Sparse Representations: Theory and Applications

Brendt Wohlberg

brendt@lanl.gov

November 5, 2014

Contents

Chapter 1

Introduction

Over the past decade, sparse representations have become an important tool in signal and image processing, and statistics, and have recently also been increasingly applied to computer vision problems. These methods now deliver state of the art performance in a wide variety of problems, including image restoration (denoising, deblurring, inpainting, etc.) and pattern recognition.

The fundamental notion of a sparse representation of a vector (representing a signal or image, for example) is simple: the signal can be exactly or approximately represented as a linear combination of only a few vectors selected from a predetermined *dictionary* set of vectors. Reconstruction of a signal from its sparse representation is a simple linear transform, but the inverse problem is a non-linear optimisation. It is not easy to give a simple explanation of why these methods are so effective, but it is reasonable, at a high level, to view sparsity as a useful way of constraining the complexity of a signal representation, which can be very generally justified by Occam's Razor.

The initial chapters of this course are based on two main sources, the excellent survey [1] by Bruckstein, Donoho, and Elad, and the textbook by Elad [2]. Subsequent chapters are based on a wider range of sources, although the section on ADMM methods is based exclusively on the very useful tutorial by Boyd et al. [3].

1.1 Linear Inverse Problems and Regularization

One of the most fundamental operations on signals represented as vectors is the determination of the magnitude, or the distance between two signals, which is usually defined as the magnitude of the difference between the two signals. The function that defines the magnitude of a vector is a norm (see Definition A.1). The most frequently used norm is the ℓ^2 (or Euclidean) norm

$$\|\mathbf{x}\|_2 = \sqrt{\sum_{k=0}^{N-1} x_k^2}$$

where $\mathbf{x} \in \mathbb{R}^N$. This norm corresponds to a scaled version of the RMSE (Root-Mean-Square Error), and is easily minimised since it is differentiable everywhere

$$\nabla_{\mathbf{x}} \|A\mathbf{x} - \mathbf{b}\|_2^2 = A^T A\mathbf{x} - A^T\mathbf{b} \; .$$

The ℓ^2 norm is a special case of the ℓ^p norm

$$\|\mathbf{x}\|_p = \left(\sum_{k=0}^{N-1} |x_k|^p \right)^{\frac{1}{p}}$$

for $p \geq 1$ (when $p < 1$ this function is not a norm).

Subject to the model $\mathbf{s} = A\mathbf{x}$, with $\mathbf{s} \in \mathbb{R}^M$, $\mathbf{x} \in \mathbb{R}^N$, and $A \in \mathbb{R}^{M \times N}$, consider the inverse problem of determining $\mathbf{x}$ given $\mathbf{s}$. If the problem is overdetermined ($M > N$), there may not exist an $\mathbf{x}$ such that $\mathbf{s} = A\mathbf{x}$. A reasonable approach is to determine the $\mathbf{x}$ that is as close as possible (in ℓ^2 norm) to matching this requirement by solving the problem

$$\arg\min_{\mathbf{x}} \frac{1}{2} \|A\mathbf{x} - \mathbf{s}\|_2^2 = (A^T A)^{-1} A^T \mathbf{s} \; .$$

If $M < N$ or $\text{rank}(A) < M$ the problem is underdetermined since A has a non-trivial null-space, so there is an entire subspace of solutions satisfying the constraint $A\mathbf{x} = \mathbf{s}$. In this case, some additional criterion must be chosen to select a unique solution from the solution subspace. The simplest choice is the minimum ℓ^2 norm solution, corresponding to the optimization problem

$$\arg\min_{\mathbf{x}} \|\mathbf{x}\|_2^2 \text{ such that } A\mathbf{x} = \mathbf{s} \; .$$

The geometry of this approach is illustrated in Fig. 1.1. More generally, one can introduce an operator B (a derivative operator if one wished to penalise the gradient of the solution, for example)

$$\arg\min_{\mathbf{x}} \|\mathbf{B}\mathbf{x}\|_2^2 \text{ such that } A\mathbf{x} = \mathbf{s} \; .$$

Figure 1.1: Geometry of the problem $\arg\min_{\mathbf{x}} \|\mathbf{x}\|_2^2$ such that $A\mathbf{x} = \mathbf{s}$

Since real signals invariably have some noise component, a more realistic model for a linear forward problem is

$$\mathbf{s} = A\mathbf{x} + \boldsymbol{\nu} \; ,$$

where $\boldsymbol{\nu}$ represents an additive noise component. In the presence of noise it is no longer reasonable to impose the constraint $A\mathbf{x} = \mathbf{s}$, suggesting the inequality-constrained form

$$\arg\min_{\mathbf{x}} \|\mathbf{B}\mathbf{x}\|_2^2 \text{ such that } \|A\mathbf{x} - \mathbf{s}\|_2 < \epsilon \; ,$$

for known error bound ϵ (derived from knowledge of the noise). It is often more convenient to consider the penalised form

$$\arg\min_{\mathbf{x}} \frac{1}{2} \|A\mathbf{x} - \mathbf{s}\|_2^2 + \frac{\lambda}{2} \|B\mathbf{x}\|_2^2 = (A^T A + \lambda B^T B)^{-1} A^T \mathbf{s} \; ,$$

where λ is the *regularisation parameter*. This optimisation problem is known as *Tikhonov Regularisation*.

1.2 Efficient Solution of Linear Systems

While solution of a system of the form $A^T A\mathbf{x} = A^T\mathbf{s}$ is mathematically simple, it can be very demanding computationally for large A. If a fast implementation of the linear operator A is available (e.g. if it corresponds to a DCT or wavelet transform), or if A is too large to represent as an explicit matrix, an iterative solver such as Conjugate Gradient (CG) [4, Ch. 10] is appropriate. (Note, though, that an implementation of the operator A^T is also required.)

In some problems we are forced to work with an explicit matrix representation of A, and there are algorithms that require repeated solution of a linear system with the same left hand side and varying right hand side. In such cases, it may be more efficient to pay a high cost upfront to decompose A, and then solve cheaply via Gaussian elimination. Given an LU decomposition [4, Sec. 3.2] $A = LU$, we can solve $A\mathbf{x} = \mathbf{b}$ as follows (Matlab notation in square brackets):

$$\begin{aligned} LU\mathbf{x} &= \mathbf{b} & \\ L(U\mathbf{x}) &= \mathbf{b} & \\ L\mathbf{y} &= \mathbf{b} & [\mathbf{y} = L\backslash\mathbf{b}] \\ U\mathbf{x} &= \mathbf{y} & [\mathbf{x} = U\backslash\mathbf{y}] \\ & & [\mathbf{x} = U\backslash(L\backslash\mathbf{b})] \ . \end{aligned}$$

In many of the algorithms we will encounter, we will need to solve problems of the form $(A^T A + \lambda I)\mathbf{x} = \mathbf{s}$. If $A \in \mathbb{R}^{M\times N}$ with $M < N$, we would rather deal with the smaller linear system $(AA^T + \lambda I)$. This can be achieved by applying the *matrix inversion lemma* or *Woodbury matrix identity*

$$(B + UCV)^{-1} = B^{-1} - B^{-1}U\left(C^{-1} + VB^{-1}U\right)^{-1}VB^{-1} \ .$$

Substituting $B = \lambda I$, $C = I$, $U = A^T$, and $V = A$, we have

$$\begin{aligned} \left(\lambda I + A^T A\right)^{-1} &= \lambda^{-1}I - \lambda^{-1}IA^T\left(I + A\lambda^{-1}IA^T\right)^{-1}A\lambda^{-1}I \\ &= \lambda^{-1}I - \lambda^{-2}A^T\left(I + \lambda^{-1}AA^T\right)^{-1}A \\ &= \lambda^{-1}I - \lambda^{-1}A^T\left(\lambda I + AA^T\right)^{-1}A \end{aligned}$$

Now, to solve $\left(A^T A + \lambda I\right)\mathbf{u} = \mathbf{b}$,

$$\begin{aligned} \mathbf{u} &= \left(A^T A + \lambda I\right)^{-1}\mathbf{b} \\ &= \left(\lambda^{-1}I - \lambda^{-1}A^T\left(\lambda I + AA^T\right)^{-1}A\right)\mathbf{b} \\ &= \lambda^{-1}\mathbf{b} - \lambda^{-1}A^T\left(\lambda I + AA^T\right)^{-1}A\mathbf{b} \end{aligned}$$

Define $C = \left(\lambda I + AA^T\right)$ so that

$$\mathbf{u} = \lambda^{-1}\mathbf{b} - \lambda^{-1}A^T C^{-1}A\mathbf{b}$$

We want to compute $\mathbf{x} = C^{-1}A\mathbf{b}$ given the LU decomposition $C = LU$,

$$\begin{aligned} C\mathbf{x} &= A\mathbf{b} \\ LU\mathbf{x} &= A\mathbf{b} \\ L(U\mathbf{x}) &= A\mathbf{b} \\ L\mathbf{y} &= A\mathbf{b} & [\mathbf{y} = L\backslash A\mathbf{b}] \\ U\mathbf{x} &= \mathbf{y} & [\mathbf{x} = U\backslash\mathbf{y}] \\ & & [\mathbf{x} = U\backslash(L\backslash A\mathbf{b})] \end{aligned}$$

then

$$\mathbf{u} = \lambda^{-1}\mathbf{b} - \lambda^{-1}A^T\mathbf{x} \, .$$

1.3 Maximum a Posteriori Estimation

It can be shown that solutions based on such regularisation functionals are equivalent to maximum a posteriori (MAP) estimation in a Bayesian statistical estimation framework (see, for example, [1, Sec. 4.1], [5, 6]). If

$$\mathbf{s} = A\mathbf{x} + \boldsymbol{\nu} \, ,$$

where $\mathbf{s}$, A, $\mathbf{x}$, and $\boldsymbol{\nu}$ are, respectively, the measured signal, the forward operator, the data to be recovered from the inverse problem, and additive noise, then the MAP estimate of $\mathbf{x}$ given $\mathbf{s}$ is

$$\hat{\mathbf{x}} = \arg\max_{\mathbf{x}} p(\mathbf{x}|\mathbf{s})$$

where p is the conditional pdf of $\mathbf{x}$ given $\mathbf{s}$. From Bayes' Theorem

$$\hat{\mathbf{x}} = \arg\max_{\mathbf{x}} \frac{p(\mathbf{s}|\mathbf{x})p(\mathbf{x})}{p(\mathbf{s})} \, ,$$

and since the location of the maximum does not depend on $p(\mathbf{s})$

$$\begin{aligned} \hat{\mathbf{x}} &= \arg\max_{\mathbf{x}} p(\mathbf{s}|\mathbf{x})p(\mathbf{x}) \\ &= \arg\min_{\mathbf{x}} -\log(p(\mathbf{s}|\mathbf{x})) - \log(p(\mathbf{x})) \, . \end{aligned}$$

We choose the following general forms of distribution (which both include a multivariate Gaussian as the special cases $p = 2, q = 2$) for the noise and for $\mathbf{x}$

$$\begin{aligned} p(\boldsymbol{\nu}) &= b \exp\left(-\beta \left\|B\boldsymbol{\nu}\right\|_p^p\right) \\ p(\mathbf{x}) &= c \exp\left(-\gamma \left\|C\mathbf{x}\right\|_q^q\right) \, . \end{aligned}$$

Now, since $\mathbf{s} = A\mathbf{x} + \boldsymbol{\nu}$ and therefore $\boldsymbol{\nu} = \mathbf{s} - A\mathbf{x}$, the conditional distribution of $\mathbf{s}$ given $\mathbf{x}$ is simply the distribution of $\boldsymbol{\nu}$ shifted by the deterministic quantity $-A\mathbf{x}$, giving

$$p(\mathbf{s}|\mathbf{x}) = b \exp\left(-\beta \left\|B(\mathbf{s} - A\mathbf{x})\right\|_p^p\right) \, .$$

Substituting in the equation above, we have

$$\begin{aligned}\hat{\mathbf{x}} &= \arg\min_{\mathbf{x}} -\log(p(\mathbf{s}|\mathbf{x})) - \log(p(\mathbf{x})) \\ &= \arg\min_{\mathbf{x}} -\log(b) - \left(-\beta \left\|B(\mathbf{s}-A\mathbf{x})\right\|_p^p\right) - \log(c) - \left(-\gamma \left\|C\mathbf{x}\right\|_q^q\right) \\ &= \arg\min_{\mathbf{x}} \beta \left\|B(\mathbf{s}-A\mathbf{x})\right\|_p^p + \gamma \left\|C\mathbf{x}\right\|_q^q \ .\end{aligned}$$

This establishes a simple correspondence between MAP estimates in a Bayesian framework and regularisation methods within a deterministic framework.

1.4 Convex Functions

The Tikhonov regularisation solution can be written in closed form, but that is not the case for the more general minimisation problems we will consider here. Convexity of the function to be minimised is a very desirable property.

A set $S \in \mathbb{R}^N$ is *convex* if the line segment connecting any pair of points in S itself lies entirely within S [7, Ch. 1]. Formally,

$$S \text{ is convex } \Leftrightarrow \ \alpha\mathbf{x} + (1-\alpha)\mathbf{y} \in S \ \ \forall \mathbf{x}, \mathbf{y} \in S, \alpha \in [0,1]$$

A function f is convex if its domain S is a convex set and

$$f(\alpha\mathbf{x} + (1-\alpha)\mathbf{y}) \leq \alpha f(\mathbf{x}) + (1-\alpha) f(\mathbf{y}) \ \ \forall \mathbf{x}, \mathbf{y} \in S, \alpha \in [0,1]$$

If a minimisation problem has a convex objective function and the feasible region is a convex set, then it is a *convex problem*, and any local solution is also a global solution [8, Sec. 4.2.2]:

Theorem 1.1. *Any local solution of a convex problem with a convex objective function and convex feasible region is also a global solution.*

Proof. If $\mathbf{x}$ is a local optimum of the problem $\arg\min_{\mathbf{x}} f(\mathbf{x})$ such that $\mathbf{x} \in C$ then there exists an ϵ such that

$$f(\mathbf{x}) = \inf_{\mathbf{x}} \left\{ f(\mathbf{t}) \mid \mathbf{t} \in C \text{ and } \left\|\mathbf{x}-\mathbf{t}\right\|_2 \leq \epsilon \right\} \ .$$

If $\mathbf{x}$ is not also a global solution, then there exists a $\mathbf{y}$ such that $f(\mathbf{y}) < f(\mathbf{x})$. In this case, however, we can construct a point $\mathbf{z} = \alpha\mathbf{y} + (1-\alpha)\mathbf{x}$, with $\alpha = \frac{\epsilon}{2\left\|\mathbf{x}-\mathbf{y}\right\|_2}$; since C is convex, $\mathbf{z} \in C$, and from the convexity of f we have

$$f(\mathbf{z}) \leq \alpha f(\mathbf{y}) + (1-\alpha) f(\mathbf{x}) < f(\mathbf{x}) \ ,$$

which contradicts the claim that $\mathbf{x}$ is locally optimal. ■

1.5 Sparse Representations

We say that $\mathbf{s} \in \mathbb{R}^M$ has a *sparse representation* $\mathbf{x} \in \mathbb{R}^N$ with respect to *dictionary* $A \in \mathbb{R}^{M \times N}$ if $A\mathbf{x} = \mathbf{s}$ and most of the elements of $\mathbf{x}$ are zero. The sparsity of $\mathbf{x}$ is measured by the ℓ^0 "norm" (it is not a true norm)

$$\|\mathbf{x}\|_0 = n(\{k|x_k \neq 0\}) \ .$$

where $n(\cdot)$ denotes the cardinality of a set. (Note that $\|\mathbf{x}\|_0 = \lim_{p \to 0} \|\mathbf{x}\|_p^p$.)

This type of representation has largely independent origins in statistics [9] and signal processing [10, 11]. The use in statistics as a regression tool is motivated as a mechanism to control overfitting and improve interpretability, while the ability to construct adaptive signal representations using overcomplete time-frequency dictionaries was a major motivation in early signal processing applications. A more recent perspective is that sparse representations represent a useful method for constructing signal models; while there is no simple explanation for their effectiveness, at a high level, sparsity can be considered a measure of simplicity, comparable with entropy. Early sparse representations made use of analytically defined dictionaries (often with a fast implementation of the corresponding linear operator) such as DCT or wavelets, but in modern applications the dictionaries are often *learned* from the data of interest.

Chapter 2

Minimum ℓ^0 Norm: Theory

As discussed in the Introduction, sparsity is defined in terms of the l^0 "norm"

$$\|\mathbf{x}\|_0 = n(\{k|x_k \neq 0\}) \ .$$

In this chapter we consider inverse problems involving this "norm", starting with the main theoretical results before moving on to some of the leading algorithms.

2.1 Uniqueness Results

We start by considering the ideal case in which our data has an exact sparse representation, and define the inverse problem, of determining the sparsest possible representation for a given data vector:

Definition 2.1. *Problem $\mathcal{P}_0$ is defined as*

$$\arg\min_{\mathbf{x}} \|\mathbf{x}\|_0 \ \ \textit{such that}\ A\mathbf{x} = \mathbf{s} \ . \tag{2.1}$$

Once we have defined this problem, the immediate questions are (i) how can we compute the sparsest representation, and (ii) under what conditions is a sparse solution unique. We address the second, theoretical question first.

In the standard (i.e. without a sparsity constraint) case, the linear system $A\mathbf{x} = \mathbf{s}$ has a unique solution if $\mathbf{s}$ is in the range space of A, and A has a trivial null-space. When we introduce a sparsity constraint, we are interested in the behaviour of subsets of columns of A, since the sparsity constraint implies that $\mathbf{s}$ can be represented using only a few of these columns. For convenience, we introduce the following notation

Definition 2.2. *$\Omega_{N,L}$ is the set of all*

$$\left(\begin{array}{c} N \\ L \end{array} \right) = \frac{N!}{L!(N-L)!}$$

distinct index subsets of size L for a dictionary of N atoms, i.e.

$$\Omega_{N,L} = \{(\omega_0, \omega_1, \ldots, \omega_{L-1}) \mid \omega_k \in \mathbb{N},\ 0 \leq \omega_k \leq N-1,\ \omega_k < \omega_{k+1}\}\ .$$

and

Definition 2.3. *$P_{\boldsymbol{\omega}}$ as the $N \times L$ matrix such that the product $AP_{\boldsymbol{\omega}}$ consists of the columns of A indexed by $\boldsymbol{\omega}$.*

Assume that A is such that any cardinality L subset of columns is linearly independent. It is easily shown [12] that if $A\mathbf{x} = \mathbf{s}$ has a solution with $\|\mathbf{x}\|_0 \leq L/2$ then that solution is the only solution with that property: if there are two such solutions $\mathbf{x}$ and $\mathbf{y}$, then $\mathbf{x} - \mathbf{y}$ lies in the null space of A, and $\|\mathbf{x} - \mathbf{y}\|_0 \leq L$, violating the assumed properties of A.

A slightly more general form of this uniqueness result is defined in terms of the *spark* of a matrix [1, Sec. 2.1.1] [2, Sec. 2.2.1] [13]

Definition 2.4. $\text{spark}(A)$ *is the cardinality of the smallest subset of columns of A that are linearly dependent, i.e.*

$$\text{spark}(A) = \min\{L \in \{1, \ldots, N\} \mid \exists \boldsymbol{\omega} \in \Omega_{M,L} \text{ s.t. } AP_{\boldsymbol{\omega}} \text{ has linearly dependent columns}\}\ .$$

Some properties of spark worth noting are [13]

- $A\mathbf{x} = 0 \quad \Rightarrow \quad \|\mathbf{x}\|_0 \geq \text{spark}(A)$ since $A\mathbf{x} = 0$ implies that $\mathbf{x}$ is non-zero on a linearly-dependent subset of the columns of A.
- Since a subset consisting of a single column cannot be linearly dependent (assuming that there are no zero columns), $\text{spark}(A) \geq 2$.
- If $A \in \mathbb{R}^{M \times N}$ then $\text{spark}(A) \leq \min\{N, \text{rank}(A) + 1\}$.

If the spark of a matrix is known, it is possible to determine that a sufficiently sparse solution with respect to that matrix is in fact the sparsest possible solution:

Theorem 2.1. *If the system $A\mathbf{x} = \mathbf{s}$ has a solution $\mathbf{x}$ such that $\|\mathbf{x}\|_0 < \text{spark}(A)/2$, then $\mathbf{x}$ is the sparsest possible solution to that system [1, Sec. 2.1.1].*

Proof. In a variation of the argument above, consider an alternative solution $\mathbf{y}$ to the linear system. Since $\mathbf{x} - \mathbf{y}$ lies in the null space of A, $\|\mathbf{x} - \mathbf{y}\|_0 \geq \text{spark}(A)$ by the definition of the spark. But $\|\mathbf{x}\|_0 + \|\mathbf{y}\|_0 \geq \|\mathbf{x} - \mathbf{y}\|_0$, and since $\|\mathbf{x}\|_0 < \text{spark}(A)/2$, it must be the case that $\|\mathbf{y}\|_0 > \text{spark}(A)/2$. ■

Unfortunately, computing the spark involves a combinatorial optimisation that is only practical for small matrices [13]. The *mutual coherence* [1, Sec. 2.1.2] [2, Sec. 2.2.2] provides an alternative approach to characterising uniqueness of a sparse representation.

Definition 2.5. *The mutual coherence of A is the maximum of the normalised inner products between any two columns of A; i.e.*

$$\mu(A) = \max_{k,l\ k\neq l} \frac{|\langle \mathbf{a}_k, \mathbf{a}_l \rangle|}{\|\mathbf{a}_k\|_2 \|\mathbf{a}_l\|_2}$$

where $\mathbf{a}_k$ is column k of A.

Some properties of the mutual coherence worth noting are [1, Sec. 2.1.2][14]

- For full-rank matrices $A \in \mathbb{R}^{M\times N}$,

$$\mu(A) \geq \sqrt{\frac{N-M}{M(N-1)}}$$

- For structured matrices $A \in \mathbb{R}^{M\times 2M}$, $A = [A_0\ A_1]$ where A_0 and A_1 are orthogonal matrices,

$$\frac{1}{\sqrt{M}} \leq \mu(A) \leq 1$$

The mutual coherence provides a lower bound for the spark [1, Sec. 2.1.2]:

Theorem 2.2. *For any matrix $A \in \mathbb{R}^{M\times N}$*

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

Proof. First, without loss of generality we can assume that the columns of A are normalised in ℓ^2, since this operation preserves both the spark and the mutual coherence. Now choose any subset $\boldsymbol{\omega} \in \Omega_{M,L}$ of cardinality L from the columns of A, define B as the resulting matrix (i.e. $B = AP_{\boldsymbol{\omega}}$), and define $G = B^T B$. From the normalisation of columns of A, $G_{kk} = 1$, and from the definition of mutual coherence, $|G_{kl}| \leq \mu(B) \leq \mu(A)$, which implies that $\sum_{l\neq k} |G_{kl}| \leq (L-1)\mu(A)$. Choosing $L < 1 + \frac{1}{\mu(A)}$ implies that $\sum_{l\neq k} |G_{kl}| < 1 = |G_{kk}|$, and G is therefore strictly diagonally dominant (see Definition A.3). Given the diagonal dominance and positivity of G_{kk}, G must be positive definite (see Theorem A.3), which in turn implies that the columns of B are linearly independent. Since $L < 1+\frac{1}{\mu(A)}$ implies that all L cardinality subsets of columns of A are linearly independent, it is only possible for a subset of columns with $L \geq 1+\frac{1}{\mu(A)}$ to be linearly dependent, and therefore $\text{spark}(A) \geq 1+\frac{1}{\mu(A)}$. ■

Unfortunately, unless $\mu(A)$ is small, this bound is not particularly informative, as illustrated by Fig. 2.1.

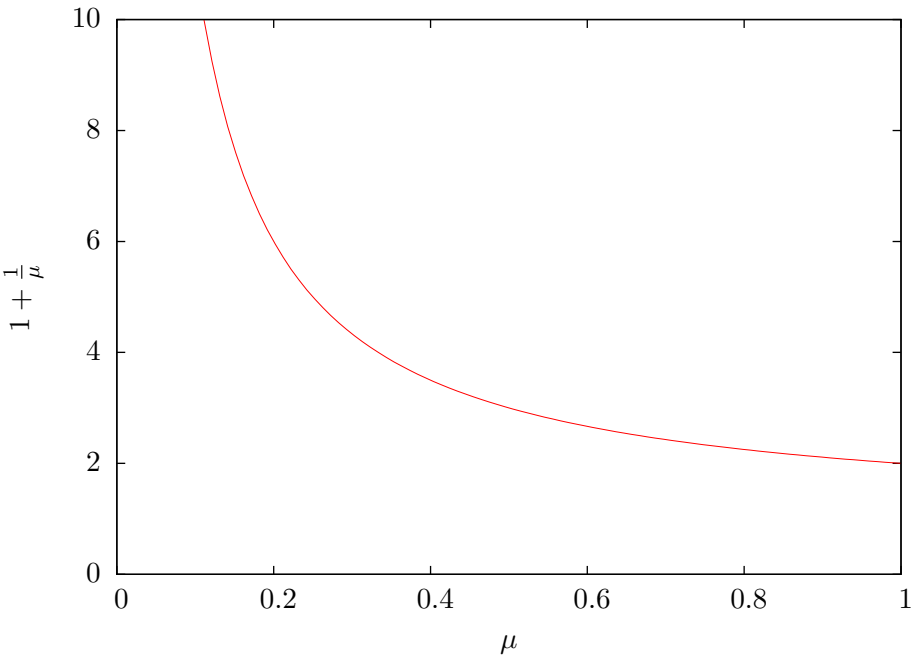


Figure 2.1: Lower bound for spark(A) plotted against $\mu(A)$

.

2.2 Stability Results

Thus far we have discussed conditions under which the linear system $A\mathbf{x} = \mathbf{s}$ has unique solutions. In any practical signal processing context, however, such exact solutions are never encountered, and it is far more useful to include noise in the model, leading to the system $A\mathbf{x}+\boldsymbol{\nu} = \mathbf{s}$, where $\boldsymbol{\nu}$ represents a noise component, so that $\mathbf{s}$ no longer has an exact representation on dictionary A. The relevant inverse problem is now:

Definition 2.6. *Define problem $\mathcal{P}_0^\epsilon$ as [2, Ch. 5]*

$$\arg\min_{\mathbf{x}} \|\mathbf{x}\|_0 \ \textit{such that}\ \|A\mathbf{x} - \mathbf{s}\|_2 \leq \epsilon\ . \tag{2.2}$$

Instead of uniqueness conditions, we will examine the stability of solution $\mathbf{x}$ given an upper bound on the noise magnitude $\|\boldsymbol{\nu}\|_2$, that is, we would like to bound the distance $\|\mathbf{x}-\mathbf{y}\|_2$ between two possible sparse solutions $\mathbf{x}$ and $\mathbf{y}$ given an upper bound on $\|\boldsymbol{\nu}\|_2$. This stability depends on the behaviour of subsets of columns (since we are considering sparse solutions) of A, which is easily characterised in terms of the Singular Value Decomposition (SVD).

Recall that the SVD of $M \times N$ matrix A is $A = U\Sigma V^T$, where U is an orthonormal $N \times N$ matrix, the columns $\mathbf{u}_k$ of which are the *left singular vectors*, V is an orthonormal $M \times M$ matrix, the columns $\mathbf{v}_k$ of which are the *right singular vectors*, and Σ is a diagonal matrix of *singular values* σ_k, $0 \leq k \leq \min\{M, N\} - 1$, ordered so that $\sigma_k \geq \sigma_{k+1}$. The maximum and minimum singular value of A are denoted as $\sigma_{\max}(A)$ and $\sigma_{\min}(A)$ respectively. Geometrically,

the singular values are the lengths of the semi-axes of the hyperellipsoid constructed as the mapping by A of the unit hypersphere in the domain space of A. Conversely, the inverses of the singular values define a hyperellipsoid in the domain space of A as the pre-image of the unit sphere in its range space.

It is easy to show (consider the constrained problems $\max_{\mathbf{u}} \|A\mathbf{u}\|_2^2$ such that $\mathbf{u} = 1$ and $\min_{\mathbf{u}} \|A\mathbf{u}\|_2^2$ such that $\mathbf{u} = 1$, and show that the extrema of the Lagrangian occur at the right singular vectors of A) that for any $A \in \mathbb{R}^{M \times N}$

$$\sigma_{\max}(A) = \sup_{\mathbf{u} \neq 0} \frac{\|A\mathbf{u}\|_2}{\|\mathbf{u}\|_2} = \max_{\|\mathbf{u}\|_2 = 1} \|A\mathbf{u}\|_2 \ ,$$

and

$$\sigma_{\min}(A) = \inf_{\mathbf{u} \neq 0} \frac{\|A\mathbf{u}\|_2}{\|\mathbf{u}\|_2} = \min_{\|\mathbf{u}\|_2 = 1} \|A\mathbf{u}\|_2 \ ,$$

so that

$$\sigma_{\min}(A) \|\mathbf{x}\|_2 \leq \|A\mathbf{x}\|_2 \leq \sigma_{\max}(A) \|\mathbf{x}\|_2 \ .$$

(Note also that $\|A\|_F^2 = \sum_k \sigma_k^2(A)$ and $\|A\|_2 = \max_{\|\mathbf{x}\|_2 = 1} \|A\mathbf{x}\|_2 = \sigma_{\max}(A)$ [4, pg. 55, pp. 70-73].) We can now define a quantity that immediately provides a bound on the stability of the sparse inverse problem involving dictionary A:

Definition 2.7. *$\zeta_L(A)$ is the smallest $\sigma_{\min}$ for all subsets of L columns selected from A, i.e.*

$$\zeta_L(A) = \min_{\boldsymbol{\omega} \in \Omega_{M,L}} \sigma_{\min}(AP_{\boldsymbol{\omega}}) \ .$$

2.2.1 Bounds based on $\zeta_L(A)$

Given the role of $\sigma_{\min}(\cdot)$ in the lower bound, we immediately have

$$\|A\mathbf{x}\|_2 \geq \zeta_L(A) \|\mathbf{x}\|_2 \quad \forall \mathbf{x} \in \mathbb{R}^N \text{ such that } \|\mathbf{x}\|_0 \leq L \ ,$$

or alternatively

$$\|AP_{\boldsymbol{\omega}}\mathbf{x}\|_2 \geq \zeta_L(A) \|\mathbf{x}\|_2 \quad \forall \mathbf{x} \in \mathbb{R}^N, \ \boldsymbol{\omega} \in \Omega_{M,L'} \text{ where } L' \leq L \ .$$

The quantity $\zeta_L(A)$ provides a measure of the stability of the linear independence of L-sized subsets of atoms of A. Given $\mathbf{s} = A\mathbf{x}$ and $\mathbf{s}' = A\mathbf{y}$, where $\mathbf{x}$ and $\mathbf{y}$ have maximum numbers of non-zero coefficients L_x and L_y respectively, ζ_L for $L = L_x + L_y$ provides a bound

$$\|\Delta\mathbf{x}\|_2 \leq \zeta_L(A)^{-1} \|\Delta\mathbf{s}\|_2$$

on the difference $\Delta\mathbf{x}$ between the two sets of coefficients in terms of the difference $\Delta\mathbf{s}$ between the two signals. Consider signal $\mathbf{s} = A\mathbf{x}$ with primary solution $\mathbf{x}$ and any $\mathbf{s} + \boldsymbol{\nu} = A\mathbf{y}$ such that $\|\boldsymbol{\nu}\|_2 \leq \epsilon$, so that $\mathbf{s} + \boldsymbol{\nu}$ is within an ϵ-ball about $\mathbf{s}$. Therefore

$$\zeta_L(A) \|\mathbf{x} - \mathbf{y}\|_2 \leq \|A(\mathbf{x} - \mathbf{y})\|_2 = \|\boldsymbol{\nu}\|_2 \leq \epsilon,$$

so that

$$\|\mathbf{x}-\mathbf{y}\|_2 \leq \zeta_L(A)^{-1}\epsilon \tag{2.3}$$

for any alternative solution $\mathbf{y}$ such that the difference between $\mathbf{x}$ and $\mathbf{y}$ has at most L non-zero coefficients [15]. This bound can be shown [15] to be the tightest possible bound that does not depend on $\mathbf{x}$. Since $\|\mathbf{x}-\mathbf{y}\|_0 \leq \|\mathbf{x}\|_0 + \|\mathbf{y}\|_0$, the bound Eq. (2.3) implies

$$\|\mathbf{x}-\mathbf{y}\|_2 \leq \zeta_L(A)^{-1}\epsilon$$

for $L = \|\mathbf{x}\|_0 + \|\mathbf{y}\|_0$. If $\mathbf{x}$ and $\mathbf{y}$ both represent $\mathbf{s}$ with some error, i.e. they are such that $\|A\mathbf{x}-\mathbf{s}\|_2 \leq \epsilon$ and $\|A\mathbf{y}-\mathbf{s}\|_2 \leq \epsilon$ then

$$\|\mathbf{x}-\mathbf{y}\|_2 \leq \frac{2\epsilon}{\zeta_L(A)} \,.$$

Given solution $\mathbf{x}$ such that $\|A\mathbf{x}-\mathbf{s}\|_2 = \epsilon$ and $\|\mathbf{x}\|_0 = L$, then for any other solution $\mathbf{y}$ that gives at least as good an approximation (i.e. $\|A\mathbf{y}-\mathbf{s}\|_2 \leq \epsilon$) and is at least a sparse (i.e. $\|\mathbf{y}\|_0 \leq L$),

$$\|\mathbf{x}-\mathbf{y}\|_2 \leq \frac{2\epsilon}{\zeta_{2L}(A)} \,.$$

If $\zeta_{2L}(A) > 0$, in the absence of noise ($\epsilon = 0$) this bound implies that $\mathbf{x} = \mathbf{y}$, i.e. that $\mathbf{x}$ is the unique solution with L or fewer non-zero coefficients.

2.2.2 Properties of $\zeta_L(A)$

The following theorems provide the main properties of $\zeta_L(A)$:

Theorem 2.3. $\zeta_L(A) \geq \zeta_{L+1}(A)$

Proof. If a column is appended to a matrix $X \in \mathbb{R}^{M\times L}$ with $L < M$, then the minimum singular value of the resulting matrix is equal to or less than that of the original matrix (see Theorem A.5). Since every subset of $L+1$ columns is obtained by adding a column to one of the subsets of L columns, it is clear that $\zeta_L(A) \geq \zeta_{L+1}(A)$. ■

An alternative way to define $\zeta_L(A)$ is therefore

$$\zeta_L(A) = \min_{\boldsymbol{\omega}\in\Omega_{M,K};\ K\leq L} \sigma_{\min}(AP_{\boldsymbol{\omega}}) \,.$$

that is, $\zeta_L(A)$ is the smallest $\sigma_{\min}$ for all subsets of L or fewer columns selected from A.

Theorem 2.4. $\zeta_1(A) = 1$ *for A with normalised columns.*

Proof. This follow immediately from the observation that the first singular value of a single-column matrix is just the ℓ^2 norm of that column. To show this, expand column matrix $\mathbf{a}$ as

$$\mathbf{a} = \left(\frac{\mathbf{a}}{\|\mathbf{a}\|_2} \; A_1 \right) \begin{pmatrix} \|\mathbf{a}\|_2 & 0 & \dots \\ 0 & 0 & \dots \\ \vdots & \vdots & \ddots \end{pmatrix} 1$$

where A_1 is a matrix with orthonormal columns that are also orthogonal to $\mathbf{a}$. ■

Theorem 2.5. *$\zeta_2(A) = \sqrt{1 - \mu(A)}$ for A with normalised columns.*

Proof. To show this, consider matrix A with normalised columns $\mathbf{a}_1$ and $\mathbf{a}_2$, with $\langle \mathbf{a}_1, \mathbf{a}_2 \rangle = \alpha$. Defining $\mathbf{b}$ such that $\langle \mathbf{a}_1, \mathbf{d} \rangle = 0$ and $\mathbf{a}_2 + \mathbf{b} = c\mathbf{a}_1$ for some scalar c, it is easy to show that $c = \alpha$ so that $\mathbf{b} = \alpha\mathbf{a}_1 - \mathbf{a}_2$, and also that $\|\mathbf{b}\|_2 = \sqrt{1 - \alpha^2}$. We now define normalised vector

$$\tilde{\mathbf{b}} = \frac{\mathbf{b}}{\sqrt{1 - \alpha^2}}$$

and matrix

$$\tilde{A} = \left(\mathbf{a}_1 \; \tilde{\mathbf{b}} \right) ,$$

so that we can write

$$A = (\mathbf{a}_1 \; \mathbf{a}_2) = \left(\mathbf{a}_1 \; \tilde{\mathbf{b}} \right) \begin{pmatrix} 1 & \alpha \\ 0 & -\sqrt{1 - \alpha^2} \end{pmatrix} = \tilde{A} \begin{pmatrix} 1 & \alpha \\ 0 & -\sqrt{1 - \alpha^2} \end{pmatrix} .$$

Now we need to compute the SVD $B = U\Sigma V^T$ of

$$B = \begin{pmatrix} 1 & \alpha \\ 0 & -\sqrt{1 - \alpha^2} \end{pmatrix} .$$

We have

$$B^T B = \begin{pmatrix} 1 & \alpha \\ \alpha & 1 \end{pmatrix}$$

and $B^T B = V\Sigma^2 V^T$ so that $B^T BV = V\Sigma^2$ and

$$\begin{pmatrix} 1 & \alpha \\ \alpha & 1 \end{pmatrix} \mathbf{v}_k = \sigma_k^2 \mathbf{v}_k \;\; k \in \{1, 2\} .$$

This can be expressed as

$$\begin{pmatrix} 1 - \sigma_k^2 & \alpha \\ \alpha & 1 - \sigma_k^2 \end{pmatrix} \mathbf{v}_k = 0 \;\; k \in \{1, 2\} ,$$

which implies (since $\mathbf{v}_1$ and $\mathbf{v}_2$ span $\mathbb{R}^2$) that [16]

$$\det \begin{pmatrix} 1 - \sigma_k^2 & \alpha \\ \alpha & 1 - \sigma_k^2 \end{pmatrix} = 0 \; k \in \{1, 2\} ,$$

so that σ_k^2 are the zeros of the polynomial $\delta^2 - 2\delta + 1 - \alpha^2$. Therefore, $\sigma_1 = \sqrt{1+\alpha}$ and $\sigma_2 = \sqrt{1-\alpha}$ and

$$\begin{pmatrix} 1 & \alpha \\ 0 & -\sqrt{1-\alpha^2} \end{pmatrix} = U \begin{pmatrix} \sqrt{1+\alpha} & 0 \\ 0 & \sqrt{1-\alpha} \end{pmatrix} V^T ,$$

allowing us to write

$$(\mathbf{a}_1 \; \mathbf{a}_2) = \left(\mathbf{a}_1 \; \tilde{\mathbf{b}}\right) U \begin{pmatrix} \sqrt{1+\alpha} & 0 \\ 0 & \sqrt{1-\alpha} \end{pmatrix} V^T .$$

Now $\left(\mathbf{a}_1 \; \tilde{\mathbf{b}}\right)$ has orthonormal columns, and U is a rotation matrix of the form

$$U = \begin{pmatrix} c & -s \\ s & c \end{pmatrix}$$

so that the product $\left(\mathbf{a}_1 \; \tilde{\mathbf{b}}\right) U$ also has orthonormal columns. Construct an $M \times (M-2)$ matrix $\tilde{U}$ with orthonormal columns that are also orthogonal to the $\mathbf{a}_1$ and $\tilde{\mathbf{b}}$, allowing us to write the SVD of A as

$$A = \left(\tilde{A}U \; \tilde{U}\right) \begin{pmatrix} \sqrt{1+\alpha} & 0 \\ 0 & \sqrt{1-\alpha} \\ 0 & 0 \\ \vdots & \vdots \end{pmatrix} V^T ,$$

demonstrating that $\sigma_2(A) = \sqrt{1-\alpha}$. Minimising $\sigma_2(A)$ over all subsets of 2 columns in computing $\zeta_L(A)$ corresponds to maximising α over all of these subsets, so that $\zeta_2(A) = \sqrt{1-\mu(A)}$ from the definition of mutual coherence. ■

Theorem 2.6.

$$\sqrt{1-(L-1)\mu(A)} \leq \zeta_L(A) \leq \sqrt{1+(L-1)\mu(A)} \tag{2.4}$$

for A with normalised columns.

Proof. Choose any subset $\boldsymbol{\omega} \in \Omega_{M,L}$ of cardinality L from the columns of A, define B as the resulting matrix (i.e. $B = AP_{\boldsymbol{\omega}}$), and define $G = B^T B$. From the normalisation of columns of A, $G_{kk} = 1$, and from the definition of mutual coherence, $|G_{kl}| \leq \mu(B) \leq \mu(A)$, which implies that $\sum_{l \neq k} |G_{kl}| \leq (L-1)\mu(A)$.

Therefore, from the Gershgorin circle theorem (see Theorem A.1), all eigenvalues $\lambda_k(G)$ of G lie in the interval $|1 - \lambda_k(G)| \leq (L-1)\mu(A)$, which implies that

$$1 - (L-1)\mu(A) \leq \lambda_k(G) \leq 1 + (L-1)\mu(A) .$$

Since $\lambda_k(G) = \lambda_k(B^T B) = \sigma_k^2(B)$, we have

$$\sqrt{1-(L-1)\mu(A)} \leq \sigma_k(B) \leq \sqrt{1+(L-1)\mu(A)} ,$$

and this bound applies, in particular, to $\sigma_{\min}$. ■

2.2.3 Relationship between $\zeta_L(A)$ and $\text{spark}_\eta(A)$

The quantity $\zeta_L(A)$ measures the stability of solutions on dictionary A given a bound on the sparsity of these solutions. Alternatively, we can define a function that measures the minimum sparsity of a solution necessary for a given bound on the stability [2, Sec. 5.2.2]:

Definition 2.8. $\text{spark}_\eta(A)$ *is the smallest* L *for which there exists* $\boldsymbol{\omega} \in \Omega_{M,L}$ *such that* $\sigma_{\min}(AP_{\boldsymbol{\omega}}) \leq \eta$, *i.e.*

$$\text{spark}_\eta(A) = \min\{L \in \{1, \ldots, N\} \mid \min_{\boldsymbol{\omega} \in \Omega_{M,L}} \sigma_{\min}(AP_{\boldsymbol{\omega}}) \leq \eta\} \, .$$

The functions $\zeta_L(A)$ and $\text{spark}_\eta(A)$ are closely related. The following results, which are immediately obvious from the definitions of $\zeta_L(A)$ and $\text{spark}_\eta(A)$, help clarify this relationship

- If $\exists \boldsymbol{\omega} \in \Omega_{M,L}$ such that $\sigma_{\min}(AP_{\boldsymbol{\omega}}) = \eta$ then $\zeta_L(A) \leq \eta$ and $\text{spark}_\eta(A) \leq L$
- If $\nexists \boldsymbol{\omega} \in \Omega_{M,L'}$, with $L' \leq L$, such that $\sigma_{\min}(AP_{\boldsymbol{\omega}}) \leq \eta$ then $\zeta_L(A) > \eta$ and $\text{spark}_\eta(A) > L$

and are useful in deriving the some of the following theorems.

Theorem 2.7. *If* $\zeta_L(A)$ *is strictly monotonic in* L *(i.e.* $\zeta_L(A) > \zeta_{L+1}(A)$*), then*

$$\text{spark}_{\zeta_L(A)}(A) = L \, .$$

If $\zeta_L(A) = \zeta_{L+1}(A)$ *for some* L, *then*

$$\text{spark}_\eta(A) = \min\{L \in \{1, \ldots, N\} \mid \zeta_L(A) \leq \eta\} \, .$$

Proof. First we consider the general case. Define L' as the smallest L such that $\zeta_{L'}(A) \leq \eta$. Therefore, there exists an $\boldsymbol{\omega} \in \Omega_{M,L'}$ such that $\sigma_{\min}(AP_{\boldsymbol{\omega}}) \leq \eta$, and there is no $\boldsymbol{\omega}' \in \Omega_{M,L'-1}$ such that $\sigma_{\min}(AP_{\boldsymbol{\omega}'}) \leq \eta$, and from the definition of $\text{spark}_\eta(A)$ we therefore have that $\text{spark}_\eta(A) = L'$. This implies that $\text{spark}_\eta(A) = \min\{L \in \{1, \ldots, N\} \mid \zeta_L(A) \leq \eta\}$.

Now we consider the case when $\zeta_{L-1}(A) \neq \zeta_L(A)$. If $\zeta_L(A) = \eta$ then $\exists \boldsymbol{\omega} \in \Omega_{M,L}$ such that $\sigma_{\min}(AP_{\boldsymbol{\omega}}) = \eta$, which implies that $\text{spark}_\eta(A) \leq L$. Now, choose L' as the largest value such that $\zeta_{L'}(A) = \eta' > \eta = \zeta_L(A)$. Since η' is the smallest value (from the definition of $\zeta_{L'}(A)$) for which $\exists \boldsymbol{\omega} \in \Omega_{M,L'}$ with $\sigma_{\min}(AP_{\boldsymbol{\omega}}) = \eta'$, we also know that $\nexists \boldsymbol{\omega}' \in \Omega_{M,L'}$ such that $\sigma_{\min}(AP_{\boldsymbol{\omega}'}) < \eta'$, which implies that $\text{spark}_{\eta'}(A) > L'$, so that we have $L \geq \text{spark}_\eta(A) \geq \text{spark}_{\eta'}(A) > L'$. If $L' = L - 1$ then $L \geq \text{spark}_\eta(A) > L - 1$ which implies that $\text{spark}_\eta(A) = L$. ■

Theorem 2.8.

$$\zeta_{\text{spark}_\eta(A)-1}(A) > \eta \geq \zeta_{\text{spark}_\eta(A)}(A)$$

Proof. If $\text{spark}_\eta(A) = L$ then $\exists \boldsymbol{\omega} \in \Omega_{M,L}$ such that $\sigma_{\min}(AP_{\boldsymbol{\omega}}) = \eta' \leq \eta$, which implies that $\zeta_L(A) \leq \eta$. Since L is the smallest value (from the definition of $\text{spark}_\eta(A)$) for which $\exists \boldsymbol{\omega} \in \Omega_{M,L}$ with $\sigma_{\min}(AP_{\boldsymbol{\omega}}) = \eta' \leq \eta$, we also know that $\nexists \boldsymbol{\omega'} \in \Omega_{M,L'}$, with $L' \leq L$, such that $\sigma_{\min}(AP_{\boldsymbol{\omega'}}) \leq \eta$, which implies that $\zeta_{L'}(A) > \eta$, and in particular $\zeta_{L-1}(A) > \eta$. We therefore have $\zeta_{L-1}(A) > \eta \geq \zeta_L(A.)$ ■

Since $\sigma_{\min}(AP_{\boldsymbol{\omega}}) = 0$ implies that the columns of $AP_{\boldsymbol{\omega}}$ are linearly dependent, $\text{spark}_0(A) = \text{spark}(A)$. It is clear from the definition that $\text{spark}_\eta(A) \geq \text{spark}_{\eta'}(A) \ \ \forall \eta' > \eta$, i.e. $\text{spark}_\eta(A)$ is monotonically decreasing in η. Furthermore, $\text{spark}_1(A) = 1$ for A with normalized columns [2, Sec. 5.2.2], and

$$1 \leq \text{spark}_\eta(A) \leq \text{spark}(A) \leq M + 1 \ \ \forall 0 \leq \eta \leq 1 \ \ .$$

2.2.4 Bounds based on $\text{spark}_\eta(A)$

The following theorem, a minor variant of [2, Lemma 5.1], gives a lower bound for the sparsity of $\mathbf{x}$ in terms of $\text{spark}_\eta(A)$ when η is bounded below by $\|A\mathbf{x}\|_2 \,/\, \|\mathbf{x}\|_2$.

Theorem 2.9.

$$\frac{\|A\mathbf{x}\|_2}{\|\mathbf{x}\|_2} \leq \eta \ \Rightarrow \ \|\mathbf{x}\|_0 \geq \text{spark}_\eta(A)$$

Proof. From the definition of $\zeta_L(A)$, we know that

$$\zeta_L(A) \leq \frac{\|A\mathbf{x}\|_2}{\|\mathbf{x}\|_2} \ \ \forall \, \|\mathbf{x}\|_0 \leq L$$

so that

$$\zeta_L(A) \leq \frac{\|A\mathbf{x}\|_2}{\|\mathbf{x}\|_2} \leq \eta \ \ \forall L \geq \|\mathbf{x}\|_0$$

Therefore $\zeta_L(A) \leq \eta \ \ \forall L \geq \|\mathbf{x}\|_0$ and in particular $\zeta_L(A) \leq \eta$ when $L = \|\mathbf{x}\|_0$. This implies that $\exists \boldsymbol{\omega} \in \Omega_{M,L}$ such that $\sigma_{\min}(AP_{\boldsymbol{\omega}}) = \eta' \leq \eta$ and therefore $\text{spark}_{\eta'}(A) \leq L$. Since we chose $L = \|x\|_0$ (and $\eta' \leq \eta$ and $\text{spark}_\eta(A)$ is monotonically decreasing) we have $\|x\|_0 \geq \text{spark}_{\eta'}(A) \geq \text{spark}_\eta(A)$. ■

The following theorem (and proof) can be seen as an extension of Theorem 2.2 (see [2, Lemma 5.2])

Theorem 2.10. *A has normalized columns and mutual-coherence $\mu(A)$ implies that*

$$\text{spark}_\eta(A) \geq \frac{1 - \eta^2}{\mu(A)} + 1$$

Proof. Choose any subset $\boldsymbol{\omega} \in \Omega_{M,L}$ of cardinality L from the columns of A, define B as the resulting matrix (i.e. $B = AP_{\boldsymbol{\omega}}$), and define $G = B^T B$. From the normalisation of columns of A, $G_{kk} = 1$, and from the definition of mutual coherence, $|G_{kl}| \leq \mu(B) \leq \mu(A)$, which implies that $\sum_{l \neq k} |G_{kl}| \leq (L-1)\mu(A)$.

Therefore, from the Gershgorin circle theorem (see Theorem A.1), all eigenvalues $\lambda_k(G)$ of G lie in the interval $|1 - \lambda_k(G)| \leq (L-1)\mu(A)$, which implies that

$$1 - (L-1)\mu(A) \leq \lambda_k(G) \leq 1 + (L-1)\mu(A) \; .$$

In particular, then,

$$1 - (L-1)\mu(A) \leq \lambda_{\min}(G)$$

where $\lambda_{\min}(G)$ is the smallest eigenvalue of G, so that

$$L \geq \frac{1 - \lambda_{\min}(G)}{\mu(A)} + 1 \; .$$

Choosing η so that $\zeta_L(A) \leq \eta$ we have $\lambda_{\min}(G) \leq \eta^2$ so that

$$L \geq \frac{1 - \eta^2}{\mu(A)} + 1 \; .$$

Since $\text{spark}_\eta(A)$ is the smallest L such that $\zeta_L(A) \leq \eta$, it must be the case that

$$\text{spark}_\eta(A) \geq \frac{1 - \eta^2}{\mu(A)} + 1 \; .$$

■

Given two sparse solutions to the same system, the following theorem [2, Lemma 5.3] provides a lower bound for the number of non-zero components in both solutions.

Theorem 2.11. *If* $\mathbf{x}$ *and* $\mathbf{y}$ *are such that* $\|A\mathbf{x} - \mathbf{s}\|_2 \leq \epsilon$ *and* $\|A\mathbf{y} - \mathbf{s}\|_2 \leq \epsilon$ *then*

$$\|\mathbf{x}\|_0 + \|\mathbf{y}\|_0 \geq \text{spark}_\eta(A)$$

for

$$\eta = \frac{2\epsilon}{\|\mathbf{x} - \mathbf{y}\|_2}$$

Proof. The geometric arrangement of vectors is illustrated in Fig. 2.2. Using the triangle inequality

$$\|A(\mathbf{x} - \mathbf{y})\|_2 = \|(A\mathbf{x} - \mathbf{s}) - (A\mathbf{y} - \mathbf{s})\|_2 \leq \|A\mathbf{x} - \mathbf{s}\|_2 + \|A\mathbf{y} - \mathbf{s}\|_2 = 2\epsilon \; .$$

Therefore

$$\frac{\|A(\mathbf{x} - \mathbf{y})\|_2}{\|\mathbf{x} - \mathbf{y}\|_2} \leq \frac{2\epsilon}{\|\mathbf{x} - \mathbf{y}\|_2} = \eta$$

and from Theorem 2.9, we have that $\|\mathbf{x} - \mathbf{y}\|_0 \geq \text{spark}_\eta(A)$, but since $\|\mathbf{x}\|_0 + \|\mathbf{y}\|_0 \geq \|\mathbf{x} - \mathbf{y}\|_0$, $\|\mathbf{x}\|_0 + \|\mathbf{y}\|_0 \geq \text{spark}_\eta(A)$. ■

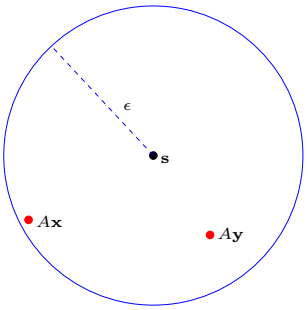


Figure 2.2: Geometry of vectors in Theorem 2.11

The following theorem [2, Theorem 5.1] provides an upper bound on the ℓ^2 distance between two sufficiently sparse solutions to the same system.

Theorem 2.12. *Given vectors* $\mathbf{x}$ *and* $\mathbf{y}$ *and distances* δ *and* ϵ*, and defining* $\eta = 2\epsilon/\delta$*, if*

$$\|A\mathbf{x} - \mathbf{s}\|_2 \leq \epsilon \quad \textit{and} \quad \|\mathbf{x}\|_0 \leq \frac{1}{2}\text{spark}_\eta(A)$$

and

$$\|A\mathbf{y} - \mathbf{s}\|_2 \leq \epsilon \quad \textit{and} \quad \|\mathbf{y}\|_0 \leq \frac{1}{2}\text{spark}_\eta(A)$$

then $\|\mathbf{x} - \mathbf{y}\|_2 \leq \delta$.

Proof. The conditions of the theorem immediately imply that

$$\text{spark}_\eta(A) \geq \|\mathbf{x}\|_0 + \|\mathbf{y}\|_0 \ ,$$

and from Theorem 2.11 we have

$$\|\mathbf{x}\|_0 + \|\mathbf{y}\|_0 \geq \text{spark}_{\eta'}(A)$$

for

$$\eta' = \frac{2\epsilon}{\|\mathbf{x} - \mathbf{y}\|_2} \ .$$

allowing us to write

$$\text{spark}_\eta(A) \geq \|\mathbf{x}\|_0 + \|\mathbf{y}\|_0 \geq \text{spark}_{\eta'}(A) \ .$$

Since $\eta' \geq \eta$ and $\text{spark}_\eta(A)$ is monotonically decreasing,

$$\frac{2\epsilon}{\|\mathbf{x} - \mathbf{y}\|_2} \geq \frac{2\epsilon}{\delta}$$

and therefore $\delta \geq \|\mathbf{x} - \mathbf{y}\|_2$. ■

The following theorem [2, Sec. 5.2.2][17] provides a bound on the ℓ^2 distance between the solution to problem $\mathcal{P}_0^\epsilon$ and any other sufficiently sparse solution to the same system:

Theorem 2.13. *For given $A, \mathbf{s}$, and ϵ, assume there exists a sparse coefficient vector $\mathbf{x}_0$ such that*

$$\|\mathbf{x}_0\|_0 < \frac{1}{2}\left(1 + \frac{1}{\mu(A)}\right)$$

and

$$\|A\mathbf{x}_0 - \mathbf{s}\|_2 \leq \epsilon \ .$$

The inequality

$$\|\mathbf{x}_0^\epsilon - \mathbf{x}_0\|_2^2 \leq \frac{4\epsilon^2}{1 - \mu(A)(2\|\mathbf{x}_0\|_0 - 1)}$$

holds for every solution $\mathbf{x}_0^\epsilon$ of $\mathcal{P}_0^\epsilon$.

Proof. First, note that since $\mathbf{x}_0^\epsilon$ is the solution of $\mathcal{P}_0^\epsilon$, it is the sparsest possible representation of $\mathbf{s}$ such that $\|A\mathbf{x}_0 - \mathbf{s}\|_2 \leq \epsilon$, and therefore $\|\mathbf{x}_0^\epsilon\|_0 \leq \|\mathbf{x}_0\|_0$. Now, if we choose an $\eta \geq 0$ such that

$$\frac{1}{2}\left(1 + \frac{1 - \eta^2}{\mu(A)}\right) \geq \|\mathbf{x}_0\|_0 \ , \tag{2.5}$$

which is guaranteed to be possible since we know that $\frac{1}{2}\left(1 + \frac{1}{\mu(A)}\right) > \|\mathbf{x}_0\|_0$, then we have

$$\frac{1}{2}\text{spark}_\eta(A) \geq \frac{1}{2}\left(1 + \frac{1 - \eta^2}{\mu(A)}\right) \geq \|\mathbf{x}_0\|_0 \geq \|\mathbf{x}_0^\epsilon\|_0 \ ,$$

where the left hand inequality is justified by Theorem 2.10. Rewriting (2.5) as an explicit bound on η gives us

$$\eta^2 \leq 1 - \mu(A)(2\|\mathbf{x}_0\|_0 - 1) \ . \tag{2.6}$$

To summarise, we have that when η conforms to the upper bound (2.6), both $\|\mathbf{x}_0^\epsilon\|_0$ and $\|\mathbf{x}_0\|_0$ are upper bounded by $\frac{1}{2}\text{spark}_\eta(A)$, so from Theorem 2.12 we have

$$\|\mathbf{x}_0^\epsilon - \mathbf{x}_0\|_2^2 \leq D^2 = \frac{4\epsilon^2}{\eta^2} \leq \frac{4\epsilon^2}{1 - \mu(A)(2\|\mathbf{x}_0\|_0 - 1)} \ .$$

■

2.2.5 Restricted Isometry Property

The restricted isometry constant [18, Def. 1.1] [19, Equ. (1.1)] or Restricted Isometry Property (RIP) [1, pg. 60] [2, Sec. 5.2.3] is defined as follows:

Definition 2.9. *The L-restricted isometry constant δ_L of A is the smallest value of δ_L such that*

$$(1 - \delta_L)\|\mathbf{x}\|_2^2 \leq \|AP_{\boldsymbol{\omega}}\mathbf{x}\|_2^2 \leq (1 + \delta_L)\|\mathbf{x}\|_2^2 \quad \forall \mathbf{x} \in \mathbb{R}^{L'},\ \boldsymbol{\omega} \in \Omega_{M,L'},\ L' \leq L \ , \tag{2.7}$$

i.e., it is the smallest δ_L such that $(1-\delta_L)\|\mathbf{x}\|_2^2 \leq \|A_{L'}\mathbf{x}\|_2^2 \leq (1+\delta_L)\|\mathbf{x}\|_2^2$ for all $\mathbf{x}$ and for all matrices $A_{L'}$ with L' columns selected from A, where $L' \leq L$.

This definition differs from that of $\zeta_L(A)$ in that it provides both a lower *and* and upper bound on the singular values of subsets of columns of A. By comparison with the definition of $\zeta_L(A)$ it is easy to see that $\sqrt{1-\delta_L(A)} \leq \zeta_L(A)$, with equality when δ_L is determined by the lower bound rather than the upper bound in (2.7).

Theorem 2.14. *A has normalized columns and mutual-coherence $\mu(A)$ implies that*

$$\delta_L(A) \leq (L-1)\mu(A) \; .$$

Proof. See [2, Sec. 5.2.3]. Choosing any $\boldsymbol{\omega} \in \Omega_{M,L}$, define $B = AP_{\boldsymbol{\omega}}$. From the properties of the SVD we have

$$\sigma_{\min}(B)\|\mathbf{x}\|_2 \leq \|B\mathbf{x}\|_2 \leq \sigma_{\max}(B)\|\mathbf{x}\|_2 \quad \forall \mathbf{x} \in \mathbb{R}^L \; ,$$

and therefore

$$\sigma_{\min}^2(B)\|\mathbf{x}\|_2^2 \leq \|B\mathbf{x}\|_2^2 \leq \sigma_{\max}^2(B)\|\mathbf{x}\|_2^2 \quad \forall \mathbf{x} \in \mathbb{R}^L \; .$$

Since $\sigma_k^2(B) = \lambda_k(B^TB)$,

$$\lambda_{\min}(B^TB)\|\mathbf{x}\|_2^2 \leq \|B\mathbf{x}\|_2^2 \leq \lambda_{\max}(B^TB)\|\mathbf{x}\|_2^2 \quad \forall \mathbf{x} \in \mathbb{R}^L \; ,$$

and from the from the Gershgorin circle theorem (see Theorems A.1 and 2.4)

$$1-(L-1)\mu(A) \leq \lambda_k(B^TB) \leq 1+(L-1)\mu(A) \; ,$$

so that

$$(1-(L-1)\mu(A))\|\mathbf{x}\|_2^2 \leq \lambda_{\min}(B^TB)\|\mathbf{x}\|_2^2 \leq \|B\mathbf{x}\|_2^2 \leq \lambda_{\max}(B^TB)\|\mathbf{x}\|_2^2 \leq (1+(L-1)\mu(A))\|\mathbf{x}\|_2^2 \; .$$

By comparing with the definition of the L-restricted isometry constant δ_L, it is clear that δ_L must be such that

$$1-(L-1)\mu(A) \leq 1-\delta_L \; , \tag{2.8}$$

or

$$\delta_L \leq (L-1)\mu(A) \; .$$

■

Directly following the derivation of the stability bound based on $\zeta_L(A)$ in Sec. 2.2.1, a corresponding bound involving δ_L can be derived as

$$\|\mathbf{x}-\mathbf{y}\|_2^2 \leq \frac{4\epsilon^2}{1-\delta_L}$$

for $L = \|\mathbf{x}\|_0 + \|\mathbf{y}\|_0$, and from Eq. (2.8), we also have

$$\|\mathbf{x}-\mathbf{y}\|_2^2 \leq \frac{4\epsilon^2}{1-\delta_L} \leq \frac{4\epsilon^2}{1-(L-1)\mu(A)} \; .$$

A small variation on this argument [2, Sec. 5.2.3] can also provide an alternative derivation of Theorem 2.13.

Chapter 3

Minimum ℓ^0 Norm: Algorithms

Since problem $\mathcal{P}_0^\epsilon$ is NP-hard [20], it is in general not possible to solve it exactly. We now introduce the family of greedy algorithms that construct a sparse representation via *forward stepwise selection.*

3.1 Matching Pursuit

Matching Pursuit (MP) is a greedy algorithm for (approximately) solving problem $\mathcal{P}_0^\epsilon$. It is based on the greedy selection of the single best dictionary element at each iteration of the algorithm. The derivation of the algorithm is simplified by first deriving a simple result regarding the selection of the best match to a specified target vector from a set of candidate vectors.

Consider the problem

$$\min_a \|a\mathbf{x} - \mathbf{y}\|_2^2 \ ,$$

for which the minimiser is easily shown (either geometrically or as a special case of the problem $\arg\min_{\mathbf{x}}(1/2)\,\|A\mathbf{x} - \mathbf{s}\|_2^2$) to be

$$\hat{a} = \frac{\mathbf{x}^T\mathbf{y}}{\mathbf{x}^T\mathbf{x}} \ .$$

After substituting and a bit of manipulation we have

$$\min_a \|a\mathbf{x} - \mathbf{y}\|_2^2 = \mathbf{y}^T\mathbf{y} - \frac{(\mathbf{x}^T\mathbf{y})^2}{\mathbf{x}^T\mathbf{x}} \ .$$

Minimising over a set of $\mathbf{x}_k$, we have

$$\min_{a,k} \|a\mathbf{x}_k - \mathbf{y}\|_2^2 = \min_k \mathbf{y}^T\mathbf{y} - \frac{(\mathbf{x}_k^T\mathbf{y})^2}{\mathbf{x}_k^T\mathbf{x}_k} \ ,$$

which can be solved by selecting k as

$$\hat{k} = \arg\max_k \frac{(\mathbf{x}_k^T\mathbf{y})^2}{\mathbf{x}_k^T\mathbf{x}_k} \ ,$$

or equivalently

$$\hat{k} = \arg\max_k \left| \frac{\mathbf{x}_k^T \mathbf{y}}{\sqrt{\mathbf{x}_k^T \mathbf{x}_k}} \right| .$$

Returning to the Matching Pursuit algorithm, we derive the procedure for a single iteration. Starting with $\mathbf{x}^{(0)} = 0$, at iteration i of the algorithm we want to perform the update (where δ_k is a vector that is zero everywhere except for index k, where it is unity)

$$\mathbf{x}^{(i+1)} = \mathbf{x}^{(i)} + x\delta_k$$

that maximally reduces the residual error $\left\| A\mathbf{x}^{(i+1)} - \mathbf{s} \right\|_2$. Therefore, we want to solve

$$\begin{aligned} \min_{\mathbf{x}^{(i+1)}} \left\| A\mathbf{x}^{(i+1)} - \mathbf{s} \right\|_2^2 &= \min_{x,k} \left\| A\mathbf{x}^{(i)} + Ax\delta_k - \mathbf{s} \right\|_2^2 \\ &= \min_{x,k} \left\| x\mathbf{a}_k - \mathbf{r}^{(i)} \right\|_2^2 , \end{aligned}$$

where $\mathbf{a}_k$ is column k of A and $\mathbf{r}^{(i)} = \mathbf{s} - A\mathbf{x}^{(i)}$. From the result above, this residual error is minimised by selecting

$$\begin{aligned} \hat{k} &= \arg\max_k \left| \frac{\mathbf{a}_k^T \mathbf{r}^{(i)}}{\sqrt{\mathbf{a}_k^T \mathbf{a}_k}} \right| \\ \hat{x} &= \frac{\mathbf{a}_{\hat{k}}^T \mathbf{r}^{(i)}}{\mathbf{a}_{\hat{k}}^T \mathbf{a}_{\hat{k}}} . \end{aligned}$$

The usual termination criteria for these iterations are either $\left\| \mathbf{r}^{(i)} \right\|_2 \leq \epsilon$ or $\left\| \mathbf{x}^{(i)} \right\|_0 \geq L$.

MP does not update the coefficients as additional columns are added to the active set. While the associated computational cost is low, the residual at each iteration is not the minimum possible residual for the current active set, which has an associated performance penalty. This penalty can be avoided by re-computing the coefficients on the current dictionary subset after each iteration, leading to Orthogonal Matching Pursuit (OMP).

3.2 Orthogonal Matching Pursuit

Since standard Matching Pursuit does not update the coefficient for a column of A once it is added to the solution, the resulting solution does not provide the minimum ℓ^2 error for the chosen set of columns of A. This can be addressed by adding an additional stage [21] at the end each iteration of MP, optimising the coefficients with respect to the currently chosen set of columns of A. Defining $A_\omega = AP_\omega$ where ω is the index set on which the solution $\mathbf{x}^{(i)}$ is non-zero, this optimisation can be expressed as

$$\tilde{\mathbf{x}}^{(i)} = \arg\min_{\mathbf{x}} \frac{1}{2} \left\| A_\omega \mathbf{x} - \mathbf{s} \right\|_2^2 ,$$

resulting in

$$\tilde{\mathbf{x}}^{(i)} = (A_\omega^T A_\omega)^{-1} A_\omega^T \mathbf{s} \; .$$

The residual $\mathbf{r}^{(i)}$ is then updated using the optimised coefficients

$$\mathbf{r}^{(i)} = \mathbf{s} - A\tilde{\mathbf{x}}^{(i)} \; .$$

Since the residual is the projection of $\mathbf{s}$ into the subspace spanned by the columns of A_ω, it is orthogonal to each of these columns, leading to the name Orthogonal Matching Pursuit for this variant. Since the residual is orthogonal to all columns of A that have already been selected, it is clear that OMP will never select the same column more than once.

It can be shown that OMP solves $\mathcal{P}_0$ when that solution is sufficiently sparse [1, Sec. 2.3.1] [2, Sec. 4.2.1] [22]

Theorem 3.1. *For given A and $\mathbf{s}$, if there is a solution $\mathbf{x}_0$ to $\mathcal{P}_0$ such that*

$$\|\mathbf{x}_0\|_0 < \frac{1}{2}\left(1 + \frac{1}{\mu(A)}\right)$$

then OMP with a threshold of 0 will find that solution.

Proof. First, we observe that the initial residual is just $\mathbf{s}$, so we can write

$$\mathbf{r}^{(0)} = \mathbf{s} = A\mathbf{x}_0 = \sum_{k \in \omega} (x_0)_k \mathbf{a}_k \; ,$$

where ω is the support of the non-zero coefficients in $\mathbf{x}_0$, and iteration i results in an update to the residual

$$\mathbf{r}^{(i)} = \mathbf{s} - A\tilde{\mathbf{x}}^{(i)} \; .$$

If the first iteration selects one of the correct columns of A (i.e. one of those with a non-zero coefficient in $\mathbf{x}_0$) then the first residual $\mathbf{r}^{(1)}$ will also be a linear combination of the columns $\{\mathbf{a}_k\}_{k \in \omega}$, and the second iteration will not select the column selected in the first iteration since each residual $\mathbf{r}^{(i)}$ is orthogonal to the columns of A with non-zero coefficients in $\mathbf{x}^{(i)}$. Repeating this argument, it is clear that OMP must select the correct subset of columns of A if we can show that each iteration it selects one of the correct columns (i.e. one of the columns of A corresponding to a non-zero coefficient in $\mathbf{x}_0$).

To show this, we start by expressing the residual $\mathbf{r}$ at an arbitrary iteration as a linear combination of the columns in the solution $\mathbf{x}_0$. We simplify the proof slightly by assuming that the columns of A are normalized, i.e. $\|\mathbf{a}_k\|_2 = 1$ for each column k of A. Without loss of generality, we also assume that the columns of A are ordered so that the solution to $\mathcal{P}_0$ has non-zero value only on the first $L = \|\mathbf{x}_0\|_0$ columns of A, and that these non-zero coefficients are ordered with decreasing absolute value, so that

$$\mathbf{r} = \sum_{l=1}^{L} x_l \mathbf{a}_l \quad \text{with} \quad |x_l| \geq |x_{l+1}| \; .$$

If OMP is to select the correct set of non-zero coefficients, each iteration should select one of the first L columns of A. If we require that

$$\left|\mathbf{a}_1^T\mathbf{r}\right| > \left|\mathbf{a}_k^T\mathbf{r}\right| \quad \forall k > L \;, \tag{3.1}$$

then OMP will certainly select the first column of A before selecting any of the columns indexed by $L+1$ or greater. This is sufficient to guarantee that OMP selects one of the first L columns of A; if there is another one of the initial L columns $\mathbf{a}_l$ such that $\left|\mathbf{a}_l^T\mathbf{r}\right| > \left|\mathbf{a}_1^T\mathbf{r}\right|$ then OMP will select that column, but the column selected remains within the required set.

Substituting the expansion $\mathbf{r} = \sum_{l=1}^{L} x_l\mathbf{a}_l$ into Eq. (3.1), this requirement can be expressed as

$$\left|\sum_{l=1}^{L} x_l\mathbf{a}_1^T\mathbf{a}_l\right| > \left|\sum_{l=1}^{L} x_l\mathbf{a}_k^T\mathbf{a}_l\right| \quad \forall k > L \;. \tag{3.2}$$

We now proceed by finding a lower bound for the left hand term in Eq. (3.2) and an upper bound for the right hand term in Eq. (3.2). If we can find conditions under which the lower bound for the left hand term is greater than the upper bound for the right hand term, then these conditions must be sufficient for the requirement Eq. (3.2) to be satisfied.

First we derive a lower bound for the left hand term, as follows

$$\begin{aligned}
\left|\sum_{l=1}^{L} x_l\mathbf{a}_1^T\mathbf{a}_l\right| &\geq \left|x_1\mathbf{a}_1^T\mathbf{a}_1\right| - \left|\sum_{l=2}^{L} x_l\mathbf{a}_1^T\mathbf{a}_l\right| \qquad \text{since } |x+y| \geq |x| - |y| \\
&= |x_1| - \left|\sum_{l=2}^{L} x_l\mathbf{a}_1^T\mathbf{a}_l\right| \qquad \text{since } \mathbf{a}_1^T\mathbf{a}_1 = 1 \\
&\geq |x_1| - \sum_{l=2}^{L} \left|x_l\mathbf{a}_1^T\mathbf{a}_l\right| \qquad \text{since } |x+y| \leq |x| + |y| \\
&= |x_1| - \sum_{l=2}^{L} |x_l| \left|\mathbf{a}_1^T\mathbf{a}_l\right| \qquad \text{since } |xy| = |x|\,|y| \\
&\geq |x_1| - \sum_{l=2}^{L} |x_l|\,\mu(A) \qquad \text{using } \mathbf{a}_l^T\mathbf{a}_l = 1 \text{ and the definition of } \mu(\cdot) \\
&\geq |x_1| - (L-1)\,|x_1|\,\mu(A) \qquad \text{since } |x_1| \geq |x_l| \text{ for } l > 1 \\
&= |x_1|\,(1 - \mu(A)(L-1)) \;.
\end{aligned}$$

Now we turn to an upper bound for the right hand term, as follows

$$\begin{aligned}
\left|\sum_{l=1}^{L} x_l\mathbf{a}_k^T\mathbf{a}_l\right| &\leq \sum_{l=1}^{L} |x_l| \left|\mathbf{a}_k^T\mathbf{a}_l\right| \\
&\leq \sum_{l=1}^{L} |x_l|\,\mu(A) \\
&\leq |x_1|\,\mu(A)L \;.
\end{aligned}$$

Setting the first bound greater than the second bound

$$|x_1|\,(1-\mu(A)(L-1)) > |x_1|\,\mu(A)L\ ,$$

it is clear after some simple re-arrangement that this inequality holds when

$$1+\mu(A) > 2\mu(A)L\ ,$$

or

$$L < \frac{1}{2}\left(1+\frac{1}{\mu(A)}\right)\ .$$

If this final condition is satisfied, then the first bound is indeed greater than the second bound, which in turn implies that the inequality Eq. (3.2) holds, which implies that OMP will select one of the initial L columns $\mathbf{a}_l$ as one of the columns of A in the solution. ■

Theorem 3.2. *For given A and $\mathbf{s}$, if there exists an $\mathbf{x}_0$ such that*

$$\|\mathbf{x}\|_0 < \frac{1}{2}\left(1+\frac{1}{\mu(A)}\right) - \frac{\epsilon}{\mu(A)c}\ ,$$

where c is the element of $\mathbf{x}_0$ that has the smallest non-zero absolute value, and

$$\|A\mathbf{x}_0 - \mathbf{s}\|_2 \leq \epsilon$$

then the solution $\mathbf{x}_1$ of OMP is such that

$$\|\mathbf{x}_1 - \mathbf{x}_0\|_2^2 \leq \frac{\epsilon^2}{1-\mu(A)(\|\mathbf{x}_0\|_0 - 1)}\ .$$

Proof. See [1, Sec. 3.2.4]. ■

Chapter 4

Minimum ℓ^1 Norm: Theory

Since minimisation with respect to the ℓ^0 "norm" is, in general, intractable, it would be convenient if we could replace it with an alternative criterion that leads to minimisation of a convex function. If we consider the family of ℓ^p norms, $p = 1$ is the obvious choice since a function such as $(1/2)\|A\mathbf{x} - \mathbf{s}\|_2^2 + \lambda \|\mathbf{x}\|_p^p$ is non-convex for $p < 1$, and does not lead to a sparse solution for $p > 1$ (see Fig. 4.1); the ℓ^1 norm can be viewed as the tightest convex relaxation of the ℓ^0 "norm" [23, Sec. II]. The widespread use of ℓ^1 minimisation methods is supported by the remarkable result that, under certain conditions, ℓ^1 minimisation leads to the same solution as ℓ^0 minimisation.

4.1 Uniqueness Results

The inverse problem of interest in this section is:

Definition 4.1. *Problem $\mathcal{P}_1$ is defined as*

$$\arg\min_{\mathbf{x}} \|\mathbf{x}\|_1 \ \text{ such that } A\mathbf{x} = \mathbf{s} \ . \tag{4.1}$$

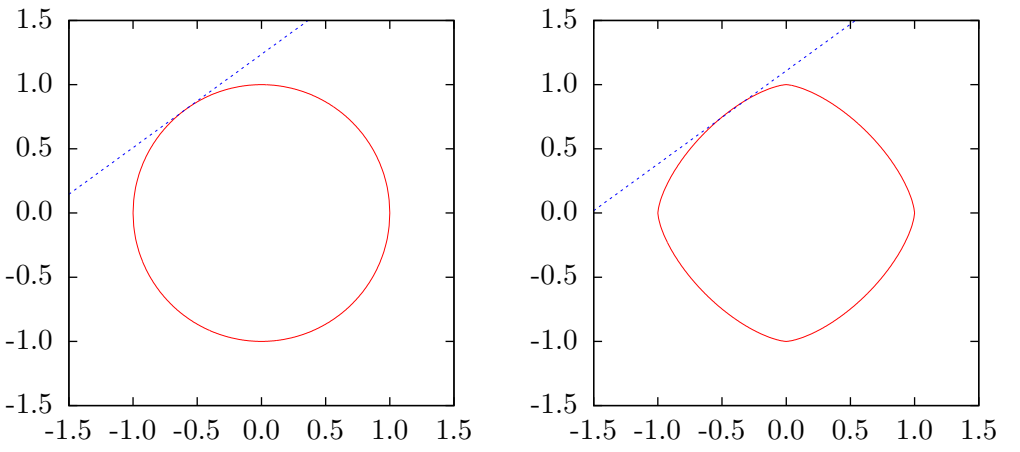
The fundamental result here is that, if the solution of $\mathcal{P}_0$ is sufficiently sparse, then solving $\mathcal{P}_1$ will give the same solution:

Theorem 4.1. *For given $A \in \mathbb{R}^{M \times N}$ (A is full rank and $M < N$) and $\mathbf{s}$, if there is a solution $\mathbf{x}$ such that $A\mathbf{x} = \mathbf{s}$, and*

$$\|\mathbf{x}\|_0 < \frac{1}{2}\left(1 + \frac{1}{\mu(A)}\right) \ ,$$

then $\mathbf{x}$ is the unique solution to both $\mathcal{P}_0$ and $\mathcal{P}_1$.

Proof. See [1, Sec. 2.3.2] [2, Sec. 4.2.3]. ■



(a) $p = 2.0$ (b) $p = 1.5$

(c) $p = 1.0$ (d) $p = 0.5$

Figure 4.1: Geometry of the problem $\arg\min_{\mathbf{x}} \|\mathbf{x}\|_p^p$ such that $A\mathbf{x} = \mathbf{s}$.

It is also possible to construct uniqueness results based on the Restricted Isometry Property Eq. (2.7).

Theorem 4.2. *For given $A \in \mathbb{R}^{M \times N}$ and $\mathbf{s}$, if there is a solution $\mathbf{x}$ such that $A\mathbf{x} = \mathbf{s}$ and $\|\mathbf{x}\|_0 \leq L$ for $L \geq 1$, and if A has restricted isometry constants such that $\delta_L + \delta_{2L} + \delta_{3L} < 1$ then $\mathbf{x}$ is the unique solution to problem $\mathcal{P}_1$.*

Proof. See [18, Theorem 1.3 and Lemma 1.1][19, Sec. 1.1] ■

A more general result based on the Restricted Isometry Property is also possible. In this case we again assume that $A\mathbf{x} = \mathbf{s}$ holds exactly, but we do not require that $\mathbf{x}$ be sparse, and instead define $\mathbf{x}_L$ as the best L-sparse approximation to $\mathbf{x}$, constructed by setting to zero all but the L entries that are largest in absolute value. The following theorem bounds the error in recovering $\mathbf{x}$ via problem $\mathcal{P}_1$ in terms of the ℓ^1 norm of the difference between $\mathbf{x}$ and its L-sparse approximation $\mathbf{x}_L$. If $\mathbf{x}$ is exactly L-sparse, then the error bound is zero and exact recovery is guaranteed.

Theorem 4.3. *Given A with restricted isometry constant $\delta_{2L} < \sqrt{2} - 1$, and $\mathbf{s}$, such that $\mathbf{s} = A\mathbf{x}$ for unknown $\mathbf{x}$, and defining $\mathbf{x}_L$ as the best L-sparse approximation to $\mathbf{x}$, and $\mathbf{x}^*$ as the solution to problem $\mathcal{P}_1$ for A and $\mathbf{s}$,*

$$\|\mathbf{x} - \mathbf{x}^*\|_1 \leq C_0 \|\mathbf{x} - \mathbf{x}_L\|_1$$
$$\|\mathbf{x} - \mathbf{x}^*\|_2 \leq \frac{C_0}{\sqrt{L}} \|\mathbf{x} - \mathbf{x}_L\|_1$$

for some constant C_0 depending only on δ_{2L}.

Proof. See [24, Theorem 1.1]. ■

4.2 Stability Results

Since we now consider the inverse problem for which an exact sparse representation is not required, the relevant inverse problem is

Definition 4.2. *Define problem $\mathcal{P}_1^\epsilon$ as*

$$\arg\min_{\mathbf{x}} \|\mathbf{x}\|_1 \ \text{ such that } \ \|A\mathbf{x} - \mathbf{s}\|_2 \leq \epsilon \ . \tag{4.2}$$

Given a sufficiently sparse solution to $\mathcal{P}_0^\epsilon$, it is possible to bound the difference between that solution and the solution found by $\mathcal{P}_1^\epsilon$:

Theorem 4.4. *For given A and $\mathbf{s}$, if there exists an $\mathbf{x}_0$ such that*

$$\|\mathbf{x}_0\|_0 < \frac{1}{4}\left(1 + \frac{1}{\mu(A)}\right)$$

and

$$\|A\mathbf{x}_0 - \mathbf{s}\|_2 \leq \epsilon$$

then the solution $\mathbf{x}_1$ to $\mathcal{P}_1^\epsilon$ is such that

$$\|\mathbf{x}_1 - \mathbf{x}_0\|_2^2 \leq \frac{4\epsilon^2}{1 - \mu(A)(4\|\mathbf{x}_0\|_0 - 1)} \,.$$

Proof. See [1, Sec. 3.2.4] [2, Sec. 5.5.1] [25, 26]. ■

Theorem 4.3 is a special case of a more general result that holds in the presence of noise.

Theorem 4.5. *Given A with restricted isometry constant $\delta_{2L} < \sqrt{2} - 1$, and $\mathbf{s}$, such that $\mathbf{s} = A\mathbf{x} + \boldsymbol{\nu}$ for unknown $\mathbf{x}$ and $\boldsymbol{\nu}$ with $\|\boldsymbol{\nu}\|_2 \leq \epsilon$, and defining $\mathbf{x}_L$ as the best L-sparse approximation to $\mathbf{x}$, and $\mathbf{x}^*$ as the solution to problem $\mathcal{P}_1^\epsilon$ for A and $\mathbf{s}$,*

$$\|\mathbf{x} - \mathbf{x}^*\|_2 \leq \frac{C_0}{\sqrt{L}}\|\mathbf{x} - \mathbf{x}_L\|_1 + C_1\epsilon$$

for some constants C_0 and C_1 depending only on δ_{2L}.

Proof. See [24, Theorem 1.2]. ■

The sufficient condition $\delta_{2L} < \sqrt{2}-1 \approx 0.4142$ of this theorem has recently been improved [27] to $\delta_{2L} < \frac{3}{4+\sqrt{6}} \approx 0.4652$.

A nunber of similar stability results based on the RIP are available; for example, see [19, Theorem 1.1 and Theorem 1.2] with sufficient condition $\delta_{3L} + 3\delta_{4L} < 2$.

Chapter 5

Minimum ℓ^1 Norm: Algorithms

There are a few different standard ℓ^1 optimisation problems, but the names adopted in the research literature are, unfortunately, often not applied consistently. Probably the least controversial with respect to naming is the *Basis Pursuit* problem

$$\arg\min_{\mathbf{x}} \|\mathbf{x}\|_1 \text{ such that } A\mathbf{x} = \mathbf{s} \; .$$

The corresponding bound constrained variant

$$\arg\min_{\mathbf{x}} \|\mathbf{x}\|_1 \text{ such that } \|A\mathbf{x} - \mathbf{s}\|_2 \leq \epsilon$$

is the most convenient in many contexts, but does not have a standard name (although it is sometimes referred to as *Basis Pursuit DeNoising* [28, 29], confusing it with a different form). The other standard problems are, adopting the names established in the publications that introduced them, the *Lasso* [9]

$$\arg\min_{\mathbf{x}} \|A\mathbf{x} - \mathbf{s}\|_2 \text{ such that } \|\mathbf{x}\|_1 \leq \tau \; ,$$

and *Basis Pursuit DeNoising* (BPDN) [11]

$$\arg\min_{\mathbf{x}} \|A\mathbf{x} - \mathbf{s}\|_2^2 + \lambda \|\mathbf{x}\|_1 \; ,$$

adopting the names established in the publications that introduced these methods, but these names are interchanged by some authors.

5.1 Linear Programming

One of the earliest solutions to the Basis Pursuit problem (i.e. $\mathcal{P}_1$) was via a mapping to the standard problem of Linear Programming (LP) [11], for which a number of algorithms are available, and which can be posed as

$$\arg\min_{\mathbf{x}} \mathbf{d}^T\mathbf{x} \text{ such that } B\mathbf{x} = \mathbf{c} \; , \mathbf{x} \geq 0 \; .$$

The primary difficulties in constructing this mapping are

- the LP problem minimises a linear combination on the solution vector, without the absolute value in the ℓ^1 norm, and
- the solution to the LP problem is positive, while the Basis Pursuit problem is not constrained in this way.

We can resolve these issues by splitting the solution into positive and negative parts, thereby doubling the size of the problem; specifically, we rewrite the constraint $A\mathbf{x} = \mathbf{s}$ as

$$\begin{pmatrix} A & -A \end{pmatrix} \begin{pmatrix} \mathbf{u} \\ \mathbf{v} \end{pmatrix} = \mathbf{s}$$

with $\mathbf{u} \geq 0$ and $\mathbf{v} \geq 0$. We can now write $\mathbf{x} = \mathbf{u} - \mathbf{v}$ and since $\mathbf{u}$ and $\mathbf{v}$ are both positive, we can write $\|\mathbf{x}\|_1 = \sum_k u_k + \sum_k v_k$, or

$$\|\mathbf{x}\|_1 = \mathbb{1}^T \begin{pmatrix} \mathbf{u} \\ \mathbf{v} \end{pmatrix} ,$$

where $\mathbb{1}$ is a column vector with unit entries. The LP problem is then solved with

$$\begin{aligned} \mathbf{b} &= \mathbf{s} \\ D &= \begin{pmatrix} A & -A \end{pmatrix} \\ \mathbf{c} &= \mathbb{1} . \end{aligned}$$

This method is not efficient due to the necessity of doubling the problem size, and is largely of historical interest.

5.2 Least Angle Regression

Least Angle Regression (LARS) [30][2, Sec. 5.3.3][31, Sec. 3.4.4] is a forward selection algorithm related to OMP. A modified version of this algorithm that includes the ability to remove a previously selected dictionary column from the solution can be shown to exactly solve the BPDN problem. A very useful property of this algorithm is that it provides the full *regularization path* for this problem, i.e. the set of all solutions for all possible choices of the regularization parameter λ. It is not a practical algorithm for high dimensional problems since the number of steps required is proportional to the dimensionality of the solution vector.

5.3 Iteratively Reweighted Least Squares

Iteratively Reweighted Least Squares (IRLS) is a relatively old method (originally developed in the statistics community for least squares fitting) for solving an ℓ^p problem by iteratively replacing it with a quadratic approximation represented by a weighted ℓ^2 problem [32, 33, 34, 35, 36]. The first application of this method in algorithms for sparse representations was the FOCUSS algorithm [37, 12, 38].

Consider minimisation of the function

$$f(\mathbf{x}) = \frac{1}{2} \|A\mathbf{x} - \mathbf{s}\|_2^2 + \frac{\lambda}{p} \|\mathbf{x}\|_p^p \ .$$

Selecting $p = 1$ gives an ℓ^1 minimisation problem, but note that it is possible to select $p < 1$, in which case the algorithm can be considered as providing a heuristic solution to the ℓ^0 problem. Observing that $\frac{d}{dx} |x| = \frac{d}{dx} \operatorname{sign}(x) x = \operatorname{sign}(x)$ when $x \neq 0$ and $\frac{d}{dx} |x|^p = p |x|^{p-1} \frac{d}{dx} |x| = p |x|^{p-1} \operatorname{sign}(x)$ when $x \neq 0$, we have (after writing $p |x|^{p-1} \operatorname{sign}(x) = p |x|^{p-2} |x| \operatorname{sign}(x) = p |x|^{p-2} x \operatorname{sign}(x) \operatorname{sign}(x) = p |x|^{p-2} x$) that

$$\nabla_{\mathbf{x}} f(\mathbf{x}) = A^T A\mathbf{x} - A^T \mathbf{s} + \lambda \mathbf{x} \odot |\mathbf{x}|^{p-2} \ \text{ where } \mathbf{x} \neq 0 \ .$$

Now, we construct a quadratic approximation to $f(\mathbf{x})$ about the point $\mathbf{x}_0$ by defining the functional with weighted ℓ^2 regularisation

$$g(\mathbf{x}) = \frac{1}{2} \|A\mathbf{x} - \mathbf{s}\|_2^2 + \frac{\lambda}{2} \left\| W^{1/2} \mathbf{x} \right\|_2^2 + \left(1 - \frac{p}{2}\right) \frac{\lambda}{p} \|\mathbf{x}_0\|_p^p \ ,$$

for which

$$\nabla_{\mathbf{x}} g(\mathbf{x}) = A^T A\mathbf{x} - A^T \mathbf{s} + \lambda W \mathbf{x} \ .$$

If we define $W = \operatorname{diag}(|\mathbf{x}_0|^{p-2})$, we observe that $f(\mathbf{x}) = g(\mathbf{x})$, and $\nabla_{\mathbf{x}} f(\mathbf{x})|_{\mathbf{x}=\mathbf{x}_0} = \nabla_{\mathbf{x}} g(\mathbf{x})|_{\mathbf{x}=\mathbf{x}_0}$, so that $g(\mathbf{x})$ is a quadratic approximation to $f(\mathbf{x})$, with the same value and gradient at $\mathbf{x}_0$. Furthermore, for $p < 2$, we have that $g(\mathbf{x}) \geq f(\mathbf{x})$; while this holds in general, here we demonstrate it for the simplest case $p = 1$.

Since $f(\mathbf{x})$ and $g(\mathbf{x})$ have the same first term, we only need to consider the rest of each functional, and since both forms are additive, we can consider just the scalar functions $s(x) = \lambda |x|$ and $t(x) = \frac{\lambda}{2} |x_0|^{-1} x^2 + \frac{\lambda}{2} |x_0|$. We show that $t(x) \geq s(x)$ by demonstrating that $t(x) - s(x)$ is positive:

$$\begin{aligned} t(x) - s(x) &= \frac{\lambda}{2} |x_0|^{-1} x^2 + \frac{\lambda}{2} |x_0| - \lambda |x| \\ &= \frac{\lambda}{2} |x_0|^{-1} \left(x^2 - 2 |x_0| |x| + |x_0|^2 \right) \\ &= \frac{\lambda}{2} |x_0|^{-1} (x - |x_0|)^2 \ . \end{aligned}$$

Now, consider the sequence of solutions constructed by iteratively constructing the quadratic approximation $g^{(i)}(\mathbf{x})$ about solution $\mathbf{x}^{(i)}$ and defining $\mathbf{x}^{(i+1)}$ as the minimiser of that $g^{(i)}(\mathbf{x})$. From the arguments above, we have [36, Appendix]

$$g^{(i)}(\mathbf{x}^{(i)}) = f(\mathbf{x}^{(i)})$$

and

$$g^{(i)}(\mathbf{x}) \geq f(\mathbf{x}) \ ,$$

and since $\mathbf{x}^{(i+1)}$ minimises $g^{(i)}(\mathbf{x})$,

$$g^{(i)}(\mathbf{x}^{(i+1)}) \leq g^{(i)}(\mathbf{x}) \ ,$$

so that we have

$$f(\mathbf{x}^{(i)}) = g^{(i)}(\mathbf{x}^{(i)}) \geq g^{(i)}(\mathbf{x}^{(i+1)}) \geq f(\mathbf{x}^{(i+1)}) \;.$$

The sequence $\left\{f(\mathbf{x}^{(i)})\right\}$ is therefore decreasing, and must converge since $f(\mathbf{x}) \geq 0 \; \forall \mathbf{x}$. A little more work [36, Appendix] is required to show that this sequence converges to the minimiser of $f(\mathbf{x})$. A single iteration of a 1-d problem is illustrated in Fig. 5.1.

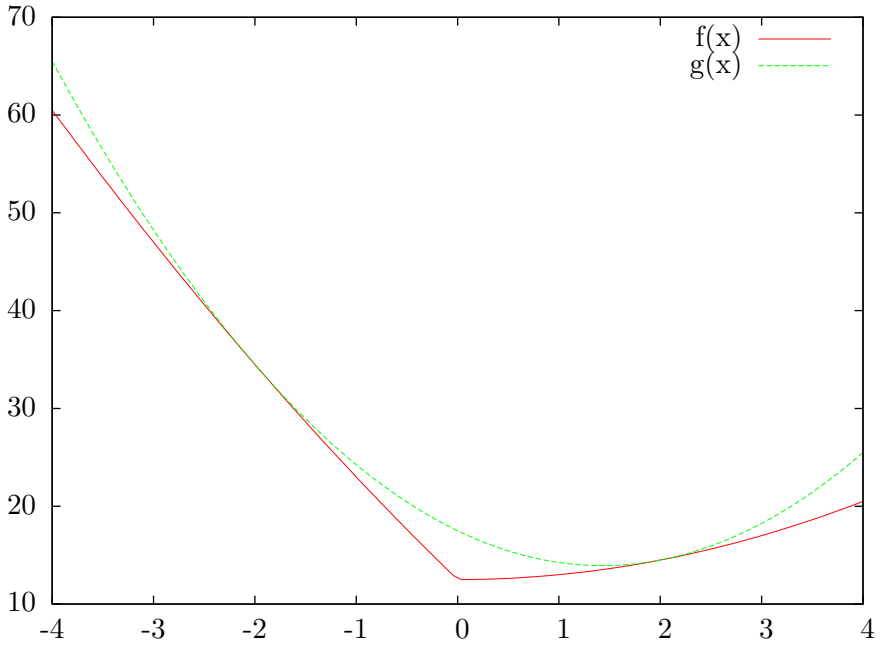


Figure 5.1: The function $f(x) = \frac{1}{2}(x-5)^2 + 5\,|x|$ and its quadratic approximation $g(x) = \frac{1}{2}(x-5)^2 + \frac{5}{2}\,|x_0|^{-1}\,x^2 + \frac{5}{2}\,|x_0|$ about the point $x_0 = -2$.

If $p < 2$, we need to prevent $W_{k,k}$ from becoming very large when $|x_k|$ is small. The obvious, and most common, solution is to define either

$$w_k = \frac{1}{|x_k|^{2-p} + \epsilon}$$

or

$$w_k = \begin{cases} |x_k|^{p-2} & \text{if } |x_k| \geq \epsilon \\ \epsilon^{p-2} & \text{if } |x_k| < \epsilon \end{cases} \;.$$

The FOCUSS algorithm [37, 12, 38] makes the variable substitution $\mathbf{z} = W^{1/2}$, so that the weighted BPDN problem becomes

$$\arg\min_{\mathbf{z}} \frac{1}{2} \left\| AW^{-1/2}\mathbf{z} - \mathbf{s} \right\|_2^2 + \frac{\lambda}{2} \|\mathbf{z}\|_2^2 \;,$$

thus avoiding the necessity of this thresholding operation.

It is worth noting that IRLS is a special case of the Majorization-Minimization (MM) approach [39], which has a similar convergence proof to that of IRLS, but is not necessarily based on a *quadratic* approximation to the function to be minimized. While the methods discussed in the remainder of this chapter are now substantially more popular than IRLS algorithms, the latter are still highly competitive in some problem contexts [40].

5.4 Proximal Methods

This section is intended to give a brief introduction to an important family of techniques that is capable of solving optimization problems involving ℓ^1 norms.

5.4.1 Proximal Gradient

Given the optimization problem $\arg\min_{\mathbf{x}} f(\mathbf{x})$ for differentiable f, one of the simplest optimisation algorithms is *gradient descent*, i.e., choosing a *step size* α, iterate

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \alpha \nabla f\left(\mathbf{x}^{(k)}\right) ,$$

where the step size might be $\alpha^{(k)}$, i.e. depending on the iteration k, which could be chosen at each iteration via a *line search* to ensure that $f(\mathbf{x}^{(k+1)}) < f(\mathbf{x}^{(k)})$. One possible derivation of the gradient descent method [41] is to construct a quadratic approximation to f at point y

$$q_\alpha(\mathbf{x}, \mathbf{y}) = \frac{1}{2\alpha} \|\mathbf{x} - \mathbf{y}\|_2^2 + (\mathbf{x} - \mathbf{y})^T \nabla f(\mathbf{y}) + f(\mathbf{y}) \tag{5.1}$$

such that $q_\alpha(\mathbf{y}, \mathbf{y}) = f(\mathbf{y})$ and $\nabla_{\mathbf{x}} q_\alpha(\mathbf{x}, \mathbf{y})|_{\mathbf{x}=\mathbf{y}} = \nabla f(\mathbf{y})$, and iterate

$$\mathbf{x}^{(k+1)} = \arg\min_{\mathbf{x}} q_\alpha(\mathbf{x}, \mathbf{x}^{(k)}) ,$$

which leads to the gradient descent scheme above since $\nabla_{\mathbf{x}} q_\alpha(\mathbf{x}, \mathbf{y}) = (\mathbf{x} - \mathbf{y})/\alpha + \nabla f(\mathbf{y})$.

Now consider the constrained problem (again with differentiable f)

$$\arg\min_{\mathbf{x} \in C} f(\mathbf{x}) \tag{5.2}$$

where C is a non-empty closed convex set [42, 43]. First, note that Eq. (5.1) can be re-expressed as

$$q_\alpha(\mathbf{x}, \mathbf{y}) = \frac{1}{2\alpha} \|\mathbf{x} - (\mathbf{y} - \alpha \nabla f(\mathbf{y}))\|_2^2 - \frac{\alpha}{2} \|\nabla f(\mathbf{y})\|_2^2 + f(\mathbf{y}) . \tag{5.3}$$

Following the derivation of the gradient descent scheme above, but now with a constraint, we have

$$\begin{aligned} \mathbf{x}^{(k+1)} &= \arg\min_{\mathbf{x} \in C} q_\alpha(\mathbf{x}, \mathbf{x}^{(k)}) \\ &= \arg\min_{\mathbf{x} \in C} \frac{1}{2} \left\| \mathbf{x} - \left(\mathbf{x}^{(k)} - \alpha \nabla f(\mathbf{x}^{(k)}) \right) \right\|_2^2 , \end{aligned}$$

omitting the constant (with respect to $\mathbf{x}$) terms from Eq. (5.3). If we define the operator projecting point $\mathbf{x}$ to the nearest point in C

$$P_C(\mathbf{x}) = \arg\min_{\mathbf{z} \in C} \frac{1}{2} \|\mathbf{z} - \mathbf{x}\|_2^2 , \tag{5.4}$$

then this iteration can be written as

$$\mathbf{x}^{(k+1)} = P_C\left(\mathbf{x}^{(k)} - \alpha\nabla f(\mathbf{x}^{(k)}\right) \ ,$$

which can be interpreted as a a gradient descent step followed by a projection onto the constraint set C, and is referred to as a *projected gradient* [42] or *gradient projection* [41] method.

Defining the *indicator function* of set C as

$$\iota_C(\mathbf{x}) = \begin{cases} 0 & \text{if } \mathbf{x} \in C \\ \infty & \text{if } \mathbf{x} \notin C \ , \end{cases}$$

the projection Eq. (5.4) can instead be written as

$$P_C(\mathbf{x}) = \arg\min_{\mathbf{y}} \frac{1}{2}\left\|\mathbf{y}-\mathbf{x}\right\|_2^2 + \iota_C(\mathbf{y}) \ . \tag{5.5}$$

Instead of the indicator function of a convex set $\iota_C(\cdot)$ in Eq. (5.5), substitute an arbitrary (we omit some technical requirements [43] here), not necessarily smooth function $g(\cdot)$ to give a generalisation of a projection. This generalisation is the *proximal mapping* of $g(\cdot)$,

$$\operatorname*{prox}_g(\mathbf{x}) = \arg\min_{\mathbf{y}} \frac{1}{2}\left\|\mathbf{y}-\mathbf{x}\right\|_2^2 + g(\mathbf{y}) \ ,$$

and it is of interest because we can often find an analytical expression for it, making it computationally efficient to implement. We now turn to our primary interest here, which is solving a problem of the form

$$\arg\min_{\mathbf{x}} f(\mathbf{x}) + g(\mathbf{x}) \ , \tag{5.6}$$

where f is smooth and g is not. A natural modification [41] of the quadratic approximation approach applied above is to define

$$\begin{aligned} q_\alpha(\mathbf{x},\mathbf{y}) =& \frac{1}{2\alpha}\left\|\mathbf{x}-\mathbf{y}\right\|_2^2 + (\mathbf{x}-\mathbf{y})^T\nabla f(\mathbf{y}) + f(\mathbf{y}) + g(\mathbf{x}) \\ =& \frac{1}{2\alpha}\left\|\mathbf{x}-(\mathbf{y}-\alpha\nabla f(\mathbf{y}))\right\|_2^2 - \frac{\alpha}{2}\left\|\nabla f(\mathbf{y})\right\|_2^2 + f(\mathbf{y}) + g(\mathbf{x}) \ , \end{aligned} \tag{5.7}$$

consisting of a quadratic approximation to f and the original g. Following the same procedure as before, we obtain the iteration

$$\begin{aligned} \mathbf{x}^{(k+1)} &= \arg\min_{\mathbf{x}} q_\alpha(\mathbf{x},\mathbf{x}^{(k)}) \\ &= \arg\min_{\mathbf{x}} \frac{1}{2\alpha}\left\|\mathbf{x}-\left(\mathbf{x}^{(k)}-\alpha\nabla f(\mathbf{x}^{(k)})\right)\right\|_2^2 + g(\mathbf{x}) \\ &= \arg\min_{\mathbf{x}} \frac{1}{2}\left\|\mathbf{x}-\left(\mathbf{x}^{(k)}-\alpha\nabla f(\mathbf{x}^{(k)})\right)\right\|_2^2 + \alpha g(\mathbf{x}) \\ &= \operatorname*{prox}_{\alpha g}\left(\mathbf{x}^{(k)}-\alpha\nabla f(\mathbf{x}^{(k)})\right) \end{aligned}$$

of the *proximal gradient* method [41] (it is also a type of *forward-backward* splitting [43]). An alternative derivation of this approach is from the fixed point equation [41, 43]

$$\mathbf{x}^* = \operatorname*{prox}_{\alpha g}\left(\mathbf{x}^* - \alpha\nabla f(\mathbf{x}^*)\right)$$

that holds for the solution $\mathbf{x}^*$ of Eq. (5.6). If f has a β-Lipshitz continuous gradient, i.e.

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_2 \leq \beta \|\mathbf{x} - \mathbf{y}\|_2 \ ,$$

then this algorithm can be shown to converge [41] when $\alpha \leq \beta^{-1}$. (If $f(\mathbf{x}) = \frac{1}{2}\|A\mathbf{x} - \mathbf{s}\|_2^2$ then it is easy to show that the gradient of f has a β-Lipshitz continuous gradient with $\beta = \left\|A^T A\right\|_2 = \lambda_{\max}(A^T A)$). Alternatively, the step size can be chosen via a backtracking approach [41, Sec. 1.4].

5.4.2 Iterative Shrinkage

We will show that the proximal mapping of $\lambda \|\mathbf{x}\|_1$ has an analytical solution called *soft thresholding* or *shrinkage*. The resulting proximal gradient algorithm is *Iterative Shrinkage* (see [2, Ch. 6] for detailed discussion of this method) or *iterative shrinkage-thresholding algorithm* (ISTA).

The more elegant approach to this derivation is via subdifferential calculus, but since we do not assume that as a prerequisite, we will take an alternative approach. First, we observe that each component of $\frac{1}{2}\|\mathbf{x} - \mathbf{s}\|_2^2 + \lambda \|\mathbf{x}\|_1 = \frac{1}{2}\sum_k (x_k - s_k)^2 + \lambda \sum_k |x_k|$ is independent, so that we can consider the scalar problem $\arg\min_x f(x)$ where

$$f(x) = \frac{1}{2}(x - s)^2 + \lambda |x| \ .$$

Ignoring, for now, the non-differentiability at $x = 0$, and noting that $|x| = \text{sign}(x)x$, we have

$$f'(x) = x - s + \lambda \,\text{sign}(x) \ \text{ where } x \neq 0 \ .$$

Setting the gradient to 0, we have, for an extremal point $\hat{x}$

$$\hat{x} = s - \lambda \,\text{sign}(\hat{x}) \ .$$

By inspection of $f(x)$ it is clear that a minimiser will always have the same sign as s, so we can write

$$\hat{x} = s - \lambda \,\text{sign}(s) = |s| \,\text{sign}(s) - \lambda \,\text{sign}(s) = \text{sign}(s)(|s| - \lambda) \ .$$

But we have not considered the possibility of a minimum at $x = 0$, and it is easy to show that $f(0) < f(\epsilon)$ if $\text{sign}(\epsilon) = \text{sign}(s)$ and $|s| < \lambda$. The minimiser of $f(x)$ is therefore $\text{sign}(s)(|s| - \lambda)$ when $\lambda \leq |s|$, or 0 when $\lambda > |s|$, which can be compactly expressed as

$$x^* = \max\left(0, \text{sign}(s)(|s| - \lambda)\right) = \text{sign}(s)\max\left(0, |s| - \lambda\right) \ .$$

This operation is usually referred to as *soft thresholding* or *shrinkage*, and is computationally very cheap. We will denote this operation on a vector by

$$\mathcal{S}_{1,\lambda}(\mathbf{u}) = \text{sign}(\mathbf{u}) \odot \max(0, |\mathbf{u}| - \lambda) \ ,$$

where scalar operations on a vector are applied element-wise.

It is also worth noting that when the dictionary is orthonormal, we can write $\|A\mathbf{x} - \mathbf{s}\|_2^2 = \left\|A(\mathbf{x} - A^T\mathbf{s})\right\|_2^2 = \left\|\mathbf{x} - A^T\mathbf{s}\right\|_2^2$ (since, for orthonormal A, $AA^T = I$ and $\|A\mathbf{x}\|_2 = \|\mathbf{x}\|_2$), so that the solution to the BPDN problem becomes a simple shrinkage, as applied in early wavelet-based denoising techniques [44].

5.4.3 Fast Proximal Gradient

Unfortunately, proximal gradient methods converge slowly [41]. The *fast proximal gradient* (or *accelerated proximal gradient*) method provides significantly faster convergence by modifying the gradient step to be based on an interpolation of the two most recent iterates [41]. There are two variants of the *fast iterative shrinkage-thresholding algorithm* (FISTA) [45]; here we show the simpler FISTA with constant stepsize, and refer the reader to [45] for further detail, including the variant with backtracking: starting with $\gamma^{(1)} = 1$ and $\mathbf{y}^{(1)} = \mathbf{x}^{(0)}$

$$\begin{aligned}\mathbf{x}^{(k)} &= \underset{\alpha g}{\text{prox}}\left(\mathbf{y}^{(k)} - \alpha \nabla f(\mathbf{y}^{(k)})\right) \\ \gamma^{(k+1)} &= \frac{1+\sqrt{1+4(\gamma^{(k)})^2}}{2} \\ \mathbf{y}^{(k+1)} &= \mathbf{x}^{(k)} + \frac{\gamma^{(k)}-1}{\gamma^{(k+1)}}(\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}) \ .\end{aligned}$$

Nesterov's Algorithm (NESTA) [46] is a different construction of an accelerated proximal gradient method.

5.5 Alternating Direction Method of Multipliers

In this section we introduce Alternating Direction Method of Multipliers (ADMM), a very versatile method for solving optimization problems involving non-smooth functions such as the ℓ^1 norm. Some background in convex optimization (see [8], for example) is required to understand all of the details, but not to make use of these techniques. Note that the Split Bregman [47] method is equivalent to ADMM for problems with linear constraints [48].

5.5.1 Motivation

The difficulty in solving

$$\arg\min_{\mathbf{x}} \frac{1}{2}\left\|A\mathbf{x}-\mathbf{s}\right\|_2^2 + \lambda \left\|\mathbf{x}\right\|_1$$

is due to the non-differentiability of the $\left\|\mathbf{x}\right\|_1$ at 0. Let us convert to an equivalent optimisation problem by introducing an auxiliary variable and a constraint

$$\arg\min_{\mathbf{x},\mathbf{y}} \frac{1}{2}\left\|A\mathbf{x}-\mathbf{s}\right\|_2^2 + \lambda \left\|\mathbf{y}\right\|_1 \quad \text{such that} \quad \mathbf{x}=\mathbf{y} \ ,$$

and deal with the constraint by introducing an additional penalty term

$$\arg\min_{\mathbf{x},\mathbf{y}} \frac{1}{2}\left\|A\mathbf{x}-\mathbf{s}\right\|_2^2 + \lambda \left\|\mathbf{y}\right\|_1 + \frac{\gamma}{2}\left\|\mathbf{x}-\mathbf{y}\right\|_2^2 \ ,$$

where we take γ to be large so that the penalty approximately enforces the constraint $\mathbf{x} = \mathbf{y}$. We tackle this new problem by alternating minimisation with respect to $\mathbf{x}$ and $\mathbf{y}$

$$\begin{aligned} \arg\min_{\mathbf{x}} \ & \frac{1}{2} \|A\mathbf{x} - \mathbf{s}\|_2^2 + \frac{\gamma}{2} \|\mathbf{x} - \mathbf{y}\|_2^2 \\ \arg\min_{\mathbf{y}} \ & \frac{\gamma}{2} \|\mathbf{y} - \mathbf{x}\|_2^2 + \lambda \|\mathbf{y}\|_1 \ . \end{aligned}$$

The $\mathbf{x}$ subproblem can be solved by solving the linear system

$$(A^T A + \gamma I)\mathbf{x} = A^T \mathbf{s} + \gamma \mathbf{y} \ ,$$

and the $\mathbf{y}$ subproblem is just the proximal mapping of $\frac{\lambda}{\gamma} \|\mathbf{x}\|_1$ from Sec. 5.4.2. Since this is very cheap to compute, we can construct an efficient algorithm by iteratively minimising the $\mathbf{x}$ and $\mathbf{y}$ subproblems, while increasing γ (so-called *continuation*). This approach has recently been applied to image reconstruction problems [49], but can be traced back much further. An undesirable property of this approach is that the $\mathbf{x}$ subproblem can become badly ill-conditioned when γ becomes large; this shortcoming motivates the main topic of this section.

5.5.2 Augmented Lagrangian

While the splitting technique is very effective, we would like an alternative to continuation for enforcing the necessary constraint on the problem with the auxiliary variable. The necessary ideas will first be introduced in the context of Augmented Lagrangian methods, before moving on to the Alternating Direction Method of Multipliers (ADMM) methods applicable to the problem with the auxiliary variable. (The remainder of this section will be heavily based on the excellent tutorial by Boyd et al. [3].)

Consider the convex optimisation problem (we require $f(\mathbf{x})$ is convex)

$$\arg\min_{\mathbf{x}} f(\mathbf{x}) \ \text{ such that } \ A\mathbf{x} = \mathbf{b} \ . \tag{5.8}$$

The basic approach to solving such a problem is to construct the *Lagrangian*

$$L(\mathbf{x}, \mathbf{y}) = f(\mathbf{x}) + \mathbf{y}^T (A\mathbf{x} - \mathbf{b}) \ , \tag{5.9}$$

where $\mathbf{y}$ is the vector of Lagrange multipliers, and solve for $\nabla_{\mathbf{y}} L(\mathbf{x}, \mathbf{y}) = 0$ and $\nabla_{\mathbf{x}} L(\mathbf{x}, \mathbf{y}) = 0$. The corresponding conditions for the optimal primal and dual variables $\mathbf{x}^*$ and $\mathbf{y}^*$ are the primal and dual feasibility conditions

$$A\mathbf{x}^* - \mathbf{b} = 0 \tag{5.10}$$

$$\nabla f(\mathbf{x}^*) + A^T \mathbf{y}^* = 0 \ . \tag{5.11}$$

We can also define the *dual function*

$$g(\mathbf{y}) = \inf_{\mathbf{x}} L(\mathbf{x}, \mathbf{y}) \ .$$

It is easy to show that

$$g(\mathbf{y}) \leq f(\mathbf{x}^*) \ \ \forall \mathbf{y}$$

where $\mathbf{x}^*$ is the optimal solution for the primal problem Eq. (5.8): if we take $\tilde{\mathbf{x}}$ to be any primal feasible vector (i.e. $A\tilde{\mathbf{x}} = \mathbf{b}$) we have

$$L(\tilde{\mathbf{x}}, \mathbf{y}) = f(\tilde{\mathbf{x}}) \ \ \forall \mathbf{y} \ ,$$

and therefore

$$\inf_{\mathbf{x}} L(\mathbf{x}, \mathbf{y}) \leq L(\tilde{\mathbf{x}}, \mathbf{y}) = f(\tilde{\mathbf{x}}) \ \ \forall \mathbf{y} \ ;$$

since this holds for all feasible $\tilde{\mathbf{x}}$, it must also hold for the optimal $\mathbf{x}^*$. The *dual problem* is defined as

$$\arg\max_{\mathbf{y}} g(\mathbf{y}) \ ,$$

and it is clear from the argument above that the solution to the dual problem $\mathbf{y}^* = \arg\max_{\mathbf{y}} g(\mathbf{y})$ also provides a lower bound

$$g(\mathbf{y}^*) \leq f(\mathbf{x}^*)$$

to the solution $\mathbf{x}^*$ to the primal problem, which is referred to as *weak duality* [8, Ch. 5]. Under certain conditions [8, Ch. 5], which are satisfied by problem Eq. (5.8) with convex $f(\mathbf{x})$, the solution to the dual problem has the same value as the solution to the primal problem

$$g(\mathbf{y}^*) = f(\mathbf{x}^*) \ ,$$

which is referred to as *strong duality*; in this case the desired solution can be obtained by solving the dual problem, which is often easier to solve, to obtain $\mathbf{y}^*$, followed by

$$\mathbf{x}^* = \arg\min_{\mathbf{x}} L(\mathbf{x}, \mathbf{y}^*) \ .$$

The *dual ascent* method iteratively solves the dual problem via gradient ascent. Given a previous solution $\mathbf{y}^{(k)}$, we can update the primal variable as

$$\mathbf{x}^{(k+1)} = \arg\min_{\mathbf{x}} L(\mathbf{x}, \mathbf{y}^{(k)}) \ .$$

We have

$$\begin{aligned} g(\mathbf{y}) =& \inf_{\mathbf{x}} L(\mathbf{x}, \mathbf{y}) = L(\mathbf{x}^{(k+1)}, \mathbf{y}) \\ =& f(\mathbf{x}^{(k+1)}) + \mathbf{y}^T(A\mathbf{x}^{(k+1)} - \mathbf{b}) \ , \end{aligned}$$

so that, assuming $g(\cdot)$ is differentiable,

$$\nabla g(\mathbf{y}) = A\mathbf{x}^{(k+1)} - \mathbf{b} \ ,$$

and a gradient ascent step is

$$\mathbf{y}^{(k+1)} = \mathbf{y}^{(k)} + \alpha^{(k)}(A\mathbf{x}^{(k+1)} - \mathbf{b}) \ .$$

This method is convenient, but the conditions for convergence are relatively restrictive [3, Ch. 2].

The *Augmented Lagrangian* method [3, Ch. 2][7, Ch. 17] makes dual ascent more robust, with convergence under less restrictive conditions. The Lagrangian Eq. (5.9) is augmented with a quadratic term, giving the Augmented Lagrangian

$$L_\rho(\mathbf{x}, \mathbf{y}) = f(\mathbf{x}) + \mathbf{y}^T(A\mathbf{x} - \mathbf{b}) + \frac{\rho}{2} \|A\mathbf{x} - \mathbf{b}\|_2^2 \ ,$$

where ρ is the penalty parameter, and can be considered to be the standard Lagrangian for the augmented problem

$$\arg\min_{\mathbf{x}} f(\mathbf{x}) + \frac{\rho}{2} \|A\mathbf{x} - \mathbf{b}\|_2^2 \ \text{ such that } \ A\mathbf{x} = \mathbf{b} \ .$$

It is obvious from this interpretation that solving via the Augmented Lagrangian gives the same solution as the original problem, since the additional term is zero when the constraint is satisfied. Applying dual ascent to this augmented problem gives the iterations of the *method of multipliers*

$$\begin{aligned} \mathbf{x}^{(k+1)} &= \arg\min_{\mathbf{x}} L_\rho(\mathbf{x}, \mathbf{y}^{(k)}) \\ \mathbf{y}^{(k+1)} &= \mathbf{y}^{(k)} + \rho(A\mathbf{x}^{(k+1)} - \mathbf{b}) \ , \end{aligned}$$

which converges under much more general conditions than dual ascent. Note that the gradient ascent step size is the same as the penalty parameter ρ, which can be explained as follows. The primal update selects $\mathbf{x}^{(k+1)}$ as the minimiser of $L_\rho(\mathbf{x}, \mathbf{y}^{(k)})$, so $[\nabla L_\rho(\cdot, \mathbf{y}^{(k)})](\mathbf{x}^{(k+1)}) = 0$. But

$$[\nabla L_\rho(\cdot, \mathbf{y}^{(k)})](\mathbf{x}^{(k+1)}) = \nabla f(\mathbf{x}^{(k+1)}) + A^T \left(\mathbf{y}^{(k)} + \rho(A\mathbf{x}^{(k+1)} - \mathbf{b})\right)$$

so selecting $\mathbf{y}^{(k+1)} = \mathbf{y}^{(k)} + \rho(A\mathbf{x}^{(k+1)} - \mathbf{b})$ gives

$$[\nabla L_\rho(\cdot, \mathbf{y}^{(k)})](\mathbf{x}^{(k+1)}) = \nabla f(\mathbf{x}^{(k+1)}) + A^T\mathbf{y}^{(k+1)} = 0 \ ,$$

so that the iterate $(\mathbf{x}^{(k+1)}, \mathbf{y}^{(k+1)})$ is dual feasible (i.e. it satisfies the dual optimality condition $\nabla f(\mathbf{x}^*) + A^T\mathbf{y}^* = 0$) for the original problem.

5.5.3 ADMM

The ADMM method [3, Ch. 3] is constructed as a variant of the method of multipliers applied to the problem

$$\arg\min_{\mathbf{x},\mathbf{y}} f(\mathbf{x}) + g(\mathbf{y}) \ \text{ such that } \ A\mathbf{x} + B\mathbf{y} = \mathbf{c} \ .$$

The augmented Lagrangian for this problem is

$$L_\rho(\mathbf{x}, \mathbf{y}, \mathbf{z}) = f(\mathbf{x}) + g(\mathbf{y}) + \mathbf{z}^T(A\mathbf{x} + B\mathbf{y} - \mathbf{c}) + \frac{\rho}{2} \|A\mathbf{x} + B\mathbf{y} - \mathbf{c}\|_2^2 \ ,$$

and the corresponding method of multipliers iterations would be

$$\begin{aligned} (\mathbf{x}^{(k+1)}, \mathbf{y}^{(k+1)}) &= \arg\min_{\mathbf{x},\mathbf{y}} L_\rho(\mathbf{x}, \mathbf{y}, \mathbf{z}^{(k)}) \\ \mathbf{z}^{(k+1)} &= \mathbf{z}^{(k)} + \rho(A\mathbf{x}^{(k+1)} + B\mathbf{y}^{(k+1)} - \mathbf{c}) \ . \end{aligned}$$

The modification introduced by ADMM is to alternate the $\mathbf{x}$ and $\mathbf{y}$ updates instead of performing them jointly (thus the *alternating direction*)

$$\begin{aligned}\mathbf{x}^{(k+1)} &= \arg\min_{\mathbf{x}} L_\rho(\mathbf{x}, \mathbf{y}^{(k)}, \mathbf{z}^{(k)}) \\ \mathbf{y}^{(k+1)} &= \arg\min_{\mathbf{y}} L_\rho(\mathbf{x}^{(k+1)}, \mathbf{y}, \mathbf{z}^{(k)}) \\ \mathbf{z}^{(k+1)} &= \mathbf{z}^{(k)} + \rho(A\mathbf{x}^{(k+1)} + B\mathbf{y}^{(k+1)} - \mathbf{c}) \ .\end{aligned}$$

It is often more convenient to work with the *scaled form* of ADMM, which is obtained by the change of variable to the *scaled dual variable* $\mathbf{u} = \rho^{-1}\mathbf{z}$. Defining the residual $\mathbf{r} = A\mathbf{x} + B\mathbf{y} - \mathbf{c}$ we have

$$L_\rho(\mathbf{x}, \mathbf{y}, \mathbf{z}) = f(\mathbf{x}) + g(\mathbf{y}) + \mathbf{z}^T\mathbf{r} + \frac{\rho}{2}\|\mathbf{r}\|_2^2 \ ,$$

and by replacing $\mathbf{z}$ with $\mathbf{u}$ we have

$$L_\rho(\mathbf{x}, \mathbf{y}, \mathbf{u}) = f(\mathbf{x}) + g(\mathbf{y}) + \rho\mathbf{u}^T\mathbf{r} + \frac{\rho}{2}\|\mathbf{r}\|_2^2 \ .$$

Noting that $\frac{1}{2}\|\mathbf{r} + \mathbf{u}\|_2^2 = \frac{1}{2}\|\mathbf{r}\|_2^2 + \mathbf{u}^T\mathbf{r} + \frac{1}{2}\|\mathbf{u}\|_2^2$, we can express this as

$$L_\rho(\mathbf{x}, \mathbf{y}, \mathbf{u}) = f(\mathbf{x}) + g(\mathbf{y}) + \frac{\rho}{2}\|\mathbf{r} + \mathbf{u}\|_2^2 - \frac{\rho}{2}\|\mathbf{u}\|_2^2 \ .$$

Since the minimisers of $L_\rho(\mathbf{x}, \mathbf{y}, \mathbf{u})$ with respect to $\mathbf{x}$ and $\mathbf{y}$ do not depend on the final $\frac{\rho}{2}\|\mathbf{u}\|_2^2$ term, we have

$$\begin{aligned}\arg\min_{\mathbf{x}} L_\rho(\mathbf{x}, \mathbf{y}, \mathbf{u}) &= \arg\min_{\mathbf{x}} f(\mathbf{x}) + \frac{\rho}{2}\|A\mathbf{x} + B\mathbf{y} - \mathbf{c} + \mathbf{u}\|_2^2 \\ \arg\min_{\mathbf{y}} L_\rho(\mathbf{x}, \mathbf{y}, \mathbf{u}) &= \arg\min_{\mathbf{y}} g(\mathbf{y}) + \frac{\rho}{2}\|A\mathbf{x} + B\mathbf{y} - \mathbf{c} + \mathbf{u}\|_2^2 \ ,\end{aligned}$$

leading to the iterations

$$\begin{aligned}\mathbf{x}^{(k+1)} &= \arg\min_{\mathbf{x}} f(\mathbf{x}) + \frac{\rho}{2}\left\|A\mathbf{x} + B\mathbf{y}^{(k)} - \mathbf{c} + \mathbf{u}^{(k)}\right\|_2^2 \\ \mathbf{y}^{(k+1)} &= \arg\min_{\mathbf{y}} g(\mathbf{y}) + \frac{\rho}{2}\left\|A\mathbf{x}^{(k+1)} + B\mathbf{y} - \mathbf{c} + \mathbf{u}^{(k)}\right\|_2^2 \\ \mathbf{u}^{(k+1)} &= \mathbf{u}^{(k)} + A\mathbf{x}^{(k+1)} + B\mathbf{y}^{(k+1)} - \mathbf{c} \ .\end{aligned}$$

If we define the residual at iteration k as $\mathbf{r}^{(k)} = A\mathbf{x}^{(k)} + B\mathbf{y}^{(k)} - \mathbf{c}$ it is clear that

$$\mathbf{u}^{(k)} = \mathbf{u}^{(0)} + \sum_{l=1}^{k} \mathbf{r}^{(l)}$$

so that $\mathbf{u}^{(k)}$ can be interpreted as a running sum of the constraint violation residuals. We refer the reader to [3, Sec. 3.1.1 - 3.3] for further detail on convergence, stopping conditions, and other details of the ADMM method.

The splitting of the BPDN problem introduced in Sec. 5.5.1 fits exactly within this framework, leading to the iterations (using the scaled form, but denoting the scaled dual variable by $\mathbf{z}$)

$$\mathbf{x}^{(k+1)} = \arg\min_{\mathbf{x}} \frac{1}{2} \|A\mathbf{x} - \mathbf{s}\|_2^2 + \frac{\rho}{2} \left\| \mathbf{x} - \mathbf{y}^{(k)} + \mathbf{z}^{(k)} \right\|_2^2 \tag{5.12}$$

$$\mathbf{y}^{(k+1)} = \arg\min_{\mathbf{y}} \lambda \|\mathbf{y}\|_1 + \frac{\rho}{2} \left\| \mathbf{x}^{(k+1)} - \mathbf{y} + \mathbf{z}^{(k)} \right\|_2^2 \tag{5.13}$$

$$\mathbf{z}^{(k+1)} = \mathbf{z}^{(k)} + \mathbf{x}^{(k+1)} - \mathbf{y}^{(k+1)} \; .$$

Since the right hand side of Eq. (5.12) is a simple quadratic problem, at the solution $\mathbf{x}^*$ we have

$$(A^T A + \rho I)\mathbf{x}^* = A^T \mathbf{s} + \rho(\mathbf{y}^{(k)} - \mathbf{z}^{(k)}) \; .$$

A very useful property here is that the left hand side of this equation is constant while ρ is constant, allowing application of some of the methods described in Sec. 1.2. For the right hand side of Eq. (5.13) we have

$$\begin{aligned} \arg\min_{\mathbf{y}} \lambda \|\mathbf{y}\|_1 + \frac{\rho}{2} \left\| \mathbf{x}^{(k+1)} - \mathbf{y} + \mathbf{z}^{(k)} \right\|_2^2 &= \arg\min_{\mathbf{y}} \frac{1}{2} \left\| \mathbf{y} - (\mathbf{x}^{(k+1)} + \mathbf{z}^{(k)}) \right\|_2^2 + \frac{\lambda}{\rho} \|\mathbf{y}\|_1 \\ &= \operatorname*{prox}_{\frac{\lambda}{\rho}\|\mathbf{y}\|_1} (\mathbf{x}^{(k+1)} + \mathbf{z}^{(k)}) = \mathcal{S}_{1,\frac{\lambda}{\rho}}(\mathbf{x}^{(k+1)} + \mathbf{z}^{(k)}) \; . \end{aligned}$$

There is evidence that this class of algorithm is faster than the proximal gradient methods [50].

5.5.4 Constrained Problems

The ADMM approach is also easily applied to constrained problems [51] such as

$$\arg\min_{\mathbf{x}} \|\mathbf{x}\|_1 \text{ such that } \|A\mathbf{x} - \mathbf{s}\|_2 \leq \epsilon \; .$$

Defining

$$C(\epsilon, A, \mathbf{b}) = \{\mathbf{x} \in \mathbb{R}^n \mid \; \|A\mathbf{x} - \mathbf{b}\|_2 \leq \epsilon\} \; ,$$

this problem can be written as

$$\arg\min_{\mathbf{x} \in C(\epsilon, A, \mathbf{s})} \|\mathbf{x}\|_1 \; ,$$

or using the indicator function to write it as an unconstrained problem

$$\arg\min_{\mathbf{x}} \|\mathbf{x}\|_1 + \iota_{C(\epsilon, A, \mathbf{s})}(\mathbf{x}) \; ,$$

or

$$\arg\min_{\mathbf{x}} \|\mathbf{x}\|_1 + \iota_{C(\epsilon, I, \mathbf{s})}(A\mathbf{x}) \; .$$

Now we introduce auxiliary variables $\mathbf{y}$ and $\mathbf{z}$ and apply variable splitting to give the equivalent problem

$$\arg\min_{\mathbf{x},\mathbf{y},\mathbf{z}} \|\mathbf{y}\|_1 + \iota_{C(\epsilon, I, \mathbf{s})}(\mathbf{z}) \text{ such that } \mathbf{y} = \mathbf{x}, \mathbf{z} = A\mathbf{x} \; .$$

We now have two variables to constrain (see [52] for a theoretical analysis of this case), but the corresponding ADMM iterations are easily derived from the Augmented Lagrangian formulation in the two-variable case, giving

$$\begin{aligned}
\mathbf{x}^{(k+1)} &= \arg\min_{\mathbf{x}} \frac{\gamma}{2} \left\| \mathbf{x} - \mathbf{y}^{(k)} + \mathbf{u}^{(k)} \right\|_2^2 + \frac{\delta}{2} \left\| A\mathbf{x} - \mathbf{z}^{(k)} + \mathbf{v}^{(k)} \right\|_2^2 && (5.14) \\
\mathbf{y}^{(k+1)} &= \arg\min_{\mathbf{y}} \|\mathbf{y}\|_1 + \frac{\gamma}{2} \left\| \mathbf{x}^{(k+1)} - \mathbf{y} + \mathbf{u}^{(k)} \right\|_2^2 && (5.15) \\
\mathbf{z}^{(k+1)} &= \arg\min_{\mathbf{z}} \iota_{C(\epsilon,I,\mathbf{s})}(\mathbf{z}) + \frac{\delta}{2} \left\| A\mathbf{x}^{(k+1)} - \mathbf{z} + \mathbf{v}^{(k)} \right\|_2^2 && (5.16) \\
\mathbf{u}^{(k+1)} &= \mathbf{u}^{(k)} + \mathbf{x}^{(k+1)} - \mathbf{y}^{(k+1)} \\
\mathbf{v}^{(k+1)} &= \mathbf{v}^{(k)} + A\mathbf{x}^{(k+1)} - \mathbf{z}^{(k+1)} \quad .
\end{aligned}$$

We have already encountered methods for solving the the right hand sides of Eq. (5.14) and Eq. (5.15), but we are left with Eq. (5.16). From the geometric interpretation of $\arg\min_{\mathbf{x}} \frac{1}{2}\|\mathbf{x} - \mathbf{b}\|_2^2 + \iota_{C(\epsilon,I,\mathbf{s})}(\mathbf{x})$ as the projection of $\mathbf{b}$ into the constraint set $C(\epsilon, I, \mathbf{s})$, which is simply an ℓ^2 ball, it is easily derived that

$$\arg\min_{\mathbf{x}} \frac{1}{2}\|\mathbf{x} - \mathbf{b}\|_2^2 + \iota_{E(\epsilon,I,\mathbf{s})}(\mathbf{x}) = \begin{cases} \mathbf{s} + \epsilon \frac{\mathbf{b}-\mathbf{s}}{\|\mathbf{b}-\mathbf{s}\|_2} & \text{if } \|\mathbf{b} - \mathbf{s}\|_2 > \epsilon \\ \mathbf{b} & \text{if } \|\mathbf{b} - \mathbf{s}\|_2 \leq \epsilon \end{cases},$$

allowing us to solve Eq. (5.16).

Chapter 6

Extended Problems

In this chapter we consider a variety of more advanced problems that are often necessary in practical applications.

6.1 Dictionary Learning

Thus far we have assumed that the dictionary is known (constructed, perhaps, from an analytically defined set of functions such as a Fourier basis or wavelets), and have concentrated on theory and algorithms surrounding the selection of the best sparse representation for a signal with respect to this fixed dictionary. We now turn to the more difficult question of optimising the dictionary for a specific problem [1, Sec. 5.2] [2, Ch. 12] [53]. Since it does not make sense to optimise the dictionary for a specific signal (the optimal dictionary would have a single column consisting of the signal itself), we optimise the dictionary to represent an entire *training set* of signals relevant to a specific problem. The first step in achieving this goal is usually to consider sparse coding within the *simultaneous sparse approximation* [54] or *multiple-measurement-vector* (MMV) [55] framework. An MMV version of BPDN, for example, can be expressed as

$$\arg\min_X \frac{1}{2} \|AX - S\|_F^2 + \lambda \|X\|_1 \ , \tag{6.1}$$

where X and S are now matrices instead of vectors, and $\|\cdot\|_F$ denotes the Frobenius norm. While it is computationally more efficient to simultaneously solve for all of the columns of X in functionals such as this with an ℓ^2 data fidelity term and ℓ^1 regularisation term (or the corresponding constrained variants), there is no coupling between these columns, which can therefore, in principle, be solved independently using one of the single-measurement-vector (SMV) algorithms we have already encountered. This is usually not done in practice since most SMV algorithms are easily extended to the MMV case.

When the dictionary is desired to be optimised to minimise such a functional, the dictionary learning problem can be expressed as the joint optimisation over the dictionary and coefficient matrices

$$\arg\min_{A,X} \frac{1}{2} \|AX - S\|_F^2 + \lambda \|X\|_1 \quad \text{such that} \quad \|\mathbf{a}_k\|_2 = 1 \ \forall k \ ,$$

where the normalisation of the columns of A is required to avoid the entries of X becoming arbitrarily small, compensated by unconstrained growth in the magnitude of the columns of A. (The exposition here is based on the unconstrained (BPDN) problem, but note that dictionary learning algorithms are often based on constrained problems such as

$$\arg\min_{A,X} \frac{1}{2} \|AX - S\|_F^2 \quad \text{such that} \quad \|X\|_1 \leq \tau \;, \|\mathbf{a}_k\|_2 = 1 \;\forall k \;.$$

) These joint optimisations can be interpreted as matrix factorization problems (such as Non-negative Matrix Factorization), finding an approximate factorisation $S \approx AX$ with the requirement that X be sparse. All of the algorithms we will consider here solve the joint optimisation problem by alternating minimisation, starting with some initial dictionary (common choices are an analytically defined dictionary or a subset of the training data) and then alternating between optimising the coefficients X and the dictionary A. The primary difference between these algorithms is in how they perform the dictionary update step.

The following results will be useful in the remainder of this section:

$$\nabla_X \frac{1}{2} \|AX - S\|_F^2 = A^T(AX - S)$$
$$\nabla_A \frac{1}{2} \|AX - S\|_F^2 = (AX - S)X^T \;.$$

(Note that the Matrix Cookbook is a valuable resource for looking up vector and matrix derivatives [56, Ch. 2].)

6.1.1 Gradient Descent

The first dictionary learning method, due to Olshausen and Field [57], alternated between solving for the coefficient matrix X and taking a gradient descent step to update A. That is, given coefficient and dictionary matrices $X^{(k)}$ and $A^{(k)}$ respectively at step k,

$$A^{(k+1)} = A^{(k)} - \eta \left(A^{(k)}X^{(k)} - S\right) \left(X^{(k)}\right)^T \;,$$

with a fixed step size η, followed by normalisation of the columns of $A^{(k+1)}$. Note that this can be interpreted within the projected gradient framework (see Sec. 5.4.1) as

$$A^{(k+1)} = P_{\mathcal{C}} \left(A^{(k)} - \eta \left(A^{(k)}X^{(k)} - S\right) \left(X^{(k)}\right)^T \right)$$

where $P_{\mathcal{C}}$ is the projection into the set $\mathcal{C} = \{A \mid \|\mathbf{a}_k\|_2 = 1 \;\forall k\}$.

6.1.2 Method of Optimal Directions

Instead of taking a gradient descent step in the direction that reduces the error $\|AX - S\|_F^2$, the Method of Optimal Directions [58] (MOD) updates the dictionary to be the least squares solution at each step

$$A^{(k+1)} = \arg\min_A \frac{1}{2} \left\| AX^{(k)} - S \right\|_F^2 = S \left(X^{(k)}\right)^T \left(X^{(k)} \left(X^{(k)}\right)^T \right)^{-1} \;.$$

6.1.3 K-SVD

Instead of alternating between updates of the coefficients and dictionary, the K-SVD [59] alternates between an update of the coefficients and a joint dictionary and coefficient update, consisting of an iteration over each atom in the dictionary, jointly updating that atom and the coefficient elements corresponding to that atom. For dictionary atom l, define the residual error if atom l were not used (i.e. all of the corresponding coefficients would be zero)

$$E_l = S - AX + \mathbf{a}_l \mathbf{x}_l^T \ ,$$

where $\mathbf{a}_l$ is column l of A, and (by some abuse of notation) $\mathbf{x}_l^T$ is row l of X. Now selecting the optimum values of $\mathbf{a}_l$ and $\mathbf{x}_l^T$ can be expressed as

$$\arg\min_{\mathbf{a},\mathbf{x}} \left\| E_l - \mathbf{a}\mathbf{x}^T \right\|_F^2 \text{ such that } \ \|\mathbf{a}\|_2 = 1 \ .$$

Since $\mathbf{a}\mathbf{x}^T$ has rank 1, the minimisers of this functional correspond to the optimum rank 1 approximation of E_l, which is given by the SVD of E_l. That is, given the SVD $E_l = U\Sigma V^T$, set $\mathbf{a} = \mathbf{u}_1$ and $\mathbf{x} = \sigma_1 \mathbf{v}_1$. (Note that a full SVD is not necessary since only the first singular value and vectors are needed.) The flaw in this approach is that this update will usually not give a sparse $\mathbf{x}$; this is remedied by defining $\hat{E}_l$ as the restriction of E_l to only those columns k for which the corresponding entries in $\mathbf{x}_l^T$ are non-zero, and applying the same procedure to $\hat{E}_l$.

6.1.4 Other Methods

Here we briefly mention a few other methods that deserve further study

- It is argued that the choice of a large step size in the gradient descent method provides better results than standard gradient descent, MOD, or K-SVD [60].
- An interesting alternative to the standard method, which consists of a dictionary update following a full optimisation of the coefficients, is to perform dictionary updates interleaved with steps of an ADMM algorithm for sparse coding [61].
- A stochastic gradient descent / block coordinate descent algorithm has been proposed for dictionary learning with very large data sets for which standard methods are not computationally feasible [62].
- Some attention has been given to the construction of *multiscale* dictionaries [63, 64].
- Instead of designing dictionaries by minimising reconstruction error, *supervised*, *task-driven*, or *discriminative* dictionary learning [65, 66, 67, 68, 69, 70, 71] attempts to optimise a dictionary for the specific task of interest, such as classification problems. Although this type of dictionary learning problem is substantially more difficult, these methods are very promising for applications that do not explicitly depend on reconstruction error.

6.2 Structured Sparsity

As we have seen, a sparse representation is one that has few non-zero coefficients, without any additional assumptions regarding the distribution[1] of those coefficients within the coefficient vector. It is, however, possible and often useful to impose additional structure on the pattern of non-zero coefficient, leading to the notion of *structured sparse representations*. The simplest example of such structure is *joint sparsity* in the MMV context (which we have already encountered when discussing dictionary learning), introduced below.

6.2.1 Joint Sparsity

As we have already discussed, the MMV versions of SMV algorithms (e.g. BPDN Eq. (6.1)) have no coupling between the columns of the solution. We would often expect, however, that related signals (i.e. the columns of the signal matrix) would have related sparse representations, in the sense that they would tend to make use of the same dictionary elements (i.e. have the same support set), in which case such coupling is desirable. Perhaps the most straightforward approach is via a relatively simple modification of OMP in which a single atom is selected to update the approximation to all of the measurement vectors at each step [72, 73, 54].

For greater flexibility, however, it would be convenient to have a regularisation term that promotes the desired coupling between solutions in the MMV context. Let us define the $\ell^{p,q}$ norm (or ℓ^p/ℓ^q mixed norm) of a matrix

$$\|X\|_{p,q} = \left(\sum_k \left(\sum_l |X_{k,l}|^p \right)^{\frac{q}{p}} \right)^{\frac{1}{q}} \;,$$

which is just the ℓ^q norm of the vector consisting of the ℓ^p norms of each row of the matrix [74]. (Do not confuse this norm with the operator norm $\sup_{\mathbf{x}\neq 0} \frac{\|A\mathbf{x}\|_p}{\|\mathbf{x}\|_q}$, which is often denoted using the same notation [54].) If we set $q = 0$, this "norm" corresponds to the number of rows of the matrix that have a non-zero ℓ^p norm; clearly a penalty on this norm will promote solutions with as few non-zero ℓ^p norm rows as possible, which corresponds to promoting solutions in which each column has its non-zero values on the same support set. As in the case of the ℓ^0 "norm", optimising with respect to the $\ell^{p,0}$ norm is not tractable, and we turn to a convex relaxation, choosing $q = 1$. This norm is just the sum (or ℓ^1 norm) of the ℓ^p norm of each row of the matrix. Since the ℓ^1 norm promotes a sparse solution, we can expect that this norm will promote a solution matrix for which the vector of ℓ^p norms of the rows is sparse, i.e. for which most rows have small or zero ℓ^p norm and only a few have large ℓ^p norm. (This is not the case for $p = 1$ since the $\ell^{1,1}$ norm is just the ℓ^1 norm of the matrix considered as a vector, with no coupling between the columns.) A number of authors have examined conditions under which $\ell^{p,1}$ norms (the $\ell^{2,1}$ norm in particular) are equivalent to $\ell^{p,0}$ norms, or give unique solutions to the corresponding reconstruction problems [73, 75, 76, 77, 78, 74, 79, 80, 81].

Here we will concentrate on the $\ell^{2,1}$ since it is the most widely used (note that the $\ell^{\infty,1}$ norm, corresponding the the sum of the maximum values in each row, has also received some

[1] The term is not used on the sense of a probability distribution here.

attention [82, 83]), and the easiest to deal with computationaly. Consider the $\ell^{2,1}$ variant of Joint BPDN

$$\arg\min_X \frac{1}{2} \|AX - S\|_F^2 + \lambda \|X\|_{2,1} \ . \tag{6.2}$$

A number of authors have derived IRLS algorithms for this problem [72, 84, 85], but here we only provide detail of an ADMM algorithm (which has been analysed in some detail by Deng et al. [86]).

Just as in Sec. 5.5.1, we can deal with this problem via a splitting approach with the introduction of an auxiliary variable

$$\arg\min_{X,Y} \frac{1}{2} \|AX - S\|_F^2 + \lambda \|Y\|_{2,1} \quad \text{such that} \quad X = Y \ .$$

Applying the ADMM framework of Sec. 5.5.3 for this constrained problem, we have the iterations

$$X^{(k+1)} = \arg\min_X \frac{1}{2} \|AX - S\|_F^2 + \frac{\rho}{2} \left\| X - Y^{(k)} + Z^{(k)} \right\|_F^2 \tag{6.3}$$

$$Y^{(k+1)} = \arg\min_Y \lambda \|Y\|_{2,1} + \frac{\rho}{2} \left\| X^{(k+1)} - Y + Z^{(k)} \right\|_F^2 \tag{6.4}$$

$$Z^{(k+1)} = Z^{(k)} + X^{(k+1)} - Y^{(k+1)} \ .$$

Since the right hand side of Eq. (6.3) is a simple quadratic problem, at the solution X^* we have

$$(A^T A + \rho I) X^* = A^T S + \rho (Y^{(k)} - Z^{(k)}) \ .$$

For the right hand side of Eq. (6.4) we have

$$\begin{aligned} \arg\min_Y \lambda \|Y\|_{2,1} + \frac{\rho}{2} \left\| X^{(k+1)} - Y + Z^{(k)} \right\|_F^2 &= \arg\min_Y \frac{1}{2} \left\| Y - (X^{(k+1)} + Z^{(k)}) \right\|_F^2 + \frac{\lambda}{\rho} \|Y\|_{2,1} \\ &= \operatorname*{prox}_{\frac{\lambda}{\rho}\|Y\|_{2,1}} (X^{(k+1)} + Z^{(k)}) \ . \end{aligned}$$

To solve this second subproblem, we need to be able to compute

$$\operatorname*{prox}_{\gamma\|\cdot\|_{2,1}} (S) = \arg\min_X \frac{1}{2} \|X - S\|_F^2 + \gamma \|X\|_{2,1} \ .$$

Observe that the rows are decoupled, so we can solve each row independently, and therefore only need to consider the problem for a vector

$$\operatorname*{prox}_{\gamma\|\cdot\|_2} (\mathbf{s}) = \arg\min_{\mathbf{x}} \frac{1}{2} \|\mathbf{x} - \mathbf{s}\|_2^2 + \gamma \|\mathbf{x}\|_2 \ .$$

Via a derivation similar to that in Sec. 5.4.2, it is not difficult to show that

$$\operatorname*{prox}_{\gamma\|\cdot\|_2} (\mathbf{u}) = \mathcal{S}_{2,\gamma}(\mathbf{u}) = \frac{\mathbf{u}}{\|\mathbf{u}\|_2} \max(0, \|\mathbf{u}\|_2 - \gamma) \ .$$

6.2.2 Group Sparsity

In contrast to joint sparsity, *group sparsity* (also referred to as *block sparsity*) is not inherently defined within the MMV framework. In this case, the model is that the vector of coefficients is not just sparse, but that the non-zero coefficients occur in pre-defined groups (or blocks). With $\mathbf{x} \in \mathbb{R}^N$, and denoting the kth group of element indices as g_k and the corresponding sub-vector of $\mathbf{x}$ as $\mathbf{x}_{g_k}$, this model corresponds to the assumption that the vector of group norms $\begin{pmatrix} \|\mathbf{x}_{g_1}\|_p & \|\mathbf{x}_{g_2}\|_p & \dots & \|\mathbf{x}_{g_N}\|_p \end{pmatrix}$ is sparse. (The simplest case computationally is when the groups g_k form a partition of the elements of $\mathbf{x}$, but overlapping groups and incomplete cover of the elements of $\mathbf{x}$ are also tractable.) As in the case of joint sparsity, a mixed norm approach provides a suitable regularisation term, but now the grouping is not naturally defined by the rows of the matrix, and the norm has to be defined with respect to the groups

$$\|\mathbf{x}\|_{p,q} = \left\| \begin{pmatrix} \|\mathbf{x}_{g_1}\|_p & \|\mathbf{x}_{g_2}\|_p & \dots & \|\mathbf{x}_{g_N}\|_p \end{pmatrix} \right\|_q \ ,$$

or for the most common case $q = 1$,

$$\|\mathbf{x}\|_{p,1} = \sum_{k=1}^{N} \|\mathbf{x}_{g_k}\|_p \ .$$

The $\ell^{2,1}$ variant of the Group BPDN problem (often referred to as the Group Lasso [87]) is

$$\arg\min_{\mathbf{x}} \|A\mathbf{x} - \mathbf{s}\|_2^2 + \lambda \|\mathbf{x}\|_{2,1} \ ,$$

which can, like the Joint BPDN problem Eq. (6.2), be efficiently solved via an ADMM algorithm [3, Sec. 6.4.2][86]. (This is not surprising, since joint sparsity can be posed as a special case of group sparsity in which each matrix row represents a group.) It is possible to define groups within an MMV framework, giving simultaneous joint and group sparsity [88, 89].

We will not go into detail here, but it is worth mentioning that group sparsity can be used to define hierarchical structure via overlapping groups [90, 91].

6.3 Compressed Sensing

Results such as Theorems 2.1 and 4.1 (and many more we have not reviewed here) show that if we have $A\mathbf{x} = \mathbf{s}$, a sufficiently sparse $\mathbf{x}$ can be recovered, given A and $\mathbf{s}$, by the solution to a problem such as $\mathcal{P}_0$ or $\mathcal{P}_1$. If $A \in \mathbb{R}^{M \times N}$ has suitable spark, mutual coherence, etc. properties, this recovery is possible even when $N \gg M$, i.e. when A is highly overcomplete. When we apply sparse representation methods we usually think of $\mathbf{s} \in \mathbb{R}^M$ as our measured signal and $\mathbf{x} \in \mathbb{R}^N$ as some *latent* generating vector on an appropriate dictionary A. Compressed Sensing [92, 93, 94, 95, 96, 97] is essentially the observation that it is possible to change perspective slightly and view $\mathbf{x}$ as a physical signal to be sampled (or perhaps more accurately, *resampled*, since it is already considered to be discrete), A as a linear sampling operator, and $\mathbf{s}$ as the sampled representation. (Sampling by taking inner products with the rows of a sampling matrix is unconventional, but it is quite feasible to implement a mechanism for

sampling physical signals for which this is an appropriate model.) If we know that the signal to be sampled, $\mathbf{x}$, is sufficiently sparse, and can construct a *measurement matrix* A with the necessary properties, then it is possible to recover the signal consisting of N values from the measurement consisting of M values, with $M \ll N$.

Much of the theory of compressed sensing has revolved around the construction of dictionaries with suitable stability properties. It turns out that random matrices represent good choices for the measurement matrix since they can be shown, under some conditions, to have, with high probability, the necessary stability properties for recovery via problem $\mathcal{P}_1$. We will discuss random Gaussian matrices, but other choices are also possible [19, Sec. 1.3], including random subsets of the columns of Fourier transform matrices, allowing application of these methods to inherently Fourier domain problems such as Magnetic Resonance Imaging.

Theorem 6.1. *If $A \in \mathbb{R}^{M \times N}$ is a random matrix with entries sampled independently from a Gaussian distribution with zero mean and variance $1/M$, then for a specified δ_L there exist constants C_1 and C_2, where C_1 can be chosen as convenient, and C_2 depends only on C_1 and δ_L, so that if*

$$L \leq C_1 \frac{M}{\log(N/L)} \tag{6.5}$$

then A has L-restricted isometry constant δ_L with probability at least $1 - 2\exp(-C_2 M)$.

Proof. See [98]. ∎

We can use this result to construct a random matrix so that, with very high probability, $\delta_{2L} < \sqrt{2} - 1$, and therefore, according to Theorem 4.5, the signal sampled using A can be recovered via problem $\mathcal{P}_1$ with the same probability. An equivalent constraint to Eq. (6.5) is

$$M \geq C_1^{-1} L \log(N/L) \ ,$$

which indicates that $O(L \log(N/L))$ measurements are sufficient to recover an L-sparse $\mathbf{x}$ measured by A.

It is, of course, unrealistically restrictive to assume that the signal of interest is itself sparse (i.e. that it is sparse on the dictionary corresponding to the identity matrix). It is trivial, however, to extend this approach to a signal $\mathbf{y}$ that is sparse on an arbitrary dictionary, Ψ, by choosing a suitable measurement matrix Φ and constructing the measurement as $\Phi\mathbf{y} = \mathbf{s}$. Since $\mathbf{y}$ has a sparse representation on dictionary Ψ, we have $\mathbf{y} = \Psi\mathbf{x}$ for some sparse $\mathbf{x}$, which can be recovered by finding the sparsest $\mathbf{x}$ such that $\Phi\Psi\mathbf{x} = \mathbf{s}$. In this more general case, the ability to recover $\mathbf{x}$ (and therefore $\mathbf{y}$) depends on the properties of the product dictionary $\Phi\Psi$. If Ψ is orthonormal then the product $\Phi\Psi$ has the same stability properties as Φ, otherwise additional arguments are necessary [99].

An additional unrealistic assumption is that the sampling can be performed without any noise. Since the conditions stated above for exact recovery apply to a stability result, the same conditions provide for a bound on the recovery error subject to noise when utilising problem $\mathcal{P}_1^\epsilon$

$$\arg\min_{\mathbf{x}} \|\mathbf{x}\|_1 \quad \text{such that} \quad \|\Phi\Psi\mathbf{x} - \mathbf{s}\|_2 \leq \epsilon \ ,$$

where the signal to be reconstructed is recovered as $\mathbf{y} = \Phi\mathbf{x}$. Recovery using the analysis form

$$\arg\min_{\mathbf{y}} \|\Psi\mathbf{y}\|_1 \quad \text{such that} \quad \|\Phi\mathbf{y} - \mathbf{s}\|_2 \leq \epsilon$$

is also possible, but there are far fewer theoretical results for this case [99, Sec. 1.3].

Compressed sensing as a research field has experienced a very rapid growth in recent years, and has driven much the the recent progress in theory and algorithms for sparse representations. Nevertheless, it is worth emphasising that compressed sensing is a subfield of sparse representations, and the relatively common practice of referring to anything involving sparse representations as compressed sensing should be avoided; if the problem at hand does not involve sensing or sampling, the technique is not compressed sensing.

6.4 Matrix Completion and Robust PCA

An interesting problem that grew out of Compressed Sensing is matrix completion: given only a small fraction of the entries of matrix D, estimate the full matrix. Clearly this is impossible without some knowledge of the structure of D. A suitable model is that D is low-rank, which corresponds to sparsity of the singular values in the SVD of D. If we define $\Omega = \{(k,l) \mid D_{k,l} \text{ is known}\}$ and the projection $P_\Omega(X)$ as the linear projection of matrix X onto the set of known elements in D, then the natural optimisation to recover D under the low-rank model is

$$\arg\min_X \text{rank}(X) \text{ such that } P_\Omega(X) = P_\Omega(D) \; .$$

Unfortunately this problem is NP-hard [100]. The tightest convex relaxation of rank is the *nuclear norm* [23, Sec. II]

$$\|X\|_* = \sum_k \sigma(X)$$

(the sum of the singular values of a matrix instead of the number of non-zero singular values), allowing us to replace the intractable problem with the convex problem [100]

$$\arg\min_X \|X\|_* \text{ such that } P_\Omega(X) - P_\Omega(D) \; .$$

It can be shown, for uniform random sampling of random orthogonal matrices, that the optimization above is capable of exactly recovering the full matrix D with high probability when only a small fraction of the entries are sampled [100, Thrm 1.1].

Under more realistic conditions that include noise in the measurements of the entries of D, the appropriate optimization becomes

$$\arg\min_X \|X\|_* \text{ such that } \|P_\Omega(X) - P_\Omega(D)\|_F \leq \epsilon \; ,$$

or the unconstrained form

$$\arg\min_X \frac{1}{2} \|P_\Omega(X) - P_\Omega(D)\|_F^2 + \lambda \|X\|_* \; .$$

It can be shown that under certain conditions, this recovery is stable in the presence of noise [23].

The proximal mapping of $\lambda\|X\|_*$ can be computed as a soft thresholding of the singular values of X, i.e. [101, Thrm 2.1]

$$\arg\min_X \frac{1}{2}\|X - Y\|_F^2 + \lambda\|X\|_* = U \operatorname{diag}(\max(0, \sigma_k - \lambda))V^T \ ,$$

where $Y = U\Sigma V^T$ and σ_k is the kth element on the diagonal of Σ. This result allows efficient algorithms for solving the optimisation problems above to be constructed using proximal methods or ADMM. Note, however, that since the SVD is computationally expensive, the proximal mapping of $\lambda\|X\|_*$ is far more expensive to compute than that of $\lambda\|X\|_1$. If the matrix in question is known to have very low rank, an iterative computation [102] of the first few singular values and vectors of the matrix can be far more efficient than a full computation of the SVD.

6.4.1 Robust PCA

We can use the ideas above to construct a version of Principal Component Analysis (PCA) that is robust to outliers. Since PCA seeks a low-rank approximation to the data, and outliers are in general sparse, we have the optimisation

$$\arg\min_{X,Y} \operatorname{rank}(X) + \lambda\|Y\|_0 \text{ such that } X + Y = D \ ,$$

and since this is intractable, we again replace it with its convex relaxation [103]

$$\arg\min_{X,Y} \|X\|_* + \lambda \|Y\|_1 \text{ such that } X + Y = D \ ,$$

which can be solved via an Augmented Lagrangian algorithm [103]. Theory gives a universally good parameter choice $\lambda = 1/\sqrt{\max(M, N)}$ for $M \times N$ matrix D.

Chapter 7

Applications

We now turn to applications of the techniques we have discussed in the previous chapters. In most (if not all) of these applications, the sparse representations play the role, either explicitly or implicitly, of a signal (or image) model. These models can be interpreted within the MAP estimation framework, as we have seen in Sec. 1.3. It is also worth noting that ℓ^1 (and other norms, but here we restrict our discussion to this specific case) regularisation methods involving a forward operator F and dictionary A can be divided into two broad classes, namely the *analysis*

$$\arg\min_{\mathbf{x}} \|F(\mathbf{x}) - \mathbf{s}\|_2^2 + \lambda \|A\mathbf{x}\|_1$$

and *synthesis*

$$\arg\min_{\mathbf{x}} \|F(A\mathbf{x}) - \mathbf{s}\|_2^2 + \lambda \|\mathbf{x}\|_1$$

frameworks. These are equivalent when A is invertible, but in general these two frameworks have substantially different properties [6]. The reader is also referred to a recent survey with a coherent discussion of sparse representations as signal models, and the role of this model in various applications [104], and also [2, Ch. 9].

At a simplified level, if the signal or image analysis problem we wish to solve can be posed as an inverse problem involving a (preferably linear) forward operator F, then we can solve it by choosing (or constructing) an appropriate dictionary A and solving one of the analysis or synthesis problems (or related constrained forms) above. Of course, each signal processing problem has its own unique features and complications which may require modifications to this simple framework, but this does not alter the conceptually simple underlying framework.

7.1 Denoising

Denoising is the simplest of all image restoration problems in the sense that the forward operator F is the identity transform. Arguably the earliest application of sparse representations in signal denoising is the wavelet soft-thresholding method [44]. If we wish to estimate $\mathbf{x}$

given $\mathbf{s} = \mathbf{x} + \boldsymbol{\nu}$, where $\boldsymbol{\nu}$ is additive white noise, the wavelet soft-thresholding estimator of $\mathbf{s}$, with A the synthesis operator for an orthonormal wavelet basis, is

$$\hat{\mathbf{x}} = A \mathcal{S}_{1,\lambda}(A^T \mathbf{s}) \; .$$

As noted at the end of Sec. 5.4.2, this is equivalent to

$$\hat{\mathbf{x}} = A \arg\min_{\mathbf{x}} \left\{ \frac{1}{2} \left\| \mathbf{x} - A^T \mathbf{s} \right\|_2^2 + \lambda \left\| \mathbf{x} \right\|_1 \right\} \; ,$$

which establishes the relationship of this method to sparse representations. When A is an overcomplete tight frame (e.g. a shift-invariant wavelet basis) we have $AA^T = I$ but not $\|A\mathbf{x}\|_2 = \|\mathbf{x}\|_2$) so that the optimisation cannot be reduced to a simple soft-thresholding in the transform domain [104]. A learned dictionary A is usually also not orthogonal, and thus also requires solution of an optimisation problem such as that above, or

$$\arg\min_{\mathbf{x}} \|\mathbf{x}\|_1 \;\; \text{such that} \;\; \|A\mathbf{x} - \mathbf{s}\|_2 \leq \epsilon \; .$$

Thus far we have assumed that the signal or image to be denoised is processed as a single unit, but when we use a learned dictionary A, it is impractical to solve the relevant optimisation problem on more than a small image patch at a time, making it necessary to break the image into patches, solve an optimisation problem on each patch, and then re-assemble the solutions into a final estimate of the denoised image. (This is a good example of the type of complication mentioned in the introduction to this chapter.) One way to formulate the problem is [105][2, Ch. 14]

$$\arg\min_{\mathbf{y}, \{\mathbf{x}_k\}_{k \in \mathcal{B}}} \frac{1}{2} \sum_{k \in \mathcal{B}} \|D\mathbf{x}_k - R_k \mathbf{y}\|_2^2 + \sum_{k \in \mathcal{B}} \mu_k \|\mathbf{x}_k\|_0 + \frac{\lambda}{2} \|\mathbf{y} - \mathbf{s}\|_2^2 \; ,$$

where $\mathbf{y}$ is the denoised image, $\mathbf{s}$ is the noisy image, D is a dictionary for image blocks, learned using the K-SVD, R_k is the operator extracting the kth block from its argument, and $\mathcal{B}$ is the set of image block indices. Solving via alternating minimisation, fixing $\mathbf{y}$ and solving for $\{\mathbf{x}_k\}_{k \in \mathcal{B}}$, and then vice versa, we have the subproblems

$$\arg\min_{\{\mathbf{x}_k\}_{k \in \mathcal{B}}} \frac{1}{2} \sum_{k \in \mathcal{B}} \|D\mathbf{x}_k - R_k \mathbf{y}\|_2^2 + \sum_{k \in \mathcal{B}} \mu_k \|\mathbf{x}\|_0$$

$$\arg\min_{\mathbf{y}} \frac{1}{2} \sum_{k \in \mathcal{B}} \|R_k \mathbf{y} - D\mathbf{x}_k\|_2^2 + \frac{\lambda}{2} \|\mathbf{y} - \mathbf{s}\|_2^2 \; .$$

The first subproblem can be solved independently for each block k via OMP, and the second subproblem has an analytical solution. This approach has been extended to colour images with non-uniform noise [106]. Improved performance has been obtained using an alternative dictionary learning algorithm [61], and via joint sparse coding of matching blocks combined with an online dictionary learning algorithm [107].

7.2 Inpainting

Image inpainting [108] is an image restoration problem that attempts to fill in a specified (usually contiguous) region of missing pixels with visually "reasonable" content so that the

inpainted region is not clearly noticeable to a human viewer. It is the simplest inverse problem other than denoising in the sense that the relevant forward operator is just a diagonal matrix with zero entries corresponding to unknown pixels and unit entries corresponding to known pixels. Defining such an operator P, and choosing an appropriate dictionary D, a solution can be constructed via an optimization such as

$$\arg\min_{\mathbf{x}} \frac{1}{2} \|PD\mathbf{x} - P\mathbf{s}\|_2^2 + \lambda \|\mathbf{x}\|_1 \ ,$$

where $\mathbf{s}$ is the image to be inpainted [109, 110]. Applying such an estimation strategy at the block level is more complicated than in the denoising problem. The most common approaches are to progressively fill in the missing region starting from the border with the known region; this can either be done following a fixed set of layers defined according to the distance from the boundary of the known region [111, 112], or adaptively in an attempt to give priority to significant image structure leading from the known region into the inpainting region [113, 114]. Alternatively, the need to choose a fill-order can be avoided by simultaneously estimating a set of overlapping blocks covering the inpainting region [115].

7.3 Deconvolution

Given an image formation model

$$\mathbf{s} = \mathbf{h} \star \mathbf{y} + \boldsymbol{\nu} \ ,$$

or equivalently

$$\mathbf{s} = H\mathbf{y} + \boldsymbol{\nu} \ ,$$

where $\mathbf{s}$ is the measured image, $\mathbf{h}$ is a convolution kernel and H is the corresponding linear operator on the entire image, $\mathbf{y}$ is the original image, and $\boldsymbol{\nu}$ is additive noise, the goal of deconvolution is to estimate the original image $\mathbf{x}$. If we also assume that $\mathbf{y}$ has a sparse representation on dictionary D

$$\mathbf{y} = D\mathbf{x}$$

then we have

$$\mathbf{s} = HD\mathbf{x} + \boldsymbol{\nu} \ ,$$

An obvious approach is to solve a problem such as [2, Ch. 10]

$$\arg\min_{\mathbf{x}} \frac{1}{2} \|HD\mathbf{x} - \mathbf{s}\|_2^2 + \lambda \|\mathbf{x}\|_1 \ ,$$

and then estimate the original image as $\hat{\mathbf{y}} = D\mathbf{x}$.

If we wish to use a learned dictionary, it becomes necessary to work with image patches. The dictionary can be trained using a large set of un-blurred patches (taken from a set of un-blurred images with similar content to $\mathbf{x}$), and the optimisation problem for each patch k can be set up as a problem such as

$$\arg\min_{\mathbf{x}_k} \frac{1}{2} \|PH_pD\mathbf{x} - PR_k\mathbf{s}\|_2^2 + \lambda \|\mathbf{x}\|_1 \ ,$$

where H_p is the convolution operator for patch-sized images, R_k is the block extraction operator (as used in the denoising method above), and P is a projection operator that projects the patch to a smaller central region (depending on the size of the convolution kernel, see Fig. 7.1) in which the convolution corresponding to the operator H_p is not affected by boundary effects. The final deconvolved image could be reconstructed as the solution to

$$\arg\min_{\mathbf{y}} \frac{1}{2} \sum_k \|R_k \mathbf{y} - D\mathbf{x}_k\|_2^2 \ ,$$

i.e. by averaging

$$\hat{\mathbf{y}} = \left(\sum_k R_k^T R_k \right)^{-1} \sum_k R_k^T D\mathbf{x}_k \ .$$

This type of approach, with some additional refinements, has been found to be quite effective [116], as has a far more complex approach involving simultaneous learning of dictionaries for blurred and un-blurred patches [117].

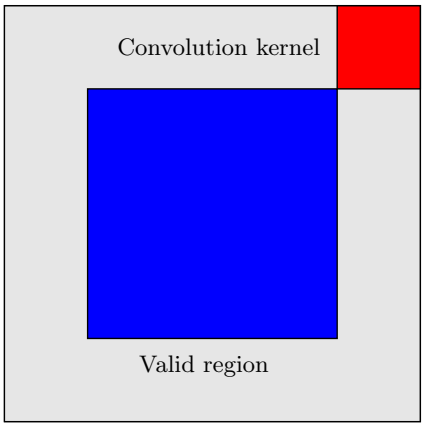


Figure 7.1: Illustration of the region of an image patch that is not influenced by boundary effects when convolving with a specific kernel.

7.4 Superresolution

In this section we consider the problem of single-image superresolution (which should not be confused with the more standard superresolution problem of estimating a high-resolution image given multiple images of the same scene). The forward model in this case is

$$\mathbf{s} = SH\mathbf{y} + \boldsymbol{\nu} \ ,$$

where S is a downsampling operator and H is the linear operator corresponding to a convolution by a lowpass filter. As in the case of deconvolution, a possible approach is via a problem such as

$$\arg\min_{\mathbf{x}} \frac{1}{2} \|SHD\mathbf{x} - \mathbf{s}\|_2^2 + \lambda \|\mathbf{x}\|_1 \ .$$

One of the most well-known block-based approaches is based on simultaneous learning of two dictionaries, one each for low and high resolution patches [118]. Given a set of corresponding low and high dimensional training patches, denoted Y_L and Y_H respectively, and each patch vector being M_L and M_H dimensional respectively, these dictionaries are constructed via the standard dictionary learning problem

$$\arg\min_{D,X} \frac{1}{2} \|DX - Y\|_2^2 + \lambda \|X\|_1$$

where

$$D = \begin{pmatrix} \frac{1}{\sqrt{M_L}} D_L \\ \frac{1}{\sqrt{M_H}} D_H \end{pmatrix} \quad Y = \begin{pmatrix} \frac{1}{\sqrt{M_L}} Y_L \\ \frac{1}{\sqrt{M_H}} Y_H \end{pmatrix} .$$

The sparse coefficients $\mathbf{x}$ of each image patch $\mathbf{y}$ are then estimated by solving

$$\arg\min_{\mathbf{x}} \|\mathbf{x}\|_1 \quad \text{such that} \quad \|W(D_L\mathbf{x} - \mathbf{y})\|_2^2 \leq \epsilon \;,$$

where W is a weighting matrix emphasising high frequencies in the ℓ^2 norm, and including some additional refinements such as another constraint on the difference between the high-resolution reconstruction of the current block, $D_H\mathbf{x}$, and the high-resolution reconstructions of previously processed blocks with which it overlaps [118].

7.5 Other

7.5.1 Structure/Texture Separation

Suppose we are given a signal

$$\mathbf{s} = \mathbf{s}_0 + \mathbf{s}_1 + \boldsymbol{\nu}$$

which is a sum of two components, $\mathbf{s}_0$ and $\mathbf{s}_0$, with distinct properties, and noise $\boldsymbol{\nu}$. If, in addition, we have two dictionaries D_0 and D_1 with the property that $\mathbf{s}_0$ is expected to have a sparse representation on D_0 and $\mathbf{s}_1$ is not, and correspondingly for D_1, then a reasonable approach to estimating the individual components $\mathbf{s}_0$ and $\mathbf{s}_0$ given $\mathbf{s}$ is to solve an optimisation such as

$$\arg\min_{(\mathbf{x}_0 \;\; \mathbf{x}_1)^T} \frac{1}{2} \left\| \begin{pmatrix} D_0 & D_1 \end{pmatrix} \begin{pmatrix} \mathbf{x}_0 \\ \mathbf{x}_1 \end{pmatrix} - \mathbf{s} \right\|_2^2 + \lambda \left\| \begin{pmatrix} \mathbf{x}_0 \\ \mathbf{x}_1 \end{pmatrix} \right\|_1$$

and then set

$$\hat{\mathbf{s}_0} = D_0\mathbf{x}_0 \qquad \hat{\mathbf{s}_1} = D_1\mathbf{x}_1 \;.$$

This approach, with some additional complications, including Total Variation (TV) regularisation of one of the components, has been used in decomposing an image into structure and texture components, using a curvelet dictionary for the structure and a DCT dictionary for the texture [119, 120]. Dictionary learning has also been proposed in this application [121].

7.5.2 Image Fusion

Sparse representations have also been applied to the problem of image fusion. One approach operates on image patches, but makes use of analytically defined dictionaries such as wavelets [122]. Given a set of K aligned images to fuse, the set of patches for each image k is extracted and concatenated as vectors to give matrix Y_k, arranged so that corresponding columns of each Y_k represent corresponding patches in the original images. OMP is then used to independently determine a set of coefficient matrices X_k such that $\|DX_k - Y_k\|_2 \leq \epsilon$. A fused coefficient matrix X_F is constructed by choosing each of its columns to be a corresponding column from one of the X_k, selecting the column that has the largest ℓ^1 norm. The fused image patch set is then reconstructed as $Y_F = DX_F$, and the final fused image is reconstructed by averaging of the image patches at their original locations in the image [122].

An alternative approach [123] first applies the K-SVD to learn a dictionary D using the concatenated Y_k as training data and then jointly constructs a set of common coefficient vectors X_C and a set of individual coefficients vectors X_k for each source image such that the coefficient matrices are sparse, and

$$\left\| \begin{pmatrix} D & D & 0 & \cdots & 0 \\ D & 0 & D & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ D & 0 & 0 & \cdots & D \end{pmatrix} \begin{pmatrix} X_C \\ X_1 \\ \vdots \\ X_K \end{pmatrix} - \begin{pmatrix} Y_1 \\ \vdots \\ Y_K \end{pmatrix} \right\|_2 \leq \epsilon \;.$$

The final fused coefficient matrix X_F is constructed as a weighted average of the common coefficient matrix X_C and the fusion, performed as before, of the individual coefficients X_k [123], and the fused image is reconstructed as before.

7.5.3 Video Background Modeling

The Robust PCA model from Sec. 6.4.1 has been found to give very good results in background/foreground separation in video data [103] – if the camera is stationary and the moving objects in the foreground only occupy a small fraction of the frame, then the background and foreground are well modeled as being low rank and sparse respectively.

7.6 Regression

One of the earliest applications of sparse representations was in statistics, as a form of regularisation applied to linear regression [31, Ch. 3]. The goal of regression is to model the relationship between a *dependent variables* y and a set of *explanatory variables* or *independent variables* x_k. In the case of linear regression, this model is of the form $y = \sum_{k=1}^{N} \beta_k x_k + \epsilon = \mathbf{x}^T \boldsymbol{\beta} + \epsilon$. Given an ensemble of M measurements y_k and corresponding $\mathbf{x}_k$, we can define

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_M \end{pmatrix} \qquad X = \begin{pmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_M^T \end{pmatrix} = \begin{pmatrix} x_{1,1} & \cdots & x_{1,N} \\ x_{2,1} & \cdots & x_{2,N} \\ \vdots & \ddots & \vdots \\ x_{M,1} & \cdots & x_{M,N} \end{pmatrix}$$

so that we can write

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon} \ .$$

Note that the interpretation of this linear model is somewhat different from those we have encountered in a signal processing context, where the column vector $\mathbf{y}$ represents a single measurement, and additional measurements become additional columns of a matrix – the representation here is transposed, in the sense that each entry in $\mathbf{y}$ is a distinct measurement, and there is a corresponding change between rows and columns of the dictionary X.

The standard approach is to solve for $\boldsymbol{\beta}$ and $\boldsymbol{\epsilon}$ via least squares, but it is often desirable to find a solution with a sparse $\boldsymbol{\beta}$, both to improve interpretability of the result by limiting dependence of the dependent variables to only a few of the explanatory variables, and to trade increased bias for decreased variance [9]. The *least absolute shrinkage and selection operator* (lasso)

$$\arg\min_{\boldsymbol{\beta}} \|X\boldsymbol{\beta} - \mathbf{y}\|_2^2 \ \text{ such that } \ \|\boldsymbol{\beta}\|_1 \leq t$$

was introduced for precisely this purpose [9]. The *elastic net*

$$\arg\min_{\boldsymbol{\beta}} \|X\boldsymbol{\beta} - \mathbf{y}\|_2^2 + \lambda_1 \|\boldsymbol{\beta}\|_1 + \lambda_2 \|\boldsymbol{\beta}\|_2^2$$

combines both ℓ^1 and ℓ^2 regularization, and is shown to improve on the performance of the lasso when grouping of variables is appropriate [124] (and has also found use in signal processing applications). Group sparsity [82, 87] has been proposed for selection of significant factors (grouped explanatory variables) in a regression model, and group sparsity with sparse groups [125] has also been considered.

While we will not go into detail, other relevant work worth mentioning includes:

- A fundamentally different approach to linear regression than that described above is to use the sparse representation of a vector as a feature vector within an expected risk minimisation framework [71].
- A connection can be made [126] between sparsity regularisation and the Support Vector Machine (SVM) [31, Sec. 12.3.6].
- It is also possible to construct a kernelised version of the LASSO [127]).

7.7 Classification

Sparse representations can be used to construct a generalisation of the well-known Nearest Neighbour classification method [128]. Given a set of training data consisting of labeled images corresponding to one of K classes, we can concatenate (after normalisation) all of the images corresponding to class k into a dictionary matrix D_k, and then concatenate all K of these dictionary matrices into a combined dictionary D. Given an unlabeled test image $\mathbf{s}$ to be classified, we assume that it will have a better sparse representation on the subdictionary D_k when k corresponds to the true class of the image than on any other subdictionary. Solving the problem

$$\arg\min_{\mathbf{x}} \|\mathbf{x}\|_1 \ \text{ such that } \ \|D\mathbf{x} - \mathbf{s}\|_2 \leq \epsilon \ ,$$

and denoting the subvector of $\mathbf{x}$ corresponding to D_k as $\mathbf{x}_k$, our estimate of the class of $\mathbf{s}$ is

$$\hat{k} = \arg\min_k \|D_k \mathbf{x}_k - \mathbf{s}\|_2 \ ,$$

and the reliability of the classification can be estimated by measuring the degree to which the non-zero coefficients are concentrated in $\mathbf{x}_k$ [128]. Occlusion of a small fraction of the test image can be dealt with my augmenting D with the identity matrix and solving

$$\arg\min_{\mathbf{x}} \|\mathbf{x}\|_1 \quad \text{such that} \quad \left\| \begin{pmatrix} D & I \end{pmatrix} \mathbf{x} - \mathbf{s} \right\|_2 \leq \epsilon \ ,$$

where the subvector of $\mathbf{x}$ corresponding to the identity represents the occlusions [128].

Numerous variations on this scheme have been considered:

- As one might expect, modifying the above problem to minimise the $\ell^{2,1}$ norm for a group sparse representation, with the groups corresponding to the classes in the training data, provides improved performance [129].
- A kernelised variant can be constructed [130].
- The class-specific dictionaries can be constructed via a dictionary learning procedure rather than simply consisting of the training data for each class [131].
- When learning class-specific dictionaries, improved performance can be obtained by including a term with a softmax cost function to encourage the correct dictionary to give a better representation than others for each labeled training example [65].
- An alternative approach to joint learning of class-specific dictionaries includes a term promoting incoherence of dictionary pairs [132]. (It is also argued that the full BPDN functional value, including ℓ^1 term, provides a better classification criterion than just the residual error.)
- Another variation [133] on the class-specific dictionary learning is the inclusion of a term based on Fisher's linear discriminant (closely related to Linear Discriminant Analysis [31, Ch. 4]).
- It has been argued that the use of sparsity-promoting regularization such as ℓ^1 regularization is not as important as originally claimed [134].

The basic method described above, and all of its variants, have the common property of basing the classification on reconstruction error of a set of class-specific dictionaries. A fundamentally different approach is to use the sparse representation itself for classification; a wide variety of methods based on this general approach has also been considered, including:

- SVM classification of a sparse representation computed using a functional that includes a Fisher's linear discriminant term [135].
- SVM classification of a convolutional sparse representation for shift-invariance [136].
- A linear classifier, constructed with a quadratic loss function, with the sparse representation as feature vector [137].
- A linear classifier, constructed with a logistic loss function [31, Ch. 4], with the sparse representation (elastic net problem) as feature vector [67, 71].

- A linear classifier on the sparse representation coefficients, with the classifier and dictionary are jointly learned via a modified "Discriminative" K-SVD [69].
- A multi-class linear classifier on the sparse representation coefficients, where the classifier and dictionary are jointly learned via a modified "Label Consistent" K-SVD [70].

Appendix A

Mathematical Background

Definition A.1. *A norm on a vector space V over $\mathbb{R}$ is a function $\|\cdot\| : V \to \mathbb{R}$ such that*

1. $\|\mathbf{v}\| \geq 0 \quad \forall \mathbf{v} \in V$
2. $\|a\mathbf{v}\| = |a| \, \|\mathbf{v}\| \quad \forall a \in \mathbb{R}, \mathbf{v} \in V$
3. $\|\mathbf{u} + \mathbf{v}\| \leq \|\mathbf{u}\| + \|\mathbf{v}\| \quad \forall \mathbf{u}, \mathbf{v} \in V$
4. $\|\mathbf{v}\| = 0 \Leftrightarrow \mathbf{v} = \mathbf{0} \quad \forall \mathbf{v} \in V$

The Gershgorin circle theorem (or Gershgorin disc theorem) provides bounds on the eigenvalues of a square matrix [4, Theorem 7.2.1] [138, Sec. 6.1] [139, Sec. 10.6]

Theorem A.1. *Every eigenvalue of a square matrix $A \in \mathbb{R}^{N \times N}$ lies within at least one* Gershgorin disc $D(A_{kk}, r_k)$*, where*

$$D(c, r) = \{x| \ |x - c| \leq r\} \ ,$$

and

$$r_k = \sum_{1 \leq l \leq N, \, l \neq k} |A_{kl}|$$

is the sum of the absolute values of off-diagonal elements for row k of A.

An equivalent statement of the theorem is that all eigenvalues of A lie within the union of the N Gershgorin discs $D(A_{kk}, r_k)$ $1 \leq k \leq N$.

Definition A.2. *A symmetric real matrix A is positive definite iff $\mathbf{z}^T A \mathbf{z} > 0$ for all $\mathbf{z} \neq 0$*

Theorem A.2. *$B = A^T A$ and B is positive definite $\Leftrightarrow$ the columns of A are linearly independent*

Definition A.3. *Matrix A is strictly diagonally dominant iff $|A_{kk}| > \sum_{l \neq k} |A_{kl}| \quad \forall k$ (see [138, Definition 6.1.9]).*

Theorem A.3. *If $A_{kk} > 0 \; \forall k$ and A is strictly diagonally dominant, then A is positive definite.*

Proof: This is easily shown using the Gershgorin circle theorem (see Theorem A.1): diagonal dominance bounds each Gershgorin disc so that it does not include zero, so that all eigenvalues are positive, and the matrix is therefore positive definite. (See [138, Theorem 6.1.10].)

Theorem A.4. *Given matrix $A \in \mathbb{R}^{M \times N}$ and matrix $\tilde{A} \in \mathbb{R}^{M \times (N-1)}$ constructed by deleting a column of A, then*

$$\sigma_k(A) \geq \sigma_k(\tilde{A}) \geq \sigma_{k+1}(A) \;\; 1 \leq k \leq \min\{M, N\} \; ,$$

where we define, for $X \in \mathbb{R}^{M \times N}$, that $\sigma_k(X) = 0$ when $k > \min\{M, N\}$.

Proof: See [140, Sec. 3.1, Corollary 3.1.3]

Theorem A.5. *If a column is removed from a matrix $A \in \mathbb{R}^{M \times N}$ where $N \leq M$, then the minimum singular value of the resulting matrix $\tilde{A} \in \mathbb{R}^{M \times (N-1)}$ is equal to or greater than that of the original matrix.*

Proof: When $N \leq M$ we have $\sigma_{\min}(A) = \sigma_N(A)$ and $\sigma_{\min}(\tilde{A}) = \sigma_{N-1}(\tilde{A})$. Substituting into the result from Theorem A.4

$$\sigma_k(A) \geq \sigma_k(\tilde{A}) \geq \sigma_{k+1}(A) \;\; 1 \leq k \leq \min\{M, N\}$$

we have

$$\sigma_{N-1}(A) \geq \sigma_{N-1}(\tilde{A}) \geq \sigma_N(A) \; ,$$

and therefore

$$\sigma_{\min}(\tilde{A}) \geq \sigma_{\min}(A) \; .$$

Note that when $N > M$ this ordering is reversed, since in this case $\sigma_{\min}(A) = \sigma_M(A)$ and $\sigma_{\min}(\tilde{A}) = \sigma_M(\tilde{A})$ so that the inequality from Theorem A.4 gives

$$\sigma_{\min}(A) \geq \sigma_{\min}(\tilde{A}) \; .$$

References

[1] A. M. Bruckstein, D. L. Donoho, and M. Elad, "From sparse solutions of systems of equations to sparse modeling of signals and images," *SIAM Review*, vol. 51, no. 1, pp. 34–81, 2009. doi:10.1137/060657704

[2] M. Elad, *Sparse and Redundant Representations.* Springer, 2010.

[3] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends in Machine Learning*, vol. 3, no. 1, pp. 1–122, 2010. doi:10.1561/2200000016

[4] G. H. Golub and C. F. V. Loan, *Matrix Computations.* Baltimore, MD, USA: The Johns Hopkins University Press, 1996.

[5] M. Elad and D. Datsenko, "Example-based regularization deployed to super-resolution reconstruction of a single image," *The Computer Journal*, 2007. doi:10.1093/comjnl/bxm008

[6] M. Elad, P. Milanfar, and R. Rubinstein, "Analysis versus synthesis in signal priors," *Inverse Problems*, vol. 23, no. 3, p. 947, 2007. doi:10.1088/0266-5611/23/3/007

[7] J. Nocedal and S. J. Wright, *Numerical Optimization*, 2nd ed. Springer, 2006.

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

[9] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996. [Online]. Available: http://www.jstor.org/stable/2346178

[10] S. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Transactions on Signal Processing*, vol. 41, no. 12, pp. 3397–3415, December 1993. doi:10.1109/78.258082

[11] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM Journal on Scientific Computing*, vol. 20, no. 1, pp. 33–61, 1998. doi:10.1137/S1064827596304010

[12] I. F. Gorodnitsky and B. D. Rao, "Sparse signal reconstruction from limited data using FOCUSS: a re-weighted minimum norm algorithm." *IEEE Transactions on Signal Processing*, vol. 45, no. 3, pp. 600–616, March 1997. doi:10.1109/78.558475

[13] D. L. Donoho and M. Elad, "Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ^1 minimization," *Proceedings of the National Academy of Sciences*, vol. 100, no. 5, pp. 2197–2202, 2003. doi:10.1073/pnas.0437847100

[14] T. Strohmer and R. W. H. Jr., "Grassmannian frames with applications to coding and communication," *Applied and Computational Harmonic Analysis*, vol. 14, no. 3, pp. 257 – 275, 2003. doi:10.1016/S1063-5203(03)00023-X

[15] B. Wohlberg, "Noise sensitivity of sparse signal representations: Reconstruction error bounds for the inverse problem," *IEEE Transactions on Signal Processing*, vol. 51, no. 12, pp. 3053–3060, Dec. 2003. doi:10.1109/TSP.2003.819006

[16] S. Qiao, S. Qiao, and X. Wang, "Computing the singular values of 2-by-2 complex matrices," Software Quality Research Laboratory, Computing and Software Department, McMaster

University, SQRL Report 5, June 2002. [Online]. Available: http://www.cas.mcmaster.ca/sqrl/papers/sqrl5.pdf

[17] D. L. Donoho, M. Elad, and V. N. Temlyakov, "Stable recovery of sparse overcomplete representations in the presence of noise," *IEEE Transactions on Information Theory*, vol. 52, no. 1, pp. 6–18, 2006. doi:10.1109/tit.2005.860430

[18] E. J. Candès and T. Tao, "Decoding by linear programming," *IEEE Transactions on Information Theory*, vol. 51, no. 12, pp. 4203–4215, 2005. doi:10.1109/tit.2005.858979

[19] E. J. Candès, J. K. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Communications on Pure and Applied Mathematics*, vol. 59, no. 8, pp. 1207–1223, 2006. doi:10.1002/cpa.20124

[20] B. K. Natarajan, "Sparse approximate solutions to linear systems," *SIAM Journal on Computing*, vol. 24, no. 2, pp. 227–234, April 1995. doi:10.1137/s0097539792240406

[21] Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad, "Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition," in *Conference Record of The Twenty-Seventh Asilomar Conference on Signals, Systems and Computers*, November 1993, pp. 40–44. doi:10.1109/acssc.1993.342465

[22] J. A. Tropp, "Greed is good: algorithmic results for sparse approximation," *IEEE Transactions on Information Theory*, vol. 50, no. 10, pp. 2231–2242, 2004. doi:10.1109/tit.2004.834793

[23] E. J. Candès and Y. Plan, "Matrix completion with noise," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 925–936, 2010. doi:10.1109/JPROC.2009.2035722

[24] E. J. Candès, "The restricted isometry property and its implications for compressed sensing," *Comptes Rendus Mathematique*, vol. 346, no. 9-10, pp. 589–592, 2008. doi:10.1016/j.crma.2008.03.014

[25] D. M. Malioutov, M. Cetin, and A. S. Willsky, "Optimal sparse representations in general overcomplete bases," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 2, 2004, pp. II793–II796. doi:10.1109/ICASSP.2004.1326377

[26] D. L. Donoho, "For most large underdetermined systems of linear equations the minimal 1-norm solution is also the sparsest solution," *Communications on Pure and Applied Mathematics*, vol. 59, no. 6, pp. 797–829, 2006. doi:10.1002/cpa.20132

[27] S. Foucart, "A note on guaranteed sparse recovery via ℓ_1-minimization," *Applied and Computational Harmonic Analysis*, vol. 29, no. 1, pp. 97–103, 2010. doi:10.1016/j.acha.2009.10.004

[28] E. van den Berg and M. P. Friedlander, "Probing the pareto frontier for basis pursuit solutions," *SIAM Journal on Scientific Computing*, vol. 31, no. 2, pp. 890–912, 2008. doi:10.1137/080714488

[29] ——, "Sparse optimization with least-squares constraints," *SIAM J. Optimization*, vol. 21, no. 4, pp. 1201–1229, 2011. doi:10.1137/100785028

[30] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *Annals of Statistics*, vol. 32, no. 2, pp. 407–499, 2004. doi:10.1214/009053604000000067

[31] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed., ser. Springer Series in Statistics. New York, NY, USA: Springer New York Inc., 2009.

[32] A. E. Beaton and J. W. Tukey, "The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data," *Technometrics*, no. 16, pp. 147–185, 1974.

[33] J. A. Scales and A. Gersztenkorn, "Robust methods in inverse theory," *Inverse Problems*, vol. 4, no. 4, pp. 1071–1091, Oct. 1988.

[34] R. Wolke and H. Schwetlick, "Iteratively reweighted least squares: Algorithms, convergence analysis, and numerical comparisons," *SIAM J. on Sci. and Stat. Comp.*, vol. 9, no. 5, pp. 907–921, Sept. 1988.

[35] S. A. Ruzinsky and E. T. Olsen, "L_1 and L_∞ minimization via a variant of Karmarkar's algo-

rithm," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 37, pp. 245–253, 1989.

[36] K. P. Bube and R. T. Langan, "Hybrid ℓ^1/ℓ^2 minimization with applications to tomography," *Geophysics*, vol. 62, no. 4, pp. 1183–1195, July-August 1997.

[37] I. F. Gorodnitsky, B. D. Rao, and J. S. George, "Source localization in magnetoencephalography using an iterative weighted minimum norm algorithm," in *Conference Record of The Twenty-Sixth Asilomar Conference on Signals, Systems and Computers, 1992*, Pacific Grove, CA, USA, 1992, pp. 167–171. doi:10.1109/ACSSC.1992.269280

[38] B. D. Rao and K. Kreutz-Delgado, "An affine scaling methodology for best basis selection," *IEEE Transactions On Signal Processing*, vol. 47, no. 1, pp. 187–200, Jan. 1999.

[39] D. Hunter and K. Lange, "A tutorial on MM algorithms," *The American Statistician*, vol. 58, no. 1, pp. 30–37, 2004.

[40] P. Rodríguez and B. Wohlberg, "A comparison of the computational performance of iteratively reweighted least squares and alternating minimization algorithms for l_1 inverse problems," in *Proceedings of IEEE International Conference on Image Processing (ICIP)*, Orlando, FL, USA, Oct. 2012, pp. 3069–3072. doi:10.1109/ICIP.2012.6467548

[41] A. Beck and M. Teboulle, "Gradient-based algorithms with applications to signal recovery problems," in *Convex Optimization in Signal Processing and Communications*, D. P. Palomar and Y. C. Eldar, Eds. Cambridge University Press, 2010, ch. 2, pp. 42–88.

[42] P. H. Calamai and J. J. Mor, "Projected gradient methods for linearly constrained problems," *Mathematical Programming*, vol. 39, pp. 93–116, 1987. doi:10.1007/BF02592073

[43] P. L. Combettes and J.-C. Pesquet, "Proximal splitting methods in signal processing," in *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, ser. Springer Optimization and Its Applications, H. H. Bauschke, R. S. Burachik, P. L. Combettes, V. Elser, D. R. Luke, and H. Wolkowicz, Eds. Springer New York, 2011, vol. 49, pp. 185–212. doi:10.1007/978-1-4419-9569-8_10

[44] D. L. Donoho, "De-noising by soft-thresholding," *IEEE Transactions on Information Theory*, vol. 41, no. 3, pp. 613–627, 1995. doi:10.1109/18.382009

[45] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM Journal on Imaging Sciences*, vol. 2, no. 1, pp. 183–202, 2009. doi:10.1137/080716542

[46] S. Becker, J. Bobin, and E. J. Candès, "Nesta: A fast and accurate first-order method for sparse recovery," *SIAM Journal on Imaging Sciences*, vol. 4, no. 1, pp. 1–39, 2011. doi:10.1137/090756855

[47] T. Goldstein and S. J. Osher, "The split bregman method for l1-regularized problems," *SIAM Journal on Imaging Sciences*, vol. 2, no. 2, pp. 323–343, 2009. doi:10.1137/080725891

[48] E. Esser, "Applications of Lagrangian-based alternating direction methods and connections to split Bregman," UCLA, CAM Report 09-31, March 2009. [Online]. Available: ftp://ftp.math.ucla.edu/pub/camreport/cam09-31.pdf

[49] Y. Wang, J. Yang, W. Yin, and Y. Zhang, "A new alternating minimization algorithm for total variation image reconstruction," *SIAM J. Img. Sci.*, vol. 1, no. 3, pp. 248–272, August 2008. doi:10.1137/080724265

[50] M. V. Afonso, J. M. Bioucas-Dias, and M. A. T. Figueiredo, "Fast image recovery using variable splitting and constrained optimization," *IEEE Transactions on Image Processing*, vol. 19, no. 9, pp. 2345–2356, 2010. doi:10.1109/tip.2010.2047910

[51] ——, "An Augmented Lagrangian approach to the constrained optimization formulation of imaging inverse problems," *IEEE Transactions on Image Processing*, vol. 20, no. 3, pp. 681–695, March 2011. doi:10.1109/tip.2010.2076294

[52] M. Hong and Z.-Q. Luo, "On the linear convergence of the alternating direction method of multipliers," arXiv, Tech. Rep. arXiv:1208.3922v3, 2013. [Online]. Available: http://arxiv.org/abs/1208.3922v3

[53] I. Tosic and P. Frossard, "Dictionary learning," *IEEE Signal Processing Magazine*, vol. 28, no. 2, pp. 27–38, March 2011. doi:10.1109/msp.2010.939537

[54] J. A. Tropp, A. C. Gilbert, and M. J. Strauss, "Algorithms for simultaneous sparse approximation. part i: Greedy pursuit," *Signal Processing*, vol. 86, no. 3, pp. 572–588, November 2006. doi:10.1016/j.sigpro.2005.05.030

[55] E. van den Berg and M. P. Friedlander, "Joint-sparse recovery from multiple measurements," Department of Computer Science, University of British Columbia, Tech. Rep. TR-2009-07, Apr. 2009.

[56] K. B. Petersen and M. S. Pedersen, "The matrix cookbook," November 2012, version 20121115. [Online]. Available: http://www2.imm.dtu.dk/pubdb/p.php?3274

[57] B. A. Olshausen and D. J. Field, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," *Nature*, vol. 381, no. 6583, pp. 607–609, 1996.

[58] K. Engan, S. O. Aase, and J. Hakon Husoy, "Method of optimal directions for frame design," in *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 5, 1999, pp. 2443–2446. doi:10.1109/icassp.1999.760624

[59] M. Aharon, M. Elad, and A. M. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Transactions on Signal Processing*, vol. 54, no. 11, pp. 4311–4322, 2006. doi:10.1109/tsp.2006.881199

[60] B. Mailhé and M. D. Plumbley, "Dictionary learning with large step gradient descent for sparse representations," in *Latent Variable Analysis and Signal Separation*, ser. Lecture Notes in Computer Science, F. J. Theis, A. Cichocki, A. Yeredor, and M. Zibulevsky, Eds. Springer Berlin Heidelberg, 2012, vol. 7191, pp. 231–238. doi:10.1007/978-3-642-28551-6_29

[61] Q. Liu, S. Wang, J. Luo, Y. Zhu, and M. Ye, "An augmented Lagrangian approach to general dictionary learning for image denoising," *Journal of Visual Communication and Image Representation*, 2012. doi:10.1016/j.jvcir.2012.04.003

[62] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online learning for matrix factorization and sparse coding," *J. Mach. Learn. Res.*, vol. 11, pp. 19–60, 2010.

[63] J. Mairal, G. Sapiro, and M. Elad, "Learning multiscale sparse representations for image and video restoration," *Multiscale Modeling & Simulation*, vol. 7, no. 1, pp. 214–241, 2008. doi:10.1137/070697653

[64] Q. Liu, J. Luo, S. Wang, M. Xiao, and M. Ye, "An augmented Lagrangian multi-scale dictionary learning algorithm," *EURASIP Journal on Advances in Signal Processing*, vol. 2011, pp. 1–16, 2011. doi:10.1186/1687-6180-2011-58

[65] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman, "Discriminative learned dictionaries for local image analysis," in *IEEE Conference on Computer Vision and Pattern Recognition, 2008. CVPR 2008.*, June 2008, pp. 1–8. doi:10.1109/cvpr.2008.4587652

[66] J. Mairal, M. Leordeanu, F. Bach, M. Hebert, and J. Ponce, "Discriminative sparse image models for class-specific edge detection and image interpretation," in *Proceedings of the 10th European Conference on Computer Vision: Part III*, ser. ECCV '08. Springer-Verlag, 2008, pp. 43–56. doi:10.1007/978-3-540-88690-7_4

[67] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman, "Supervised dictionary learning," in *Advances in Neural Information Processing Systems 21*, D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, Eds. Neural Information Processing Systems Foundation, 2008, pp. 1033–1040. [Online]. Available: http://books.nips.cc/papers/files/nips21/NIPS2008_0775.pdf

[68] P. Sprechmann and G. Sapiro, "Dictionary learning and sparse coding for unsupervised clustering," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference*

on, March 2010, pp. 2042–2045. doi:10.1109/icassp.2010.5494985

[69] Q. Zhang and B. Li, "Discriminative K-SVD for dictionary learning in face recognition," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, 2010, pp. 2691–2698. doi:10.1109/cvpr.2010.5539989

[70] Z. Jiang, Z. Lin, and L. S. Davis, "Learning a discriminative dictionary for sparse coding via label consistent K-SVD," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011, pp. 1697–1704. doi:10.1109/cvpr.2011.5995354

[71] J. Mairal, F. Bach, and J. Ponce, "Task-driven dictionary learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 4, pp. 791–804, April 2012. doi:10.1109/tpami.2011.156

[72] S. F. Cotter, B. D. Rao, K. Engan, and K. Kreutz-Delgado, "Sparse solutions to linear inverse problems with multiple measurement vectors," *IEEE Transactions on Signal Processing*, vol. 53, no. 7, pp. 2477–2488, July 2005. doi:10.1109/tsp.2005.849172

[73] J. Chen and X. Huo, "Theoretical results on sparse representations of multiple-measurement vectors," *IEEE Transactions on Signal Processing*, vol. 54, no. 12, pp. 4634–4643, 2006. doi:10.1109/tsp.2006.881263

[74] E. van den Berg and M. P. Friedlander, "Theoretical and empirical results for recovery from multiple measurements," *IEEE Transactions on Information Theory*, vol. 56, no. 5, pp. 2516–2527, May 2010. doi:10.1109/tit.2010.2043876

[75] M. Stojnic, F. Parvaresh, and B. Hassibi, "On the reconstruction of block-sparse signals with an optimal number of measurements," *IEEE Transactions on Signal Processing*, vol. 57, no. 8, pp. 3075–3085, 2009. doi:10.1109/tsp.2009.2020754

[76] Y. C. Eldar and M. Mishali, "Robust recovery of signals from a structured union of subspaces," *IEEE Transactions on Information Theory*, vol. 55, no. 11, pp. 5302–5316, 2009. doi:10.1109/tit.2009.2030471

[77] J. Huang and T. Zhang, "The benefit of group sparsity," *The Annals of Statistics*, vol. 38, no. 4, pp. 1978–2004, 2010. doi:10.1214/09-aos778

[78] Y. C. Eldar and H. Rauhut, "Average case analysis of multichannel sparse recovery using convex relaxation," *IEEE Transactions on Information Theory*, vol. 56, no. 1, pp. 505–519, 2010. doi:10.1109/tit.2009.2034789

[79] X. Lv, G. Bi, and C. Wan, "The group lasso for stable recovery of block-sparse signal representations," *IEEE Transactions on Signal Processing*, vol. 59, no. 4, pp. 1371–1382, April 2011. doi:10.1109/tsp.2011.2105478

[80] M. E. Davies and Y. C. Eldar, "Rank awareness in joint sparse recovery," *IEEE Transactions on Information Theory*, vol. 58, no. 2, pp. 1135–1146, 2012. doi:10.1109/tit.2011.2173722

[81] E. Elhamifar and R. Vidal, "Block-sparse recovery via convex optimization," *IEEE Transactions on Signal Processing*, vol. 60, no. 8, pp. 4094–4107, Aug. 2012. doi:10.1109/tsp.2012.2196694

[82] B. A. Turlach, W. N. Venables, and S. J. Wright, "Simultaneous variable selection," *Technometrics*, vol. 47, no. 3, pp. 349–363, 2005. doi:10.1198/004017005000000139

[83] A. Quattoni, X. Carreras, M. Collins, and T. Darrell, "An efficient projection for l1, regularization," in *Proceedings of the 26th Annual International Conference on Machine Learning*, ser. ICML '09. New York, NY, USA: ACM, 2009, pp. 857–864. doi:10.1145/1553374.1553484

[84] A. Rakotomamonjy, "Algorithms for Multiple Basis Pursuit Denoising," in *SPARS'09 - Signal Processing with Adaptive Sparse Structured Representations*, R. Gribonval, Ed., April 2009. [Online]. Available: http://hal.inria.fr/docs/00/36/95/35/PDF/64.pdf

[85] M. Kowalski, "Sparse regression using mixed norms," *Applied and Computational Harmonic Analysis*, vol. 27, no. 3, pp. 303–324, 2009. doi:10.1016/j.acha.2009.05.006

[86] W. Deng, W. Yin, and Y. Zhang, "Group sparse optimization by alternating direction method,"

Department of Computational and Applied Mathematics, Rice University, Tech. Rep., 2011.

[87] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 68, no. 1, pp. 49–67, 2006. doi:10.1111/j.1467-9868.2005.00532.x

[88] P. Sprechmann, I. Ramírez, G. Sapiro, and Y. C. Eldar, "Collaborative hierarchical sparse modeling," *CoRR*, vol. abs/1003.0400, 2010. [Online]. Available: http://arxiv.org/abs/1003.0400

[89] ——, "C-HiLasso: A collaborative hierarchical sparse modeling framework," *IEEE Transactions on Signal Processing*, vol. 59, no. 9, pp. 4183–4198, 2011. doi:10.1109/tsp.2011.2157912

[90] P. Zhao, G. Rocha, and B. Yu, "The composite absolute penalties family for grouped and hierarchical variable selection," *The Annals of Statistics*, vol. 37, no. 6A, pp. 3468–3497, 2009. doi:10.1214/07-aos584

[91] R. Jenatton, J. Mairal, G. Obozinski, and F. Bach, "Proximal methods for hierarchical sparse coding," *J. Mach. Learn. Res.*, vol. 12, pp. 2297–2334, July 2011. [Online]. Available: http://dl.acm.org/citation.cfm?id=1953048.2021074

[92] E. J. Candès, "Compressive sampling," in *Proceedings of the International Congress of Mathematicians*, August 2006, pp. 1433–1452. doi:10.4171/022-3/69

[93] M. Fornasier and H. Rauhut, "Compressive sensing," in *Handbook of Mathematical Methods in Imaging*, O. Scherzer, Ed. Springer, 2011, pp. 187–229.

[94] L. Jacques and P. Vandergheynst, "Compressed sensing: when sparsity meets sampling," in *Optical and Digital Image Processing*. Wiley-VCH, 2011, pp. 507–527. doi:10.1002/9783527635245.ch23

[95] M. F. Duarte and Y. C. Eldar, "Structured compressed sensing: From theory to applications," *IEEE Transactions on Signal Processing*, vol. 59, no. 9, pp. 4053–4085, 2011. doi:10.1109/tsp.2011.2161982

[96] G. Kutyniok, "Compressed sensing: Theory and applications," *CoRR*, vol. abs/1203.3815, 2012.

[97] M. A. Davenport, M. F. Duarte, Y. C. Eldar, and G. Kutyniok, "Introduction to compressed sensing," in *Compressed Sensing: Theory and Applications*, Y. C. Eldar and G. Kutyniok, Eds. Cambridge University Press, 2012, ch. 1, pp. 1–68.

[98] R. G. Baraniuk, M. A. Davenport, R. DeVore, and M. B. Wakin, "A simple proof of the restricted isometry property for random matrices," *Constructive Approximation*, vol. 28, pp. 253–263, 2008. doi:10.1007/s00365-007-9003-x

[99] E. J. Candès, Y. C. Eldar, D. Needell, and P. Randall, "Compressed sensing with coherent and redundant dictionaries," *Applied and Computational Harmonic Analysis*, vol. 31, no. 1, pp. 59–73, 2011. doi:10.1016/j.acha.2010.10.002

[100] E. J. Candès and B. Recht, "Exact matrix completion via convex optimization," *Foundations of Computational Mathematics*, vol. 9, pp. 717–772, 2009. doi:10.1007/s10208-009-9045-5

[101] J.-F. Cai, E. J. Candès, and Z. Shen, "A singular value thresholding algorithm for matrix completion," *SIAM Journal on Optimization*, vol. 20, no. 4, pp. 1956–1982, 2010. doi:10.1137/080738970

[102] R. M. Larsen, "PROPACK," functions for computing the singular value decomposition of large and sparse or structured matrices, available from http://sun.stanford.edu/~rmunk/PROPACK.

[103] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?" *J. ACM*, vol. 58, pp. 11:1–11:37, June 2011. doi:10.1145/1970392.1970395

[104] M. Elad, M. A. T. Figueiredo, and Y. Ma, "On the role of sparse and redundant representation in image processing," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 972–982, June 2010. doi:10.1109/jproc.2009.2037655

[105] M. Elad and M. Aharon, "Image denoising via sparse and redundant representations over learned dictionaries," *IEEE Transactions on Image Processing*, vol. 15, no. 12, pp. 3736–3745, December 2006. doi:10.1109/tip.2006.881969

[106] J. Mairal, M. Elad, and G. Sapiro, "Sparse representation for color image restoration," *IEEE Transactions on Image Processing*, vol. 17, no. 1, pp. 53–69, 2008. doi:10.1109/tip.2007.911828

[107] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman, "Non-local sparse models for image restoration," in *IEEE 12th International Conference on Computer Vision*, 2009, pp. 2272–2279. doi:10.1109/iccv.2009.5459452

[108] M. Bertalmio, G. Sapiro, V. Caselles, and C. Ballester, "Image inpainting," in *Proc. ACM SIGGRAPH*, 2000, pp. 417–424. doi:10.1145/344779.344972

[109] M. Elad, J.-L. Starck, P. Querre, and D. L. Donoho, "Simultaneous cartoon and texture image inpainting using morphological component analysis (MCA)," *Applied and Computational Harmonic Analysis*, vol. 19, pp. 340–358, 2005.

[110] J. Fadili, J.-L. Starck, and F. Murtagh, "Inpainting and zooming using sparse representations," *The Computer Journal*, 2007. doi:10.1093/comjnl/bxm055

[111] O. G. Guleryuz, "Nonlinear approximation based image recovery using adaptive sparse reconstructions and iterated denoising: Part i - theory," *IEEE Transactions on Image Processing*, vol. 15, no. 3, pp. 539–554, 2006. doi:10.1109/tip.2005.863057

[112] ——, "Nonlinear approximation based image recovery using adaptive sparse reconstructions and iterated denoising: Part ii - adaptive algorithms," *IEEE Transactions on Image Processing*, vol. 15, no. 3, pp. 555–571, 2006. doi:10.1109/tip.2005.863055

[113] B. Shen, W. Hu, Y. Zhang, and Y.-J. Zhang, "Image inpainting via sparse representation," in *IEEE International Conference on Acoustics, Speech and Signal Processing, 2009. ICASSP 2009*, April 2009, pp. 697–700. doi:10.1109/icassp.2009.4959679

[114] Z. Xu and J. Sun, "Image inpainting by patch propagation using patch sparsity," *IEEE Transactions on Image Processing*, vol. 19, no. 5, pp. 1153–1165, 2010. doi:10.1109/tip.2010.2042098

[115] B. Wohlberg, "Inpainting by joint optimization of linear combinations of exemplars," *IEEE Signal Processing Letters*, vol. 18, no. 1, pp. 75–78, Jan. 2011. doi:10.1109/LSP.2010.2095842

[116] Y. Lou, A. Bertozzi, and S. Soatto, "Direct sparse deblurring," *Journal of Mathematical Imaging and Vision*, vol. 39, pp. 1–12, 2011. doi:10.1007/s10851-010-0220-8

[117] F. Couzinie-Devy, J. Mairal, F. Bach, and J. Ponce, "Dictionary learning for deblurring and digital zoom," *International Journal of Computer Vision*, 2011, preprint. [Online]. Available: http://arxiv.org/abs/1110.0957

[118] J. Yang, J. Wright, T. S. Huang, and Y. Ma, "Image super-resolution via sparse representation," *IEEE Transactions on Image Processing*, vol. 19, no. 11, pp. 2861–2873, 2010. doi:10.1109/tip.2010.2050625

[119] J.-L. Starck, M. Elad, and D. L. Donoho, "Image decomposition: Separation of texture from piecewise smooth content," *Proceedings of SPIE*, vol. 5207, no. 2, pp. 571–582, 2003.

[120] ——, "Image decomposition via the combination of sparse representation and a variational approach," *IEEE Transaction on Image Processing*, vol. 14, no. 10, pp. 1570–1582, October 2005.

[121] G. Peyré, J. Fadili, and J.-L. Starck, "Learning adapted dictionaries for geometry and texture separation," in *SPIE Wavelet XII*, August 2007.

[122] B. Yang and S. Li, "Multifocus image fusion and restoration with sparse representation," *IEEE Transactions on Instrumentation and Measurement*, vol. 59, no. 4, pp. 884–892, April 2010. doi:10.1109/tim.2009.2026612

[123] N. Yu, T. Qiu, F. Bi, and A. Wang, "Image features extraction and fusion based on joint sparse representation," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 5, pp. 1074–1082, 2011. doi:10.1109/jstsp.2011.2112332

[124] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no. 2, pp. 301–320, 2005.

doi:10.1111/j.1467-9868.2005.00503.x

[125] N. Simon, J. Friedman, T. Hastie, and R. Tibshirani, "A sparse-group lasso," *Journal of Computational and Graphical Statistics*, 2012. doi:10.1080/10618600.2012.681250

[126] F. Girosi, "An equivalence between sparse approximation and support vector machines," *Neural Computation*, vol. 10, no. 6, pp. 1455–1480, August 1998.

[127] V. Roth, "The generalized lasso," *IEEE Transactions on Neural Networks*, vol. 15, no. 1, pp. 16–28, 2004. doi:10.1109/tnn.2003.809398

[128] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 210–227, February 2009. doi:10.1109/tpami.2008.79

[129] E. Elhamifar and R. Vidal, "Robust classification using structured sparse representation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2011, pp. 1873–1879. doi:10.1109/cvpr.2011.5995664

[130] L. Zhang, W.-D. Zhou, P.-C. Chang, J. Liu, Z. Yan, T. Wang, and F.-Z. Li, "Kernel sparse representation-based classifier," *IEEE Transactions on Signal Processing*, vol. 60, no. 4, pp. 1684–1695, 2012. doi:10.1109/tsp.2011.2179539

[131] M. Yang, L. Zhang, J. Yang, and D. Zhang, "Metaface learning for sparse representation based face recognition," in *IEEE International Conference on Image Processing (ICIP)*, 2010, pp. 1601–1604. doi:10.1109/icip.2010.5652363

[132] I. Ramírez, P. Sprechmann, and G. Sapiro, "Classification and clustering via dictionary learning with structured incoherence and shared features," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010, pp. 3501–3508. doi:10.1109/cvpr.2010.5539964

[133] M. Yang, L. Zhang, X. Feng, and D. Zhang, "Fisher discrimination dictionary learning for sparse representation," in *IEEE International Conference on Computer Vision (ICCV)*, 2011, pp. 543–550. doi:10.1109/iccv.2011.6126286

[134] L. Zhang, M. Yang, and X. Feng, "Sparse representation or collaborative representation: Which helps face recognition?" in *IEEE International Conference on Computer Vision (ICCV)*, 2011, pp. 471–478. doi:10.1109/iccv.2011.6126277

[135] K. Huang and S. Aviyente, "Sparse representation for signal classification," in *Advances in Neural Information Processing Systems 19*, B. Schölkopf, J. C. Platt, and T. Hofmann, Eds. Cambridge, MA: MIT Press, 2006, pp. 609–616.

[136] R. Grosse, R. Raina, H. Kwong, and A. Y. Ng, "Shift-invariant sparse coding for audio classification," in *Proceedings of the Twenty-third Conference on Uncertainty in Artificial Intelligence (UAI)*, 2007.

[137] D.-S. Pham and S. Venkatesh, "Joint learning and dictionary construction for pattern recognition," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, 2008, pp. 1–8. doi:10.1109/cvpr.2008.4587408

[138] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge University Press, 1985.

[139] P. Lancaster, *Theory of Matrices*, 2nd ed. New York, NY, USA: Academic Press, 1985.

[140] R. A. Horn and C. R. Johnson, *Topics in Matrix Analysis*. Cambridge University Press, 1991.