COMPRESSED SENSING AND RELATED LEARNING PROBLEMS

YINGZHEN LI

Submitted for the degree of Bachelor of Science
Department of Mathematics, Sun Yat-sen University
Supervised by Prof. Haizhang Zhang

MAY 2013

Abstract

Classical sensing algorithms represent the raw signals in some domain constructed by basis functions. They sample the original data with the Nyquist rate which guarantees the exact recovery, and use fast computing processors to reconstruct the signals for further analysis. However, with the speed on demand, slow sensors become the main limitation of signal processing, which cannot be solved merely by developing faster sensing devices. Noticing this obstacle, compressed sensing then came to the research world and have attracted focuses from mathematicians, statisticians and computer scientists. With better understanding of the sparsity in the dataset, it acquires samples with a much lower sampling rate than the Nyquist's, and solves the l_1-minimization problem for signal reconstruction. Significant topics include the sensing method, number of samples we need and fast reconstruction for l_1-optimizations without loss of stability. By now researchers have introduced theories yielding high-quality recoveries as well as better measurement matrix and dictionary constructions, which are close-related to machine learning: it views the recovery as a kind of prediction and discusses the optimization problem from the probability perspective. Applications such as the advanced MRI proved the efficiency of compressed sensing methods, resulting in the emerging studies of it and related learning techniques.

Keywords: compressed sensing, machine learning, data sparsity, sparse recovery, l_1-minimization

Contents

List of Figures

List of Symbols

f	the original signal
Φ, Φ^{Ω}	measurement (sensing) matrix
$\mathbf{y}$	observed measurement vector
m	the number of measurements
n	length of the original signal
f^*	optimal recovery of the original signal
k, S	degree of sparsity (k-sparse or S-sparse)
$H(\Phi)$	the Hilbert space constructed by the column vectors of Φ
$\mathcal{M}(\Gamma, \Gamma')$	mutual incoherence of matrices Γ and Γ'
T	support set of a vector (f), i.e. the index set of non-zero elements
δ_k, δ_S	the k (or S)-restricted isometry constant (RIC)
$\theta_{S,S'}$	the S, S'-restricted isometry constant
Ω	index set of Φ generated by some distribution
$\mathcal{B}(\Phi)$	the basis set (population) in which Φ is generated from
$\mu(\Phi)$	the coherence parameter of the basis population $\mathcal{B}(\Phi)$
$\mathbf{z}$	noise of the signal
Ψ	the dictionary for sparse representation
$\mathbf{x}$	the sparse representation of the signal over a dictionary
Ψ^*	optimal construction of the dictionary
$\mathbf{x}^*$	optimal recovery of the sparse representation
$f^*(t), \Psi^*(t)$	the temporal optimal recoveries in the tth step

Chapter 1

Introduction

We humans observe different types of signals, including sound, images, waves and sensor data, in our daily life. Widespread signal processing technologies combine researches in systems engineering, electrical engineering and applied mathematics to analyse these signals to supply various demands. Typical researches of signal processing include signal sampling, quality improvement (e.g. noise reduction), signal compression for long-distance transmission and signal reconstruction at the reception side [29]. Also machine learning people are interested in feature extraction (or feature learning) from raw signals, such as image understanding in computer vision and speech recognition in natural language processing.

It is infeasible to process continuous signals directly in polynomial time. Notice the need of tractable signal processing, people instead try to deal with sampled signals that are considered as defined in discrete time spaces, or, represent the continuous signals with some finite set of basis signals (functions) $\{\phi_1, \phi_2, ..., \phi_k\}$ (e.g. the Fourier Transformation). Thanks to the Nyquist-Shannon sampling theorem, we can guarantee the recovery of original signal with band-limit no greater than half of sampling bandwidth. However, it is still computational expensive when sampling from high frequency signals like high-resolution images. Especially in some specific circumstances, such as the MR imaging, signals need to be processed in a relatively short time, which request for a faster algorithm with good quality preserved.

On the other hand, people have observed the big data explosion with sparsity. The data set can be sparse itself or after some appropriate transformations. The first type of sparsity is common in social network datasets, with a lot of missing information (such as personal info, messages and relationships). Image processing often takes advantage of the later one, by representing the original image in the frequency domain and filtering out the secondary informations. This compression works very well and

benefits the storage and transmission of large images. However, prevailing techniques of compression handle the image with processors after sensor sampling, and there exists a conflict between fast processors and slow sensors. Sensors can hardly sample the raw signals of such high rates, then the compression and recovery performed by high-speed processors will result in unsatisfactory.

1.1 Emerging Research: Compressed Sensing

To overcome the limit of relatively slow sensors, David L. Donoho raised an idea of compressed sensing (CS) in 2006 [12], which then became the focus of researches in signal processing and related fields. Researchers thought that sensors should sample the requested informations directly from the raw data, i.e. the sampling and compression are performed simultaneously, then the original signal is reconstructed by relatively few samples. From the mathematical point of view, it is equivalent to solving an underdetermined linear system, slightly different from the prevailing methods where the system is overdetermined. This new research topic called for better algorithms of sampling and reconstruction instead of improving the speed of sensors, hence it attracted mathematics, statistics and computer scientists to work on. Candès et al. theoretically proved the feasibility of the underdetermined linear system under the restricted isometry property (RIP) [8].

As introduced the foundation behind compressed sensing methods is to solve an underdetermined linear system:

$$\mathbf{y} = \Phi f. \tag{1.1}$$

It also utilizes techniques of classical signal decomposition that represent the original signal f by a set of measurement functions (the same as the base functions in Fourier transform and other methods) Φ and coefficients $\mathbf{y}$ we observed. Different from direct recovery by the pseudo-inverse approach, researchers defined an optimization problem 'Basis Pursuit' to get the best recovery indicating sparsity. Fast algorithms to find solutions are also discussed, including the converts to linear programming and LASSO problems.

One may wonder if it is also applicable to the dataset with latent sparsity even though CS methods work well in order to achieve a sparse recovery. In fact we do not know whether there exists a sparse representation at a glance, and it is also with low probability that we can just pick one of the common-used bases such as wavelets and curvelets then observe that sparsity over it. But after the development of 'dictionary'

theory we surprisingly see that with the assumption of latent sparsity, compressed sensing techniques also helps recover the original signal, which makes the research of CS more valuable. People also investigate in the construction of dictionaries, aiming at better recoveries.

1.2 Machine Learning: Heuristic Processing of Data

The concept of machine learning appeared in about 1959, by Arthur Samuel's definition that "Machine learning is a field of study that allows computers to learn without being explicitly programmed". As a branch of artificial intelligence studies, it emphasizes more on the construction of systems (or algorithms) to process (to learn from) data. Since 1980s machine learning has become a research focus of the AI field, aiming at processing datasets for recognition, classification and generalization tasks with proper algorithms, where their efficiencies have been theoretically proved.

When observing the explosion of data, people began to notice that it's intractable and time-wasting to process such large datasets by hand. Hence mathematicians, statisticians and computer science people are interested in seeking an automatic approach to handle them, with fast computers in which the capability of their computation is still emerging. Data mining as the main application world of machine learning techniques helps analysts extracting useful and meaningful informations from the sea of data, especially that of the internet which has billions of people access to. This two areas of research utilize many same algorithms and has a lot of topics overlapped, however the slightly difference is that machine learning focuses on the prediction side more, while data mining emphasizes on the knowledge discovery from the dataset (and there exists another term 'knowledge discovery and data mining' abbreviated as KDD). Since machine learning and data mining people interact and perform research together frequently, here we tend to focus on the machine learning algorithms which are also well-applied in data mining works. Furthermore, in the next chapters we show that the term 'recovery' in compressed sensing can be viewed as 'prediction' in machine learning, which implies close relationship between these two fields.

1.3 Organization of the Thesis

This thesis contains 5 chapters and is organized as follows. Other than the first chapter which gives an introduction, Chapter 2 reviews the background of signal processing, including the classical sensing theories, definition of sparsity and applications

we are familiar with. Chapter 3 discusses the theories, properties and algorithms of compressed sensing, as well as advanced CS techniques applied to image processing. Chapter 4 focuses on the learning problems in compressed sensing such as the construction of the measurement matrix, dictionary theories, looking the optimization problem from the probabilistic perspective and some machine learning techniques to solve the underdetermined system. Finally Chapter 5 concludes the thesis and points out some probable future researches.

Chapter 2

Background

Signal processing is among the hotspots of the AI studies. It contains topics of sampling, representation, compression, reconstruction and so on, and becomes an interdisciplinary study with overlaps of biology, medical researches and natural language processing. We give a brief review of these researches by introducing the classical sampling theorem and illustrating the signal decomposition tasks. In practise we are interested in the rich informations contained in the original signals, so we expect to recover the signal with better representation of those informations.

Sparsity attracts focuses from both the researchers and engineers. We live in a world filled with lots of data which are sparse (according to our need). Analysis without pre-processing of the sparsity causes the waste of time and storage, furthermore it may return worse results when the important information is masked by some noises. How to treat with the sparsity then becomes another crucial issue in the studies of datasets, aiming at better learning from them with less expenses.

2.1 Classical Sampling Theories

In signal processing, sampling reduces the continuous signals into discrete signals, which are represented by a discrete set of points in the time-value or space-value plane. Previous researches in sampling theories tried to figure out some better methods for discrete signal transformation which can avoid information loss. Also for faster transmission and less storage, signal compression is beyond the most important topics in signal processing researches. Applications based on these classical theories have brought the prosperity of modern technologies, such as *magnetic resonance imaging* (MRI) in medical use and pop MPEG coding designation.

2.1.1 Nyquist-Shannon Sampling Theorem

A significant results presented by Claude Shannon pointed out the best sampling rate when applying undersampling methods in certain circumstances [26]:

Theorem 2.1.1. (Nyquist-Shannon Sampling Theorem) *Suppose a signal $f = f(t)$ contains no frequencies higher than the bandlimit f_c, to perfectly reconstruct the original signal $f(t)$, the sampling rate f_s should be at least twice of the bandlimit, i.e.*

$$f_s \geq 2f_c. \tag{2.1}$$

In Shannon's proof of Theorem 2.1.1, he raised an example of signal reconstruction by sinc functions, which then derived as

$$f(t) = \sum_{n=-\infty}^{\infty} x_n \frac{\sin \pi(2Wt - n)}{\pi(2Wt - n)} \tag{2.2}$$

where x_n indicates the sample value. **Theorem 2.1.1** provides a sufficient condition for band-limited signal's sampling and reconstruction.

2.1.2 Signal Decomposition

A main research field in signal processing is signal decomposition, which aims at decomposing an original signal into combinations of basis signals (e.g. triangle functions) then performing operations on the coefficients. Fourier analysis has become the common method of signal decomposition since the mid 20th century, especially the discrete version is widely applied in audio and image processing. For instance, the 2-dimensions Fourier Transform in image processing is defined as:

$$F(u, v) = \frac{1}{n} \sum_{x=0}^{n-1} \sum_{y=0}^{n-1} f(x, y) exp[\frac{-2\pi i(ux + vy)}{n}]. \tag{2.3}$$

With the inverse transform defined similar as (2.3), it can recover the decomposed image with acceptable loss. Also several techniques like edge detection utilize the Fourier transform, which operates on the frequency domain then reconstruct the image with edges shown. In recent years Fourier analysis has been supplemented by other approaches, such as the most notable wavelets analysis, but it still contributes to the theoretical evaluation of the new tools as well as continuing in applications.

2.1.3 Signal Compression

Signal compression is the key technology for file storage and transmission. With lossless compression, a signal is compressed by some method which guarantees a perfect recovery. However, in many circumstances the loss of some secondary informations is acceptable, such as the compression of x-ray signal which is continuous. Lossy compression algorithms are important for the rapid transmission of voice and image data. In previous years researchers and engineers worked on revising existing methods of signal compressions, and excellent technologies such as the HDTV standard have been successfully applied in television signal transmissions.

2.1.4 MRI & MPEG: Well-known Applications

Magnetic Resonance Imaging (MRI) is a medical imaging technique which can visualize internal structures of the human body. It samples resonance signals produced by aligned protons in the water molecules. Magnetic field causes protons at different locations to generate signals in different frequencies, which allow spatial structures to be reconstructed using Fourier analysis [19]:

$$I(\omega_1, \omega_2) = \sum_{all\ \ sample\ \ points\ \ (t_1,t_2)} f(t_1, t_2) exp[-2\pi i(\omega_1 t_1 + \omega_2 t_2)]. \tag{2.4}$$

The output $I(\omega_1, \omega_2)$ represents the reconstructed image, which is expected to be identical to the original one. The coefficients $f(t_1, t_2)$ stands for the image in the frequency domain, which implies the important structures and knowledges for recovery.

The Moving Picture Experts Group (MPEG) [22] works on standards for audio and video compression and transmission. They released the MPEG-1 standard in 1993, which coded moving pictures with audio associated for file storage. During these years, they revised the old MPEG standards and released new versions, for example the MPEG-4, by adding technologies and interfaces to deal with multimedia and interactions.

2.2 Disabilities of Classical Sampling

Significant problems impact the application of the Nyquist-Shannon sampling theorem. In many situations we do not know the bandlimit f_c in advance (we often sample from raw signals then measure it), then to reduce distortions the sampling rate f_s is set to be large, which is defined as oversampling that samples many points

unnecessary for the reconstruction. Also the frequencies of signals are not consistent, then sampling at the same rate caused the waste of time and storage: it gets too much samples for a perfect recovery. Sampling with adaptive rate helps deal with this problem, however the adaptation still needs improvements for better reconstruction of the original signal. The third problem comes from the increase in data to sample. In MRI sampling (2.4) if the number of sample points (t_1, t_2) is too small, the reconstruction of the original image $I(\omega_1, \omega_2)$ is impossible. Solutions include sampling of large coefficients in the frequency domain, but previously there's little prior knowledge to indicate the positions of them hence we do not know how to assign (t_1, t_2).

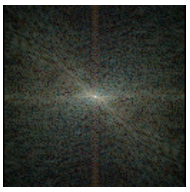
In the state-of-the-art methods of signal compression, fast processors embedded in computers can easily analyse and pick out the constitutes with important informations. However, the development of sensor's computation can hardly catch up with the processor's, so it's still computational expensive to collect raw signals, especially with high sampling rates. Also due to the limitation of the sensor, there exists some distortion when sampling from raw data. Hence the computer processing and recovery are not as accurate as the compression theory proved.

2.3 Data with Sparsity

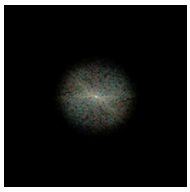
The original concept of sparsity comes from matrix analysis, in which a matrix is called sparse if only few of its elements hold non-zero values. In modern data analysis researchers often view or handle the collected data in matrix structures, with a lot of unknown (or unavailable) data setted as zeros. From this perspective, we can observe the sparsity from various datasets, especially from the social network with large amount of informations remain unknown. Twitter and microblog platforms collect and save their users' data, and their analysts construct matrices populated primarily with zeros to represent non-provided contents.

In signal processing researchers also look into the sparsity with some operations revealing it. Discrete Fourier transform is often used to process and compress images (e.g. the JPEG format of image), and after transformation people can observe the sparsity of large element if viewing the image of the frequency domain as a matrix. To compress the image with good recovery or improve the quality of the original one, small coefficients (sometimes a threshold is set to pick them out) or secondary elements are filtered out then substituted by zero, i.e. the data are expressed by a sparse matrix (see Figure 2.1).

(a) original



(b) spectrum



(c) filtered spectrum

(d) recovery

Figure 2.1: Fourier transform and recovery after filtering [28]. The original image (a) is transformed and expressed in the frequency domain (represented in Figure (b)). The low-pass filter cuts off all the high-frequency signals containing secondary informations (c) to reconstruct the image (d), which looks smoother.

Also in many circumstance the signal we sampled is sparse itself. Suppose we have a photo of the sky which was took in a sunny day, in which a large part of it is coloured in blue. When storing this photo as a BMP format image file (which store a matrix containing all the pixels of the image) lots of elements are assigned the same value that indicates the blue color when showed on the screen. MRI image comes as another instance, for that in practical use the image of human internal body also indicates sparsity. Especially, doctors often focus on the pathological regions and neglect other healthy plants in order to reach a more precise diagnose.

In short, sparsity helps people understand the structure and extract the main information from raw data. We can see in many circumstances only a small amount of data indicate useful informations for human use, or the data can be transformed to a sparse structure that represents the message more efficiently. From many applications in machine learning and data mining we know that it's crucial to develop related techniques dealing with sparse dataset, especially when the time of 'Big Data' comes with the size of dataset growing up exponentially.

Chapter 3

Compressed Sensing

Compressed sensing (compressive sensing, compressive sampling, or sparse sampling) provides new approaches to sample and recover a signal, which is sparse or compressed in some domain after transformation. It advocates for the simultaneous sampling and compression, which also allow the entire signal to be represented by relatively few measurement functions (basis signals) and coefficients. The quality of its recovery is commendable or even better than previous techniques to some extent, with very high probabilities.

3.1 Basic Idea: Theories for Sparsity

In classical sampling methods as **Theorem 2.1.1** shows, sensors sample the raw signal (which is continuous) with high rates. However, the bandlimit f_c tends to be large in practical use, hence the sampling rate f_s should be increased, which challenges nowadays sensors enormously. Instead, we obtain the idea from Fourier analysis[1]:

$$f = \sum_k y_k \phi_k. \tag{3.1}$$

ϕ_k is called the measurement function. In Fourier transform ϕ_k is the sinusoid function, which is orthogonal to other measurement functions with different subscripts. From orthogonal measurement functions the measurement y_k is computed by $y_k = \langle f, \phi_k \rangle$.

In general, we write a $m \times n$ matrix $\Phi = (\phi_1, \phi_2, ..., \phi_n)$ to represent the measurement matrix (or sensing matrix). We can also represent the measurement y_k by linear

[1]In the remain content we denote f as the raw signal and only consider the discrete circumstance.

transformation and solve that linear system to recover the original signal f:

$$\mathbf{y} = \Phi f. \tag{3.2}$$

The measurement matrix Φ is a $m \times n$ matrix, and the linear system (3.2) is underdetermined since $m < n$, so recovering f with (3.1) doesn't work. Instead of solving the linear system directly, it is better to solve an optimization problem

$$f^* = \arg\min_f ||f||_p \ s.t. \ \mathbf{y} = \Phi f. \tag{3.3}$$

Here the l_p-norm is defined as $||x||_p = (\sum_i |x_i|^p)^{\frac{1}{p}}$ (when $p \geq 1$) , and normally we set $0 \leq p \leq 2$.

3.1.1 Disabilities of l_2-Norm and l_0-Norm

Equation (3.2) defines a solution plane P in which the solution of (3.3) lies on. In previous practises researchers used the Euclid Norm ($p = 2$) in (3.3). However, the l_2-norm doesn't work well with the observation of data sparsity, since the ball often touches the solution plane at some point containing dense elements.

It is straightforward that the l_0-norm indicates sparsity directly:

$$||\mathbf{x}||_0 = |\{x_i | x_i \neq 0\}|. \tag{3.4}$$

But solving the optimization problem (3.3) with l_0-norm is NP-hard, which means that we cannot get the precise solution in polynomial time[2]. To illustrate this we introduce a formal definition of *k-sparse* vectors:

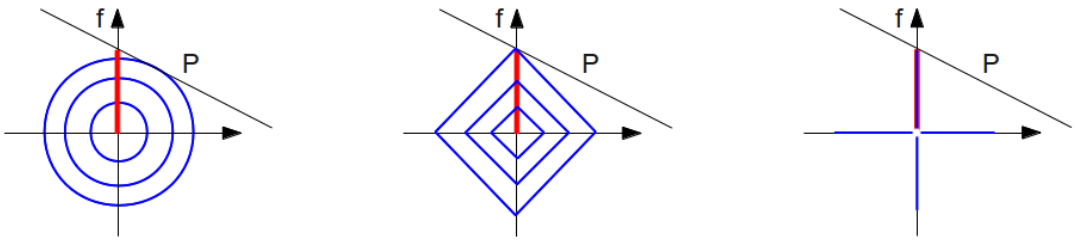
Definition 3.1.1. *For an arbitrary $k \in \mathbb{N}$, a vector $\boldsymbol{x} = (x_1, x_2, ..., x_n)$ ($k \ll n$) is said to be k-sparse if it satisfies*

$$||\boldsymbol{x}||_0 \leq k. \tag{3.5}$$

Then solving the optimization problem includes two major missions – figuring out the positions of non-zero elements, as well as the values of them (notice that we do not suppose an order here). Finding positions of k non-zero elements in a n-vector requires $O(n^k)$ computation, and finding the solution in the second step needs non-polynomial time. Moreover, the solution (if assign $p = 0$) is not unique, so it has a risk that we recover a totally different signal without any notifications.

[2]Suppose there's a NP-complete problem Q which can be reduced to (3.3) in polynomial time, we cannot guarantee that solution also satisfies (3.3) too.

To deal with the sparsity problem in less time and get stable recovery, in [11] the l_1-norm was called into use (*Basis Pursuit*). Section 3.1.2 explains why solving (3.3) with l_1-norm is equivalent to optimizing the l_0-norm under some conditions. Also we consider other norms with $0 < p < 1$, which is called the *Focal Underdetermined System Solver* (FOCUSS). We can easily figure out that although its similarity of true sparsity is better, the optimization problem may be non-convex then result in local-minima. Furthermore, when $1 < p < 2$ the ball also touches the solution plane at non-sparse points.



(a) $p = 2$ (b) $p = 1$ (c) $p = 0$

Figure 3.1: Balls with different norms. The original signal f is sparse (red line) which lies on the solution plane P. The optimization is equivalent to find the ball which the plane P is tangent to. In (a) the solution f^* (the tangent point) is different from f evidently, while in (b) and (c) they return perfect recoveries.

3.1.2 Why l_1-norm Indicates Sparsity?

In **Definition 3.1.1** we expand the definition of sparsity by using the l_0-norm. However as Section 3.1.1 indicated, solving the optimization problem (3.3) with $p = 0$ is NP-hard. Instead we use the l_1-norm and guarantee the same solution as l_0-norm minimization by introducing the concept of *mutual incoherence*:

Definition 3.1.2. *Suppose the matrices* $\Gamma = (\gamma_1, \gamma_2, ..., \gamma_n)$ *and* $\Gamma' = (\gamma'_1, \gamma'_2, ..., \gamma'_n)$ *satisfy* $||\gamma_i||_2 = ||\gamma'_i||_2 = 1$ *for any* i. *Then the mutual incoherence of* Γ *is computed by*

$$\mathcal{M}(\Gamma, \Gamma') = \max_{i,j} |\gamma_i^T \gamma'_j|. \tag{3.6}$$

For convenience we call the optimization problem (3.3) with $p = 0$ the problem (P_0) and that with $p = 1$ the problem (P_1). If the solutions of these two problems coincides, then the l_1-norm indicates sparsity and can be used to accelerate the recovery. In practice the sensing matrix Φ is not orthogonal, and here we assume the

matrix $\Phi = [\Gamma, \Gamma']$ is composed by two orthgonal system in $\mathbb{R}^{n/2}$ (n is an even number). Fortunately, in [13] David L. Donoho and X. Huo proved the coincidence with the restriction:

Theorem 3.1.1. (Donoho and Huo) *Solving the optimization problem* (P_0) *is equivalent to solve* (P_1) *if the solution* f^* *satisfies*

$$||f^*||_0 \leq \frac{1}{2}(1 + \frac{1}{\mathcal{M}(\Gamma, \Gamma')}). \tag{3.7}$$

Here the solution satisfies $\boldsymbol{y} = [\Gamma, \Gamma']f^*$.

Proof. Assume $\mathbf{y} = \Gamma f = \Gamma' g$, i.e. $\mathbf{y}$ is the sparse representations of signals f and g over the $m \times \frac{n}{2}$ measurement matrix Γ and Γ' respectively. Without loss of generality we assume both Γ and Γ' are orthogonal and $\mathbf{y}^T\mathbf{y} = 1$, then

$$1 = \mathbf{y}^T\mathbf{y} = f^T\Gamma^T\Gamma' g = \sum_{i=1}^{n/2}\sum_{j=1}^{n/2} f_i(\gamma_i^T\gamma_j')g_j \leq \sum_{i=1}^{n/2}\sum_{j=1}^{n/2} f_i\mathcal{M}(\Gamma, \Gamma')g_j.$$

From the orthogonality of the measurement matrix Γ and using the *Parseval Equality*

$$1 = \mathbf{y}^T\mathbf{y} = \sum_{i=1}^{n/2} f_i^2 = \sum_{j=1}^{n/2} g_j^2.$$

Now assume that f and g are sparse, and denote the index sets of non-zero elements as $I = \{i|f_i \neq 0\}$ and $J = \{j|g_j \neq 0\}$ respectively (also called the *support*), we can represent $\mathbf{y}^T\mathbf{y}$ by

$$\begin{aligned}\mathbf{y}^T\mathbf{y} &= \sum_{i\in I}\sum_{j\in J} f_i(\gamma_i^T\gamma_j')f_j \\ &\leq \sum_{i\in I}\sum_{j\in J} |f_i|\mathcal{M}(\Gamma, \Gamma')|f_j| \\ &= \mathcal{M}(\Gamma, \Gamma')\sum_{i\in I}|f_i|\sum_{j\in J}|g_j|.\end{aligned}$$

To indicate the upper bound let's instead solve these optimization problems

$$\begin{aligned}\arg\max_f \sum_{i\in I} f_i \ s.t. f_i > 0, ||f||_2 = 1, \\ \arg\max_g \sum_{j\in J} g_j \ s.t. g_j > 0, ||g||_2 = 1.\end{aligned} \tag{3.8}$$

Using the *Lagrange multiplier* we can solve (3.8) as

$$f_i = \frac{1}{\sqrt{||f||_0}} \ (\forall i \in I), \sum_{i \in I} f_i = \sqrt{||f||_0},$$
$$g_j = \frac{1}{\sqrt{||g||_0}} \ (\forall j \in J), \sum_{j \in J} g_j = \sqrt{||g||_0}.$$

Returning to the inequality we have

$$\frac{1}{\mathcal{M}(\Gamma, \Gamma')} \leq \sum_{i \in I} \sum_{j \in J} |f_i||g_j| \leq \sqrt{||f||_0||g||_0},$$

and by the arithmetic means inequality and $\mathcal{M}(\Gamma, \Gamma') \leq 1$

$$||f||_0 + ||g||_0 \geq \frac{2}{\mathcal{M}(\Gamma, \Gamma')} \geq 1 + \frac{1}{\mathcal{M}(\Gamma, \Gamma')}.$$

Now consider the unique recovery of $\mathbf{y}$. If there exist another solutions $f^{*'} = [f'^T, g'^T]^T$, then we have $\Gamma(f - f') = \Gamma'(g' - g) = \mathbf{y}'$, where $\mathbf{y}'$ is some vector. So we can also obtain

$$||f - f'||_0 + ||g - g'||_0 \geq \frac{2}{\mathcal{M}(\Gamma, \Gamma')}.$$

Without loss of generality we assume the l_0-norm of f^* and $f^{*'}$ are bounded by F:

$$\frac{2}{\mathcal{M}(\Gamma, \Gamma')} \leq ||f - f'||_0 + ||g - g'||_0 = ||f^* - f^{*'}||_0 < 2F.$$

Hence we achieve a looser restriction of uniqueness ($\mathcal{M}(\Gamma, \Gamma') \leq 1$):

$$||f^*||_0 \leq \frac{1}{\mathcal{M}(\Gamma, \Gamma')} \tag{3.9}$$

(otherwise there may exits non-unique solutions of the optimization problem), and we complete the proof. □

Corollary 3.1.2. (Elad and Brunkstein [14]) *The solutions of* (P_0) *and* (P_1) *still coincide if*

$$||f^*||_0 \leq \frac{\sqrt{2} - \frac{1}{2}}{\mathcal{M}(\Gamma, \Gamma')}. \tag{3.10}$$

The proof of Corollary 3.1.2 is not provided here, which can be found in the original paper. By the restriction of the l_0-norm we guarantee the coincidence of the solutions of (P_0) and (P_1).

3.1.3 Combination Methods: the LASSO

Though we have observed a large amount of sparse signals in practical use, there still exist some signals without evident sparsity. Especially we take the image as an example, it may be dense in one or some small regions. For these images the lower bound of the raw signal's l_0-norm in **Theorem 3.1.1** and **Theorem 3.1.2** need to be augmented, hence we advocate for a random measurement matrix (details in Section 3.2 and the next chapter) to reduce $\mathcal{M}(\Gamma, \Gamma')$. Also it costs more expensive computation as previously indicated. In the meanwhile there are amount of polluted signals containing large elements that do no good to the recovery (details in Section 3.3). To address these problems we introduce the *Least Absolute Shrinkage and Selection Operator* (LASSO), which can help reduce the energy (or the noise) of the signal and restrict the solution to be sparse [6]:

$$f^* = \arg\min_f ||\mathbf{y} - \Phi f||_2 + \lambda ||f||_1 \tag{3.11}$$

where λ indicates the proportion of the l_1-constraints. Or we can perform this by giving an upper bound k of the l_1-norm

$$f^* = \arg\min_f ||\mathbf{y} - \Phi f||_2 \ s.t. ||f||_1 \leq k. \tag{3.12}$$

The LASSO can better solve the recovery of polluted signal, which is discussed in Section 3.3.3.

LASSO techniques can be viewed as the l_2-optimization with l_1-penalty, for example, minimizing the error $||\mathbf{y} - \Phi f||_2$ such that $||f||_1$ is with the upper bound of some constant C. In the next sections we will see that under some constraints the l_2-minimization problem intend to choose the optimal solution indicating sparse representations, and here we give a geometric view as Figure 3.2 shows.

3.2 Algorithms of Compressed Sensing

Since the heuristic behind compressed sensing is to solve an underdetermined linear system, one may suspect that if there exists a unique and stable recovery of the original signal. From linear algebra we know that since $Rank\ \Phi < dim(f)$, if no other restrictions there will exist a subspace in which all its vectors satisfy the underdetermined system. Fortunately, the unique (and also stable or even exact) recovery is achievable with some restrictions of the sensing matrix Φ. We introduce

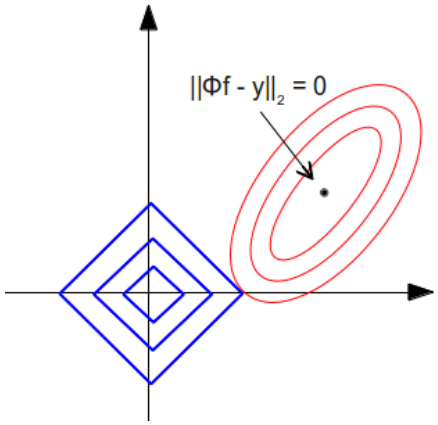


Figure 3.2: The geometry of the LASSO methods. The contours of the target function $||\mathbf{y} - \Phi f||_2$ (in red) intersect with the regularizer $||f||_1$ contours (in blue), and we pick the intersection point (lies on the axis which indicates sparsity) as the solution f^*.

these restrictions in this section, and surprisingly they're so weak that even some measurement matrix with random entries could satisfy them, in which details will be explained in the next chapter. Here we focus on the theories with a given sensing matrix Φ satisfying them as well as the algorithms of recovery computation, while the construction of Φ (related to learning) will be also discussed later.

3.2.1 Restricted Isometry Property

In (3.2) researchers focused on the construction of the measurement matrix Φ, which may be not orthogonal. To gain the equivalence of problem (P_0) and (P_1) with weak restrictions, [9] introduced the definition of the *restricted isometry constant*:

Definition 3.2.1. (Restricted Isometry Constant) *Let Φ be the $m \times n$ measurement matrix. For any integer $1 \leq S \leq n$, we define the S-restricted isometry constant δ_S be the smallest real number such that for all $k \leq S$ and k-sparse f with support $T \subset \{1, 2, ..., n\}$,*

$$(1 - \delta_S)||f||_2^2 \leq ||\Phi_T f_T||_2^2 \leq (1 + \delta_S)||f||_2^2. \tag{3.13}$$

Here Φ_T stands for the matrix formed by the column vectors with indexes in T.

Similarly we define the S, S'-restricted isometry constant $\theta_{S,S'}$ for $S + S' \leq n$ that $\theta_{S,S'}$ is the smallest number satisfying

$$|\langle \Phi_T f_T, \Phi_{T'} f'_{T'} \rangle| \leq \theta_{S,S'} ||f||_2 ||f'||_2 \tag{3.14}$$

for all $k \leq S$, $k' \leq S'$ and k-sparse vector f, k'-sparse f' with disjoint supports T and

T', *respectively.*

The numbers δ_S and θ_S indicate how close the measurement system Φ behaves like an orthonormal system when picking out vectors without no greater than S non-zero elements randomly, from the original space $\mathbb{R}^n$. Also without loss of generality it is straightforward to deduce that

$$|\langle \Phi_T f_T, \Phi_{T'} f'_{T'} \rangle| \leq \delta_{S+S'} ||f||_2 ||f'||_2 \tag{3.15}$$

by assuming $||f||_2 = ||f'||_2 = 1$ and applying the definition of the $S + S'$-restricted isometry constant.

With the definition of the *restricted isometry constant* the *restricted isometry property* (RIP) goes like this:

Definition 3.2.2. (Restricted Isometry Property) *The measurement matrix* Φ *is said to obey the restricted isometry property of order* S *if there exists a* S*-restricted isometry constant* $\delta_S \in (0, 1)$.

Here we introduce the uniqueness of recovery then discuss the equivalence of two optimization problems with restrictions of the RIP.

Theorem 3.2.1. (Uniqueness of Reconstruction) *Suppose* $\delta_{2S} < 1$*, then for all* k*-sparse* ($k < S$) *signal* f *with support* T *(i.e.* $|T| = k < S$*), the index set* T *and signal* f *can be reconstructed uniquely from the measurement* $\boldsymbol{y}$ *and measurement matrix* Φ.

Proof. If invalid, suppose there's a different representation of $\mathbf{y} = \Phi_{T'} f'$ in which f' has support T' satisfying $|T'| < S$ and $T \cap T' = \emptyset$. Then from (3.13) we have

$$(1 - \delta_{2S}) ||(f - f')||_2^2 \leq ||\Phi_{T \cup T'} (f - f')_{T \cup T'}||_2^2 = 0.$$

Since $\delta_{2S} < 1$ we have $f = f'$, $T = T'$ and reach a contradiction.

Now we assume $T \cap T' \neq \emptyset$, then we have some vector $\mathbf{y}'$ satisfying

$$\mathbf{y}' = \Phi_T (f - f'_{T' \cap T})_T = \Phi_{T' \setminus T} f'_{T' \setminus T}$$

where $f - f'_{T' \cap T}$ has support T and $f'_{T' \setminus T}$ has support $T' \setminus T$ respectively. So with similar manner we have $f = f'$, $T = T'$, also a contradiction. □

Before discussing the equivalence problem we first introduce Lemma 3.2.2:

Lemma 3.2.2. *Suppose $H(\Phi) = span(\{\phi_1, \phi_2, ..., \phi_n\})$ is the Hilbert space spanned by the measurement vectors, then if $\delta_{2S} < 1$, for any f with support T there exists a vector w such that*

$$\begin{aligned} \langle w, \phi_j \rangle_{H(\Phi)} &= \text{sgn}(f_j), \ \forall j \in T, \\ |\langle w, \phi_j \rangle_{H(\Phi)}| &< 1, \forall j \notin T. \end{aligned} \tag{3.16}$$

Theorem 3.2.3. (Equivalence of Problem (P_0) and (P_1)) *Suppose $S \geq 1$ and $\delta_S + \theta_S + \theta_{S,2S} < 1$, Then for any f and its support T satisfying $|T| \leq S$, f is the unique solution of the optimization problem*

$$f^* = \arg\min_x ||x||_1 \ s.t. \ \boldsymbol{y} = \Phi x \tag{3.17}$$

for any measurement matrix Φ and the measurement $\boldsymbol{y} = \Phi f$.

Proof. $\delta_S + \theta_S + \theta_{S,2S} < 1$ implies $\delta_{2S} < 1$. We can apply **Lemma 3.2.2** then gain the vector w. Suppose one of the solution of (3.17) is f', since f also satisfies $\mathbf{y} = \Phi f$ we have $||f'||_1 \leq ||f||_1$. Now we prove the l_1-norm of f' is no lower than $||f||_1$:

$$\begin{aligned} ||f'||_1 &= \sum_{j \in T} |f_j + (f'_j - f_j)| + \sum_{j \notin T} |f'_j| \\ &\geq \sum_{j \in T} sgn(f_j)(f_j + f'_j - f_j) + \sum_{j \notin T} f'_j \langle w, \phi_j \rangle_{H(\Phi)} \\ &= \sum_{j \in T} |f_j| + \sum_{j \in T} \langle w, \phi_j \rangle_{H(\Phi)} (f'_j - f_j) + \sum_{j \notin T} \langle w, \phi_j \rangle_{H(\Phi)} f'_j \\ &= \sum_{j \in T} |f_j| + \langle w, \sum_{j=1}^{n} f'_j \phi_j - \sum_{j \in T} f_j \phi_j \rangle_{H(\Phi)} \\ &= \sum_{j \in T} |f_j| + \langle w, \mathbf{y} - \mathbf{y} \rangle_{H(\Phi)} \\ &= ||f||_1. \end{aligned}$$

Also we have $|\langle w, \phi_j \rangle| < 1$ for all $j \notin T$, then $f'_j = 0$ $(\forall j \notin T)$ to guarantee that $||f'||_1 = ||f||_1$. Finally we apply **Theorem 3.2.1** $(\delta_{2S} < 1)$ and get $f' = f$, i.e. solving (P_1) is equivalent to solve (P_0). □

Corollary 3.2.4. (Noiseless Recovery [4]) *The conclusion of* **Theorem 3.2.3** *still holds if $\delta_{2S} \leq \sqrt{2} - 1$.*

We do not provide the proof here since it's similar to that of the noisy circumstance, which is more general and will be discussed in Section 3.3.1.

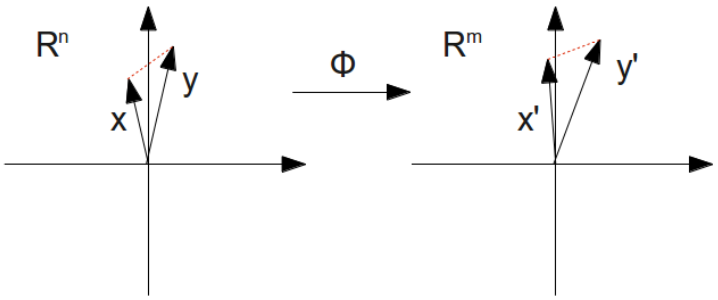


Figure 3.3: The geometry of the RIP. In the left figure x and y are some arbitrary *S-sparse* vectors, and after projection Φ which obeys RIP with RIC $\delta_{2S} < 1$ the transformed x' and y' keep the Euclid distance $||x'-y'||_2 \approx ||x-y||_2$ (red dashed line). Note that $x'-y'$ is at most $2S$-*sparse* in which $\delta_{2S} < 1$ guarantees the preservation of dataset geometry to some extent. Also the crucial notice is that this good condition of 'geometry preservation' only validates for *S-sparse* vectors in $\mathbb{R}^n$.

Remark. (of **Corollary 3.2.4**) In practise we often extract the S-largest elements of f and eliminate others (substituted by zero). For convenience we denote this vector as f_S. Then the recovery indeed contains bounded errors:

$$\begin{aligned} ||f^* - f||_1 &\leq C_0||f_S^* - f||_1, \\ ||f^* - f||_2 &\leq C_0 S^{-\frac{1}{2}}||f_S^* - f||_2. \end{aligned} \tag{3.18}$$

For those S-sparse vectors the recovery is exact as **Corollary 3.2.4** described.

3.2.2 Stable Recovery

In this section we introduce that the RIP indicates the *Uniform Uncertainty Principle* (UUP), and combining with the *Exact Reconstruction Principle* (ERP) we can figure out how stable the recovery is. We skip the discussion of the construction of measurement matrix Φ, which is presented in Chapter 4.

Given an n-dimensional Hilbert space H with basis functions $\{\phi_i \in \mathbb{R}^n\}$, we want to form a matrix by randomly selecting n basis functions ϕ_i (denoted as $[\phi_1, \phi_2, ..., \phi_n]$) then picking out m rows of that matrix with index set Ω and denote it as Φ^Ω:

$$\Omega = \{i|X_i = 1\},\ X_i \sim Bernoulli(\tau)\ i.i.d.$$

in which τ satisfies $E(|\Omega|) = n\tau = m$. Note that in previous sections Φ was the $m \times n$ measurement matrix, which can be viewed that the index set Ω is given in advance then assign $\Phi := \Phi^\Omega$. In Section 3.2.2, Section 3.2.3 and the MP algorithm in Section

3.2.4 we use the symbol Φ^Ω to indicate the constructed measurement matrix and Φ the matrix formed by vectors in H, while in other sections we still use Φ to represent the sensing matrix.

Definition 3.2.3. (Uniform Uncertainty Principle [10]) *A measurement matrix Φ^Ω obeys the uniform uncertainty principle with oversampling factor λ if for every sufficiently small value $\alpha > 0$, there exists a fixed constant $p > 0$ such that the statement following is true with probability no less then $1 - O(n^{-p/\alpha})$: for any support T of f which obeys*

$$|T| = k \leq \alpha \times \frac{m}{\lambda}, \tag{3.19}$$

the matrix Φ^Ω satisfies

$$\frac{1}{2}\frac{m}{n}||f||_2^2 \leq ||\Phi^\Omega f||_2^2 \leq \frac{3}{2}\frac{m}{n}||f||_2^2. \tag{3.20}$$

Remark. In [10] it provided the explanation why **Definition 3.2.3** is referred as the 'uniform uncertainty principle'. The 'uncertainty' illustrates that '(with overwhelming probability) a sparse signal f cannot be concentrated in frequency on Ω regardless of T, unless m is comparable to n.' On the other hand, terminology 'uniform' stands for that 'with overwhelming probability, we obtain the estimate (3.20) for all sets T obeying (3.19)'.

Also the constants $(\frac{1}{2}, \frac{3}{2})$ can be replaced by any pair (a, b) with $a > 0$ and $b < \infty$. In fact it is equivalent to the RIP expression by setting $a = \frac{n}{m}(1 - \delta_S)$, $b = \frac{n}{m}(1 + \delta_S)$ with $\delta_S < 1$.

Definition 3.2.4. (Exact Reconstruction Principle [10]) *We say the measurement matrix Φ^Ω satisfies the exact reconstruction principle with oversampling factor λ if for all sufficiently small $\alpha > 0$, each fixed T satisfying* (3.19) *and each 'sign' vector σ on T that $|\sigma(i)| = 1$, there exists a vector p at least with the same probability in* **Definition 3.2.3** *such that*
(1) $p_i = \sigma_i,\ \forall i \in T$;
(2) $\exists v$ s.t. $p = (\Phi^\Omega)^T v$;
(3) $|p_i| \leq \frac{1}{2}, \forall i \notin T$.

The stable recovery of the original signal is guaranteed by the UUP and the ERP in the meaning of minimizing the error level in l_2-norm. We introduce the theorem as well as lemmas needed for proof here, and the proofs are provided in [10] (so we do not discuss here). In fact **Lemma 3.2.4** is another version of **Lemma 3.2.5**, so the proof of l_1-norm stability is similar by proving the coincidence of the original signal's

and optimal signal's l_1-norms then the coincidence of their supports and elements. Note that if without noises then the recovery is exact.

Lemma 3.2.5. (Extension) *Assume Φ^Ω obeys the UUP, then with probability no less then $1-O(n^{-p/\alpha})$, for all index set T obeying* (3.19) *and signal f which support is T (i.e. $f \in l_2(\Omega)$), there exists another signal f' such that:*
(1) $f'_i = f_i$, $\forall i \in T$;
(2) $\exists w \in l_2(\Omega)$ s.t. $f' = (\Phi^\Omega)^T w$;
(3) $\forall T' \subset \{1, 2, ..., n\}$, $||f'_{T'}||_2 \leq C(1 + \frac{\lambda|T'|}{\alpha m})^{\frac{1}{2}}||f||_2$.
In (3) $f'_{T'}$ represents the vector composed by the elements of f' indexed in T' with order preserved.

Lemma 3.2.6. (ERP Recovery) *Assume Φ^Ω obeys the ERP, then for all $f = f_0 + h$ in which f_0 has the support T obeying the assumption of the UUP* (3.19)*, with probability no less then $1 - O(n^{-p/\alpha})$ the optimal solution f^* of (P_1) satisfies*

$$||(f^*)^T \mathbf{1}_{T^c}||_1 \leq 4||h||_1 \tag{3.21}$$

Here we consider the distortion h as a kind of 'structural noises', which is different from the random noises common in the sampling procedure. We will discuss the sampling with random noises in the next section and now focus on the stable recovery guaranteed by the sensing matrix which obeys the UUP and the ERP.

Theorem 3.2.7. (Recovery with UUP and ERP) *Assume Φ^Ω satisfies the UUP and the ERP with oversampling factors λ_1 and λ_2 respectively. Let $\lambda = \max(\lambda_1, \lambda_2)$ and $m \geq \lambda$. Then given a constant R, suppose a signal f in which the nth-largest coefficient of $|\Phi^\Omega f|$ (denoted as $|\Phi^\Omega f|_{(n)}$)satisfies*
(1) $|\Phi^\Omega f|_{(n)} \leq R \cdot n^{-1/p}$ for a fixed $0 < p < 1$
(2) or $||f||_1 \leq R$ when $p = 1$.
Assign $r := 1/p - 1/2$. Then for all sufficiently small α, the solution f^* of (P_1) (with sensing matrix Φ^Ω) satisfies

$$||f - f^*|| \leq C_{p,\alpha} \cdot R \cdot (\frac{m}{\lambda})^{-r} \tag{3.22}$$

with probability no less then $1 - O(n^{-p/\alpha})$. Here the constant $C_{p,\alpha}$ depends on the choices of those parameters.

How to interpret the relationship between the RIP and the UUP? Actually the RIP is a more abstract version of the UUP, i.e. those matrices Φ^{Ω} satisfying the RIP also satisfy the UUP with probability almost one. Also Φ^{Ω} will also satisfies the ERP if we have $\delta_S + \theta_{S,2S} < 1$ and the typical vector w in **Lemma 3.2.2** obeying

$$|\langle w, \phi_j \rangle| \leq \frac{\theta_S}{1 - \delta_S - \theta_{S,2S}\sqrt{S}} ||f||_2, \tag{3.23}$$

which is automatically satisfied in **Theorem 3.2.3** [9]. So with the RIP (or the UUP+ERP) and the assumption in **Theorem 3.2.3**, the error level of reconstruction can be controlled as **Theorem 3.2.7** described.

3.2.3 Number of Samples for Stable Recovery

Compressed sensing algorithms aims at recovering the original signals from few collected data, i.e. the measurement matrix Φ has dimensions $m \ll n$. But how to assign m in order to ensure the stable recovery given n and assumed f is *k-sparse*? Surprisingly, the measurement matrix Φ works well if it is randomly constructed. To explain this, we introduce the **coherence parameter** with the assumption that Φ is composed by n vectors randomly selected from some basis set (population), and we denote such a set as $\mathcal{B}(\Phi)$[3]

$$\mu(\Phi) = \underset{a \in \mathcal{B}(\Phi)}{supp} (\max_j |a_j|^2) \tag{3.24}$$

and the restriction of *isotropy property*

$$E[aa^T] = I, \ \forall a \in \mathcal{B}(\Phi), \tag{3.25}$$

then we have the theorem:

Theorem 3.2.8. (Noiseless incoherent sampling) *If the signal f is k-sparse and the basis population $\mathcal{B}(\Phi)$ satisfies the isotropy property, then for any $\beta > 0$, with probability at least $1 - 5/n - e^{-\beta}$ the signal can be perfectly recovered if*

$$m \geq C(1 + \beta)\mu(\Phi) k \log n. \tag{3.26}$$

where C is some positive constant.

[3]There may exist several sets satisfying this definition.

Some simple interpretation of **Theorem 3.2.8** goes like this. Assume n is known and intuitively k is related to the empirical distribution of such a kind of signals, the only method to decrease the lower bound of m is by modifying $\mu(\Phi)$ (actually β implies the accuracy of the recovery). In fact Φ is composed by the basis functions randomly sampled from $\mathcal{B}(\Phi)$, hence Φ indicates the structure and distribution of that basis population. The theorem clarifies the effectiveness of random sensing matrix, which relieves us from the burden of feature selection. Detailed proof of the theorem can be found in [6], another explanation of this is also available in Section 4.1.1.

3.2.4 Solving the Underdetermined Linear System

While the exact recovery of the original signal is guaranteed by the RIP, another mission in the research of compressed sensing is to find a fast algorithm to solve (P_0). Although previous techniques like the simplex method can work, recently the *Matching Pursuit* (MP) emerges as the state-of-the-art approach. MP aims at finding the 'best matching' measurements of high-dimensional data over an over-complete dictionary[4], where in compressed sensing the dictionary is exactly the measurement matrix Φ^{Ω}. The crucial points are finding the set Ω then figuring out the measurements $\mathbf{y}$ under Φ^{Ω}, which can be solved by the greedy algorithm implied by the MP methods. Notice that this method does not guarantee the exact reconstruction as previously explained. In practice, the prevailing methods of reconstruction include the *orthogonal matching pursuit* (OMP) algorithm [21]:

Problems of the OMP include the disability to yield accurate (or with acceptable error level ϵ) recoveries (see the last 5 lines in Algorithm 1). To deal with this researchers instead focus on the fast algorithms of *Basis Pursuit* (BP) which is exactly the (P_1) problem. In [11] it proved that (P_1) can be solved by linear programming (LP). Consider the formal expression of LP

$$\begin{aligned} \mathbf{g}^* = \arg\min_{\mathbf{g}} & \mathbf{c}^T\mathbf{g} \ s.t. \ A\mathbf{g} \ = \mathbf{b} \\ c_i(\mathbf{g}_i) \leq 0, \ & i = 1, 2, ..., q \\ \mathbf{g}_i \geq 0, \ & \mathbf{g} \in \mathbb{R}^s; \end{aligned} \tag{3.27}$$

in which the inequality constraint $c_i()$ is linear. Hence we can convert the BP problem

[4]The concept of dictionary will be discussed in Chapter 4.

Algorithm 1 Orthogonal Matching Pursuit for Compressed Sensing

initialize Φ, k, $\mathbf{y}$, ϵ
$t \leftarrow 1$, $r_0 \leftarrow \mathbf{y}$, $\Omega = \varnothing$
while $t \leq k$ **do**
  $\lambda_t \leftarrow \arg\max_j |\langle \phi_j, r_{t-1} \rangle|$
  $\Omega \leftarrow \Omega \cup \{\lambda_t\}$
  $f^*(t) \leftarrow \arg\min_f ||\mathbf{y} - \Phi^\Omega f||_2$
  $\mathbf{y}(t) \leftarrow \Phi^\Omega f^*(t)$
  $r(t) \leftarrow \mathbf{y} - \mathbf{y}(t)$
  $t \leftarrow t + 1$
end while
if $||r(t)||_2 \leq \epsilon$ **then**
  return $f^*(t)$
else
  return 'No solution found.'
end if

into the style of LP by assigning

$$s := 2m,\ A := (\Phi, -\Phi),\ \mathbf{b} := \mathbf{y},\ \mathbf{c} = (\mathbf{1}_n, \mathbf{1}_n),\ \mathbf{g} := (\mathbf{u}, \mathbf{v}), \tag{3.28}$$

then solve (P_1) by $f = \mathbf{u} - \mathbf{v}$. The '$l_1$-Magic' software [7] provided a *primal-dual log-barrier LP* algorithm to solve (3.27) with parameters (3.28). By utilizing the *Karush-Kuhn-Tucker* conditions in which the optimal solution $\mathbf{g}$ of LP and that of its dual problem $\mathbf{v}^*$ should satisfy:

$$\begin{aligned} \mathbf{c} + A^T\mathbf{v}^* + \sum_i \lambda_i^* \bigtriangledown c_i(\mathbf{g}^*) &= 0 \\ \lambda_i^* c_i(\mathbf{g}^*) &= 0 \\ A\mathbf{g}^* &= \mathbf{b} \\ c_i(\mathbf{g}^*) &\leq 0 \end{aligned} \tag{3.29}$$

where λ_i^* are the Lagrange multipliers (in fact KKT is a kind of generalized Lagrange multipliers methods). In practise we loose the equality condition $\lambda_i^* c_i(\mathbf{g}^*) = 0$ to $\lambda_i^* c_i(\mathbf{g}^*) = -1/\tau^t$ in the tth iteration and compute $(\mathbf{g}^*, \mathbf{v}^*, \lambda^*)$ by newton methods. Denote $(\mathbf{g}^*(t), \mathbf{v}^*(t), \lambda^*(t))$ as the solution in the tth iteration and we have the Algorithm 2.

The trick is to obtain $(\Delta\mathbf{g}, \Delta\mathbf{v}, \Delta\lambda)$ which yields $r_\tau(\mathbf{g}^*(t) + \Delta\mathbf{g}, \mathbf{v}^*(t) + \Delta\mathbf{v}, \lambda^*(t) + \Delta\lambda) = 0$. We use the gradient methods and approximate $(\Delta\mathbf{g}, \Delta\mathbf{v}, \Delta\lambda)$ with the

Algorithm 2 Primal-Dual Log-Barrier LP Algorithm

initialize $t \leftarrow 1$, $(\mathbf{g}^*(t), \mathbf{v}^*(t), \lambda^*(t))$, s
while $stop(\mathbf{g}^*(t), \mathbf{v}^*(t), \lambda^*(t)) = False$ **do**
  compute the residual column vector $r_\tau(\mathbf{g}^*(t), \mathbf{v}^*(t), \lambda^*(t)) = (r_{dual}, r_{cent}, r_{pri})^T$:

$$r_{dual} \leftarrow \mathbf{c} + A^T\mathbf{v}^*(t) + \sum_i (\lambda^*(t))_i \nabla c_i(\mathbf{g}^*(t))$$

$$r_{cent} \leftarrow -\sum_i (\lambda^*(t))_i c_i(\mathbf{g}^*(t)) - (1/\tau)\mathbf{1}$$

$$r_{pri} \leftarrow A\mathbf{g}^*(t) - \mathbf{b}$$

  compute the Jacobian matrix $J_{r_\tau}(\mathbf{g}^*(t), \mathbf{v}^*(t), \lambda^*(t))$
  compute $(\Delta\mathbf{g}, \Delta\mathbf{v}, \Delta\lambda)$ by solving

$$J_{r_\tau}(\mathbf{g}^*(t), \mathbf{v}^*(t), \lambda^*(t))(\Delta\mathbf{g}, \Delta\mathbf{v}, \Delta\lambda)^T = -r_\tau(\mathbf{g}^*(t), \mathbf{v}^*(t), \lambda^*(t))$$

  update: $(\mathbf{g}^*(t+1), \mathbf{v}^*(t+1), \lambda^*(t+1)) \leftarrow (\mathbf{g}^*(t), \mathbf{v}^*(t), \lambda^*(t)) + s(\Delta\mathbf{g}, \Delta\mathbf{v}, \Delta\lambda)$
  if $\exists i$ *s.t.* $(\lambda^*(t+1))_i < 0$ *or* $c_i(\mathbf{g}^*(t+1)) > 0$ **then**
    adjust s s.t. $\forall i$ $(\lambda^*(t+1))_i \geq 0$, $c_i(\mathbf{g}^*(t+1)) \leq 0$
    revise $(\mathbf{g}^*(t+1), \mathbf{v}^*(t+1), \lambda^*(t+1)) \leftarrow (\mathbf{g}^*(t), \mathbf{v}^*(t), \lambda^*(t)) + s(\Delta\mathbf{g}, \Delta\mathbf{v}, \Delta\lambda)$
  end if
  $t \leftarrow t + 1$
end while
return $(\mathbf{g}^*(t), \mathbf{v}^*(t), \lambda^*(t))$

Taylor expansion

$$r_\tau(\mathbf{g}^*(t)+\Delta\mathbf{g},\mathbf{v}^*(t)+\Delta\mathbf{v},\lambda^*(t)+\Delta\lambda) \approx r_\tau(\mathbf{g}^*(t),\mathbf{v}^*(t),\lambda^*(t))+J_{r_\tau}(\mathbf{g}^*(t),\mathbf{v}^*(t),\lambda^*(t))(\Delta\mathbf{g},\Delta\mathbf{v},\Delta\lambda)^T. \tag{3.30}$$

Since the constraints are linear we have the Jacobian matrix

$$J_{r_\tau}(\mathbf{g}^*(t),\mathbf{v}^*(t),\lambda^*(t))=\begin{pmatrix} 0 & A^T & C^T \\ -\Lambda(t)C & 0 & -C'(t) \\ A & 0 & 0 \end{pmatrix}$$

where $C=(\nabla c_1(),\nabla c_2(),...,\nabla c_q())^T$, $C'(t)=diag(c_1(\mathbf{g}^*(t)),c_2(\mathbf{g}^*(t)),...,c_q(\mathbf{g}^*(t)))$ and $\Lambda(t)=diag((\lambda^*(t))_1,(\lambda^*(t))_2,...,(\lambda^*(t))_q)$. The function $stop(\mathbf{g}^*(t),\mathbf{v}^*(t),\lambda^*(t))$ indicates the termination condition such as the requirement of the recovery satisfying $||\mathbf{y}-\Phi f||_2\leq\epsilon$, where ϵ is some threshold indicating the acceptable error level. The step length (or the learning rate) $s\in(0,1]$ should guarantee the fulfilment of the constraints in (3.27) as well as the efficient decrease of the residuals, i.e. for some constant α

$$||r_\tau(\mathbf{g}^*(t)+s\Delta\mathbf{g},\mathbf{v}^*(t)+s\Delta\mathbf{v},\lambda^*(t)+s\Delta\lambda)||_2\leq(1-\alpha s)||r_\tau(\mathbf{g}^*(t),\mathbf{v}^*(t),\lambda^*(t))||_2.$$

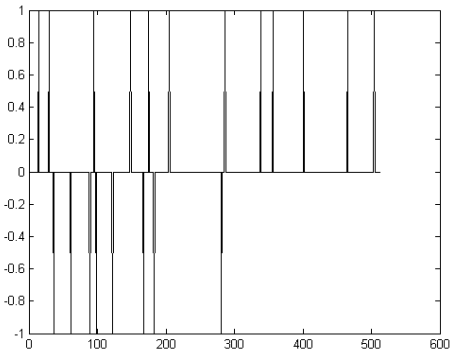
Figure 3.4 illustrates the comparison of the accuracy between BP and energy minimization methods.

There exists some other algorithms to solve BP problems fast and smoothly. These include (cite). Also some improvement of OMP aiming at solving (P_0) include StOMP, ROMP and CoSaMP.

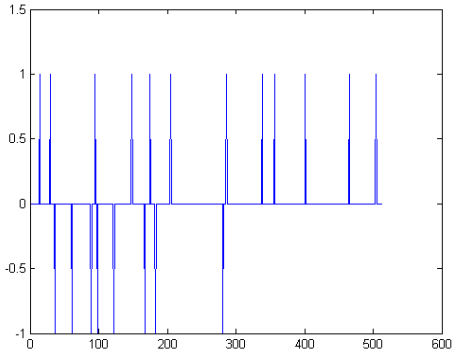
3.3 Recovery of Polluted Sampling

In the application of compressed sensing, we sample the signal in which the noise cannot be filtered out perfectly. Hence our goal now is to expand the result in Section 3.2 to this more practical situation. We revise (3.2) to represent the polluted measurements as

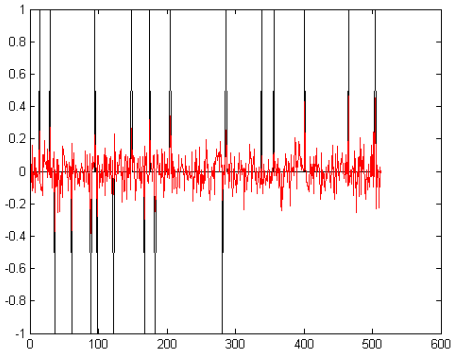
$$\mathbf{y}=\Phi f+\mathbf{z} \tag{3.31}$$



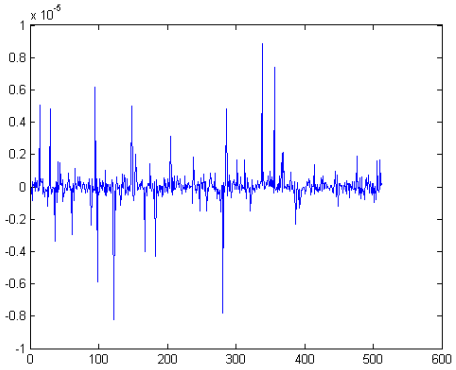
(a) original



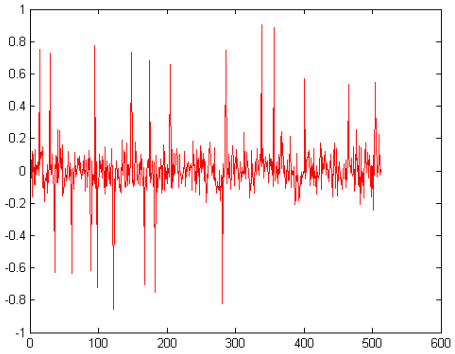
(b) recovery ($p = 1$)



(c) recovery ($p = 2$)



(d) error ($p = 1$)



(e) error ($p = 2$)

Figure 3.4: Comparing l_1 and l_2 recoveries. By running the matlab code [7] we test the accuracy of l_1-recovery compared to l_2-optimization. The original signal (black spikes) in (a) has 512 elements with only 20 spikes ranging from 0 to 1. Blue spikes in (b) and (d) stand for the recovery from sparse measurements with $m = 120$ and randomly generated Φ according to the Gaussian distribution. Red spikes in (c) and (e) represent the reconstructed signal with minimized energy (in l_2-norm). We can easily see that the recovery minimizing the energy runs far away from the original that the error level $||f^* - f|| = 3.7739$, while with l_1-norm minimization it returns nearly perfect reconstruction and $||f^* - f||_2 = 2.5061e^{-5}$.

where $\mathbf{z}$ is the noise we sampled and $dim(\mathbf{y}) = dim(\mathbf{z})$. Then the optimization problem should be revised as

$$f^* = \arg\min_f ||f||_p \ s.t. \ ||\mathbf{y} - \Phi f||_2 \leq \epsilon. \tag{3.32}$$

We denote the problem (3.32) as (P_1') with $p = 1$ and (P_0') with $p = 0$.

Fortunately, with revision adding the noise, the 3 properties introduced before still hold then with random sensing matrix they also guarantee the equivalence of (P_0') and (P_1') as well as the recovery with acceptable error.

3.3.1 Noisy Recovery

With the RIP holds we introduce the recovery from a signal with noise:

Theorem 3.3.1. (Noisy Recovery [4]) *Assume* $\delta_{2k} < \sqrt{2} - 1$ *and* $||\boldsymbol{z}||_2 \leq \epsilon$, *then the solution of* (3.32) *satisfies*

$$||f^* - f||_2 \leq C_0 k^{-\frac{1}{2}} ||f - f_k||_2 + C_1 \epsilon \tag{3.33}$$

with C_0 *the same as that of the* **Remark** *of* **Corollary 3.2.4** *and another constant* C_1.

We give the abbreviated proof here and detailed version is available in [8]. Figure 3.5 illustrates the geometry of l_1 noisy recovery.

Proof. Denote $h = f^* - f$, then we can divide the set $\{1, 2, ..., n\}$ into subsets $T_0, T_1, ...$ of size at most k s.t. $h = \sum_i h_{T_i}$, T_0 contains indexes of the k-largest coefficients of f and T_i contains locations of the k-largest coefficients of $h_{(T_0 \cup T_1 \cup ... \cup T_{i-1})^c}$. Then we can proof this 3 statement (see [4]) that

$$\sum_{i \geq 2} ||h_{T_i}||_2 \leq k^{-1/2} ||h_{T_0^c}||_1 \tag{3.34}$$

$$||h_{T_0^c}||_1 \leq ||h_{T_0}||_1 + 2||f_{T_0^c}||_1, \tag{3.35}$$

$$||h_{(T_0 \cup T_1)^c}||_2 \leq ||h_{T_0}||_2 + 2k^{-1/2} ||f - f_k||_1. \tag{3.36}$$

Then we apply the RIP to $h_{T_0 \cup T_1}$ and from the inequality $||\Phi h||_2 \leq ||\Phi(f^* - \mathbf{y})||_2 +$

$||\Phi(\mathbf{y} - f)||_2$ we have

$$\begin{aligned} ||\Phi h_{T_0 \cup T_1}||_2^2 &= \langle \Phi h_{T_0 \cup T_1}, \Phi h \rangle - \langle \Phi h_{T_0 \cup T_1}, \sum_{i \geq 2} \Phi h_{T_i} \rangle \\ &\leq ||\Phi h_{T_0 \cup T_1}||_2 ||\Phi h||_2 - \langle \Phi h_{T_0 \cup T_1}, \sum_{i \geq 2} \Phi h_{T_i} \rangle \\ &\leq 2\epsilon \sqrt{1 + \delta_{2k}} ||h_{T_0 \cup T_1}||_2 - \langle \Phi h_{T_0 \cup T_1}, \sum_{i \geq 2} \Phi h_{T_i} \rangle. \end{aligned} \tag{3.37}$$

Moreover from (3.15) we know that $|\langle \Phi h_{T_i}, \Phi h_{T_j} \rangle| \leq \delta_{2k} ||h_{T_i}||_2 ||h_{T_j}||_2$. Note that $h_{T_0 \cup T_1} = h_{T_0} + h_{T_1}$ since $T_0 \cap T_1 = \varnothing$, and the inequality $||h_{T_0}||_2 + ||h_{T_1}||_2 \leq \sqrt{2} ||h_{T_0 \cup T_1}||_2$ also holds. We apply the RIP to $h_{T_0 \cup T_1}$ and the inequality (3.34)

$$\begin{aligned} (1 - \delta_{2k}) ||h_{T_0 \cup T_1}||_2^2 &\leq ||\Phi h_{T_0 \cup T_1}||_2^2 \\ &= \langle \Phi h_{T_0 \cup T_1}, \Phi h \rangle - \langle \Phi (h_{T_0} + h_{T_1}), \sum_{i \geq 2} \Phi h_{T_i} \rangle \\ &\leq 2\epsilon \sqrt{1 + \delta_{2k}} ||h_{T_0 \cup T_1}||_2 + \delta_{2k} \sum_{i \geq 2} ||h_{T_0}||_2 ||h_{T_i}||_2 + \delta_{2k} \sum_{i \geq 2} ||h_{T_1}||_2 ||h_{T_i}||_2 \\ &\leq (2\epsilon \sqrt{1 + \delta_{2k}} + \sqrt{2} \delta_{2k} \sum_{i \geq 2} ||h_{T_i}||_2) ||h_{T_0 \cup T_1}||_2 \\ &\leq (2\epsilon \sqrt{1 + \delta_{2k}} + \sqrt{2} \delta_{2k} k^{-1/2} ||h_{T_0^c}||_1) ||h_{T_0 \cup T_1}||_2. \end{aligned}$$

So we can get the upper bound of $||h_{T_0 \cup T_1}||_2$ by applying (3.35) that

$$\begin{aligned} ||h_{T_0 \cup T_1}||_2 &\leq \alpha\epsilon + \rho k^{-1/2} ||h_{T_0^c}||_1 \\ &\leq \alpha\epsilon + \rho k^{-1/2} ||h_{T_0 \cup T_1}||_2 + 2\rho k^{-1/2} ||f - f_k||_1 \\ &\leq (1 - \rho)^{-1} (\alpha\epsilon + \rho k^{-1/2} ||f - f_k||_1), \\ \alpha = \frac{2\sqrt{1 + \delta_{2k}}}{1 - \delta_{2k}}, &\qquad \rho = \frac{\sqrt{2}\delta_{2k}}{1 - \delta_{2k}}. \end{aligned}$$

Finally we conclude our proof by applying (3.36)

$$\begin{aligned} ||h||_2 &\leq ||h_{T_0 \cup T_1}||_2 + ||h_{(T_0 \cup T_1)^c}||_2 \\ &\leq 2||h_{T_0 \cup T_1}||_2 + 2k^{-1/2} ||f - f_k||_1 \\ &\leq 2\frac{1 + \rho}{1 - \rho} k^{-1/2} ||f - f_k||_1 + \frac{2\alpha}{1 - \rho}\epsilon, \end{aligned}$$

which is we want to show ($C_0 := 2\frac{1+\rho}{1-\rho}$ and $C_1 := \frac{2\alpha}{1-\rho}$). □

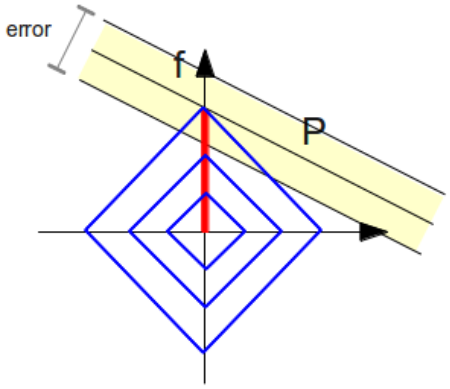


Figure 3.5: Ball touches the solution plane with noises. P is the solution plane (the center line), and the shaded region represents the drift of P controlled by $\mathbf{z}$. Solutions lie in the intersection of the vertical axis and the shaded region, which are not far away from the original signal f.

3.3.2 Weak Restricted Isometry Property

In the definition of the **coherence parameter** (lower bound) $\mu(\Phi)$ (see Section 3.2.1), the uniformly bound may not be deterministic (i.e. $\mu(\Phi) = \infty$), so we raise the definition and denote the smallest value μ as the '**near bound**' which satisfies

$$E[n^{-1}||\phi_i||_2^2 \mathbf{1}_{A^c}] \leq \frac{1}{20} n^{-\frac{3}{2}},$$
$$P(A^c) \leq (mn)^{-1}$$

in which A is the event '$\mu(\Phi) = \mu$'. Notice that even $\mu(\Phi)$ exists the definition of μ is still valid, so with this definition we can introduce the *weak restricted isometry property* [6]:

Theorem 3.3.2. (Weak RIP [6]) *Let T an index set as before with $|T| = k$. Then given $\delta > 0$ and $k' \in \mathbb{N}$, if*

$$m \geq C_\delta \beta \mu(\Phi) \max(k \log(k\mu), k' \log n \log^2 k' \log(k' \mu \log n)) \tag{3.38}$$

then with probability at least $1 - 5e^{-\beta}$, for any index set $|T'| < k'$ and signal f with support $T \cup T'$ the statement below is true:

$$(1 - \delta)||f||_2^2 \leq ||\Phi f||_2^2 \leq (1 + \delta)||f||_2^2. \tag{3.39}$$

3.3.3 LASSO Methods Guaranteed by the Weak RIP

In Section 3.2.4 we introduced the primal-dual methods to solve (P_1). The problem (P_1') is slightly different from (P_1) since the noisy observation $\mathbf{y}$ is sampled from the signal f, so we consider the LASSO method (see Section 3.1.3). We raise the example of signals polluted by Gaussian noises then discuss the recoveries using LASSO methods, which are guaranteed by the weak RIP. Our goal is to recover f which satisfies

$$\mathbf{y} = \Phi f + \sigma \mathbf{z}, \tag{3.40}$$

where $\mathbf{z}$ is a vector containing standard Gaussian entries (i.e. $\mathbf{z} \sim \mathcal{N}(0, I)$) and σ indicates the 'amplitude' of the noises. Then the l_2-minimization problem with l_1-regularizer goes like this:

$$f^* = \arg\min_f \frac{1}{2}||\Phi f - \mathbf{y}||_2^2 + \lambda\sigma||f||_1, \tag{3.41}$$

where λ implicates the ratio of the l_1-regularizer effects. In practise we restrict the sensing matrix to be 'nearly normalized' by dividing the observation $\mathbf{y}$ and parameters Φ and *sigma* with $\sqrt{m}$, then solve (3.41) with these updated factors. Now we introduce the stable recovery theorem of the LASSO algorithm [6].

Theorem 3.3.3. *Given $\beta > 0$, with probability at least $1 - \frac{6}{n} - 6e^{-\beta}$, the solution f^* of (3.41) with $\lambda = 10\sqrt{\log n}$ satisfies*

$$\begin{aligned} ||f^* - f||_2 &\leq \min_{1\leq k\leq \bar{k}} C(1+\alpha)[\frac{||f - f_k||_1}{\sqrt{k}} + \sigma\sqrt{\frac{k\log n}{m}}], \\ ||f^* - f||_1 &\leq \min_{1\leq k\leq \bar{k}} C(1+\alpha)[||f - f_k||_1 + k\sigma\sqrt{\frac{\log n}{m}}], \end{aligned} \tag{3.42}$$

where m satisfying **Theorem 3.2.8** *with respected to all $\bar{k}$-sparse vectors. Here $\alpha = \sqrt{\frac{(1+\beta)k\mu\log n\log m\log^2 k}{m}}$ and $\bar{k}$ indicates the upper bound on allowable sparsity levels k that still lead to stable recovery.*

Proofs of the weak RIP and **Theorem 3.3.3** are also available in [6] so we skip them and instead focus on the algorithms solving LASSO analysis, e.g. an online learning algorithm utilizing the homotopy [16]. Here we give an abbreviated description of it.

The key feature of the online algorithm is to compute the support of the optimal solution $f^*(t) \in \mathbb{R}^n$. To explain this we denote $\Phi^t = \Phi(1, 2, ..., t; \cdot)$ which contains the

1st to the tth rows of Φ, and with similar definition we have the notation $\mathbf{y}^t$. Also we notice that since the number of samples sensed varies, we need to adjust $\sigma_t = \sigma/\sqrt{t}$ in each iteration. Now we rewrite the problem (3.41) as

$$f^*(t) = \arg\min_f \frac{1}{2}||\Phi^t f - \mathbf{y}^t||_2^2 + \lambda\sigma_t||f||_1. \tag{3.43}$$

It can be solved by the gradient methods. Since we have the optimal solution $f^*(t)$ satisfying

$$(\Phi^t)^T(\Phi^t f^*(t) - \mathbf{y}^t) + \lambda\sigma_t \mathrm{sgn}(\mathrm{f}^*(\mathrm{t})) = 0,$$

we rewrite $[f^*(t)]^T = ([f^*_{T_t}(t)]^T, \mathbf{0}^T)^T$ and compute the solution of (3.43) in the tth iteration by

$$f^*_{T_t}(t) = [(\Phi^t_{T_t})^T\Phi^t_{T_t}]^{-1}[(\Phi^t_{T_t})^T\mathbf{y}^t - \lambda\sigma_t \mathrm{sgn}(f^*_{T_t}(t))]. \tag{3.44}$$

One may ask that how to compute $sgn(f^*_{T_t}(t))$ in (3.44) without revealing of $f^*_{T_t}(t)$. In fact, online learning provides knowledges about the optimal solution since $f^*(t-1)$ and $f^*(t)$ are co-related. We define the function

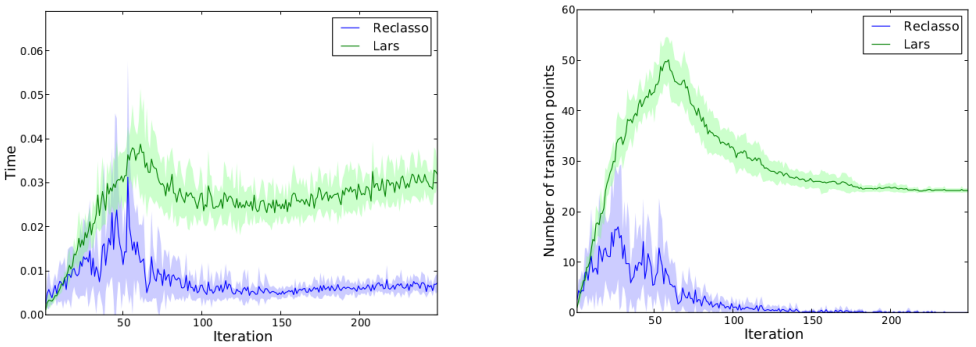
$$f^*(r,\mu) = \arg\min_g \frac{1}{2}|| \begin{pmatrix} \Phi^t \\ r(\Phi(t+1,\cdot))^T \end{pmatrix} g - \begin{pmatrix} \mathbf{y}^t \\ r\mathbf{y}_{t+1} \end{pmatrix} ||_2^2 + \lambda\mu||g||_1 \tag{3.45}$$

then it is straightforward that $f^*(t) = f^*(0,\sigma_t)$ and $f^*(t+1) = f^*(1,\sigma_{t+1})$. The function $f^*(r,\mu)$ behaves smoothly in the r-domain ($[0,1]$), and surprisingly we can define the concept of *transition point* set $\{r'\}$ such that for any r'_1 and the next transition point r'_2, they satisfy that $\forall r \in [r'_1, r'_2)$, $f^*(r,\mu)$ has support and sign coincide with those of $f^*(r'_1,\mu)$ but different from those of $f^*(r'_2,\mu)$. With this observation we have the Algorithm 3 computing $f^*(t+1)$ from $f^*(t)$, with complexity $O(km^2)$.

Details of the update of T and *sign* can be find in the original paper of the algorithm [16] (Reclasso), also the computation of *temp* in it is nearly the same as (3.44) of step $t+1$ except the substitution of σ_{t+1} with some constant μ, which is determined by t. We test the algorithm by adjusting its source code and show the results in Figure 3.6, with comparison of the *least angle regression* (LARS), another homotopy algorithm.

Algorithm 3 A Homotopy Online Algorithm of LASSO Optimization

initialize Φ, k, $\mathbf{y}$
$t \leftarrow 1$, compute $f^*(t)$ and support T // $T := T_t$
while $t \leq m$ **do**
  $temp \leftarrow f^*(0, \sigma_{t+1})$ and update T
  $sign \leftarrow \text{sgn}(f^*(0, \sigma_{t+1}))_T$
  find the first transition point $r' > 0$
  while $r' \leq 1$ **do**
    $temp \leftarrow f^*(r', \sigma_{t+1})$
    update T and $sign \leftarrow \text{sgn}(temp)_T$
    $r' \leftarrow$ the next transition point
  end while // now we get $T_{t+1} = T$ and $sgn(f^*(t+1)) = sign$
  $f_T^*(t+1) \leftarrow [(\Phi_T^{t+1})^T \Phi_T^{t+1}]^{-1}[(\Phi_T^{t+1})^T \mathbf{y}^{t+1} - \lambda \sigma_{t+1} sign]$ // $f^*(1, \sigma_{t+1})$
  $t \leftarrow t + 1$
end while
return $f^*(t)$



(a) running time of each $\mathbf{y}_t$ (b) #transition points of each iteration

Figure 3.6: Comparing the Reclasso algorithm with LARS. By inputting 250 measurements $\mathbf{y}_t$ ($t \leq 250$), though the reconstruction of both algorithms are the same, the Reclasso returned the recovery much faster than the LARS. This is resulted from the less transition points in each iteration.

3.4 Applications

3.4.1 Single-Pixel Camera

The most promising applications of compressed sensing is the *single-pixel camera* (Compressive Imaging (CI) camera) invented by researchers at Rice University [27]. It incorporated a mirror array controlled by 'pseudorandom' measurement bases generated by some algorithm with a random seed, as well as a single or multiple photodiode optical sensor. The optical processor detects incoherent image measurements $\mathbf{y}$, then the computing device reconstructs the original image f as the compressed sensing theory presents. Figure 3.7 shows the structure of the single-pixel camera and Figure 3.8 presents the result of a simple experiment performed by the researchers.

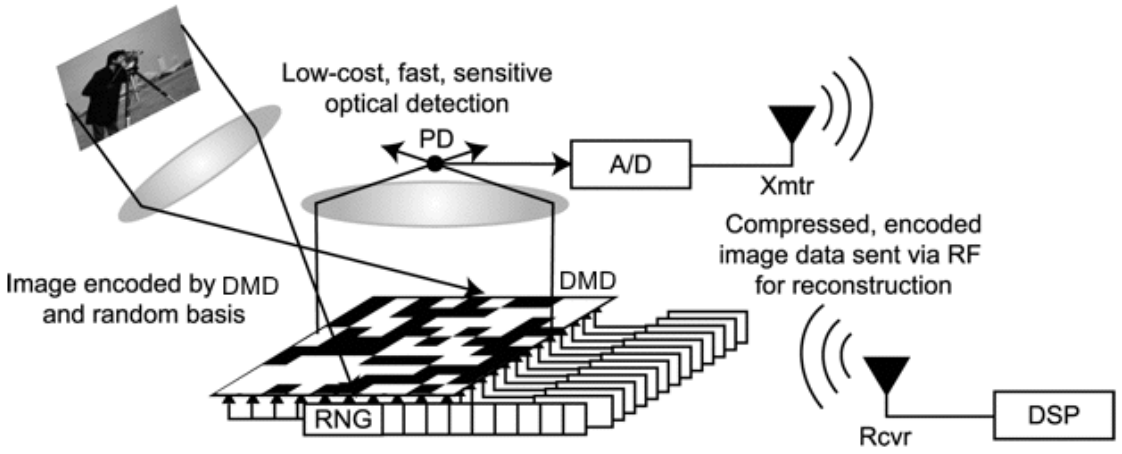


Figure 3.7: Compressive Imaging camera block diagram [27]. Image is reflected off a digital micro-mirror device (DMD) array controlled by the pseudorandom pattern Φ generated by the random number generators (RNG). These pattern produce voltages at the single photodiode (PD) that corresponds to $\mathbf{y}$. After transmission we can reconstruct a sparse approximation to the desired image from received $\mathbf{y}$ with algorithms described before.

3.4.2 Improving MRI detection of Pathological Regions

Several medical imaging researchers have proposed approaches to boost the speed of MRI. However, they can hardly break the major obstacle of classical methods that the sampling rate obeys the restriction of the Nyquist-Shannon Sampling Theorem, thus with at least the Nyquist rate sampling still costs a relatively long time. Fortunately, recent compressed sensing techniques take advantage of the sparsity in the

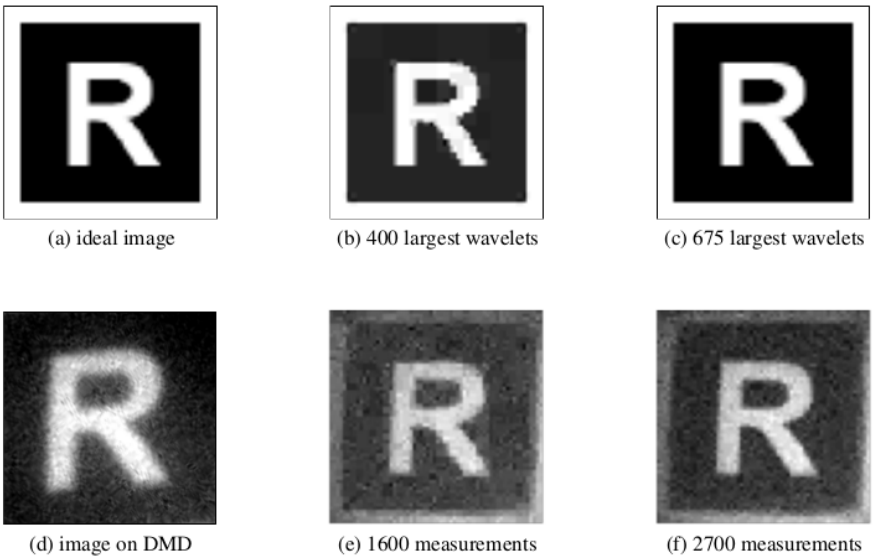


Figure 3.8: Compressed sensing v.s. wavelet decomposition [27]. Random matrices are generated for measurements of the DMD image (d), then the sparse reconstructions are obtained (see (e) and (f)). It is computational expensive of the standard method even though it only picks a small set of the largest wavelets to compute high-quality recoveries ((b) and (c)).

MR images, in which the sampling rate is much lower.

In [19] a compressed sensing MRI is introduced (CS-MRI). Designing a CS scheme for MRI requests for finding an easily-sampled subset of the frequency domain that is incoherent and reveals the sparsity. Though CS theory advocates for a completely random subset of k-space (very low coherence), sampling that subset is generally impractical with observation of the hardware and physiological constraints. But fortunately, most energy in MR imagery is concentrated close to the center of the k-space with very high density. Hence this new technique samples raw signals randomly but concentrated near that center, which is fast and incoherent. The reconstruction applies l_1-analysis methods which minimize the l_1-norm of the sparse representation over some dictionary (see Chapter 4), and gains nearly as precise as the recovery of Nyquist sampling. It clearly captures the dominating informations, and improves the quality of MR images by filtering out secondary noises (Figure 3.9).

CS-MRI is still in its infancy with a lot of crucial problems unsolved. The optimization of sampling trajectories, sparse and incoherent transforms of that trajectories as well as fast and accurate reconstructions still call for further studies. Signal processing and medical study communities have a major opportunity to develop theoretical and practical techniques improving or accompanying the CS-MRI, which shows

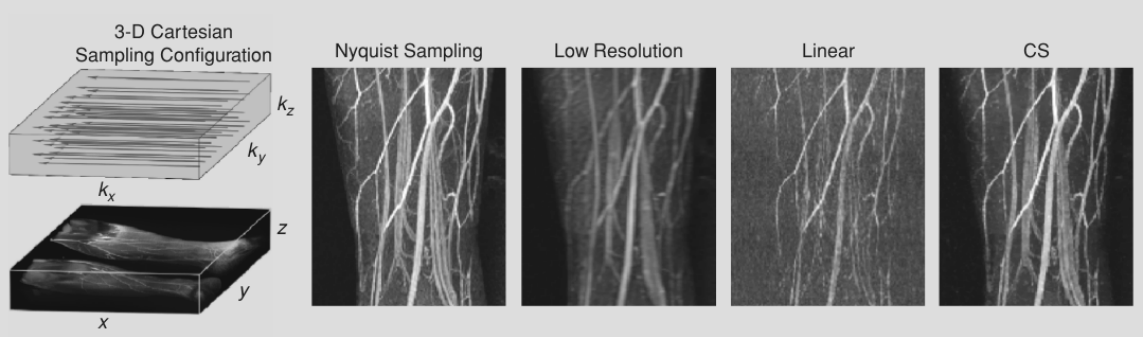


Figure 3.9: Applying CS techniques to MRI. CS methods are applied to a 3-D Cartesian contrast-enhanced angiography, the most common scheme in clinical practise. The classical approach acquires equispaced parallel lines in the k-space, while CS takes a pseudo-random subset (here approx. 10%) of those lines. Even that CS can recover most blood vessel information revealed by Nyquist sampling, and achieve better noise reduction and higher resolution.

a promising future of medical diagnoses in clinics.

Chapter 4

Learning from Sparsity

Machine Learning aims at processing datasets for recognition, classification and generalization tasks with proper algorithms. It focuses on known properties learned from the training data then utilizes them for prediction. Usually the training data outnumbers the known features (both in number of samples and dimensions), hence obtaining a low-dimensional representation of a dataset becomes an important component of the machine learning process. In this way machine learning and compressed sensing is co-related, since compressed sensing algorithms 'predict' the original signals from relatively few 'known features' – measurement $\mathbf{y}$.

Improvements of feature extraction are among the main researches in machine learning, and so does the measurement matrix construction with respect to compressed sensing. Also researchers discuss other topics such as local/global feature detection and faster machine learning algorithms for reconstructions. Especially, here we consider the learning problems of high-dimensional dataset with observation of sparsity (or latent sparsity, see Section 4.1.2), which is common in signal processing problems.

4.1 Best Measurement Matrices for Recovery

Exponentially growth of dataset dimensions calls for better algorithms to process high-dimensional data, in which the majority of them include projections to low-dimensional spaces. The goal of related research is to construct the projector (i.e. the sensing matrix) Φ that can preserve local and global features of the original dataset then guarantee a better reconstruction. Approaches includes adaptive projections and random projections, and in compressed sensing the latter method is often used

that for example with Gaussian entries

$$\Phi_{ij} \sim \mathcal{N}(0, \frac{1}{n}). \tag{4.1}$$

4.1.1 Why Random Projections Work?

The measurement matrix Φ in compressed sensing is equivalent to the projector in dimension reduction, and surprisingly researchers have figured out that using random matrices as the projector works well with some restrictions of projected space dimension m, which is briefly explained in Section 3.2.3. We describe the *Johnson-Lindenstrauss* Lemma [18] that helps understand the claim in general circumstance.

Lemma 4.1.1. (Johnson-Lindenstrauss) *Assume $\epsilon \in (0, 1)$, then for every set Q, if $m \gtrsim O(\ln |Q|/\epsilon^2)$, there exists a Lipschitz mapping $\mathcal{F} : \mathbb{R}^n \to \mathbb{R}^m$ such that for any $u, v \in Q$,*

$$(1 - \epsilon)||u - v||_2^2 \leq ||\mathcal{F}(u) - \mathcal{F}(v)||_2^2 \leq (1 + \epsilon)||u - v||_2^2. \tag{4.2}$$

In [1] it proved that the projector $\mathcal{F}$ can be the $m \times n$ random matrix Φ. Summary of the proof includes establishing

$$\begin{aligned} E(||\Phi x||_2^2) &= ||x||_2^2 \\ P(||\Phi x||_2^2 - ||x||_2^2 \geq \epsilon||x||_2^2) &\leq 2e^{-mC_\epsilon} \end{aligned} \tag{4.3}$$

then applying these two results to complete the proof by assuming $||u - v||_2 \leq 1$ without loss of generality. The second equation of (4.3) is called the **concentration inequality**.

With the random projector Φ satisfying **Lemma 4.1.1** we can illustrate the effectiveness of random measurement matrix in compressed sensing by introducing the theorem below [3].

Lemma 4.1.2. *Let Φ be a $m \times n$ matrix generated by some arbitrary distribution that satisfies the* **concentration inequality** (4.3)*. Then for all k-sparse signal f and $\delta \in (0, 1)$ we have*

$$(1 - \delta)||f||_2 \leq ||\Phi f||_2 \leq (1 + \delta)||f||_2 \tag{4.4}$$

with overwhelm probability at least $1 - 2(12/\delta)^k e^{-mC_{\delta/2}}$.

Proof. Without loss of generality we can assume $||f||_2 = 1$ since Φ is a linear projector. Then for any $\delta < 1$ we can construct the δ-net Q of $\{f | ||f||_2 = 1, ||f||_0 = k\}$ in which

$|Q| \leq (12/\delta)^k$. By assigning ϵ in Lemma 4.1.1 with $\delta/2$ and applying the union bound of difference in Q and (4.3) we have the RIP

$$(1-\delta/2)||f_Q||_2 \leq ||\Phi f_Q||_2^2 \leq (1+\delta/2)||f_Q||_2, \forall f_Q \in Q$$

with probability $1 - 2(12/\delta)^k e^{-mC_{\delta/2}}$.

Now we assume a lower bound A as the smallest value satisfying

$$||\Phi f||_2 \leq (1+A)||f||_2$$

for any *k-sparse* signal f, and the goal is to prove $A \leq \delta$. From the property of the ϵ-net in functional analysis and $||f_Q|| = 1$ we know

$$||\Phi f||_2 \leq ||\Phi f_Q||_2 + ||\Phi(f - f_Q)||_2 \leq (1+\delta/2) + (1+A)\delta/4.$$

Since A is the smallest value (the lower bound) we have $(1+A) \leq (1+\delta/2)+(1+A)\delta/4$, then it is sufficient to get $A \leq \frac{3\delta}{4(1-\delta/4)} \leq \delta$. Also since $A \leq \delta$ we have

$$||\Phi f||_2 \geq ||\Phi f_Q||_2 - ||\Phi(f - f_Q)||_2 \geq (1-\delta/2) - (1+\delta)\delta/4 \geq 1-\delta$$

and complete the proof. □

Theorem 4.1.3. *If given n, m, $0 < \delta < 1$ and the $m \times n$ random matrix Φ satisfying* (4.3)*, then there exists some constants c_δ and c'_δ such that the* RIP *holds for Φ when $k \leq c_\delta m/\log(n/k)$ with probability $1 - 2e^{mc'_\delta}$.*

Proof. With **Lemma 4.1.2** holds the proof is intuitive by noticing that there are C_n^k subspaces of $k-sparse$ vectors in f and $C_n^k \leq (en/k)^k$. Given a positive value c_δ and $k \leq c_\delta m/\log(n/k)$, the overwhelm probability of the conclusion validation is

$$\begin{aligned} probability &\geq 1 - 2(en/k)^k(12/\delta)^k e^{-mC_{\delta/2}} \\ &= 1 - 2exp[-mC_{\delta/2} + k(\log(en/k) + \log(12/\delta))] \\ &\geq 1 - 2exp[-mC_{\delta/2} + c_\delta m(1 + (1 + \frac{\log(12/\delta))}{\log(n/k)})], \end{aligned}$$

and by assigning $c'_\delta \leq C_{\delta/2} - c_\delta(1 + \frac{1+\log(12/\delta)}{\log(n/k)})$ we complete the proof. □

The proofs of **Lemma 4.1.2** and **Theorem 4.1.3** clearly indicate that the RIP and the *Johnson-Lindenstrauss* Lemma is equivalent: the RIP is a straightforward consequence of the *Johnson-Lindenstrauss* Lemma, in which any random matrix Φ

obeying (4.3) will also satisfies the RIP with very high probability. The *Johnson-Lindenstrauss* Lemma then can be used as a simple verification of a random matrix satisfying the RIP and benefiting the measurements. Furthermore, random matrix satisfying the **concentration inequality** tends to obey the RIP with overwhelm probability, and **Lemma 4.1.2** shows the probability of perfect recovery with respect to the *restricted isometry constant* δ. Combined with **Corollary 3.2.4**, if $\delta \leq \sqrt{2}-1$, then all k satisfying

$$k \leq \frac{1}{2} c_\delta m / \log(n/2k)$$

will return perfect recoveries with high probability. Also if we have some prior knowledge of k, we can figure out the approximate number of samples we need:

$$m \gtrsim kC(\delta)\log(n/k)$$

where $C(\delta)$ is a constant controlled by δ, and it indicates the probability of stable (or even perfect) recovery. Figure 4.1 provides an example of 1-D signal sensing.

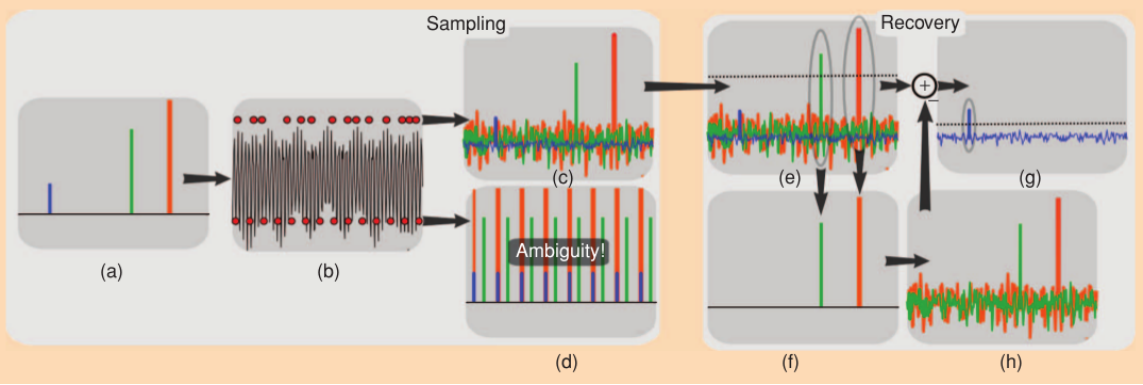


Figure 4.1: Pseudo-random sensing of 1-D signals [19]. A signal (a) is transformed into some domain (b) indicating sparse structures. If sensing from this domain with equispaced undersampling (the down red point line in (b)) it will return alias reconstruction (d). On the other hand, pseudo-random sampling (the top red point line in (b)) will return recovery with acceptable noises (c). By picking out signals with overwhelm amplitudes ((e) and (f)) then filtering out their noises (h), we can reach the separation of the sparse signals ((f) and (g)) as we want.

4.1.2 Dictionary for Latent Sparsity

We often pay attention to find the structures implicating the sparsity of datasets, which is better for representation, storage and processing. In general, the original signal, f, is not sparse at a glance. However, there may exist a sparse representation

over the dictionary Ψ such that the coefficient vector $\mathbf{x}$ is *k-sparse*:

$$f = \Psi\mathbf{x}. \tag{4.5}$$

The idea of using dictionaries is popular in recent machine learning researches, which advocates for better representations of the original data. Also the $n \times d$ matrix Ψ is often overcomplete ($d > n$). Since the best orthogonal base matrix indicating data sparsity can hardly be found in some practical tasks, working with overcomplete dictionary provides flexibility and convenience to construct relatively 'sparse' representations. Section 4.3 discusses the construction of such dictionaries and here we assume Ψ is given and fixed. Then the measurement $\mathbf{y}$ given by (3.2) is revised as

$$\mathbf{y} = \Phi\Psi\mathbf{x}, \tag{4.6}$$

where $\Phi\Psi$ has d columns and $\mathbf{x} \in \mathbb{R}^d$.

To apply compressed sensing algorithms we want the matrix $\Phi\Psi \in \mathbb{R}^{m\times d}$ to satisfy the RIP. Since the dictionary is pre-designed the structure of Φ determines whether $\Phi\Psi$ satisfies the RIP or not. **Lemma 4.1.2** and **Theorem 4.1.3** also provides implications of such structure. With the assumption that Ψ is orthogonal and preserves the l_2-norm of $\mathbf{x}$ (i.e. $||f||_2 = ||\mathbf{x}||_2$), we can construct the δ-net of f (see the proof of Lemma 4.1.2) by the points in $span(\psi_{i_1}, \psi_{i_2}, ..., \psi_{i_k})$ in which ψ_{i_j} is a randomly picked column vector of Ψ. Then the validation of the proof still holds and by revising the inequality we have

$$(1-\delta)||\mathbf{x}||_2 \leq ||\Phi\Psi\mathbf{x}||_2 \leq (1+\delta)||\mathbf{x}||_2 \tag{4.7}$$

which is what we want.

Another change of CS methods with respected to $\mathbf{x}$ is the number of samples we need for reconstruction. In **Theorem 3.2.8** we have the evaluation of sampling numbers m. Here we also have a similar conclusion with respect to the representation $\mathbf{x}$ that the number of samples needed for $\mathbf{x}$'s recovery is

$$m \geq C(1+\beta)\mu(\Phi\Psi)k\log d \tag{4.8}$$

if $\mathbf{x}$ is *k-sparse*. Here the coherence $\mu(\Phi\Psi)$ is computed by searching vectors in some basis set $\mathcal{B}(\Phi\Psi)$ in which $\Phi\Psi$ is generated from, and with the knowledge of Ψ we can suppose such a basis population in advance. Also We do not propose a lower bound of m but just claim the warranty of recovery with such enough samples. Noticing the effectiveness of random projectors (see Section 4.1.1), without loss of generality

we assume the measurement matrix Φ has i.i.d. Gaussian entries. We specify the preciseness of recovery with the theorem as follows:

Theorem 4.1.4. *If Φ is a Gaussian matrix with $m \gtrsim k\log(d/k)$, then the solution f^* of (P_1) yields*

$$||f^* - f|| \leq C\frac{||\Psi^+ f^* - (\Psi^+ f^*)_k||_1}{\sqrt{k}} \tag{4.9}$$

with constant C. $(\Psi^+ f^*)_k$ represents the vector of zeros except the k-largest elements. If $\boldsymbol{x} = \Psi^+ f$ is itself k-sparse *the recovery is exact.*

Details will be discussed by introducing similar theorems with respect to the noisy recovery (we can view the noiseless situation as a special case of it in which $||\mathbf{z}||_2 = 0$). In general Φ is driven by distributions other than Gaussian or even not randomly generated. Section 4.1.3 describes a detailed version of the RIP adaptive to Ψ, where this restriction of the measurement matrix also guarantees the recovery with less level of error.

Remark. We assume the dictionary Ψ to be orthonomal in order to prove the preservation of the RIP. However, the incoherence of the column vectors is unnecessary even though this will violate the exact recovery of $\mathbf{x}$, since the ultimate goal is to reach the reconstruction of $f = \Psi\mathbf{x}$. Also the claim of the increase of #measurements m above may not hold since we do not care the precise computation of $\mathbf{x}$ seriously.

4.1.3 Measurement Matrix Adaptive to the Dictionary

To illustrate the reconstruction problem in detail we first look into the optimization task with respect to $\mathbf{x}$ and still abbreviated as (P_1) (l_1-analysis):

$$f^* = \arg\min_f ||\Psi^+ f||_1 \ s.t. \ ||\mathbf{y} - \Phi f||_2 \leq \epsilon. \tag{4.10}$$

As previously indicated we do not concern about the exact value of $\mathbf{x} = \Psi^+ f$ and instead we compute the original signal directly. Here ϵ is an upper bound of the noise level $||\mathbf{z}||_2$, if $\mathbf{y} = \Phi f + \mathbf{z}$ is the polluted measurements. The noise level $||\mathbf{z}||_2 = 0$ means that the measurements are noiseless then it is possible to achieve exact recovery.

Then we specify the *restricted isometry property* adaptive to the dictionary Ψ [5]:

Definition 4.1.1. (D-RIP) *Let Σ_k be the union of all subspaces spanned by all subsets*

of k column vectors of Ψ, i.e.

$$\Sigma_k = \bigcup_{|T|=k} H(\Psi_T),\ T \subset \{1, 2, 3, ..., d\}.$$

We say the matrix Φ obeys the restricted isometry property adaptive to Ψ with parameter δ_k if for any $f \in \Sigma_k$ the following statement holds:

$$(1 - \delta_k)||f||_2^2 \leq ||\Phi f||_2^2 \leq (1 + \delta_k)||f||_2^2. \tag{4.11}$$

The D-RIP indicates a natural extension of the RIP by adding descriptions of f, which we want to recover. Since f varies in the 'union space' Σ_k determined by the dictionary, proper choice of Ψ is fundamental for the perfect reconstruction. We will discuss how to set Ψ later and focus on the general case of **Theorem 4.1.4**.

How to select measurement matrix Φ adaptive to Ψ is the main concern of this section. Fortunately we can easily figure out that many random compressed sensing matrices satisfying the D-RIP with very high probability. We describe this observation by proving the Lemma as follows.

Lemma 4.1.5. ([25]) *Assume Φ a random matrix satisfying the RIP with constant δ_k and the concentration inequality*

$$P(||\Phi f||_2^2 - ||f||_2^2 \geq \epsilon||f||_2^2) \leq 2e^{-c\frac{m}{2}\epsilon^2} \tag{4.12}$$

for all $f \in \mathbb{R}^n$ and constants c and $0 < \epsilon < 1/3$. Then for any $T \subset \{1, 2, 3, ..., d\}$ of size k, if given a constant $\delta \in (0, 1)$ and we set $\nu := \delta + \delta_k + \delta_k\delta$, the claim

$$(1 - \nu)||\boldsymbol{x}||_2^2 \leq ||\Phi\Psi_T\boldsymbol{x}||_2^2 \leq (1 - \nu)||\boldsymbol{x}||_2^2 \tag{4.13}$$

holds with overwhelm probability

$$1 - 2(1 + \frac{12}{\delta})^k e^{-\frac{c}{9}\delta^2 m}$$

The proof is similar as that of **Lemma 4.1.2**. We construct an ϵ-net of $\mathbf{x}$, obtain ν satisfying (4.13) for all points in that ϵ-net, then expand this result to $\mathbf{x}$ and complete the proof by comparing the lower bound of ν and $\delta + \delta_k + \delta_k\delta$. Examples of random matrix satisfying D-RIP include those with Gaussian, subgaussian, or Bernoulli entries, which also yield $m \gtrsim k \log(d/k)$.

Theorem 4.1.6. *Let Ψ be an arbitrary tight dictionary such that $\Psi^+\Psi = I_d$ and Φ be the measurement matrix satisfying the restricted isometry property adaptive to Ψ with $\delta_{2k} < 0.08$, then the solution f^* of (P_1) also satisfies*

$$||f^* - f|| \leq C_1\epsilon + C_2\frac{||\Psi^+ f^* - (\Psi^+ f^*)_k||_1}{\sqrt{k}} \tag{4.14}$$

for some constant C_1 and C_2 only depends of δ_{2k}.

The proof is nearly the same as that of **Theorem 3.3.1** (can be found in [5]). It indicates that the recovery with l_1-analysis is accurate if $\mathbf{x} = \Psi^+ f$ is sparse. However, it also provides a warning that if not the case it may not guarantee the perfect recovery. In many practical circumstances we can hardly find a good dictionary such that there exists a sparse representation of f over Ψ. However, it is easier to find a set of dictionaries $\{\Psi_1, \Psi_2, ...\Psi_l\}$ in which they can help decompose the original signal and get the sparse representations, i.e.

$$\begin{aligned} f &= f_1 + f_2 + ... + f_l, \\ \mathbf{x}_i &= \Psi_i^T f_i \ (k_i - sparse), \\ k &= \max_i \ k_i. \end{aligned}$$

Then the l_1-optimization problem (4.10) is revised as

$$(f_1^*, f_2^*, ..., f_l^*) = \arg\min_{(f_1, f_2, ..., f_l)} \sum_i ||\Psi_i^T f_i||_1 \ s.t. \ ||\mathbf{y} - \Phi\sum_i f_i||_2 \leq \epsilon. \tag{4.15}$$

Examples includes the mixed data of wavelets, curvelets and other signals.

Though in the **Remark** of Section 4.1.2 we mentioned that it is no need to recover the sparse representation, the *Basis Pursuit* method modelling $\mathbf{x}$ directly also works for the reconstruction of $f = \Psi\mathbf{x}$, which is called the l_1-synthesis:

$$\mathbf{x}^* = \arg\min_{\mathbf{x}} ||\mathbf{x}||_1 \ s.t. \ ||\mathbf{y} - \Phi\Psi\mathbf{x}||_2 \leq \epsilon. \tag{4.16}$$

The heuristic behind the *Basis Pursuit* here is totally different from that of (4.10), which models f instead. When enlarging the size of Ψ, solving (4.10) takes more time since the searching space Σ_k grows exponentially. However, if that size is relatively small, the risk of assigning a wrong non-zero coefficient to $\mathbf{x}$ also increases because the significance of each elements grows enormously [15].

4.2 Learning for Sparse Coding

Machine learning people study probabilistic methods for prediction tasks since they believe that the output prediction function is equivalent some probability. While few of them take advantages from the frequentist probability, a lot of researchers focus on the *Bayesian statistics* or inferences. We give a review of the fundamental *Bayes' theorem* for introduction.

Theorem 4.2.1. (Bayes' Theorem) *For events A and B, the conditional probability $P(A|B)$ is given as follows:*

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)}. \tag{4.17}$$

Here comes a brief interpretation from Bayesian's view. The event A is called the *proposition*, e.g. the happening of pneumonia. In clinical studies, one who is sick with pneumonia often suffers from fever, which can be understood as an *evidence* B. The *prior* $P(A)$ and *likelihood* $P(B|A)$ are easy to compute by doing statistics. Then we may be interested in the probability that if the fever is caused by pneumonia, which, in the mathematical expression, is the *posterior* $P(A|B)$. Thanks to the Bayes' theorem we can easily figure out this probability, and in practice machine learning people apply the condition as follows more often:

$$P(A|B) \propto P(A)P(B|A). \tag{4.18}$$

4.2.1 Maximizing the Posteriori

We can consider the recovery from the probabilistic perspective that the solution maximizes the posterior. Take the noisy recovery as example: with the sensing matrix Φ fixed, we can observe the samples $\mathbf{y} = \Phi f + \mathbf{z}$ as previously indicated. Hence we can consider the original signal f as the proposition and have the prior $P(f)$. Straightforwardly $\mathbf{y}$ is the evidence, and with the linear projection Φ as well as the noise $\mathbf{z}$ obeying some distribution, the likelihood $P(\mathbf{y}|f)$ can be evaluated. Having these assumptions, we can gain the recovery with *maximum a posteriori* techniques (MAP, a standard approach in machine learning), and we often take the log form

$$\begin{aligned} f^* &= \arg\max_f \log P(f|\mathbf{y};\Phi) \\ &= \arg\max_f (\log P(\mathbf{y}|f;\Phi) + \log P(f)). \end{aligned} \tag{4.19}$$

Now we claim that the optimization problem (4.19) is equivalent to the initial problem (3.3) and provide a brief proof with assumption of independent Gaussian noises, i.e. $\mathbf{z} \sim \mathcal{N}(0, \sigma^2)$ (if consider the noisy version). Since the measurement matrix Φ is known, with the assumption of f we can easily figure out that

$$\begin{aligned} \log P(\mathbf{y}|f;\Phi) &= \log P(\mathbf{y} - \Phi f|f;\Phi) \\ &= \log P(\mathbf{z}|f;\Phi) \\ &= \log P(\mathbf{z}) \\ &\propto -||\mathbf{z}||_2^2 \\ &= -||\mathbf{y} - \Phi f||_2^2, \end{aligned} \tag{4.20}$$

which is exactly the error level of reconstruction. On the other hand, we can define a function $d_p(f)$ which tend to measure the sparsity of f, with p indicating the l_p-norm. Without loss of generality the prior of f is assumed as

$$P(f) = \frac{1}{Z} e^{-\lambda_p d_p(f)} \tag{4.21}$$

where Z is some normalization factor ensuring $P(\cdot)$ a valid probability function. A common choice of $d_p(f)$ is just the l_p-norm of f, i.e. $d_p(f) = ||f||_p$ [23]. Then (4.19) can be rewrote as

$$\begin{aligned} f^* &= \arg\max_f (\log P(\mathbf{y}|f;\Phi) + \log P(f)) \\ &= \arg\min_f ||\mathbf{y} - \Phi f||_2^2 + \lambda ||f||_p \end{aligned} \tag{4.22}$$

with the constant $\lambda \propto \lambda_p$. This is equivalent to the two optimization problems as follows, with some constant ϵ:

$$f^* = \arg\min_f ||f||_p \ s.t. \ ||\mathbf{y} - \Phi f||_2 \leq \epsilon, \tag{4.23}$$

$$f^* = \arg\min_f ||\mathbf{y} - \Phi f||_2 \ s.t. \ ||f||_p \leq \epsilon. \tag{4.24}$$

Intuitively if $\epsilon \to 0$ (4.23) is exactly the l_p-norm optimization problem (3.3). Also consider the factor p, (4.23) is problem (P_0') (or (P_0) with $\epsilon \to 0$) if $p = 0$, as well as problem (P_1') and (P_1) with $p = 1$ respectively. Furthermore, (4.24) is the standard LASSO task if $p = 1$. These minimizations reveal sparsity if Φ obeys the RIP, hence the understanding of sparse recovery from MAP perspective helps connect compressed sensing theories with machine learning methods.

4.2.2 Maximizing the Likelihood

Choosing a good dictionary for data representation can also be solved with probabilistic methods. Given the raw signals f, sparse representation problem requests for a better representation $\mathbf{x}$ of it to reveal the latent sparsity. Unlike the MAP approach described before we want to solve this problem by *maximum likelihood* techniques. Assume we have the sparse vector $\mathbf{x}$, the likelihood of f's appearance is $P(f|\mathbf{x}, \Psi)$. In practise $\mathbf{x}$ is what we want to compute then the likelihood with respect to the dictionary Ψ is

$$\begin{aligned} P(f|\Psi) &= \int_{\mathbf{x}\in\mathbb{R}^d} P(f, \mathbf{x}|\Psi) \\ &= \int_{\mathbf{x}\in\mathbb{R}^d} P(f|\mathbf{x}, \Psi)P(\mathbf{x}). \end{aligned} \tag{4.25}$$

Also consider the noisy version of (4.5)

$$f = \Psi\mathbf{x} + \mathbf{v} \tag{4.26}$$

where $\mathbf{v}$ is also some independent Gaussian noise that $\mathbf{v} \sim \mathcal{N}(0, \sigma'^2)$. Similarly by assuming Ψ is given we have

$$P(f|\mathbf{x}, \Psi) = \frac{1}{Z} e^{-\frac{1}{2\sigma'^2}||f-\Psi\mathbf{x}||_2^2}, \tag{4.27}$$

which clearly shows the relationship between the likelihood and the error level of sparse recovery. In order to indicate the sparsity we still obtain the same assumption of the prior $P(\mathbf{x}) \propto e^{-\lambda_1||\mathbf{x}||_1}$ as $P(f)$ in Section 4.2.1, then (4.25) is computed by

$$\begin{aligned} P(f|\Psi) &= \int_{\mathbf{x}\in\mathbb{R}^d} P(f|\mathbf{x}, \Psi)P(\mathbf{x}) \\ &= \frac{1}{Z} \int_{\mathbf{x}\in\mathbb{R}^d} e^{-\frac{1}{2\sigma'^2}||f-\Psi\mathbf{x}||_2^2 - \lambda_1||\mathbf{x}||_1}, \end{aligned} \tag{4.28}$$

again it contains structures indicating the LASSO optimization. Inspired by the standard LASSO problem that l_1-norm is constrained by some upper bound ϵ, we can then restrict the domain of $\mathbf{x}$ to a d-dimensional ball of l_1-norm $B_1(0, \epsilon) = \{\mathbf{x}|||\mathbf{x}||_1 \leq \epsilon\}$. With these assumptions we introduce the construction of dictionary Ψ which

maximize the sum of likelihoods given a set of signals $\{f\}$

$$\begin{aligned}\Psi^* &= \arg\max_{\Psi} \sum_f P(f|\Psi) \\ &= \arg\max_{\Psi} \sum_f \int_{\mathbf{x}\in B_1(0,\epsilon)} e^{-\frac{1}{2\sigma'^2}||f-\Psi\mathbf{x}||_2^2-\lambda_1||\mathbf{x}||_1}.\end{aligned} \tag{4.29}$$

Furthermore, to reduce the computation in [24] the integration of the representation $\mathbf{x}$ in $B_1(0,\epsilon)$ is substituted by the search of some $\mathbf{x} \in B_1(0,\epsilon)$ with respect to each f which maximize $P(f,\mathbf{x}|\Psi)$ in the l_1-ball. We denote the searching result as $\mathbf{x}_f$ then have a much more simple computation as follows, resulting in the largest log-likelihood:

$$\begin{aligned}\Psi^* &= \arg\max_{\Psi} \sum_f e^{-\frac{1}{2\sigma'^2}||f-\Psi\mathbf{x}_f||_2^2-\lambda_1||\mathbf{x}_f||_1} \\ &= \arg\min_{\Psi} \sum_f ||f-\Psi\mathbf{x}_f||_2^2 + \lambda||\mathbf{x}_f||_1\end{aligned} \tag{4.30}$$

with constant λ indicating how 'strict' the l_1-penalty is. We can also transform (4.30) into the standard LASSO problem, which omitted the part of l_1-norm in the target function since we have already introduced the bound of it:

$$\Psi^* = \arg\min_{\Psi} \sum_f ||f-\Psi\mathbf{x}_f||_2^2 \ s.t. ||\mathbf{x}_f||_1 \leq \epsilon. \tag{4.31}$$

A simple approach to solve (4.31) is again an online method. Assume the signal set $\{f\} = \{f_1, f_2, ..., f_N\}$ and we denote $\Psi^*(t)$ the optimal solution in the tth step. Similar to the LASSO Algorithm 3 described before we have the online learning Algorithm 4.

Algorithm 4 Online Learning for Likelihood Maximization

```
initialize Ψ*(1), ε, P(x), t ← 1
while t ≤ N do
    i ← 1
    while i ≤ t do
        x_{f_i} ← arg max_{x∈B_1(0,ε)} P(f_i, x|Ψ*(t))
        i ← i + 1
    end while
    Ψ*(t+1) ← arg min_Ψ Σ_{i=1}^{t} ||f_i − Ψx_{f_i}||_2^2
    t ← t + 1
end while
return Ψ*(t)
```

Also we can use gradient descent methods to compute $\mathbf{x}_{f_i}$ and $\Psi^*(t+1)$ [2]. Assume the prior $P(\mathbf{x})$ is smooth in $B_1(0,\epsilon)$, then $P(f,\mathbf{x}|\Psi^*(t))$ is derivable with respect to $\mathbf{x}$ so we can catch the poles and get $\mathbf{x}_{f_i}$. In addition we claim that the object function (4.30) is smooth, so the update with some learning rate η is straightforward:

$$\Psi^*(t+1) = \Psi^*(t) - \eta \sum_{i=1}^{t} (\Psi^*(t)\mathbf{x}_{f_i} - f_i)\mathbf{x}_{f_i}^T. \tag{4.32}$$

In fact with this type of update Algorithm 4 is exactly the *stochastic gradient descent* methods common in machine learning, which again shows the close relationship between machine learning and sparse representation.

In standard tests we assume that the dictionary is composed by uniform vectors. However, this constraints can even be stricter to rich good sparse coding with also sparse dictionary. In [20] it introduced two kinds of constraints that

$$D = \{\Psi | ||\psi_i||_2^2 + \gamma||\psi_i||_1 \leq 1\} \tag{4.33}$$

$$D = \{\Psi | (1-\gamma)||\psi_i||_2^2 + \gamma||\psi_i||_1 \leq 1\} \tag{4.34}$$

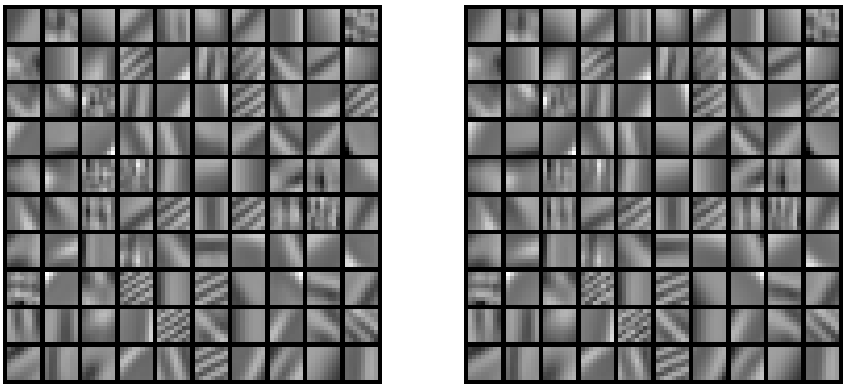
where γ represented the proportion of the l_1-penalty. In Figure 4.2 we compare the test results of Algorithm 4 with/without these dictionary constraints and figure out that they reach nearly the same performance.

4.3 Constructing Dictionaries for Sparse Representations

Recent activities in sparse representation research focus on the construction of the overcomplete dictionary for the sake of achieving better descriptions of sparsity. This dictionary can be set as a pre-specified set of basis functions, or adaptively designed to fit the features and structures of a certain dataset. Choosing the former one (pre-specified dictionary) is appealing because it is simpler to construct if with a lot of prior knowledges. If properly designed, this simple dictionary can benefit the speed up of representation computing as well as the recovery by pseudo-invert techniques. Examples include wavelets, curvelets and sinc functions. However, such dictionaries can only fit a few types of signals, in which we need to figure out the best one for processing given different sets of signal data. Hence in this section we consider the latter approach based on learning algorithms.

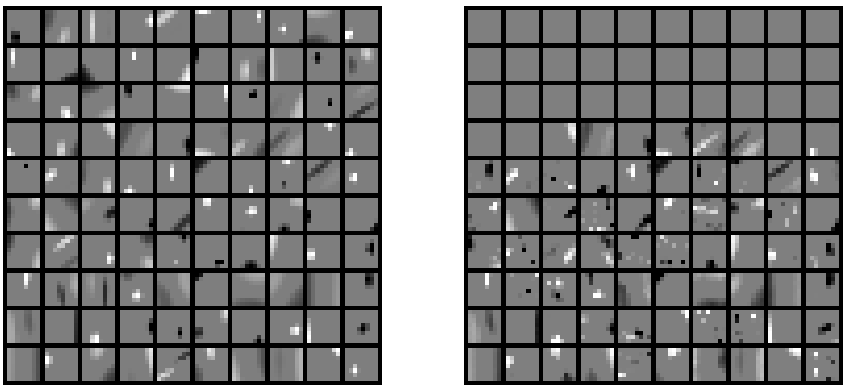
(a) original

(b) standard, target = 0.425630



(c) left-half, target = 0.426019

(d) right-half, target = 0.4257720



(e) constraint (4.33), target = 0.470179

(f) constraint (4.34), target = 0.467948

Figure 4.2: Maximum likelihood learning of the dictionary. We used related functions of the INRIA's 'SPAM-python' package [17] when coding for this test. The target function of (4.31) takes values around 0.42 if without the dictionary constraints, and we notice that using only half of the image achieves (see (c) and (d)) nearly the same results. Learning with those constraints (4.33) and (4.34) achieve a slightly worse performance, but the sparsity of the dictionary are significantly improved.

Researchers have raised approaches to learn the dictionary Ψ that yields sparse representations for the training signals. These approaches save time of looking for suited dictionaries for dataset – the analysis itself reveals specific informations of the dataset – which can also be viewed as a part of learning. Even the computational complexity will be higher, with growing computing capabilities we believe that the adaptive dictionary will outperform the pre-determined one.

4.3.1 K-SVD: Decomposition with Sparsity Restrictions

In [2] another method utilized the *singular vector decomposition* (SVD) techniques to learn the dictionary. We rewrite the (P_0) task to reveal its details. Consider the matrix $\mathbf{Y}$ with N column vectors $\{\mathbf{y}_i\}$ and assume Φ is fixed, we want to obtain the best dictionary such that all the sparse representation are at most k-sparse:

$$\Psi^* = \arg\min_{\Psi,\mathbf{X}} ||\mathbf{Y} - \Phi\Psi\mathbf{X}||_F^2 \; s.t. \; \forall i \; ||\mathbf{x}_i||_0 \leq k, \tag{4.35}$$

where $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, ...\mathbf{x}_N)$ represents some possible sparse coding of $\mathbf{Y}$. From the knowledge of linear algebra we can rewrite the target function of (4.35)

$$\begin{aligned} ||\mathbf{Y} - \Phi\Psi\mathbf{X}||_F^2 &= ||\mathbf{Y} - \Phi\sum_{i=1}^{d}\Psi_{(i)}\mathbf{X}^i||_F^2 \\ &= ||(\mathbf{Y} - \Phi\sum_{i\neq j}\Psi_{(i)}\mathbf{X}^i) - \Phi\Psi_{(j)}\mathbf{X}^j||_F^2 \end{aligned} \tag{4.36}$$

where we use $\mathbf{X}^i$ to denote the ith row vector of $\mathbf{X}$, and $\Psi = \sum_{i=1}^{d}\Psi_{(i)}$. Define $E_j := \mathbf{Y} - \Phi\sum_{i\neq j}\Psi_{(i)}\mathbf{X}^i$, then we have our goal of minimizing $||E_j - \Phi\Psi_{(j)}\mathbf{X}^j||_F^2$, which is easy to achieve by performing SVD decomposition of E_j. However, it cannot yield the k-sparsity of the updated $\mathbf{X}^j$ since it may return some dense vectors. A remedy is to discard all the zero-valued elements of $\mathbf{X}^j$ and gain a shortened vector $\tilde{\mathbf{X}}^j$ with no non-zero entries. By adjusting E_j by eliminating all column vectors except those with indices of non-zero elements in $\mathbf{X}^j$ (denoted as $\tilde{E}_j$) we get an equivalent target function

$$\Psi^* = \arg\min_{\Psi} ||\tilde{E}_j - \Phi\Psi_{(j)}\tilde{\mathbf{X}}^j||_F^2 \tag{4.37}$$

then solving it by SVD returns satisfying decomposition $\tilde{E}_j = U\Sigma V^T$. We define the update $\Psi^*_{(j)} := \Phi^+ U_1$ in which Φ^+ is the pseudo-inverse matrix and $\tilde{\mathbf{X}}^{j*} := \Sigma_{11}V_1$, by applying this method we get a part of optimal dictionary in one iteration (see Algorithm 5).

Algorithm 5 K-Singular Vector Decomposition for Dictionary Learning

initialize Y, Φ, $\Psi^*(1)$, ϵ, k, $t \leftarrow 1$, $error > \epsilon$
while $error > \epsilon$ **do**
  recover $\mathbf{X}$ from $\mathbf{Y}$ over $\Psi^*(t)$ // $\mathbf{x}_i$ is at most k-sparse
  $j \leftarrow 1$
  while $j \leq d$ **do**
    $E_j \leftarrow \mathbf{Y} - \Phi \sum_{i \neq j} \Psi_{(i)} \mathbf{X}^i$
    compute $\tilde{E}_j$, $\tilde{\mathbf{X}}^j$
    perform SVD and gain $\tilde{E}_j = U \Sigma V^T$
    $\Psi^*(t+1)_{(j)} \leftarrow \Phi^+ U_1$, $\tilde{\mathbf{X}}^{j*} \leftarrow \Sigma_{11} V_1$
    $j \leftarrow j + 1$
  end while
  $t \leftarrow t + 1$
  $error \leftarrow ||\tilde{E}_j - \Phi \Psi^*(t)_{(j)} \tilde{\mathbf{X}}_j||_2^2$
end while
return $\Psi^*(t)$

Remark. One may question about the choice of sparse representations since in each iteration the algorithm keeps $\text{sgn}(\mathbf{X})$ unchanged. However, in the reconstruction step we apply some MP or BP algorithms such as those we introduced before, so $\mathbf{X}$ varies according to the change of $\Psi^*(t)$. Further more, with t grows $\text{sgn}(\mathbf{X})$ tends to converge in order to achieve smaller error level.

4.3.2 Can EM algorithms Work?

We may concern about the stability of the methods introduced as above since they acquire only one sample with the largest probability in each step. In general, expectation provides a stable evaluation, and it's reasonable that this indicates the structure of the parameters roundly. Actually we can view the reconstruction problem as the expectation task then compute the optimal dictionary (given a fixed sensing matrix Φ)

$$\begin{aligned} \Psi^* &= \arg\max_{\Psi} E_{\mathbf{y}}[\mathbf{1}_{\mathbf{y}'=\mathbf{y}}|\Psi] \\ &= \arg\min_{\Psi} E_{\mathbf{y}}[||\mathbf{y} - \mathbf{y}'||_2^2|\Psi]. \end{aligned} \tag{4.38}$$

This learning problem can be addressed by the *expectation-maximization* algorithm, a standard approach in machine learning. Abbreviated as the EM algorithm, it provides an iterative method for finding the optimal solution of the MAP or likelihood maximization tasks. Different from taking the hidden variables yielding the maximum probability, it averages the computation over the whole probability space. Generally

speaking, each iteration alternates between performing an expectation (E) step, which computes the expectation of the loss function (or the target function we proposed) using the current estimate for the parameters, and a maximization (M) step, which computes parameters optimizing that function (in a probabilistic view it maximizes the expected likelihood or posterior) gained in the E step.

We briefly introduce the computation in each iteration. Assume in tth epoch the result of last epoch $\Psi^*(t)$ is ready:

- first we compute the average $\bar{\mathbf{y}}_{t+1} = \frac{1}{t+1}\sum_{i=1}^{t+1}\mathbf{y}_i$
- E-step: compute the average sparse representation $\bar{\mathbf{x}}_{t+1} = recovery(\bar{\mathbf{y}}_{t+1}, \Phi, \Psi^*(t))$;
- M-step: compute Ψ^* to minimize $||\bar{\mathbf{y}}_{t+1} - \mathbf{y}'||_2^2$, here $\mathbf{y}' = \Phi\Psi\bar{\mathbf{x}}_{t+1}$ and the optimal Ψ is what we want. Then assign $\Psi^*(t+1) := \Psi^*$.

Here we notice that, in the E-step, since the dictionary is assigned as $\Psi^*(t)$ (and fixed), the computation figures out the expectation of recovery $\bar{\mathbf{x}}_t$ is straight forward (e.g. by pseudo-inverse methods). The idea behind these EM algorithm is to minimize the average reconstruction error, and we utilize the average sparse representation to help estimate it. Moreover, this algorithm is also applicable if we apply the noise $\mathbf{v}$ to the M-step, by stochastically adding a noise vector or taking its expectation as well. In [20] it presented a similar algorithm except that the E-step computed the sparse representation $\mathbf{x}_{t+1}$ of the new observation $\mathbf{y}_{t+1}$ over the dictionary $\Psi^*(t)$ gained in the last M-step. Taking advantage of the previously proved equivalence of the LASSO and the maximum likelihood, it drove the M-step by solving the l_2-minimization with l_1-penalize, which is more tractable and computed by some fast algorithms like LARS.

Chapter 5

Conclusion & Future Research

The purpose of this thesis is to present a review of compressed sensing researches as well as its related learning works. We have proposed the classical sampling theory briefly and pointed out its disabilities, especially facing the increasing demands for fast processing. Compressed sensing techniques aims at overcoming these obstacles by raising a brand new theory of sampling, which was discussed in this thesis with proofs of the main results. Also fast algorithms for signal recovery were introduced with mathematical details, as well as some successful applications in the real world. On the other hand, CS studies also presents opportunities of machine learning researches, and these were discussed in the previous chapters too. We have viewed the recovery as a kind of prediction work, and discussed the optimization problems from the probabilistic perspective. Learning models for the construction of dictionary were also introduced, and we have presented the simulation results of some algorithms in this thesis.

Traditional signal sensing methods take advantage of the *Nyquist-Shannon sampling theorem*, which guarantees the perfect recovery with high-enough sampling rate. However, the request of faster processing severely challenges the speed of sensors, which are relatively slower than the computing advices for further analysis. Due to the limitations of cost and design, sensors can hardly be improved to sample high-rate signals with the Nyquist rate. But there exists another approach: why not try to figure out the smallest number of samples we need that 'with overwhelm probability' these samples can return a perfect recovery? Then it comes to the compressed sensing research, which discusses advanced methods to reconstruct the original signals from a relatively few samples with random sensing matrices and appropriate dictionaries. These ideas are so revolutionary that catch the eyes of mathematicians, statisticians and computer scientists, as well as the other people such as medical researchers since

its successful application of MR imaging systems.

Sparsity is also a main research topic in machine learning, which aims at extracting knowledges from the training data and returning precise predictions. Researchers have discussed the equivalence between the CS recovery and the ML prediction, showing the connections between machine learning and compressed sensing. With this claim, classical algorithms such as *maximum a posteriori* (MAP) or likelihood maximization can be applied to the recovery tasks. Also online learning methods, e.g. the *stochastic gradient descent* algorithm, speed up the computation of the optimal solutions, which as well provide approaches of dictionary constructions.

5.1 Benefiting from Sparsity

Why compressed sensing methods work so well? This is based on the assumption of signal sparsity, which can be discovered directly or over a specific dictionary. It states that signals can be compressed and sensed at the same time, using linear projections to convert the raw signals into a low-dimensional subspace. The *restricted isometry property* then proposes constraints on the sensing matrices (the projectors), which yields little distortion of the geometric structure of the *k-sparse* raw signals. This restriction can be weakened with a little decrease of perfect-recovery probability (even though the RIP is not so strict in practise), and surprisingly we find out that many random matrices such as those with Gaussian entries satisfy both. A brief explanation of it goes like that random sensing matrices tends to get less coherence of the constructed dictionary, then the quality of reconstruction becomes better with the same numbers of samples. The brilliant theoretical results also show that the *Basis Pursuit* (l_1-optimization) is equivalent to the l_0-optimization problems under some restrictions, which is more tractable then the NP-hard minimization (P_0).

In a nutshell, a successful compressed sensing application has 3 requirement:

- Transform sparsity. The signal has a sparse representation in some known domain. Or we can assume the existence of latent sparsity then figure out some appropriate dictionaries to reveal it. The key point of sparsity discovery is the proper choice of dictionaries, however in practise that sparse representation itself is no need to compute.

- Incoherence sensing. The sampling from k-sparse signal f should be 'noise like', i.e. measurement matrix with random entries can achieve better representation of the sparsity in the transform domain. From the mathematical point of view,

small coherence $\mu(\Phi\Psi)$ indicates the incoherence of the sensing matrix to the dictionary, which can result in fewer samples we need for recovery.

- 'Non-linear' recovery. A reconstruction should yield the sparsity of signal representation as well as its consistency simultaneously. Since directly modelling (P_0) is NP-hard (the MP algorithm cannot guarantee the recovery), (P_1) is considered as substitution, which is 'non-linear' at a glance: in some cases it can be addressed by linear programming.

Other important topics in compressed sensing research include the fast algorithms for *Basis Pursuit* and the construction of the dictionaries, calling for the ideas from other perspectives.

5.2 Learning for Better Reconstructions

One can hardly apply random dictionaries to the raw signals in order to reveal the latent sparsity, which invokes machine learning researches. Sparsity itself is a significant research topic in the machine learning world. Since compressed sensing takes advantage of sparsity, it also presents connections between these two fields. More than the construction of dictionaries, online learning methods such as the *stochastic gradient descent* algorithms also benefits the acceleration of computations. In this thesis we have shown the online methods and explained the mathematics behind them, from the perspective of probability.

- We have re-introduced the problem (P_0) and (P_1) with the concept of dictionary. Then the D-RIP was discussed, which showed the constraints of the sensing matrix Φ with dictionary Ψ given. Also we have given another view of the random matrix's effectiveness inspired by the *Johnson-Lindenstrauss* Lemma.

- In machine learning tasks the prediction can be interpreted as computing some value proportional to some likelihood with respect to the given input. Hence one can learn the parameters behind the probability by maximizing that likelihood (generative) or applying the MAP algorithm (discriminative). We have proposed an analogy between recovery and prediction by showing that these two types of learning have the same target functions as that of (P_1).

- We have shown algorithms for dictionary construction. These included the formal matrix decomposition methods with the restriction of k non-zero entries

(K-SVD), as well as the EM algorithm. Both of them are iterative methods, one returns part of the solution then sums up as the result, while another one takes in the observations one by one and updates the dictionary to be temporally optimal.

5.3 Future Researches

Compressed sensing is an emerging field of research with beautiful theories and successful applications. Some open questions still require further researches, which include as follows.

- Can the algorithms proposed keep high performance when dealing with high-dimensional data? Images (or even 3-D images) with higher resolution and other high-dimensional signals present difficulties for lossless compressions, and we do care about if researchers can find out any advanced CS methods to process them.

- Can the distribution which drives the random sensing matrix Φ be learned from the dataset? The **coherence parameter** $\mu(\Phi)$ (or $\mu(\Phi\Psi)$) controls the lower bound of the number of samples m. Since smaller m results in more satisfactory in practical uses, we are wondering if we can figure out the better distribution of Φ with respect to a specific dataset.

- Are there any restrictions equivalent to the RIP, or even weaker than it (except the weak RIP)? The RIP in fact indicates that the linear transformation of the *k-sparse* vector is nearly orthogonal, and we do not know if and to which extent this constraint can be loosed. Besides we are also interested in other explanations of the RIP.

- Are there any other approaches of dictionary learning, except the MAP, EM algorithms and their combination methods? Recently there's a new topic (*deep learning*) of machine learning which advocates for feature learning, relieving the researchers from devoting lot of time on the design of measurements. This new approach works well in computer vision applications as well as the speech processing, which are also related to signal processing researches. Also *Bayesian nonparametrics* views the distributions on an infinite-dimensional space of functions, which may provide a generalized method for CS recovery (since Φ is randomly generated we may prefer not to assume a prior in advance).

We are looking forward to seeing the researchers of compressed sensing, machine learning and related application fields interact with each other, and confident of the success of their collaborations.

References

[1] Achlioptas, D. Database-friendly random projections. *Proceedings of the twentieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems* (2001), 274–281.

[2] Aharon, M., Elad, M., and Bruckstein, A. K -svd : An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on Signal Processing 54* (2006), 4311–4322.

[3] Baraniuk, R., Davenport, M., DeVore, R., and Wakin, M. A simple proof of the restricted isometry property for random matrices. *Constructive Approximation 28* (2008), 253–263.

[4] Candès, E. J. The restricted isometry property and its implications for compressed sensing. *Comptes Rendus Mathematique 346* (2008), 589 – 592.

[5] Candès, E. J., Eldar, Y. C., Needell, D., and Randall, P. Compressed sensing with coherent and redundant dictionaries. *Applied and Computational Harmonic Analysis 31* (2011), 59–73.

[6] Candès, E. J., and Plan, Y. A probabilistic and ripless theory of compressed sensing. *IEEE Transactions on Information Theory 57* (2011), 7235–7254.

[7] Candès, E. J., and Romberg, J. K. l_1-magic: Recovery of sparse signals via convex programming. http://www-stat.stanford.edu/∼candes/l1magic/. Software & Notes.

[8] Candès, E. J., Romberg, J. K., and Tao, T. Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics 59* (2006), 1207–1223.

[9] Candès, E. J., and Tao, T. Decoding by linear programming. *IEEE Transactions on Information Theory 51* (2005), 4203–4215.

[10] Candès, E. J., and Tao, T. Near-optimal signal recovery from random projections: Universal encoding strategies? *IEEE Transactions on Information Theory 52* (2006), 5406–5425.

[11] Chen, S. S., Donoho, D. L., and Saunders, M. A. Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing 20* (1998), 33–61.

[12] Donoho, D. L. Compressed sensing. *IEEE Transactions on Information Theory 52* (2006), 1289–1306.

[13] Donoho, D. L., and Huo, X. Uncertainty principles and ideal atomic decomposition. *IEEE Transactions on Information Theory 47* (1999), 2845–2862.

[14] Elad, M., and Bruckstein, A. M. A generalized uncertainty principle and sparse representation in pairs of bases. *IEEE Transactions on Information Theory 48* (2002), 2558–2567.

[15] Elad, M., Milanfar, P., and Rubinstein, R. Analysis versus synthesis in signal priors. *Inverse Problems 23* (2007), 947–968.

[16] Garrigues, P. J., and El Ghaoui, L. An homotopy algorithm for the lasso with online observations. *Proceedings of the Conference on Neural Information Processing Systems* (2008), 489–496.

[17] INRIA. Sparse modeling software. http://spams-devel.gforge.inria.fr/. Software & Notes.

[18] Johnson, W. B., and Lindenstrauss, J. Extensions of lipschitz mappings into a hilbert space. *Conference in modern analysis and probability* (1984), 189–206.

[19] Lustig, M., Donoho, D. L., Santos, J. M., and Pauly, J. M. Compressed sensing mri. *Signal Processing Magazine, IEEE 25* (2008), 72–82.

[20] Mairal, J., Bach, F., Ponce, J., and Sapiro, G. Online dictionary learning for sparse coding. *Proceedings of the 26th Annual International Conference on Machine Learning* (2009), 689–696.

[21] Majumdar, A., Krishnan, N., Pillai, S., and Velmurugan, R. Extensions to orthogonal matching pursuit for compressed sensing. *National Conference on Communications* (2011), 1–5.

[22] MPEG. moving picture experts group. http://www.mpeg.org/.

[23] Murray, J. F., and Kreutz-Delgado, K. Sparse image coding using learned overcomplete dictionaries. *Machine Learning for Signal Processing, 2004. Proceedings of the 2004 14th IEEE Signal Processing Society Workshop* (2004), 579–588.

[24] Olshausen, B. A., and Field, D. J. Natural image statistics and efficient coding. *Network: Computation in Neural Systems 7* (1996), 333–339.

[25] Rauhut, H., Schnass, K., and Vandergheynst, P. Compressed sensing and redundant dictionaries. *IEEE Transactions on Information Theory 54* (2008), 2210–2219.

[26] Shannon, C. E. Communication in the presence of noise. *Proceedings of the Institute of Radio Engineering 37* (1949), 10–21.

[27] Takhar, D., Laska, J. N., Wakin, M. B., Duarte, M. F., Baron, D., Sarvotham, S., Kelly, K. F., and Baraniuk, R. G. A new compressive imaging camera architecture using optical-domain compression. *Proceedings of Computational Imaging IV at SPIE Electronic Imaging* (2006), 43–52.

[28] Weinhaus, F. Fourier transforms. http://www.imagemagick.org/Usage/fourier/.

[29] Wikipedia. Signal processing. http://en.wikipedia.org/wiki/Signal_processing.

Acknowledgements

It's my pleasure to do innovative researches with talented and energetic people. I felt extremely fortunate with 4-year undergraduate study and research in the department of mathematics, Sun Yat-sen University, with outstanding professors, graduates and undergraduates. I apologize here first for not mentioning the name of everyone, since there are too many of those who I owe thanks to.

First and foremost, I would like to thank Prof. Haizhang Zhang, who has been an excellent supervisor of this thesis. Compressed sensing is a quiet new research area, which I am unfamiliar with even though it relates to the machine learning, which will be my research focus in my PhD studies. However, he cheered me up, provided helps in choosing this topic and discussing it in machine learning perspective.

With my deepest gratitude, I would like to thank Prof. Guocan Feng, who advised me in machine learning researches and recommended me to both MLSS and PhD studies in the famous 'Oxbridge' – I couldn't have got those admissions if without his kindly references. Though I prefer the research style with great freedom, he always provided me both high-level inspirations and also warmest helps when in need, which I am extremely grateful for.

My happy exploration in different disciplines, including *electro-spinning*, owes much to my advisor Dr. Zhangqi Feng and other talented members in the biomedical lab. Also I would like to extend special thanks to Dr. Lei Zhang and Mr. Zhihong Huang for advising me in data mining. In addition, I am grateful to Dr. Marco Cuturi's help and Martin Ratajczak's suggestions during the MLSS at Kyoto University, as well as Dr. Hugo Larochelle's comments of deep learning and NLP. Without their helps I wouldn't have decided to pursue PhD studies in machine learning.

I enjoy the energetic life as well as doing research. My friends, though not aware of my research, have given me a wonderful and fantastic undergraduate life, which I am really thankful for. Finally, I would like to thank my family – Dad, Mum and Yinghong – for all their support and patience down these years, and want to dedicate this thesis to my grandfather, who cannot attend my commencement.