory that we have presented so far is very satisfying, because it offers the theoretical framework and conditions that guarantee uniqueness of a sparse solution to an underdetermined system of equations. Now we know that under certain conditions, the solution, which we obtain by solving the convex ℓ_1 minimization task, is the (unique) sparsest one. However, from a practical point of view, the theory, which is based on mutual coherence, does not tell the whole story and falls short in predicting what happens in practice. Experimental evidence suggests that the range of sparsity levels, for which the ℓ_0 and ℓ_1 tasks give the same solution, is much wider than the range guaranteed by the mutual coherence bound. Hence, there is a lot of theoretical happening in order to improve this bound. A detailed discussion is beyond the scope of this book. In the next section, we will present one of these bounds, because it is the one that currently dominates the scene. For more details and a related discussion, the interested reader may consult, for example, [39, 49, 50].

9.7.2 THE RESTRICTED ISOMETRY PROPERTY (RIP)

Definition 9.3. For each integer $k = 1, 2, \ldots$, define the *isometry constant* δ_k of an $N \times l$ matrix X as the *smallest* number such that

$$\boxed{(1-\delta_k)\,\|\boldsymbol{\theta}\|_2^2 \leq \|X\boldsymbol{\theta}\|_2^2 \leq (1+\delta_k)\,\|\boldsymbol{\theta}\|_2^2: \quad \text{The RIP Condition,}} \tag{9.29}$$

holds true for *all* k-sparse vectors $\boldsymbol{\theta}$.

This definition was introduced in [19]. We loosely say that matrix X obeys the RIP of order k if δ_k is not too close to one. When this property holds true, it implies that the Euclidean norm of $\boldsymbol{\theta}$ is approximately *preserved*, after projecting it onto the rows of X. Obviously, if matrix X was orthonormal then $\delta_k = 0$. Of course, because we are dealing with nonsquare matrices this is not possible. However, the closer δ_k is to zero, the closer to orthonormal *all* subsets of k columns of X are. Another viewpoint of (9.29) is that X preserves Euclidean distances between k-sparse vectors. Let us consider two k-sparse vectors, $\boldsymbol{\theta}_1$, $\boldsymbol{\theta}_2$ and apply (9.29) to their difference $\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2$, which, in general, is a $2k$-sparse vector. Then we obtain

$$(1-\delta_{2k})\,\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_2^2 \leq \|X(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2)\|_2^2 \leq (1+\delta_{2k})\,\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_2^2\,. \tag{9.30}$$

Thus, when δ_{2k} is small enough, the Euclidean distance is preserved after projection in the lower dimensional observations' space. In words, if the RIP holds true, this means that searching for a sparse vector in the lower dimensional subspace, $\mathbb{R}^N$, formed by the observations, and not in the original l-dimensional space, one can still recover the vector since distances are preserved and the target vector is not "confused" with others. After projection onto the rows of X, the discriminatory power of the method is retained. It is interesting to point out that the RIP is also related to the condition number of the Grammian matrix. In [6, 19], it is pointed out that if X_r denotes the matrix that results by considering only r of the columns of X, then the RIP in (9.29) is equivalent with requiring the respective Grammian, $X_r^{\mathrm{T}} X_r$, $r \leq k$, to have its eigenvalues within the interval $[1-\delta_k, 1+\delta_k]$. Hence, the more well conditioned the matrix, the better we dig out the information hidden in the lower dimensional space.

Theorem 9.3. *Assume that for some k,* $\delta_{2k} < \sqrt{2} - 1$. *Then the solution to the* ℓ_1 *minimizer of* (9.21)*, denoted as* $\boldsymbol{\theta}_*$*, satisfies the following two conditions:*

$$\|\boldsymbol{\theta} - \boldsymbol{\theta}_*\|_1 \leq C_0\,\|\boldsymbol{\theta} - \boldsymbol{\theta}_k\|_1\,, \tag{9.31}$$

and

$$\|\boldsymbol{\theta} - \boldsymbol{\theta}_*\|_2 \leq C_0 k^{-\frac{1}{2}}\,\|\boldsymbol{\theta} - \boldsymbol{\theta}_k\|_1\,, \tag{9.32}$$

for some constant C_0. *In the previously stated formulas,* $\boldsymbol{\theta}$ *is the true (target) vector that generates the observations in* (9.21) *and* $\boldsymbol{\theta}_k$ *is the vector that results from* $\boldsymbol{\theta}$ *if we keep its k largest components and set the rest equal to zero [18, 19, 22, 23].*

Hence, if the true vector is a sparse one, that is, $\boldsymbol{\theta} = \boldsymbol{\theta}_k$, then the ℓ_1 minimizer recovers the (unique) exact value. On the other hand, if the true vector is not a sparse one, then the minimizer results in a solution whose accuracy is dictated by a genie-aided procedure that knew in advance the locations of the k largest components of $\boldsymbol{\theta}$. This is a groundbreaking result. Moreover, it is deterministic; it is always true and not with high probability. Note that the isometry property of order $2k$ is used, because at the heart of the method lies our desire to preserve the norm of the differences between vectors.

Let us now focus on the case where there is a k-sparse vector that generates the observations, that is, $\boldsymbol{\theta} = \boldsymbol{\theta}_k$. Then it is shown in [18] that the condition $\delta_{2k} < 1$ guarantees that the ℓ_0 minimizer has a unique

k-sparse solution. In other words, in order to get the equivalence between the ℓ_1 and ℓ_0 minimizers, the range of values for δ_{2k} has to be decreased to $\delta_{2k} < \sqrt{2} - 1$, according to Theorem 9.3. This sounds reasonable. If we relax the criterion and use ℓ_1 instead of ℓ_0, then the sensing matrix has to be more carefully constructed. Although I will not provide the proofs of these theorems here, because their formulation is well beyond the scope of this book, it is interesting to follow what happens if $\delta_{2k} = 1$. This will give us a flavor of the essence behind the proofs. If $\delta_{2k} = 1$, the left-hand side term in (9.30) becomes zero. In this case, there may exist two k-sparse vectors $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2$ such that $X(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2) = \mathbf{0}$, or $X\boldsymbol{\theta}_1 = X\boldsymbol{\theta}_2$. Thus, it is not possible to recover all k-sparse vectors, after projecting them in the observations space, by any method.

The previous argument also establishes a connection between RIP and the spark of a matrix. Indeed, if $\delta_{2k} < 1$, this guarantees that any number of columns of X up to $2k$ are linearly independent, because for any $2k$-sparse $\boldsymbol{\theta}$, (9.29) guarantees that $\|X\boldsymbol{\theta}\|_2 > 0$. This implies that $\text{spark}(X) > 2k$. A connection between RIP and the coherence is established in [16], where it is shown that if X has coherence $\mu(X)$, and unit norm columns, then X satisfies the RIP of order k with δ_k, where $\delta_k \leq (k-1)\mu(X)$.

Constructing matrices that obey the RIP of order k

It is apparent from our previous discussion that the higher the value of k, for which the RIP property of a matrix, X, holds true, the better, since a larger range of sparsity levels can be handled. Hence, a main goal toward this direction is to construct such matrices. It turns out that verifying the RIP for a matrix of a general structure is a difficult task. This reminds us of the spark of the matrix, which is also a difficult task to compute. However, it turns out that for a certain class of random matrices, the RIP can be established in an affordable way. Thus, constructing such sensing matrices has dominated the scene of related research. We will present a few examples of such matrices, which are also very popular in practice, without going into details of the proofs, because this is beyond the scope of this book. The interested reader may find this information in the related references.

Perhaps the most well-known example of a random matrix is the Gaussian one, where the entries $X(i,j)$ of the sensing matrix are i.i.d. realizations from a Gaussian pdf $\mathcal{N}(0, \frac{1}{N})$. Another popular example of such matrices is constructed by sampling i.i.d. entries from Bernoulli, or related, distributions

$$X(i,j) = \begin{cases} \frac{1}{\sqrt{N}}, & \text{with probability } \frac{1}{2}, \\ -\frac{1}{\sqrt{N}}, & \text{with probability } \frac{1}{2}, \end{cases}$$

or

$$X(i,j) = \begin{cases} +\sqrt{\frac{3}{N}}, & \text{with probability } \frac{1}{6}, \\ 0, & \text{with probability } \frac{2}{3}, \\ -\sqrt{\frac{3}{N}}, & \text{with probability } \frac{1}{6}. \end{cases}$$

Finally, one can adopt the uniform distribution and construct the columns of X by sampling uniformly at random on the unit sphere in $\mathbb{R}^N$. It turns out that such matrices obey the RIP of order k with overwhelming probability, provided that the number of observations, N, satisfies the inequality

$$N \geq Ck \ln(l/k), \tag{9.33}$$

where C is some constant, which depends on the isometry constant δ_k. In words, having such a matrix at our disposal, one can recover a k-sparse vector from $N < l$ observations, where N is larger than the sparsity level by an amount controlled by the inequality (9.33). More on these issues can be obtained from, for example, [6, 67].

Besides random matrices, one can construct other matrices that obey the RIP. One such example includes the partial Fourier matrices, which are formed by selecting uniformly at random N rows drawn from the $l \times l$ DFT matrix. Although the required number of samples for the RIP to be satisfied may be larger than the bound in (9.33) (see [79]), Fourier-based sensing matrices offer certain computational advantages when it comes to storage ($\mathcal{O}(N \ln l)$) and matrix-vector products ($\mathcal{O}(l \ln l)$), [20]. In [56], the case of random Toeplitz sensing matrices, containing statistical dependencies across rows is considered and it is shown that they can also satisfy the RIP with high probability. This is of particular importance in signal processing and communications applications, where it is very common for a system to be excited in its input via a time series, hence independence between successive input rows cannot be assumed. In [44, 76], the case of separable matrices is considered where the sensing matrix is the result of a Kronecker product of matrices, which satisfy the RIP individually. Such matrices are of interest for multidimensional signals, in order to exploit the sparsity structure along each one of the involved dimensions. For example, such signals may occur while trying to "encode" information associated with an event whose activity spreads across the temporal, spectral, spatial, and other domains.

In spite of their theoretical elegance, the derived bounds that determine the number of the required observations for certain sparsity levels fall short of the experimental evidence (e.g., [39]). In practice, a rule of thumb is to use N of the order of $3k$-$5k$ [18]. For large values of l, compared to the sparsity level, the analysis in [38] suggests that we can recover most sparse signals when $N \approx 2k \ln(l/N)$. In an effort to overcome the shortcomings associated with the RIP, a number of other techniques have been proposed (e.g., [11, 30, 39, 84]). Furthermore, in specific applications, the use of an empirical study may be a more appropriate path.

Note that, in principle, the minimum number of observations that are required to recover a k-sparse vector from $N < l$ observations is $N \geq 2k$. Indeed, in the spirit of the discussion after Theorem 9.3, the main requirement that a sensing matrix must fulfill is not to map two different k-sparse vectors to the same measurement vector $\boldsymbol{y}$. Otherwise, one can never recover both vectors from their (common) observations. If we have $2k$ observations and a sensing matrix that guarantees that any $2k$ columns are linearly independent, then the previously stated requirement is satisfied. However, the bounds on the number of observations set in order for the respective matrices to satisfy the RIP are larger. This is because RIP accounts also for the stability of the recovery process. We will come to this issue in Section 9.9, where we talk about *stable* embeddings.

9.8 ROBUST SPARSE SIGNAL RECOVERY FROM NOISY MEASUREMENTS

In the previous section, our focus was on recovering a sparse solution from an underdetermined system of equations. In the formulation of the problem, we assumed that there is no noise in the obtained observations. Having acquired some experience and insight from a simpler scenario, we now turn our attention to the more realistic task, where uncertainties come into the scene. One type of uncertainty may be due to the presence of noise, and our observations' model comes back to the standard regression form

$$\boldsymbol{y} = X\boldsymbol{\theta} + \boldsymbol{\eta}, \tag{9.34}$$

where X is our familiar nonsquare $N \times l$ matrix. A sparsity-aware formulation for recovering $\boldsymbol{\theta}$ from (9.34) can be cast as

$$\begin{aligned} \min_{\boldsymbol{\theta} \in \mathbb{R}^l} \quad & \|\boldsymbol{\theta}\|_1 \\ \text{s.t.} \quad & \|\boldsymbol{y} - X\boldsymbol{\theta}\|_2^2 \leq \epsilon, \end{aligned} \tag{9.35}$$

which coincides with the LASSO task given in (9.8). Such a formulation implicitly assumes that the noise is bounded and the respective range of values is controlled by ϵ. One can consider a number of different variants. For example, one possibility would be to minimize the $\|\cdot\|_0$ norm instead of the $\|\cdot\|_1$, albeit losing the computational elegance of the latter. An alternative route would be to replace the Euclidean norm in the constraints with another one.

Besides the presence of noise, one could see the previous formulation from a different perspective. The unknown parameter vector, $\boldsymbol{\theta}$, may not be exactly sparse, but it may consist of a few large components, while the rest are small and close to, yet not necessarily equal to, zero. Such a model misfit can be accommodated by allowing a deviation of $\boldsymbol{y}$ from $X\boldsymbol{\theta}$.

In this relaxed setting of a sparse solution recovery, the notions of uniqueness and equivalence concerning the ℓ_0 and ℓ_1 solutions no longer apply. Instead, the issue that now gains importance is that of *stability* of the solution. To this end, we focus on the computationally attractive ℓ_1 task. The counterpart of Theorem 9.3 is now expressed as follows.

Theorem 9.4. *Assume that the sensing matrix, X, obeys the RIP with $\delta_{2k} < \sqrt{2} - 1$, for some k. Then the solution $\boldsymbol{\theta}_*$ of (9.35) satisfies the following ([22, 23]),*

$$\|\boldsymbol{\theta} - \boldsymbol{\theta}_*\|_2 \leq C_0 k^{-\frac{1}{2}} \|\boldsymbol{\theta} - \boldsymbol{\theta}_k\|_1 + C_1 \sqrt{\epsilon}, \tag{9.36}$$

for some constants C_1, C_0, and $\boldsymbol{\theta}_k$ as defined in Theorem 9.3.

This is also an elegant result. If the model is exact and $\epsilon = 0$ we obtain (9.32). If not, the higher the uncertainty (noise) term in the model, the higher our ambiguity about the solution. Note, also, that the ambiguity about the solution depends on how far the true model is from $\boldsymbol{\theta}_k$. If the true model is k-sparse, the first term on the right-hand side of the inequality is zero. The values of C_1, C_0 depend on δ_{2k} but they are small, for example, close to five or six, [23].

The important conclusion here is that *the LASSO formulation for solving inverse problems (which in general, as we noted in Chapter 3, tend to be ill-conditioned) is a stable one and the noise is not amplified excessively during the recovery process.*

9.9 COMPRESSED SENSING: THE GLORY OF RANDOMNESS

The way in which this chapter was deployed followed, more or less, the sequence of developments that took place during the evolution of the sparsity-aware parameter estimation field. We intentionally made an effort to follow such a path, because this is also indicative of how science evolves in most cases. The starting point had a rather strong mathematical flavor: to develop conditions for the solution of an underdetermined linear system of equations, under the sparsity constraint and in a mathematically tractable way, that is, using convex optimization. In the end, the accumulation of a sequence of individual contributions revealed that the solution can be (uniquely) recovered if the unknown quantity is sensed via randomly chosen data samples. This development has, in turn, given birth to a new field

with strong theoretical interest as well as with an enormous impact on practical applications. This new emerged area is known as *compressed sensing* or *compressive sampling* (CS). Although CS builds around the LASSO and basis pursuit (and variants of them, as we will soon see), it has changed our view on how to sense and process signals efficiently.

Compressed sensing

In compressed sensing, the goal is to directly acquire as few samples as possible that encode the minimum information needed to obtain a compressed signal representation. In order to demonstrate this, let us return to the data compression example discussed in Section 9.4. There, it was commented that the "classical" approach to compression was rather unorthodox, in the sense that first all (i.e., a number of l) samples of the signal are used, and then they are processed to obtain l transformed values, from which only a small subset is used for coding. In the CS setting, the procedure changes to the following one.

Let X be an $N \times l$ sensing matrix, which is applied to the (unknown) signal vector, $\boldsymbol{s}$, in order to obtain the observations, $\boldsymbol{y}$, and Ψ be the dictionary matrix that describes the domain where the signal $\boldsymbol{s}$ accepts a sparse representation, that is,

$$\begin{aligned} \boldsymbol{s} &= \Psi\boldsymbol{\theta}, \\ \boldsymbol{y} &= X\boldsymbol{s}. \end{aligned} \tag{9.37}$$

Assuming that at most k of the components of $\boldsymbol{\theta}$ are nonzero, this can be obtained by the following optimization task

$$\begin{aligned} \min_{\boldsymbol{\theta}\in\mathbb{R}^l} \quad & \|\boldsymbol{\theta}\|_1 \\ \text{s.t.} \quad & \boldsymbol{y} = X\Psi\boldsymbol{\theta}, \end{aligned} \tag{9.38}$$

provided that the combined matrix $X\Psi$ complies with the RIP, and the number of observations, N, is large enough, as dictated by the bound in (9.33). Note that $\boldsymbol{s}$ needs not be stored and can be obtained any time, once $\boldsymbol{\theta}$ is known. Moreover, as we will soon discuss, there are techniques that allow observations, y_n, $n = 1, 2, \ldots, N$, to be acquired directly from an analog signal $s(t)$, prior to obtaining its sample (vector) version, $\boldsymbol{s}$! Thus, from such a perspective, CS fuses the data acquisition and the compression steps together.

There are different ways to obtain a sensing matrix, X, that leads to a product $X\Psi$, which satisfies the RIP. It can be shown (Problem 9.19) that if Ψ is orthonormal and X is a random matrix, which is constructed as discussed at the end of Section 9.7.2, then the product $X\Psi$ obeys the RIP, provided that (9.33) is satisfied. An alternative way to obtain a combined matrix that respects the RIP is to consider another orthonormal matrix Φ, whose columns have low coherence with the columns of Ψ (coherence between two matrices is defined in (9.25), where now, the place of $\mathbf{x}_i$ is taken by a column of Φ and that of $\mathbf{x}_j$ by a column of Ψ). For example, Φ could be the DFT matrix and $\Psi = I$, or vice versa. Then choose N rows of Φ uniformly at random to form X in (9.37). In other words, for such a case, the sensing matrix can be written as $R\Phi$, where R is an $N \times l$ matrix that extracts N rows uniformly at random. The notion of incoherence (low coherence) between the sensing and the basis matrices is closely related to RIP. The more incoherent the two matrices, the less the number of the required observations for the RIP to hold (e.g., [21, 79]). Another way to view incoherence is that the rows of Φ cannot be sparsely represented in terms of the columns of Ψ. It turns out that if the sensing matrix X is a random one,

formed as has already been described in Section 9.7.2, then the RIP and the incoherence with any Ψ are satisfied with high probability.

It gets even better when we say that all the previously stated philosophy can be extended to the more general type of signals, which are not necessarily sparse or sparsely represented in terms of the atoms of a dictionary, and they are known as *compressible*. A signal vector is said to be compressible if its expansion in terms of a basis consists of just a few large coefficients θ_i and the rest are small. In other words, the signal vector is *approximately* sparse in some basis. Obviously, this is the most interesting case in practice, where exact sparsity is scarcely (if ever) met. Reformulating the arguments used in Section 9.8, the CS task for this case can be cast as

$$\begin{aligned} \min_{\boldsymbol{\theta}\in\mathbb{R}^l} \quad & \|\boldsymbol{\theta}\|_1 \\ \text{s.t.} \quad & \|\boldsymbol{y} - X\Psi\boldsymbol{\theta}\|_2^2 \leq \epsilon, \end{aligned} \tag{9.39}$$

and everything that has been said in Section 9.8 is also valid for this case, if in place of X we consider the product $X\Psi$.

Remarks 9.9.

- An important property in compressed sensing is that the sensing matrix, which provides the observations, may be chosen independently on the matrix Ψ, that is, the basis/dictionary in which the signal is sparsely represented. In other words, the sensing matrix can be "universal" and can be used to provide the observations for reconstructing any sparse or sparsely represented signal in any dictionary, provided RIP is not violated.
- Each measurement, y_n, is the result of an inner product of the signal vector with a row, $\boldsymbol{x}_n^{\mathrm{T}}$, of the sensing matrix, X. Assuming that the signal vector, $\boldsymbol{s}$, is the result of a sampling process on an analog signal, $s(t)$, then y_n can be directly obtained, to a good approximation, by taking the inner product (integral) of $s(t)$ with a sensing waveform, $x_n(t)$, that corresponds to $\boldsymbol{x}_n$. For example, if X is formed by ± 1, as described in Section 9.7.2, then the configuration shown in Figure 9.10 results to y_n. An important aspect of this approach, besides avoiding computing and storing the l components of $\boldsymbol{s}$, is that multiplying by ± 1 is a relatively easy operation. It is equivalent with changing the polarity of the signal and it can be implemented by employing inverters and mixers. It is a process that can be performed, in practice, at much higher rates than sampling. The

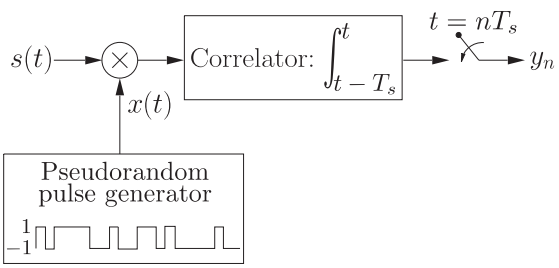

FIGURE 9.10

Sampling an analog signal $s(t)$ in order to generate the measurement y_n at the time instant n. The sampling period T_s is much lower than that required by the Nyquist sampling.

sampling system shown in Figure 9.10 is referred to as *random demodulator*, [58, 90]. It is one among the popular analog-to-digital (A/D) conversion architectures, which exploit the CS rationale in order to sample at rates much lower than those required for classical sampling. We will come back to this soon.

One of the very first CS-based acquisition systems was an imaging system called the *one pixel camera* [83], which followed an approach resembling the conventional digital CS. According to this, light of an image of interest is projected onto a random base generated by a micromirror device. A sequence of projected images is collected by a *single photodiode* and used for the reconstruction of the full image using conventional CS techniques. This was among the most catalytic examples that spread the rumor about the practical power of CS. CS is an example of common wisdom: "There is nothing more practical than a good theory!"

9.9.1 DIMENSIONALITY REDUCTION AND STABLE EMBEDDINGS

We will now shed light on what we have said so far in this chapter from a different point of view. In both cases, either when the unknown quantity was a k-sparse vector in a high-dimensional space, $\mathbb{R}^l$, or when the signal $\boldsymbol{s}$ was (approximately) sparsely represented in some dictionary ($\boldsymbol{s} = \Psi\boldsymbol{\theta}$), we chose to work in a lower dimensional space ($\mathbb{R}^N$), that is, the space of the observations, $\boldsymbol{y}$. This is a typical task of dimensionality reduction, see, Chapter 19. The main task in any (linear) dimensionality reduction technique is to choose the proper matrix X, that dictates the projection to the lower dimensional space. In general, there is always a loss of information by projecting from $\mathbb{R}^l$ to $\mathbb{R}^N$, with $N < l$, in the sense that we cannot recover any vector, $\boldsymbol{\theta}_l \in \mathbb{R}^l$, from its projection $\boldsymbol{\theta}_N \in \mathbb{R}^N$. Indeed, take any vector $\boldsymbol{\theta}_{l-N} \in \text{null}(X)$, that lies in the $(l-N)$-dimensional null space of the (full row rank) X (see Section 9.5). Then, all vectors $\boldsymbol{\theta}_l + \boldsymbol{\theta}_{l-N} \in \mathbb{R}^l$ share the same projection in $\mathbb{R}^N$. However, what we have discovered in this chapter is that if the original vector is sparse, then we can recover it exactly. This is because all the k-sparse vectors do not lie anywhere in $\mathbb{R}^l$, but rather in a subset of it, that is, in a *union of subspaces*, each one having dimensionality k. If the signal $\boldsymbol{s}$ is sparse in some dictionary Ψ, then one has to search for it in the union of all possible k-dimensional subspaces of $\mathbb{R}^l$, which are spanned by k-column vectors from Ψ [8, 62]. Of course, even in this case, where sparse vectors are involved, no projection can guarantee unique recovery. The guarantee is provided if the projection in the lower dimensional space is a *stable embedding*. A stable embedding in a lower dimensional space must guarantee that if $\boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_2$, then their projections also remain different. Yet this is not enough. A stable embedding must guarantee that distances are (approximately) preserved; that is, vectors that lie far apart in the high-dimensional space have projections that also lie far apart. Such a property guarantees robustness to noise. The sufficient conditions, which have been derived and discussed throughout this chapter, and guarantee the recovery of a sparse vector lying in $\mathbb{R}^l$ from its projections in $\mathbb{R}^N$, are conditions that guarantee stable embeddings. The RIP and the associated bound on N provide a condition on X that leads to stable embeddings. We commented on this norm-preserving property of RIP in the related section. The interesting fact that came from the theory is that we can achieve such stable embeddings via random projection matrices.

Random projections for dimensionality reduction are not new and have extensively been used in pattern recognition, clustering, and data mining (see, e.g., [1, 13, 34, 82, 86]). The advent of the big data era resparked the interest in random projection-aided data analysis algorithms (e.g., [55, 81]) for two major reasons. The first is that data processing is computationally lighter in the lower dimensional

space, because it involves operations with matrices or vectors represented with fewer parameters. Moreover, the projection of the data to lower dimensional spaces can be realized via well-structured matrices in computational cost significantly lower compared to that implied by general matrix-vector multiplications [29, 42]. The reduced computational power required by these methods renders them appealing when dealing with excessively large data volumes. The second reason is that there exist randomized algorithms, which access the data matrix a (usually fixed) number of times that is much smaller than the number of accesses performed by ordinary methods [28, 55]. This is very important whenever the full amount of data does not fit in fast memory and has to be accessed in parts from slow memory devices, such as hard discs. In such cases, the computational time is often dominated by the cost of memory access.

The spirit underlying compressed sensing has been exploited in the context of pattern recognition too. In this application, one need not return to the original high-dimensional space, after the information-digging activity in the low-dimensional subspace. Since the focus in pattern recognition is to identify the class of an object/pattern, this can be performed in the observations subspace, provided that there is no class-related information loss. In [17], it is shown, using compressed sensing arguments, that if the data is approximately linearly separable in the original high-dimensional space and the data has a sparse representation, even in an unknown basis, then projecting randomly in the observations subspace retains the structure of linear separability.

Manifold learning is another area where random projections have been recently applied. A manifold is, in general, a nonlinear k-dimensional surface, embedded in a higher dimensional (ambient) space. For example, the surface of a sphere is a two-dimensional manifold in a three-dimensional space. In [7, 95], the compressed sensing rationale is extended to signal vectors that live along a k-dimensional submanifold of the space $\mathbb{R}^l$. It is shown that if choosing a matrix, X, to project and a sufficient number, N, of observations, then the corresponding submanifold has a stable embedding in the observations subspace, under the projection matrix, X; that is, pairwise Euclidean and geodesic distances are approximately preserved after the projection mapping. More on these issues can be found in the given references and in, for example, [8]. We will come to the manifold learning task in Chapter 19.

9.9.2 SUB-NYQUIST SAMPLING: ANALOG-TO-INFORMATION CONVERSION

In our discussion in the Remarks presented before, we touched on a very important issue—that of going from the analog domain to the discrete one. The topic of analog-to-digital (A/D) conversion has been at the forefront of research and technology since the seminal works of Shannon, Nyquist, Whittaker, and Kotelnikof were published, see, for example, [91] for a thorough related review. We all know that if the highest frequency of an analog signal, $s(t)$, is less than $F/2$, then Shannon's theorem suggests that no loss of information is achieved if the signal is sampled, at least, at the Nyquist rate of $F = 1/T$, where T is the corresponding sampling period, and the signal can be perfectly recovered by its samples

$$s(t) = \sum_n s(nT)\,\text{sinc}(Ft - n),$$

where $\text{sinc}(\cdot)$ is the sampling function

$$\text{sinc}(t) = \frac{\sin(\pi t)}{\pi t}.$$

While this has been the driving force behind the development of signal acquisition devices, the increasing complexity of emerging applications demands increasingly higher sampling rates that cannot be accommodated by today's hardware technology. This is the case, for example, in wideband communications, where conversion speeds, as dictated by Shannon's bound, have become more and more difficult to obtain. Consequently, alternatives to high rate sampling are attracting strong interest, with the goal of reducing the sampling rate by exploiting the *underlying structure* of the signals at hand. For example, in many applications, the signal comprises a few frequencies or bands, see Figure 9.11 for an illustration. In such cases, sampling at the Nyquist rate is inefficient. This is an old problem investigated by a number of authors, leading to techniques that allow low rate sampling whenever the locations of the nonzero bands in the frequency spectrum are known (see, e.g., [61, 92, 93]). CS theory has inspired research to study cases where the locations (carrier frequencies) of the bands are not known a priori. A typical application of this kind, of high practical interest, lies within the field of cognitive radio (e.g., [68, 87, 102]).

The process of sampling an analog signal with a rate lower than the Nyquist one is referred to as *analog-to-information* sampling or *sub-Nyquist* sampling. Two are the most popular CS-based A/D converters. The first is the *random demodulator* (RD), which was first presented in [58] and later improved and theoretically developed in [90]. RD in its basic configuration is shown in Figure 9.10, and it is designed for acquiring at sub-Nyquist rates sparse multitone signals, that is, signals having a sparse DFT. This implies that the signal comprises a few frequency components, but these components are constrained to correspond to integral frequencies. This limitation was pointed out in [90], and potential solutions have been sought according to the general framework proposed in [24] and/or the heuristic approach described in [45]. Moreover, more elaborate RD designs, such as the *random-modulation pre-integrator* (RMPI) [101], have the potential to deal with signals that are sparse in any domain.

Another CS-based sub-Nyquist sampling strategy that has received much attention is the *modulated wideband converter* (MWC), [68, 69, 71]. The MWC is very efficient in acquiring multiband signals such as the one depicted in Figure 9.11. This concept has also been extended to accommodate

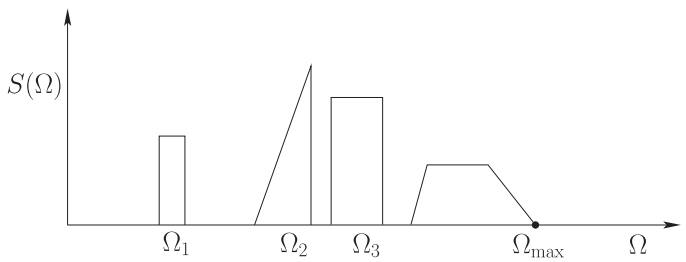

FIGURE 9.11

The Fourier transform of an analog signal, $s(t)$, which is sparse in the frequency domain; only a limited number of frequency bands contribute to its spectrum content $S(\Omega)$, where Ω stands for the angular frequency. Nyquist's theory guarantees that sampling at a frequency larger than or equal to twice the maximum $\Omega_{\max}$ is sufficient to recover the original analog signal. However, this theory does not exploit information related to the sparse structure of the signal in the frequency domain.

signals with different characteristics, such as signals consisting of short pulses [66]. An in-depth investigation, which sheds light on the similarities and differences between the RD and the MWC sampling architectures, can be found in [59].

Note that both RD and MWC sample the signal uniformly in time. In [96], a different approach is adopted, leading to much easier implementations. In particular, the preprocessing stage is avoided and nonuniformly spread in time samples are acquired directly from the raw signal. In total, less samples are obtained compared to the Nyquist sampling. Then, CS-based reconstruction is mobilized in order to recover the signal under consideration based on the values of the samples and the time information. Like in the basic RD case, the nonuniform sampling approach is suitable for signals sparse in the DFT basis. From a practical point of view, there are still a number of hardware implementation-related issues that more or less concern all the approaches above and need to be solved (see, e.g., [9, 25, 63]).

An alternative path to sub-Nyquist sampling embraces a different class of analog signals known as *multipulse* signals, that is, signals that consist of a stream of short pulses. Sparsity now refers to the time domain, and such signals may not even be bandlimited. Signals of this type can be met in a number of applications, such as in radar, ultrasound, bioimaging, and neuronal signal processing (see, e.g., [41]). An approach known as *finite rate of innovation sampling* passes an analog signal having k degrees of freedom per second through a linear time invariant filter, and then samples at a rate of $2k$ samples per second. Reconstruction is performed via rooting a high-order polynomial (see, e.g., [12, 94] and the references therein). In [66], the task of sub-Nyquist sampling is treated using CS theory arguments and an expansion in terms of Gabor functions; the signal is assumed to consist of a sum of a few pulses of finite duration, yet of unknown shape and time positions.

The task of sparsity-aware learning in the analog domain is still in its early stages, and there is a lot of ongoing activity; more on this topic can be obtained in [43, 51, 70] and the references therein.

Example 9.5. We are given a set of $N = 20$ observations stacked in the $\boldsymbol{y} \in \mathbb{R}^N$ vector. These were taken by applying a sensing matrix X on an "unknown" vector in $\mathbb{R}^{50}$, which is known to be sparse with $k = 5$ nonzero components; the location of these nonzero components in the unknown vector is not known. The sensing matrix was a random matrix with elements drawn from a normal distribution $\mathcal{N}(0, 1)$, and then the columns were normalized to unit norm. There are two scenarios for the measurements. In the first one, we are given the exact measurements, while in the second one, white Gaussian noise of variance $\sigma^2 = 0.025$ was added.

In order to recover the unknown sparse vector, the compressive sampling matching pursuit (CoSaMP, Chapter 10) algorithm was used for both scenarios.

The results are shown in Figure 9.12a and b for the noiseless and noisy scenarios, respectively. The values of the true unknown vector $\boldsymbol{\theta}$ are represented with black stems topped with open circles. Note that all but five of them are zero. In Figure 9.12a, exact recovery of the unknown values is succeeded; the estimated values of $\theta_i, i = 1, 2 \ldots, 50$, are indicated with squares in red color. In the noisy case of Figure 9.12b, the resulting estimates, which are denoted with squares, deviate from the correct values. Note that estimated values very close to zero ($|\theta| \leq 0.01$) have been omitted from the figure in order to facilitate visualization. In both figures, the stemmed gray-filled circles correspond to the minimum ℓ_2 norm LS solution. The advantages of adopting a sparsity-promoting approach to recover the solution are obvious. The CoSaMP algorithm was provided with the exact number of sparsity. The reader is advised to reproduce the example and play with different values of the parameters and see how results are affected.

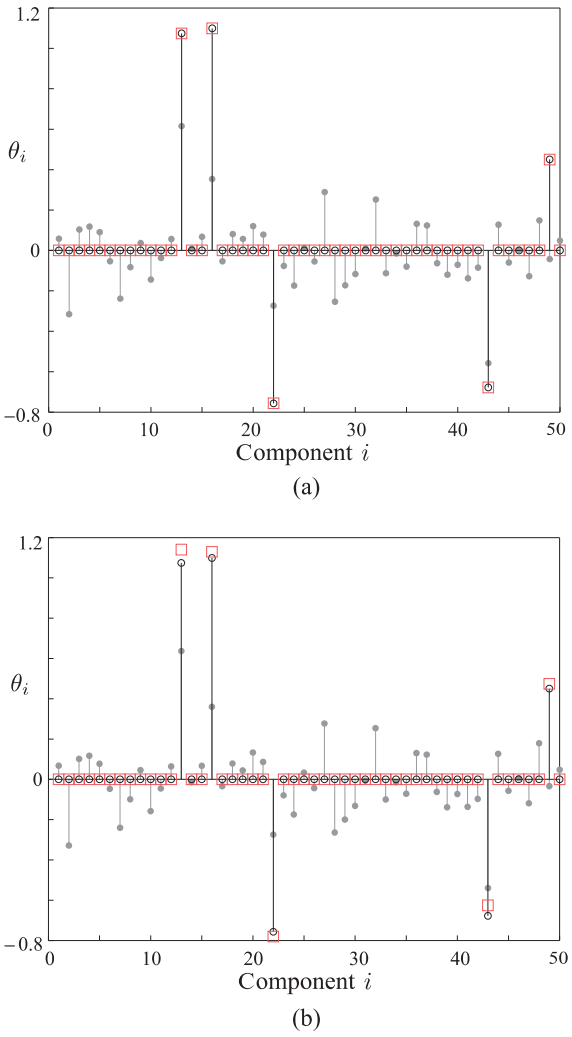

FIGURE 9.12

(a) Noiseless case. The values of the true vector, which generated the data for Example 9.5, are shown with stems topped with open circles. The recovered points are shown with squares. An exact recovery of the signal has been obtained. The stems topped with gray-filled circles correspond to the minimum Euclidean norm LS solution. (b) This figure corresponds to the noisy counterpart of that in (a). In the presence of noise, exact recovery is not possible and the higher the variance of the noise, the less accurate the results.

9.10 A CASE STUDY: IMAGE DE-NOISING

We have already discussed compressed sensing (CS) as a notable application of sparsity-aware learning. Although CS has acquired a lot of fame, a number of classical signal processing and machine learning tasks lend themselves to efficient modeling via sparsity-related arguments. Two typical examples are:

- *De-noising*: The problem in signal de-noising is that instead of the actual signal samples, $\tilde{\boldsymbol{y}}$, a noisy version of the corresponding observations, $\boldsymbol{y}$, are available; that is, $\boldsymbol{y} = \tilde{\boldsymbol{y}} + \boldsymbol{\eta}$, where $\boldsymbol{\eta}$ is the vector of noise samples. Under the sparse modeling framework, the unknown signal $\tilde{\boldsymbol{y}}$ is modeled as a sparse representation in terms of a specific known dictionary Ψ, that is, $\tilde{\boldsymbol{y}} = \Psi\boldsymbol{\theta}$. Moreover, the dictionary is allowed to be redundant (overcomplete). Then, the de-noising procedure is realized in two steps.

  First, an estimate of the sparse representation vector, $\boldsymbol{\theta}$, is obtained via the ℓ_0 norm minimizer or via any LASSO formulation, for example,

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^l} \|\boldsymbol{\theta}\|_1, \tag{9.40}$$

$$\text{s.t.} \quad \|\boldsymbol{y} - \Psi\boldsymbol{\theta}\|_2^2 \leq \epsilon. \tag{9.41}$$

  Second, the estimate of the true signal is computed as $\hat{\boldsymbol{y}} = \Psi\hat{\boldsymbol{\theta}}$. In Chapter 19, we will study the case where the dictionary is not fixed and known, but is estimated from the data.
- *Linear inverse problems*: Such problems, which come under the more general umbrella of what is known as *signal restoration*, go one step beyond de-noising. Now, the available observations are *distorted* as well as noisy versions of the true signal samples; that is, $\boldsymbol{y} = H\tilde{\boldsymbol{y}} + \boldsymbol{\eta}$, where H is a known linear operator. For example, H may correspond to the blurring point spread function of an image, as discussed in Chapter 4. Then, similar to the de-noising example, assuming that the original signal samples can be efficiently represented in terms of an overcomplete dictionary, $\hat{\boldsymbol{\theta}}$ is estimated, via any sparsity promoting method, using $H\Psi$ in place of Ψ in (9.41), and the estimate of the true signal is obtained as $\hat{\boldsymbol{y}} = \Psi\hat{\boldsymbol{\theta}}$.

  Besides de-blurring, other applications that fall under this formulation include image inpainting, if H represents the corresponding sampling mask; inverse-Radon transform in tomography, if H comprises the set of parallel projections, and so on. See, for example, [49] for more details on this topic.

In this case study, the image de-noising task, based on the sparse and redundant formulation as discussed above, is explored. Our starting point is the 256×256 image shown in Figure 9.13a. In the sequel, the image is corrupted by zero mean Gaussian noise leading to the noisy version of Figure 9.13b, corresponding to peak signal-to-noise ratio (PSNR) equal to 22 dB, which is defined as

$$\text{PSNR} = 20\log_{10}\left(\frac{m_I}{\sqrt{\text{MSE}}}\right), \tag{9.42}$$

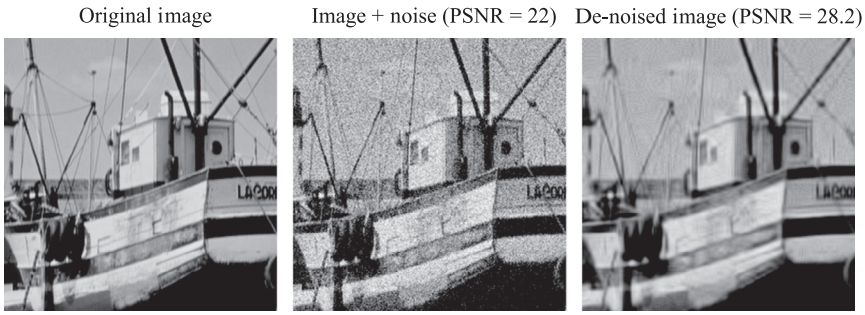

FIGURE 9.13

De-noising based on sparse and redundant representations.

where m_I is the maximum pixel value of the image and $\text{MSE} = \frac{1}{N_p}\left\|I - \tilde{I}\right\|_F^2$, with I and $\tilde{I}$ being the noisy and original image matrices, N_p is equal to the total number of pixels and the Frobenius norm for matrices has been employed.

De-noising could be applied to the full image at once. However, a more efficient practice with respect to memory consumption is to split the image to patches of size much smaller than that of the image; for our case, we chose 12×12 patches. Then, de-noising is performed to each patch separately as follows: The ith patch image is reshaped in lexicographic order forming a 1-D vector, $\boldsymbol{y}_i \in \mathbb{R}^{144}$. We assume that each one of the patches can be reproduced in terms of an overcomplete dictionary, as discussed before; hence, the de-noising task is equivalently formulated around (9.40)-(9.41). Denote by $\tilde{\boldsymbol{y}}_i$ the ith patch of the noise-free image. What is left is to choose a dictionary Ψ, which sparsely represents the $\tilde{\boldsymbol{y}}_i$, and then solve for sparse $\boldsymbol{\theta}$ according to (9.40)-(9.41).

It is known that images often exhibit sparse DCT transforms, so an appropriate choice for the dictionary is to fill the columns of Ψ with atoms of a redundant 2D-DCT reshaped in lexicographic order [49]. Here, 196 such atoms were used. There is a standard way to develop such a dictionary given the dimensionality of the image, and it is described in Exercise 9.22. The *same* dictionary is used for *all* patches. The atoms of the dictionary, reshaped to form 12×12 blocks, are depicted in Figure 9.14.

A question that naturally arises is how many patches to use. A straightforward approach is to tile the patches side by side in order to cover the whole extent of the image. This is feasible, however, it is likely to result in blocking effects at the edges of several patches. A better practice is to let the patches overlap. During the reconstruction phase ($\hat{\boldsymbol{y}} = \Psi\hat{\boldsymbol{\theta}}$), because each pixel is covered by more than one patch, the final value of each pixel is taken as the average of the corresponding predicted values from all the involved patches. The results of this method, for our case, are shown in Figure 9.13c. The attained PSNR is higher than 28 dB.

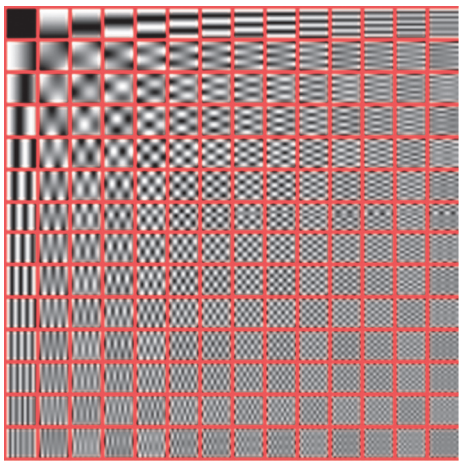

FIGURE 9.14

2D-DCT Dictionary atoms, corresponding to 12×12 patch size.

PROBLEMS

9.1 If $x_i,\ y_i,\ i = 1, 2, \ldots, l$, are real numbers, then prove the Cauchy-Schwarz inequality:

$$\left(\sum_{i=1}^{l} x_i y_i\right)^2 \leq \left(\sum_{i=1}^{l} x_i^2\right)\left(\sum_{i=1}^{l} y_i^2\right).$$

9.2 Prove that the ℓ_2 (Euclidean) norm is a true norm, that is, it satisfies the four conditions that define a norm.

Hint. To prove the triangle inequality, use the Cauchy-Schwarz inequality.

9.3 Prove that any function that is a norm is also a convex function.

9.4 Show Young's inequality for nonnegative real numbers a and b,

$$ab \leq \frac{a^p}{p} + \frac{b^q}{q},$$

for $\infty > p > 1$ and $\infty > q > 1$ such that

$$\frac{1}{p} + \frac{1}{q} = 1.$$

9.5 Prove Holder's inequality for ℓ_p norms,

$$||\boldsymbol{x}^{\mathrm{T}}\boldsymbol{y}||_1 = \sum_{i=1}^{l} |x_i y_i| \leq ||\boldsymbol{x}||_p ||\boldsymbol{y}||_q = \left(\sum_{i=1}^{l} |x_i|^p\right)^{1/p} \left(\sum_{i=1}^{q} |y_i|^q\right)^{1/q},$$

for $p \geq 1$ and $q \geq 1$ such that

$$\frac{1}{p} + \frac{1}{q} = 1.$$

Hint. Use Young's inequality.

9.6 Prove Minkowski's inequality,

$$\left(\sum_{i=1}^{l}(|x_i| + |y_i|)^p\right)^{1/p} \leq \left(\sum_{i=1}^{l}|x_i|^p\right)^{1/p} + \left(\sum_{i=1}^{l}|y_i|^p\right)^{1/p},$$

for $p \geq 1$.

Hint. Use Holder's inequality together with the identity

$$(|a| + |b|)^p = (|a| + |b|)^{p-1}|a| + (|a| + |b|)^{p-1}|b|.$$

9.7 Prove that for $p \geq 1$, the ℓ_p norm is a true norm.

9.8 Use a counterexample to show that any ℓ_p norm for $0 < p < 1$ is not a true norm and it violates the triangle inequality.

9.9 Show that the null space of a full row rank $N \times l$ matrix X is a subspace of dimensionality N, for $N < l$.

9.10 Show, using Lagrange multipliers, that the ℓ_2 minimizer in (9.18), accepts the closed form solution

$$\hat{\boldsymbol{\theta}} = X^{\mathrm{T}}\left(XX^{\mathrm{T}}\right)^{-1}\boldsymbol{y}.$$

9.11 Show that the necessary and sufficient condition for a $\boldsymbol{\theta}$ to be a minimizer of

$$\begin{aligned} &\text{minimize} && ||\boldsymbol{\theta}||_1 \\ &\text{subject to} && X\boldsymbol{\theta} = \boldsymbol{y}, \end{aligned}$$

is the following

$$\left|\sum_{i:\theta_i \neq 0} \text{sign}(\theta_i) z_i\right| \leq \sum_{i:\theta_i = 0} |z_i|, \ \forall \boldsymbol{z} \in \text{null}(X),$$

where null(X) is the null space of X. Moreover, if the minimizer is unique the previous inequality becomes a strict one.

9.12 Prove that if the ℓ_1 norm minimizer is unique, then the number of its components, which are identically zero, must be at least as large as the dimensionality of the null space of the corresponding input matrix.

9.13 Show that the ℓ_1 norm is a convex function (as all norms), yet it is not strictly convex. In contrast, the squared Euclidean norm is a strictly convex function.

9.14 Construct in the five-dimensional space a matrix that has (a) rank equal to five and spark equal to four, (b) rank equal to five and spark equal to three, and (c) rank and spark equal to four.

9.15 Let X be a full row rank $N \times l$ matrix, with $l > N$. Derive the Welch bound for the mutual coherence $\mu(X)$,

$$\mu(X) \geq \sqrt{\frac{l - N}{N(l-1)}}. \tag{9.43}$$

9.16 Let X be an $N \times l$ matrix. Then prove that its spark is bounded as

$$\text{spark}(X) \geq 1 + \frac{1}{\mu(X)},$$

where $\mu(X)$ is the mutual coherence of the matrix.
Hint. Consider the Gram matrix $X^{\mathrm{T}}X$ and the following theorem, concerning positive definite matrices: An $m \times m$ matrix A is positive definite if

$$|A(i,i)| > \sum_{j=1, j\neq i}^{m} |A(i,j)|, \ \forall i = 1, 2, \ldots, m,$$

see, for example, [57].

9.17 Show that if the underdetermined system of equations $\boldsymbol{y} = X\boldsymbol{\theta}$ accepts a solution such that

$$||\boldsymbol{\theta}||_0 < \frac{1}{2}\left(1 + \frac{1}{\mu(X)}\right),$$

then the ℓ_1 minimizer is equivalent to the ℓ_0 one. Assume that the columns of X are normalized.

9.18 Prove that if the RIP of order k is valid for a matrix X and $\delta_k < 1$, then any $m < k$ columns of X are necessarily linearly independent.

9.19 Show that if X satisfies the RIP of order k and some isometry constant δ_k so does the product $X\Psi$ if Ψ is an orthonormal matrix.

MATLAB Exercises

9.20 Consider an unknown 2-sparse vector $\boldsymbol{\theta}_o$, which when measured with the following sensing matrix

$$X = \begin{bmatrix} 0.5 & 2 & 1.5 \\ 2 & 2.3 & 3.5 \end{bmatrix},$$

that is, of $\boldsymbol{y} = X\boldsymbol{\theta}_o$, gives $\boldsymbol{y} = [1.25,\ 3.75]^{\mathrm{T}}$. Perform the following tasks in MATLAB:
(a) Based on the pseudo-inverse of X, compute $\hat{\boldsymbol{\theta}}_2$, which is the ℓ_2 norm minimized solution, (9.18). Next, check that this solution $\hat{\boldsymbol{\theta}}_2$ leads to zero estimation error (up to machine precision). Is $\hat{\boldsymbol{\theta}}_2$ a 2-sparse vector such as the true unknown vector $\boldsymbol{\theta}_o$, and if it is not, how is it possible to lead to zero estimation error? (b) Solve the ℓ_0 minimization task described in (9.20) (exhaustive search) for all possible 1- and 2-sparse solutions and get the best one, $\hat{\boldsymbol{\theta}}_o$. Does $\hat{\boldsymbol{\theta}}_o$ lead to zero estimation error (up to machine precision)? (c) Compute and compare the ℓ_2 norms of $\hat{\boldsymbol{\theta}}_2$ and $\hat{\boldsymbol{\theta}}_o$. Which is the smaller one? Was this result expected?

9.21 Generate in MATLAB a sparse vector $\boldsymbol{\theta} \in \mathbb{R}^l$, $l = 100$, with its first 5 components taking random values drawn from a normal distribution, $\mathcal{N}(0, 1)$ and the rest being equal to zero. Build, also, a sensing matrix X with $N = 30$ rows having samples normally distributed $\mathcal{N}(0, \frac{1}{\sqrt{N}})$, in order to get 30 observations based on the linear regression model $\boldsymbol{y} = X\boldsymbol{\theta}$. Then perform the following tasks: (a) Use the function "solvelasso.m"[5], or any other LASSO implementation you prefer, in order to reconstruct $\boldsymbol{\theta}$ from $\boldsymbol{y}$ and X. (b) Repeat the experiment with different realizations of X in order to compute the probability of correct reconstruction

[5] It can be found in the SparseLab MATLAB toolbox, which is freely available from http://sparselab.stanford.edu/

(assume the reconstruction is exact when $||\boldsymbol{y} - X\boldsymbol{\theta}||_2 < 10^{-8}$). (c) Construct another sensing matrix X having $N = 30$ rows taken uniformly at random from the $l \times l$ DCT matrix, which can be obtained via the built-in MATLAB function "dctmtx.m". Compute the probability of reconstruction when this DCT-based sensing matrix is used and confirm that results similar to those in question (b) are obtained. (d) Repeat the same experiment with matrices of the form

$$X(i,j) = \begin{cases} +\sqrt{\frac{\sqrt{p}}{N}}, & \text{with probability } \frac{1}{2\sqrt{p}}, \\ 0, & \text{with probability } 1 - \frac{1}{\sqrt{p}}, \\ -\sqrt{\frac{\sqrt{p}}{N}}, & \text{with probability } \frac{1}{2\sqrt{p}}, \end{cases}$$

for p equal to 1, 9, 25, 36, 64 (make sure that at each row and each column of X has at least a nonzero component). Give an explanation why the probability of reconstruction falls as p increases (observe that both the sensing matrix and the unknown vector are sparse).

9.22 This exercise reproduces the de-noising results of the case study in Section 9.10, where the image depicting the boat can be downloaded from the book website. First, extract from the image all the possible sliding patches of size 12×12 using the im2col.m Matlab function. Confirm that $(256 - 12 + 1)^2 = 60{,}025$ patches in total are obtained. Next, a dictionary in which all the patches are sparsely represented needs to be designed.

Specifically, the dictionary atoms are going to be those corresponding to the 2D redundant DCT transform, which are obtained as follows [49]:

a) Consider vectors $\boldsymbol{d}_i = [d_{i,1}, d_{i,2}, \ldots, d_{i,12}]^{\mathrm{T}}$, $i = 0, \ldots, 13$, being the sampled sinusoids of the form

$$d_{i,t+1} = \cos\left(\frac{t\pi i}{14}\right), \quad t = 0, \ldots, 11.$$

Then make a (12×14) matrix $\bar{D}$, having as columns the vectors $\boldsymbol{d}_i$ normalized to unit norm. D resembles a redundant DCT matrix.

b) construct the $(12^2 \times 14^2)$ dictionary Ψ according to $\Psi = D \bigotimes D$, where $\bigotimes$ denotes Kronecker product. Built in this way, the resulting atoms correspond to atoms related to the overcomplete 2D-DCT transform [49].

As a next step, de-noise each image patch separately. In particular, assuming that $\boldsymbol{y}_i$ is the ith patch reshaped in column vector, use the function "solvelasso.m"[6], or any other suitable algorithm you prefer, in order to estimate a sparse vector $\boldsymbol{\theta}_i$, $\in \mathbb{R}^{196}$ and obtain the corresponding de-noised vector as $\hat{\boldsymbol{y}}_i = \Psi\boldsymbol{\theta}_i$. Finally, average the values of the overlapped patches in order to form the full de-noised image.

[6] It can be found in the SparseLab MATLAB toolbox, which is freely available from http://sparselab.stanford.edu/

REFERENCES

[1] D. Achlioptas, Database-friendly random projections, in: Proceedings of the Symposium on Principles of Database Systems (PODS), ACM Press, 2001, pp. 274-281.

[2] A. Antoniadis, Wavelet methods in statistics: some recent developments and their applications, Stat. Surv. 1 (2007) 16-55.

[3] J. Arenas-Garcia, A.R. Figueiras-Vidal, Adaptive combination of proportionate filters for sparse echo cancellation, IEEE Trans. Audio Speech Language Process. 17(6) (2009) 1087-1098.

[4] S. Ariyavisitakul, N.R. Sollenberger, L.J. Greenstein, Tap-selectable decision feedback equalization, IEEE Trans. Commun. 45(12) (1997) 1498-1500.

[5] W.U. Bajwa, J. Haupt, A.M. Sayeed, R. Nowak, Compressed channel sensing: a new approach to estimating sparse multipath channels, Proc. IEEE 98(6) (2010) 1058-1076.

[6] R.G. Baraniuk, M. Davenport, R. DeVore, M.B. Wakin, A simple proof of the restricted isometry property for random matrices, Construct. Approximat. 28 (2008) 253-263.

[7] R. Baraniuk, M. Wakin, Random projections of smooth manifolds, Foundat. Comput. Math. 9(1) (2009) 51-77.

[8] R. Baraniuk, V. Cevher, M. Wakin, Low-dimensional models for dimensionality reduction and signal recovery: a geometric perspective, Proc. IEEE 98(6) (2010) 959-971.

[9] S. Becker, Practical compressed sensing: modern data acquisition and signal processing, Ph.D. thesis, Caltech, 2011.

[10] J. Benesty, T. Gansler, D.R. Morgan, M.M. Sondhi, S.L. Gay, Advances in Network and Acoustic Echo Cancellation, Springer-Verlag, Berlin, 2001.

[11] P. Bickel, Y. Ritov, A. Tsybakov, Simultaneous analysis of LASSO and Dantzig selector, Ann. Stat. 37(4) (2009) 1705-1732.

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

[13] A. Blum, Random projection, margins, kernels and feature selection, in: Lecture Notes on Computer Science (LNCS), 2006, pp. 52-68.

[14] S. Boyd, L. Vandenberghe, Convex Optimization, Cambridge University Press, 2004.

[15] A.M. Bruckstein, D.L. Donoho, M. Elad, From sparse solutions of systems of equations to sparse modeling of signals and images, SIAM Rev. 51(1) (2009) 34-81.

[16] T.T. Cai, G. Xu, J. Zhang, On recovery of sparse signals via ℓ_1 minimization, IEEE Trans. Informat. Theory 55(7) (2009) 3388-3397.

[17] R. Calderbank, S. Jeafarpour, R. Schapire, Compressed learning: Universal sparse dimensionality reduction and learning in the measurement domain, Tech. Rep., Rice University, 2009.

[18] E.J. Candès, J. Romberg, Practical signal recovery from random projections, in: Proceedings of the SPIE 17th Annual Symposium on Electronic Imaging, Bellingham, WA, 2005.

[19] E.J. Candès, T. Tao, Decoding by linear programming, IEEE Trans. Informat. Theory 51(12) (2005) 4203-4215.

[20] E. Candès, J. Romberg, T. Tao, Robust uncertainty principles: exact signal reconstruction from highly incomplete Fourier information, IEEE Trans. Informat. Theory 52(2) (2006) 489-509.

[21] E. Candès, T. Tao, Near optimal signal recovery from random projections: Universal encoding strategies, IEEE Trans. Informat. Theory 52(12) (2006) 5406-5425.

[22] E.J. Candès, J. Romberg, T. Tao, Stable recovery from incomplete and inaccurate measurements, Commun. Pure Appl. Math. 59(8) (2006) 1207-1223.

[23] E.J. Candès, M.B. Wakin, An introduction to compressive sampling, IEEE Signal Process. Mag. 25(2) (2008) 21-30.

[24] E.J. Candès, Y.C. Eldar, D. Needell, P. Randall, Compressed sensing with coherent and redundant dictionaries, Appl. Comput. Harmonic Anal. 31(1) (2011) 59-73.

[25] F. Chen, A.P. Chandrakasan, V.M. Stojanovic, Design and analysis of hardware efficient compressed sensing architectures for compression in wireless sensors, IEEE Trans. Solid State Circuits 47(3) (2012) 744-756.
[26] S. Chen, D.L. Donoho, M. Saunders, Atomic decomposition by basis pursuit, SIAM J. Sci. Comput. 20(1) (1998) 33-61.
[27] J.F. Claerbout, F. Muir, Robust modeling with erratic data, Geophysics 38(5) (1973) 826-844.
[28] K.L. Clarkson, D.P. Woodruff, Numerical linear algebra in the streaming model, in: Proceedings of the 41st annual ACM symposium on Theory of computing, ACM, 2009, pp. 205-214.
[29] K.L. Clarkson, D.P. Woodruff, Low rank approximation and regression in input sparsity time, in: Proceedings of the 45th annual ACM symposium on Symposium on theory of computing, ACM, 2013, pp. 81-90.
[30] A. Cohen, W. Dahmen, R. DeVore, Compressed sensing and best k-term approximation, J. Amer. Math. Soc 22(1) (2009) 211-231.
[31] R.R. Coifman, M.V. Wickerhauser, Entropy-based algorithms for best basis selection, IEEE Trans. Informat. Theory 38(2) (1992) 713-718.
[32] S.F. Cotter, B.D. Rao, Matching pursuit based decision-feedback equalizers, in: IEEE Conference on Acoustics, Speech and Signal Processing (ICASSP), Istanbul, Turkey, 2000.
[33] G.B. Dantzig, Linear Programming and Extensions, Princeton University Press, Princeton, NJ, 1963.
[34] S. Dasgupta, Experiments with random projections, in: Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence, Morgan-Kaufmann, San Francisco, CA, USA, 2000, pp. 143-151.
[35] I. Daubechies, Time-frequency localization operators: a geometric phase space approach, IEEE Trans. Informat. Theory 34(4) (1988) 605-612.
[36] D.L. Donoho, X. Huo, Uncertainty principles and ideal atomic decomposition, IEEE Trans. Informat. Theory 47(7) (2001) 2845-2862.
[37] D.L. Donoho, M. Elad, Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ_1 minimization, in: Proceedings of National Academy of Sciences, 2003, pp. 2197-2202.
[38] D.L. Donoho, J. Tanner, Counting faces of randomly projected polytopes when the projection radically lowers dimension, Tech. Rep. 2006-11, Stanford University, 2006.
[39] D.L. Donoho, J. Tanner, Precise undersampling theorems, Proc. IEEE 98(6) (2010) 913-924.
[40] D.L. Donoho, B.F. Logan, Signal recovery and the large sieve, SIAM J. Appl. Math. 52(2) (1992) 577-591.
[41] P.L. Dragotti, M. Vetterli, T. Blu, Sampling moments and reconstructing signals of finite rate of innovation: Shannon meets Strang-Fix, IEEE Trans. Signal Process. 55(5) (2007) 1741-1757.
[42] P. Drineas, M.W. Mahoney, S. Muthukrishnan,T. Sarlós, Faster least squares approximation, Numer. Math. 117(2) (2011) 219-249.
[43] M.F. Duarte, Y. Eldar, Structured compressed sensing: from theory to applications, IEEE Trans. Signal Process. 59(9) (2011) 4053-4085.
[44] M.F. Duarte, R.G. Baraniuk, Kronecker compressive sensing, IEEE Trans. Image Process. 21(2) (2012) 494-504.
[45] M.F. Duarte, R.G. Baraniuk, Spectral compressive sensing, Appl. Comput. Harmonic Anal. 35(1) (2013) 111-129.
[46] D. Eiwen, G. Taubock, F. Hlawatsch, H.G. Feichtinger, Group sparsity methods for compressive channel estimation in doubly dispersive multicarrier systems, in: Proceedings IEEE SPAWC, Marrakech, Morocco, June 2010.
[47] D. Eiwen, G. Taubock, F. Hlawatsch, H. Rauhut, N. Czink, Multichannel-compressive estimation of doubly selective channels in MIMO-OFDM systems: Exploiting and enhancing joint sparsity, in: Proceedings International Conference on Acoustics, Speech and Signal Processing (ICASSP), Dallas, TX, 2010.
[48] M. Elad, A.M. Bruckstein, A generalized uncertainty principle and sparse representations in pairs of bases, IEEE Trans. Informat. Theory 48(9) (2002) 2558-2567.

[49] M. Elad, Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing, Springer, 2010.
[50] Y.C. Eldar, G. Kutyniok, Compressed Sensing: Theory and Applications, Cambridge University Press, 2012.
[51] Y.C. Eldar, Sampling Theory: Beyond Bandlimited Systems, Cambridge University Press, 2014.
[52] M. Ghosh, Blind decision feedback equalization for terrestrial television receivers, Proc. IEEE 86(10) (1998) 2070-2081.
[53] I.F. Gorodnitsky, B.D. Rao, Sparse signal reconstruction from limited data using FOCUSS: a re-weighted minimum norm algorithm, IEEE Trans. Signal Process. 45(3) (1997) 600-614.
[54] R. Gribonval, M. Nielsen, Sparse decompositions in unions of bases, IEEE Trans. Informat. Theory 49(12) (2003) 3320-3325.
[55] N. Halko, P.G. Martinsson, J.A. Tropp, Finding structure with randomness: probabilistic algorithms for constructing approximate matrix decompositions, SIAM Rev. 53(2) (2011) 217-288.
[56] J. Haupt, W.U. Bajwa, G. Raz, R. Nowak, Toeplitz compressed sensing matrices with applications to sparse channel estimation, IEEE Trans. Informat. Theory 56(11) (2010) 5862-5875.
[57] R.A. Horn, C.R. Johnson, Matrix Analysis, Cambridge University Press, New York, 1985.
[58] S. Kirolos, J.N. Laska, M.B. Wakin, M.F. Duarte, D. Baron, T. Ragheb, Y. Massoud, R.G. Baraniuk, Analog to information conversion via random demodulation, in: Proceedings of the IEEE Dallas/CAS Workshop on Design, Applications, Integration and Software, Dallas, USA, 2006, pp. 71-74.
[59] M. Lexa, M. Davies, J. Thompson, Reconciling compressive sampling systems for spectrally sparse continuous-time signals, IEEE Trans. Signal Process. 60(1) (2012) 155-171.
[60] S. Lin, D.C. Constello Jr., Error Control Coding: Fundamentals and Applications, Prentice Hall, 1983.
[61] Y.-P. Lin, P.P. Vaidyanathan, Periodically nonuniform sampling of bandpass signals, IEEE Trans. Circuits Syst. II 45(3) (1998) 340-351.
[62] Y.M. Lu, M.N. Do, Sampling signals from a union of subspaces, IEEE Signal Process. Mag. 25(2) (2008) 41-47.
[63] P. Maechler, N. Felber, H. Kaeslin, A. Burg, Hardware-efficient random sampling of Fourier-sparse signals, in: Proceedings of the IEEE International Symposium on Circuits and Systems (ISCAS), 2012.
[64] A. Maleki, L. Anitori, Z. Yang, R. Baraniuk, Asymptotic analysis of complex LASSO via complex approximate message passing (CAMP), IEEE Trans. Informat. Theory, 59(7) (2013) 4290-4308.
[65] S. Mallat, S. Zhang, Matching pursuit in a time-frequency dictionary, IEEE Trans. Signal Process. 41 (1993) 3397-3415.
[66] E. Matusiak, Y.C. Eldar, Sub-Nyquist sampling of short pulses, IEEE Trans. Signal Process. 60(3) (2012) 1134-1148.
[67] S. Mendelson, A. Pajor, N. Tomczak-Jaegermann, Uniform uncertainty principle for Bernoulli and sub-Gaussian ensembles, Construct. Approximat. 28 (2008) 277-289.
[68] M. Mishali, Y.C. Eldar, A. Elron, Xampling: analog data compression, in: Proceedings Data Compression Conference, Snowbird, Utah, USA, 2010.
[69] M. Mishali, Y. Eldar, From theory to practice: sub-Nyquist sampling of sparse wideband analog signals, IEEE J. Selected Topics Signal Process. 4(2) (2010) 375-391.
[70] M. Mishali, Y.C. Eldar, Sub-Nyquist sampling, IEEE Signal Process. Mag. 28(6) (2011) 98-124.
[71] M. Mishali, Y.C. Eldar, A. Elron, Xampling: signal acquisition and processing in union of subspaces, IEEE Trans. Signal Process. 59(10) (2011) 4719-4734.
[72] B.K. Natarajan, Sparse approximate solutions to linear systems, SIAM J. Comput. 24 (1995) 227-234.
[73] P.A. Naylor, J. Cui, M. Brookes, Adaptive algorithms for sparse echo cancellation, Signal Process. 86 (2004) 1182-1192.

[74] A.M. Pinkus, On ℓ_1-Approximation, Cambridge Tracts in Mathematics, vol. 93, Cambridge University Press, 1989.
[75] Q. Qiu, V.M. Patel, P. Turaga, R. Chellappa, Domain adaptive dictionary learning, in: Proceedings of the European Conference on Computer Vision (ECCV), Florence, Italy, 2012.
[76] Y. Rivenson, A. Stern, Compressed imaging with a separable sensing operator, IEEE Signal Process. Lett. 16(6) (2009) 449-452.
[77] A. Rondogiannis, K. Berberidis, Efficient decision feedback equalization for sparse wireless channels, IEEE Trans. Wireless Commun. 2(3) (2003) 570-581.
[78] R. Rubinstein, A. Bruckstein, M. Elad, Dictionaries for sparse representation modeling, Proceed. IEEE 98(6) (2010) 1045-1057.
[79] M. Rudelson, R. Vershynin, On sparse reconstruction from Fourier and Gaussian measurements, Commun. Pure Appl. Math. 61(8) (2008) 1025-1045.
[80] F. Santosa, W.W. Symes, Linear inversion of band limited reflection seismograms, SIAM J. Sci. Comput. 7(4) (1986) 1307-1330.
[81] T. Sarlos, Improved approximation algorithms for large matrices via random projections, in: Foundations of Computer Science, 2006. FOCS'06. 47th Annual IEEE Symposium on, IEEE, 2006, pp. 143-152.
[82] P. Saurabh, C. Boutsidis, M. Magdon-Ismail, P. Drineas, Random projections for support vector machines, in: Proceedings 16th International Conference on Artificial Intelligence and Statistics (AISTATS) Scottsdale, AZ, USA, 2013.
[83] D. Takhar, V. Bansal, M. Wakin, M. Duarte, D. Baron, K.F. Kelly, R.G. Baraniuk, A compressed sensing camera: New theory and an implementation using digital micromirrors, in: Proceedings on Computational Imaging (SPIE), San Jose, CA, 2006.
[84] G. Tang, A. Nehorai, Performance analysis of sparse recovery based on constrained minimal singular values, IEEE Trans. Signal Process. 59(12) (2011) 5734-5745.
[85] H.L. Taylor, S.C. Banks, J.F. McCoy, Deconvolution with the ℓ_1 norm, Geophysics 44(1) (1979) 39-52.
[86] S. Theodoridis, K. Koutroumbas, Pattern Recognition, fourth ed., Academic Press, 2009.
[87] Z. Tian, G.B. Giannakis, Compressed sensing for wideband cognitive radios, in: Proceedings of the IEEE Conference on Acoustics, Speech and Signal Processing (ICASSP), 2007, pp. 1357-1360.
[88] R. Tibshirani, Regression shrinkage and selection via the LASSO, J. Royal. Statist. Soc. B. 58(1) (1996) 267-288.
[89] I. Tosić, P. Frossard, Dictionary learning, IEEE Signal Process. Mag. 28(2) (2011) 27-38.
[90] J.A. Tropp, J.N. Laska, M.F. Duarte, J.K. Romberg, G. Baraniuk, Beyond Nyquist: efficient sampling of sparse bandlimited signals, IEEE Trans. Informat. Theory 56(1) (2010) 520-544.
[91] M. Unser, Sampling: 50 years after Shannon, Proc. IEEE 88(4) (2000) 569-587.
[92] R.G. Vaughan, N.L. Scott, D.R. White, The theory of bandpass sampling, IEEE Trans. Signal Process. 39(9) (1991) 1973-1984.
[93] R. Venkataramani, Y. Bresler, Perfect reconstruction formulas and bounds on aliasing error in sub-Nyquist nonuniform sampling of multiband signals, IEEE Trans. Informat. Theory 46(6) (2000) 2173-2183.
[94] M. Vetterli, P. Marzilliano, T. Blu, Sampling signals with finite rate of innovation, IEEE Trans. Signal Process. 50(6) (2002) 1417-1428.
[95] M. Wakin, Manifold-based signal recovery and parameter estimation from compressive measurements, 2008. preprint: http://arxiv.org/abs/1002.1247.
[96] M. Wakin, S. Becker, E. Nakamura, M. Grant, E. Sovero, D. Ching, J. Yoo, J. Romberg, A. Emami-Neyestanak, E. Candes, A non-uniform sampler for wideband spectrally-sparse environments, IEEE Trans. on Emerging and Selected Topics in Circuits and Systems 2(3) (2012) 516-529.
[97] L.R. Welch, Lower bounds on the maximum cross correlation of signals, IEEE Trans. Informat. Theory 20(3) (1974) 397-399.

[98] S. Wright, R. Nowak, M. Figueiredo, Sparse reconstruction by separable approximation, IEEE Trans. Signal Process. 57(7) (2009) 2479-2493.
[99] M. Yaghoobi, L. Daudet, M. Davies Parametric dictionary design for sparse coding, IEEE Trans. Signal Process. 57(12) (2009) 4800-4810.
[100] Y. Ye, Interior Point Methods: Theory and Analysis, Wiley, New York, 1997.
[101] J. Yoo, S. Becker, M. Monge, M. Loh, E. Candès, A. Emami-Neyestanak, Design and implementation of a fully integrated compressed-sensing signal acquisition system, in: 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), March 2012, pp. 5325-5328.
[102] Z. Yu, S. Hoyos, B.M. Sadler, Mixed-signal parallel compressed sensing and reception for cognitive radio, in: Proceedings IEEE Conference on Acoustics, Speech and Signal Processing (ICASSP), 2008, pp. 3861-3864.