Compressive Sensing Chapter

Thomas Owen Kealy

April 29, 2016

Contents

0.1 Introduction and Preliminaries

Compressive sensing is a modern signal acquisition technique in which randomness is used as an effective sampling strategy. In practice many signals encountered 'in the wild' can be fully specified by much fewer bits than required by the Nyquist sampling theorem. This is either a natural property of the signals, for example images have large areas of similar pixels, or as a conscious design choice, for example training sequences in communication transmissions. These signals are not statistically white, and so these signals may be compressed (to save on storage). For example, lossy image compression algorithms can reduce the size of a stored image to about 1% of the size required by Nyquist sampling.

Whilst this vein of research has been extraordinarily successful, it poses the question: if the reconstruction algorithm is able to reconstruct the signal from this compressed representation, why collect all the data in the first place, when most of the information can be thrown away?

Compressed Sensing answers these questions, by way of providing an alternative signal acquisition method to the Nyquist theorem. Specifically, situations are considered where fewer samples are collected than traditional sensing schemes. That is, in contrast to Nyquist sampling, Compressive Sensing is a method of measuring the informative parts of a signal directly without acquiring unessential information at the same time.

Signals which are compressible, are signals whose information content is smaller than the ambient dimension they are acquired in. Such signals have representations in which they are sparse (i.e. the most of the co-efficients in that representation are zero, or close to zero). For example,

1. A sine wave at frequency ω is defined as a single spike in the frequency domain yet has an infinite support in the time domain

2. An image will have values for every pixel, yet the wavelet decomposition of the image will typically only have a few non-zero coefficients

Informally, CS posits that for s-sparse signals $\in \mathbb{R}^n$ - signals with s non-zero amplitudes at unknown locations) - $\mathcal{O}(s \log n)$ measurements are sufficient to exactly reconstruct the signal.

In practice this can be far fewer samples than conventional sampling schemes. For example a megapixel image requires 1,000,000 Nyquist samples, but can be perfectly recovered from 96,000 compressive samples in the wavelet domain [**?**].

As in classical sensing, the measurements are acquired linearly:

$$y_i = \langle \alpha, \psi_i \rangle \tag{0.1.0.1}$$

where, y_i is the i^{th} sample, $\alpha \in \mathbb{R}^n$ is the signal, and ψ_i is the i^{th} sensing vector.

We require that sensing vectors satisfy two technical conditions (described in detail below): an Isotropy property, which means that components of the sensing vectors have unit variance and are uncorrelated, and an Incoherence property, which means that sensing vectors are almost orthogonal. Once the set of measurements have been taken, the signal may be reconstructed from a simple linear program.

0.1.1 RIP and Stable Embeddings

A high-dimensional signal is sparse, if most of the coefficients x_i in the linear expansion

$$\alpha = \sum_{i=1}^{n} x_i \phi_i \tag{0.1.1.2}$$

are zero, where $x \in \mathbb{R}^n$, $\alpha \in \mathbb{R}$, and ϕ_i are a set of basis functions of $\mathbb{R}^n$. We can make the notion of sparsity precise by defining Σ_s as the set of s-sparse signals in $\mathbb{R}^n$:

Definition 0.1.1.

$$\Sigma_s = \{x \in \mathbb{R}^n : |\text{supp}\,(x)\,| \leq s\} \tag{0.1.1.3}$$

where $\text{supp}\,(x)$ *is the set of indices on which* x *is non-zero.*

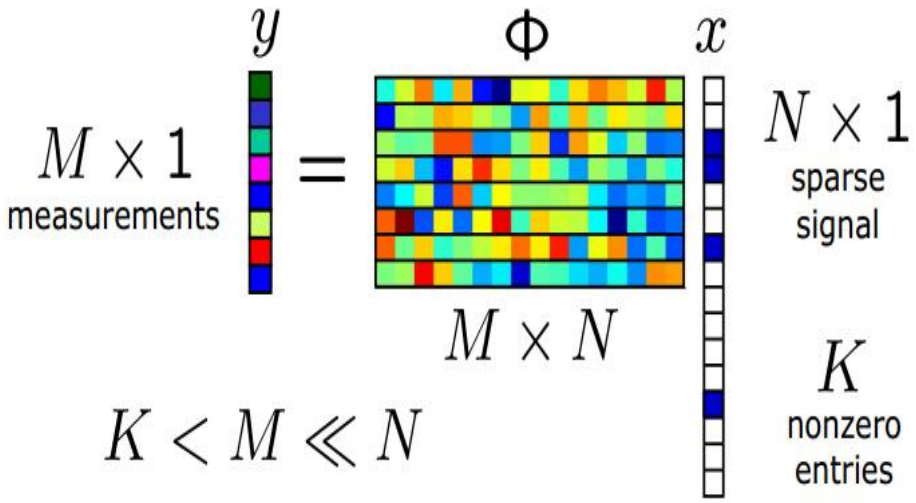

Figure 0.1.1: A visualisation of the Compressive Sensing problem as an under-determined system

We may not be able to directly obtain these coefficients, as we may not posses an appropriate measuring device or one may not exist, or there is

considerable uncertainty about where the non-zero coefficients are. Yet we still are able to measure correlations between the signal and some waveforms ψ_k i.e.

$$y_k = \langle \alpha, \psi_k \rangle \;\; k = 1 \dots m \tag{0.1.1.4}$$

Given a signal $\alpha \in \mathbb{R}^n$, a matrix $A \in \mathbb{R}^{m \times n}$, with $m \ll n$, we can acquire the signal via the set of linear measurements:

$$y = Ax \tag{0.1.1.5}$$

where in this case A represents the sampling system (i.e the columns of A are the products of the two bases ψ, ϕ). In contrast to classical sensing, which requires that $m = n$ for there to be no loss of information, it is possible to reconstruct x from an under-determined set of measurements as long as x is sparse in some basis.

There are two conditions the matrix A needs to satisfy for recovery below Nyquist rates:

1. Restricted Isometry Property.

2. Incoherence between sensing and signal bases.

Definition 0.1.2 (RIP)**.** *We say that a matrix A satisifes the RIP of order s if there exists a $\delta \in (0, 1)$ such that for all $x \in \Sigma_s$:*

$$(1 - \delta) \, ||x||_2^2 \leq ||Ax||_2^2 \leq (1 + \delta) \, ||x||_2^2 \tag{0.1.1.6}$$

i.e. A approximately preserves the lengths of all s-sparse vectors in $\mathbb{R}^n$.

Remark 0.1.3. *Although the matrix A is not square, the RIP (0.1.2) ensures that $A^T A$ is close to the identity, and so A behaves approximately as if it were orthogonal. This is formalised in the following lemma from [15]:*

Lemma 0.1.4. *Let A be a matrix which satisfies the RIP of order $2s$ with RIP constant δ. Then for two disjoint subsets $I, J \subset [n]$ each of size at most s, and for any vector $u \in \mathbb{R}^n$:*

$$\langle Au_I, Au_J \rangle \leq \delta \, ||u_I||_2^2 \, ||u_J||_2^2 \tag{0.1.1.7}$$

where u_I is the vector with component u_i if $i \in I$ and zero elsewhere.

Remark 0.1.5 (Information Preservation)**.** *A necessary condition to recover all s-sparse vectors from the measurements Ax is that $Ax_1 \neq Ax_2$ for any pair $x \not= x_2$, $x_1, x_2 \in \Sigma_s$, which is equivalent to $||A(x_1 - x_2)||_2^2 > 0$.*

This is guaranteed as long as A satisfies the RIP of order $2s$ with constant δ - as the vector $x_1 - x_2$ will hve at most $2s$ non-zero entries, and so will be distinguishable after multiplication with A. To complete the argument take $x = x_1 - x_2$ in definition (0.1.2), guaranteeing $||A(x_1 - x_2)||_2^2 > 0$, and requiring the RIP order of A to be $2s$.

Remark 0.1.6 (Stability)**.** *We also require that the dimensionality reduction of compressed sensing is the preservation of relative distances: that is if x_1 and x_2 are far apart in $\mathbb{R}^n$ then their projections Ax_1 and Ax_2 are far apart in $\mathbb{R}^m$. This will guarantee that the dimensionality reduction is robust to noise.*

A requirement on the matrix A that satisfies both of these conditions is the following:

Definition 0.1.7 (δ-stable embedding)**.** *We say that a mapping is a δ-stable embedding of $U, V \subset \mathbb{R}^n$ if*

$$(1 - \delta) \, ||u - v||_2^2 \leq ||Au - Av||_2^2 \leq (1 + \delta) \, ||u - v||_2^2 \tag{0.1.1.8}$$

for all $u \in U$ and $v \in V$.

Remark 0.1.8. *Note that a matrix A, satisfying the RIP of order $2s$ is a δ-stable embedding of Σ_s, Σ_s.*

Remark 0.1.9. *Definition 0.1.7 has a simple interpretation: the matrix A must approximately preserve Euclidean distances between all points in the signal model Σ_s.*

Given that we know a basis in which our signal is sparse, ϕ, how do we choose ψ, so that we can accomplish this sensing task? In classical sensing, we choose ψ_k to be the set of T_s-spaced delta functions (or equivalently the set of $1/T_s$ spaced delta functions in the frequency domain). A simple set of ψ_k would be to choose a (random) subset of the delta functions above.

In general, we seek waveforms in which the signals' representation would be dense.

Definition 0.1.10 (Incoherence)**.** *A pair of bases is said to be incoherent if the largest projection of two elements between the sensing (ψ) and representation (ϕ) basis is in the set $[1, \sqrt{n}]$, where n is the dimension of the signal. The coherence of a set of bases is denoted by μ.*

Examples of pairs of incoherent bases are:

- Time and Fourier bases: Let $\Phi = \mathbf{I}_n$ be the canonical basis and $\Psi = \mathbf{F}$ with $\psi_i = n^{-12} e^{i\omega k}$ be the Fourier basis, then $\mu(\phi, \psi) = 1$. This corresponds to the classical sampling strategy in time or space.

- Consider the basis Φ to have only entries in a single row, then the coherence between Φ and any fixed basis Ψ will be $\sqrt{n}$.

- Random matrices are incoherent with any fixed basis Ψ. We can choose Φ by creating n orthonormal vectors from n vectors sampled independently and uniformly on the unit sphere. With high probability $\mu = \sqrt{n \log n}$. This extends to matrices whose rows are created by sampling independent Gaussian or Bernoulli random vectors.

This implies that sensing with incoherent systems is good (in the sine wave example above it would be better to sample randomly in the time domain as opposed to the frequency domain), and efficient mechanisms ought to acquire correlations with random waveforms (e.g. white noise).

Theorem [5] Fix a signal $\mathrm{f} \in \mathbb{R}^n$ with a sparse coefficient basis, x_i in ϕ. Then a reconstruction from m random measurements in ψ is possible with probability $1 - \delta$ if:

$$m \geq C\mu^2(\phi, \psi) S \log\left(\frac{n}{\delta}\right) \tag{0.1.1.9}$$

where $\mu(\phi, \psi)$ is the coherence of the two bases, and S is the number of non-zero entries on the support of the signal.

0.1.2 Random Matrix Constructions

To construct matrices satisfying definition 0.1.7, given m, n we generate A by A_{ij} being i.i.d random variables from distributions with the following conditions [7]

Condition 1 (Norm preservation)**.** $\mathbb{E}A_{ij}^2 = \frac{1}{m}$

Condition 2 (sub-Gaussian)**.** $\mathbb{E}\left(e^{A_{ij}t}\right) \leq e^{C^2t^2/2}$

Random variables A_{ij} satisfying conditions (1) and (2) satisfy the following concentration inequality [1]:

Lemma 0.1.11 (sub-Gaussian)**.**

$$\mathbb{P}\left(\mid ||Ax||_2^2 - ||x||_2^2 \mid \geq \varepsilon ||x||_2^2\right) \leq 2e^{-cM\varepsilon^2} \tag{0.1.2.10}$$

Then in [1] the following theorem is proved:

Theorem 0.1.12. *Suppose that m, n and $0 < \delta < 1$ are given. If the probability distribution generating A satisfies condition* (0.1.2.10)*, then there exist constants c_1, c_2 depending only on δ such that the RIP* (0.1.2) *holds for A with the prescribed δ and any $s \leq \frac{c_1 n}{\log n/s}$ with probability $\geq 1 - 2e^{-c_2 n}$*

For example, if we take $A_{ij} \sim \mathcal{N}(0, 1/m)$, then the matrix A will satisfy the RIP

0.1.3 Wishart Matrices

Let $\{X_i\}_{i=1}^r$ be a set of i.i.d $1 \times p$ random vectors drawn from the multivariate normal distribution with mean 0 and covariance matrix H.

$$X_i = \left(x_1^{(i)}, \ldots, x_p^{(i)}\right) \sim N(0, H) \tag{0.1.3.11}$$

We form the matrix X by concatenating the r random vectors into a $r \times p$ matrix.

Definition 0.1.13 (Wishart Matrix)**.** *Let*

$$W = \sum_{j=1}^{r} X_j X_j^T = XX^T \tag{0.1.3.12}$$

Then $W \in \mathbb{R}^{r\times r}$ has the Wishart distribution with parameters

$$W_r(H, p) \tag{0.1.3.13}$$

where p is the number of degrees of freedom.

Remark 0.1.14. *This distribution is a generalisation of the Chi-squared distribution: let $p = H = 1$.*

Theorem 0.1.15 (Expected Value)**.**

$$\mathbb{E}(W) = rH \tag{0.1.3.14}$$

Proof.

$$\begin{aligned}
\mathbb{E}(W) &= \mathbb{E}\left(\sum_{j=1}^{r} X_j X_j^T\right) \\
&= \sum_{j=1}^{r} \mathbb{E}\left(X_j X_j^T\right) \\
&= \sum_{j=1}^{r} \left(\mathrm{Var}(X_j) + \mathbb{E}(X_j)\mathbb{E}(X_j^T)\right) \\
&= rH
\end{aligned}$$

Where the last line follows as X_j is drawn from a distribution with zero mean. □

Remark 0.1.16. *The matrix $M = A^T A$, where A is constructed by the methods from section 0.1.2, will have a Wishart distribution. In particular, it will have $\mathbb{E}M = \frac{1}{m} I_n$*

The joint distribution of the eigenvalues is given by [12]:

$$p(\lambda_1, \ldots, \lambda_r) = c_r \prod_{i=1}^{r} e^{-\lambda_i} \prod_{i<j} (\lambda_i - \lambda_j)^2 \tag{0.1.3.15}$$

The eigenvectors are uniform on the unit sphere in $\mathbb{R}^r$.

0.1.4 Reconstruction Algorithms

Compressive sensing places the computational load on reconstructing the Nyquist samples x, from the set of compressive samples y. This is in contrast to traditional sensing, where the heavy lifting computationally is done by the process with discretises the continuous signal to create the digital samples.

Many recovery algorithms have been proposed, and all are based upon minimising some functional of the data. Generally, this is based upon two terms: a data fidelity term, minimising the discrepancy between the reconstruction and the true data, and regularisation term, biasing the reconstruction towards a class of solutions with desirable properties, for example sparsity. Typically the squared error $\frac{1}{2}\left|\left|y - Ax\right|\right|_2^2$ is chosen as the data fidelity term, whilst a number of regularisation terms have been introduced in the literature.

A particularly important functional is:

$$\arg\min_x \left|\left|x\right|\right|_1 \text{ s.t } y = Ax \tag{0.1.4.16}$$

known as Basis Pursuit [6], with the following program known as the LASSO [17] as a noisy generalisation:

$$\arg\min_x \frac{1}{2}\left|\left|Ax - y\right|\right|_2^2 + \lambda\left|\left|x\right|\right|_1 \tag{0.1.4.17}$$

The statistical properties of LASSO have been well studied. The program performs, both regularisation and variable selection: the parameter λ trades of data fidelity and sparsity with higher values of λ leading to sparser solutions.

The LASSO shares several features with Ridge regression, and the Non-negative garotte (used for subset regression). It can be shown [10], that the solution to (0.1.4.17) can be written as:

$$\hat{x} = S_\lambda\left(x^{OLS}\right) = x^{OLS}\text{sign}\left(x_i - \lambda\right) \tag{0.1.4.18}$$

where $x^{OLS} = (A^TA)^{-1}A^Ty$, whereas the solution to Ridge regression:

$$\arg\min_x \frac{1}{2}\left|\left|Ax - y\right|\right|_2^2 + \lambda\left|\left|x\right|\right|_2^2 \tag{0.1.4.19}$$

can be written as:

$$\hat{x} = (1+\lambda)^{-1} x^{OLS} \tag{0.1.4.20}$$

and the solution to the best subset regression:

$$\arg\min_{x} \frac{1}{2} ||Ax - y||_2^2 + \lambda ||x||_0 \tag{0.1.4.21}$$

where $||x||_0 = \{|i| : x_i \neq 0\}$, can be written as:

$$\hat{x} = H_\lambda \left(x^{OLS}\right) = x^{OLS} \mathbb{I} \left(|x^{OLS}| > \lambda\right) \tag{0.1.4.22}$$

where $\mathbb{I}$ is the indicator function. From (0.1.4.22) and (0.1.4.20), we can see that the solution to (0.1.4.17), (0.1.4.18), translates coefficients towards zero by a constant factor, and set coefficients to zero if they are too small; thus the LASSO is able to perform both model selection (choosing relevant covariates) and regularisation (shrinking model coefficients).

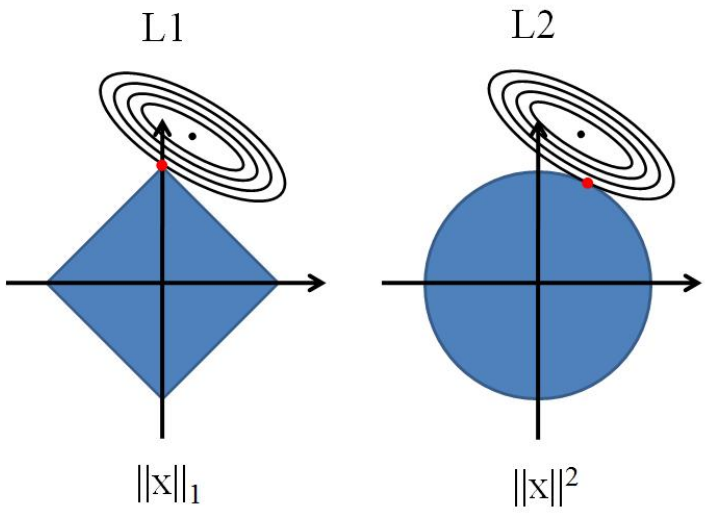

Figure 0.1.2: Solutions to the Compressive Sensing optimisation problem intersect the l_1 norm the points where all components (but one) of the vector are zero (i.e. it is sparsity promoting) [16]

Figure (0.1.2), provides a graphical demonstration of why the LASSO promotes sparse solutions. (0.1.4.17) can also be thought of as the best

convex approximation of the ℓ_0 problem (0.1.4.21), as the ℓ_1-norm is the convex hull of the points defined by $||x||_p$ for $p < 1$ as $p \to 0$.

Other examples of regularisers are:

- Elastic Net: This estimator is a blend of both (0.1.4.17) and (0.1.4.17), found by minimising:

$$\arg\min_x \frac{1}{2} ||Ax - y||_2^2 + \lambda ||x||_2^2 + \mu ||x||_1 \tag{0.1.4.23}$$

  The estimate has the benefits of both Ridge and Lasso regression: feature selection from the LASSO, and regularisation for numerical stability (useful in the under-determined case we consider here) from Ridge regression. The Elastic-net will outperform the LASSO when there is a high degree of collinearity between coefficients of the true solution.

- TV regularisation

$$\arg\min_x \frac{1}{2} ||Ax - y||_2^2 + \lambda ||\nabla x||_1 \tag{0.1.4.24}$$

  This type of regularisation is used when preserving edges whilst simultaneously de-noising a signal is required. It is used extensively in image processing, where signals exhibit large flat patches alongside large discontinuities between groups of pixels.

- Candes and Tao in [4], propose an alternative functional:

$$\min_{x \in \mathbb{R}^n} ||x||_1 \text{ s.t. } \left|\left|A^T(Ax - y)\right|\right|_\infty \leq t\sigma \tag{0.1.4.25}$$

  with $t = c\sqrt{2 \log n}$. Similarly to the LASSO this functional selects sparse vectors consistent with the data, in the sense that the residual $r = y - Ax$ is smaller than the maximum amount of noise present. In [4] it was shown that the l_2 error of the solution is within a factor of $\log n$ of the ideal l_2 error. More recent work by Bikel, Ritov, and Tsybakov, [3], has shown that the LASSO enjoys similar properties.

Broadly reconstruction algorithms fall into three classes: convex-optimisations/linear programming, greedy methods, and Bayesian methods. Convex optimisation methods offer better performance, measured in terms of reconstruction accuracy, at the cost of greater computational complexity. Greedy methods are relatively simpler, but don't have the reconstruction guarantees of convex algorithms. Bayesian methods offer the best reconstruction guarantees, as well as uncertainty estimates about the quality of reconstruction, but come with considerable computational complexity.

A recent greedy method, Approximate Message Passing (AMP), is a blend of both greedy and Bayesian methods [8].

Convex methods cast the optimisation objective either as a linear program with linear constraints, or as a second order cone program with quadratic constraints. Both of these types of program can be solved with first order interior point methods. However, their practical application to compressive sensing problems is limited due to their polynomial dependence upon the signal dimension and the number of constraints.

Compressive Sensing poses a few difficulties for convex optimisation based methods. In particular, many of the unconstrained objectives are non-smooth meaning methods based upon descent down a smooth gradient are inapplicable.

To overcome these difficulties, a series of algorithms originally proposed for wavelet-based image de-noising have been applied to CS, known as iterative shrinkage methods. These have the desirable property that they boil down to matrix-vector multiplications and component-wise thresholding.

Iterative shrinkage algorithms replace searching for a minimal facet of a complex polytope by a iteratively denoised gradient descent. The choice of the (component-wise) denoiser is dependent upon the regulaiser used in 0.1.4.17. These algorithms have an interpretation as Expectation-Maximisation [9] - where the E-step is performed as gradient descent, and the M-step is the application of the denoiser.

Greedy methods are another family of solutions to 0.1.4.17. They offer reduced computational complexity with correspondingly worse reconstruction quality and poorer guarantees on sparsity and undersampling than convex algorithms. Examples of this type are Orthogonal Matching Pursuit

1: **procedure** IST$(y, A, \mu, \tau, \varepsilon)$
2: $x^0 = 0$
3: **while** $||x^t - x^{t-1}||_2^2 \leq \varepsilon$ **do**
4: $x^{t+1} \leftarrow S_{\mu\tau}\left(x^t + \tau A^T z^t\right)$
5: $z^t \leftarrow y - Ax^t$
6: **end while**
7: **return** x^{t+1}
8: **end procedure**

Figure 0.1.3: The Iterative Soft Thresholding Algorithm

(OMP)[18], and Compressive Sensing Orthogonal Matching Pursuit (CoSAMP).

The Greedy family of algorithms abandons exhaustive searches of the solution space in favour of locally optimal single term updates. They proceed by approximating the solution by some active set of columns from the sensinng matrix A and solving a restricted least-squares problem at each (in the case of OMP). This guarantees a maximal reduction in l-2 error in each iteration.

Despite their computational simplicity, greedy algorithms have several drawbacks. Primarily they do not come with stable recovery guarantees, and they require a larger number of samples to recover the signal when compared to Bayesian and Convex recovery algorithms. Also, due to their greedy nature, these algorithms are not guaranteed to converge: in fact it can be shown that there exist k-sparse vectors and sensing matrices A such that OMP fails to converge in k iterations [19].

Bayesian methods reformulate the optimisation problem into an inference problem. These methods come with a unified theory, and standard methods to produce solutions. The theory is able to handle hyper-parameters in a unified way, provides a flexible modelling framework, and is able to provide desirable statistical quantities such as the uncertainty inherent in the prediction.

Based on the discussion above we can represent the compressive sensing measurements as:

procedure OMP(y, A, K, ε)
    $x^0 = 0$, $r = y$, $\Omega = \varnothing$, $i = 0$
    while $||x^t - x^{t-1}||_2^2 \leq \varepsilon$ **do**
        $i \leftarrow i + 1$
        $b \leftarrow A^T r$
        $\Omega \leftarrow \Omega \cup \text{supp}\left(H_1\left(b\right)\right)$
        $x \restriction_{\Omega} \leftarrow A^T \restriction_{\Omega} x$
        $x \restriction_{\Omega^c} \leftarrow 0$
        $b \leftarrow y - Ax$
    end while
    return x
end procedure

Figure 0.1.4: The OMP recovery algorithm

procedure AMP(y, A, ε)
    $x^0 = 0$, $z^0 = A^T y$
    while $||x^t - x^{t-1}||_2^2 \leq \varepsilon$ **do**
        $x^{t+1} \leftarrow S_{\mu\tau}\left(x^t - \tau + A^T z^t\right)$
        $z^{t+1} \leftarrow y - Ax^t + \frac{||x||_0}{m} z^t$
    end while
    return x^{t+1}
end procedure

Figure 0.1.5: The AMP recovery algorithm

$$y = Ax \tag{0.1.4.26}$$

where A is a $K \times N$ matrix which is the product of the measurement and sparse bases described earlier.

Note that the measurements may be noisy, with the measurement noise represented by a zero mean Gaussian distribution and unknown variance σ^2:

$$y = Ax + n \tag{0.1.4.27}$$

Where **n** is the vector representing the vector of noise, and has the same support as the measurements.

Previous sections have shown how the weights x may be found through optimisation methods such as basis pursuit or greedy algorithms. Here, an alternative Bayesian model is described.

From 0.1.4 we have a Gaussian likelihood model:

$$p\left(y \mid z,\sigma^2\right) = (2\pi\sigma^2)^{-K/2} \exp\left(-\frac{1}{2\sigma^2}\|y - Ax|_2^2\right) \tag{0.1.4.28}$$

The above has converted the CS problem of inverting sparse weight **w** into a linear regression problem with a constraint (prior) that **w** is sparse.

To seek the full posterior distribution over **w** and σ^2, we can chose a sparsity promoting prior. A popular sparseness prior is the Laplace density functions:

$$p\left(x \mid \lambda\right) = \left(\frac{\lambda}{2}\right)^N exp{-\lambda \sum_{i=1}^{N} |x_i|} \tag{0.1.4.29}$$

Note that the solution the convex optimisation problem **??** corresponds to a maximum *a posteriori* estimate for w using this prior. I.e this prior is equivalent to using the l_1 norm as an optimisation function (see figure 0.1.6 [16]).

The full posterior distribution on w and σ^2 may be realised, by using a hierarchical prior instead. To do this, define a zero-mean Gaussian prior on

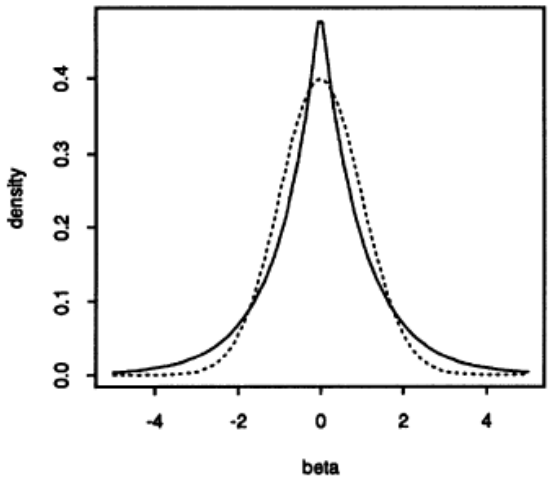

Figure 0.1.6: The Laplace (l_1-norm, bold line) and Normal (l_2-norm, dotted line) densities. Note that the Laplace density is sparsity promoting as it penalises solutions away from zero more than the Gaussian density. [16]

each element of w:

$$p\left(w \mid a\right) = \prod_{i=1}^{N} \mathbb{N}\left(w_i \mid 0, \alpha_i^{-}1\right) \tag{0.1.4.30}$$

where α is the precision of the distribution. A gamma prior is then imposed on α:

$$p\left(\alpha \mid a, b\right) = \prod_{i=1}^{N} \Gamma\left(\alpha_i \mid a, b\right) \tag{0.1.4.31}$$

The overall prior is found by marginalising over the hyperparameters:

$$p\left(w \mid a, b\right) = \prod_{i=1}^{N} \int_0^{\infty} \mathbb{N}\left(w_i \mid 0, \alpha_i^{-}1\right) \Gamma\left(\alpha_i \mid a, b\right) \tag{0.1.4.32}$$

This integral can be done analytically and is a Student-t distribution. Choosing the parameters a, b appropriately we can make the Student-t distribution peak strongly around $w_i = 0$ i.e. sparsifying. This process can be repeated for the noise variance σ^2. The hierarchical model for this process

is shown in 0.1.7. This model, and other CS models which not necessarily have closed form solutions, can be solved via belief-propagation [2], or via Monte-Carlo methods.

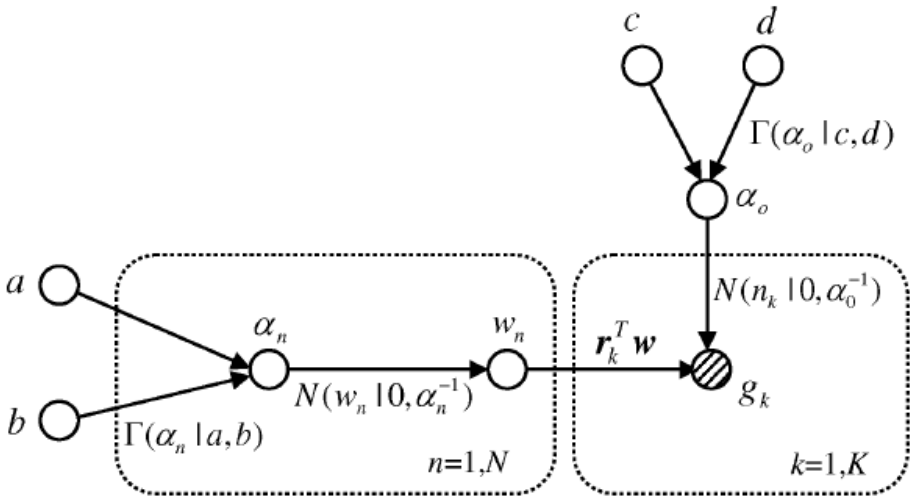

Figure 0.1.7: The hierarchical model for the Bayesian CS formulation [11]

However, as with all methodologies, Bayesian algorithms have their drawbacks. Most notable is the use of the most computationally complex recovery algorithms. In particular MCMC methods suffer in high dimensional settings, such as those considered in compressive sensing. There has been an active line of work to address this: most notably Hamiltonian Monte Carlo - an MCMC sampling method designed to follow the typical set of the posterior density.

Belief propagation (BP) [20] is a popular iterative algorithm, offering improved reconstruction quality and undersampling performance. However, it is a computationally complex algorithm. It is also difficult to implement. Approximate message passing (0.1.5) solves this issue by blending BP and (IT).

The algorithm proceeds like iterative thresholding, but computes an adjusted residual at each stage. The extra term comes from a first order approximation to the messages passed by BP [13].

The choice of prior is key in Bayesian inference, as it encodes all knowledge about the problem. Penalising the least-squares estimate with the ℓ_1 norm, has a Bayesian interpretation as a Laplace distribution prior. This is distinct from the LASSO however, as this Bayesian model does not set coefficients to zero exactly - instead it sets many coefficients to be small. Other priors have been suggested in the literature, such as the Horseshoe prior, which do set covariates to zero. These do not have the desirable property of leading to closed form posteriors (as shown above).

0.2 Compressive Sensing Architechtures

0.2.1 Modulated Wideband Converter

We consider a radio environment with a single primary user (PU) and a network of J nodes collaboratively trying to sense and reconstruct the PU signal, either in a fully distributed manner (by local communication), or by transmitting measurements to a fusion centre which then solves the linear system.

We try to sense and reconstruct a wideband signal, divided into L channels. We have a (connected) network of J $(= 50)$ nodes placed uniformly at random within the square $[0, 1] \times [0, 1]$. This is the same model, as in [21]. The calculations which follow are taken from [21] as well.

The nodes individually take measurements (as in [14]) by mixing the incoming analogue signal $x(t)$ with a mixing function $p_i(t)$ aliasing the spectrum. $x(t)$ is assumed to be bandlimited and composed of up to k uncorrelated transmissions over the L possible narrowband channels - i.e. the signal is k-sparse.

The mixing functions - which are independent for each node - are required to be periodic, with period T_p. Since p_i is periodic it has Fourier expansion:

$$p_i(t) = \sum_{l=-\infty}^{\infty} c_{il} \exp\left(jlt\frac{2\pi}{T_p}\right) \tag{0.2.1.33}$$

The c_{il} are the Fourier coefficients of the expansion and are defined in the standard manner. The result of the mixing procedure in channel i is therefore

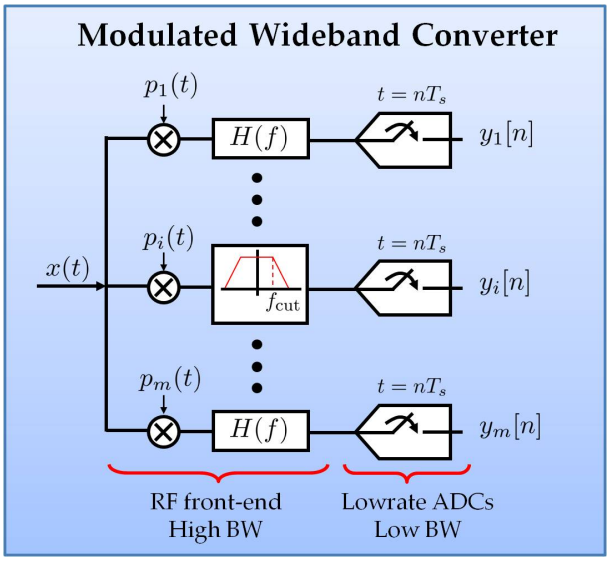

Figure 0.2.8: Mse vs SNR for the sensing model, with AWGN only, showing the performance of distributed and centralised solvers

the product xp_i, with Fourier transform (we denote the Fourier Transform of x by $X(\cdot)$):

$$\begin{aligned} X_i(f) &= \int_{-\infty}^{\infty} x(t)\, p_i(t)\, dt \\ &= \sum_{l=-\infty}^{\infty} c_{il} X(f - l f_p) \end{aligned} \tag{0.2.1.34}$$

(We insert the Fourier series for p_i, then exchange the sum and integral). The output of this mixing process then, is a linear combination of shifted copies of $X(f)$, with at most $\lceil f_N Y Q / f_p \rceil$ terms since $X(f)$ is zero outside its support (we have assumed this Nyquist frequency exists, even though we never sample at that rate).

This process is repeated in parallel at each node so that each band in x appears in baseband.

Once the mixing process has been completed the signal in each channel is low-pass filtered and sampled at a rate $f_s \geq f_p$. In the frequency domain

this is a ideal rectangle function, so the output of a single channel is:

$$Y_i\left(e^{j2\pi fT_s}\right) = \sum_{l=-L_0}^{+L_0} c_{il} X\left(f - lf_p\right) \tag{0.2.1.35}$$

since frequencies outside of $[-f_s/2, f_s/2]$ will filtered out. L_0 is the smallest integer number of non-zero contributions in $X(f)$ over $[-f_s/2, f_s/2]$ - at most $\lceil f_N YQ/f_p \rceil$ if we choose $f_s = f_p$. These relations can be written in matrix form as:

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{w} \tag{0.2.1.36}$$

where $\mathbf{y}$ contains the output of the measurement process, and $\mathbf{A}$ is a product matrix of the mixing functions, their Fourier coefficients, a partial Fourier Matrix, and a matrix of channel coefficients. $\mathbf{x}$ is the vector of unknown samples of $x(t)$.

i.e. $\mathbf{A}$ can be written:

$$\mathbf{A}^{m\times L} = \mathbf{S}^{m\times L}\mathbf{F}^{L\times L}\mathbf{D}^{L\times L}\mathbf{H}^{L\times L} \tag{0.2.1.37}$$

The system 0.2.1.36 can then be solved (in the sense of finding the sparse vector $\mathbf{x}$ by convex optimisation via minimising the objective function:

$$\frac{1}{2}\|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2^2 + \lambda\|\mathbf{x}\|_1 \tag{0.2.1.38}$$

where λ is a parameter chosen to promote sparsity. Larger λ means sparser $\mathbf{x}$.

0.2.2 Random Demodulator

We assume that the analogue signal $x(t)$ is comprised of a finite number of components from some arbitrary dictionary $\psi_n(t)$:

$$x(t) = \sum_{n=1}^{N} \alpha_n \psi_n(t) \tag{0.2.2.39}$$

The signal is said to be sparse when there are only a few non-zero α_n. The dictionary elements ψ_n may have a relatively high bandwidth, but the signal itself will have only a few degrees of freedom.

The signal acquisition method proposed consists of three stages (all analogue processing): demodulation, filtering and uniform sampling.

Initially, the signal is modulated by a pseudo-random sequence $p_c(t)$, which alternates at frequencies at (or above) the Nyquist frequency of $x(t)$. The signal is then filtered, through a filter with impulse response $h(t)$, before being sampled at rate $\mathcal{M}$ with a traditional ADC.

The output of this system, $y[m]$, can be related to the input $x(t)$ via a linear transformation of the coefficient vector α_n.

To find the transformation A, first consider the output of $y[m]$, which is the result of convolution and demodulation followed by sampling at rate $\mathcal{M}$:

$$y[m] = \int_{-\infty}^{\infty} x(\tau)\, p_c(\tau)\, h(t-\tau)\, |_{t=m\mathcal{M}}\, d\tau \tag{0.2.2.40}$$

and by expanding $x(t) = \sum_{n=1}^{N} \alpha_n \psi_n(t)$:

$$y[m] = \sum_{n=1}^{N} \alpha_n \int_{-\infty}^{\infty} \psi_n(t)\, p_c(\tau)\, h(m\mathcal{M}-\tau)\, d\tau \tag{0.2.2.41}$$

we see that the output can be written as:

$$y = Ax \tag{0.2.2.42}$$

with

$$A_{m,n} = \int_{-\infty}^{\infty} \psi_n(t)\, p_c(\tau)\, h(m\mathcal{M}-\tau)\, d\tau \tag{0.2.2.43}$$

Bibliography

[1] Richard Baraniuk, Mark Davenport, Ronald DeVore, and Michael Wakin. A simple proof of the restricted isometry property for random matrices. *Constructive Approximation*, 28(3):253–263, 2008.

[2] Dror Baron. Bayesian compressive sensing via belief propagation. *Signal Processing, IEEE ...*, 58(1):269–280, 2010.

[3] Peter J Bickel, Ya'acov Ritov, and Alexandre B Tsybakov. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, pages 1705–1732, 2009.

[4] Emmanuel Candes and Terence Tao. The dantzig selector: statistical estimation when p is much larger than n. *The Annals of Statistics*, pages 2313–2351, 2007.

[5] Emmanuel J Candes, Justin Romberg, and Terence Tao. Robust Uncertainty Principles : Exact Signal Frequency Information. 52(2):489–509, 2006.

[6] Scott Shaobing Chen, David L. Donoho, and Michael a. Saunders. Atomic Decomposition by Basis Pursuit. *SIAM Journal on Scientific Computing*, 20(1):33–61, January 1998.

[7] Mark Davenport, Petros T Boufounos, Michael B Wakin, Richard G Baraniuk, et al. Signal processing with compressive measurements. *Selected Topics in Signal Processing, IEEE Journal of*, 4(2):445–460, 2010.

[8] David L Donoho, Arian Maleki, and Andrea Montanari. Message-passing algorithms for compressed sensing. *Proceedings of the National Academy of Sciences*, 106(45):18914–18919, 2009.

[9] Mário AT Figueiredo and Robert D Nowak. An em algorithm for wavelet-based image restoration. *Image Processing, IEEE Transactions on*, 12(8):906–916, 2003.

[10] Trevor Hastie, Robert Tibshirani, Jerome Friedman, and James Franklin. *The elements of statistical learning: data mining, inference and prediction*, volume 27. Springer, 2005.

[11] Shihao Ji, Ya Xue, and Lawrence Carin. Bayesian Compressive Sensing. *IEEE Transactions on Signal Processing*, 56(6):2346–2356, June 2008.

[12] Oliver Leveque. Random matrices and communication systems an unexpected journey: There and back again.

[13] Christopher A Metzler, Arian Maleki, and Richard G Baraniuk. From denoising to compressed sensing. *arXiv preprint arXiv:1406.4175*, 2014.

[14] Moshe Mishali and Yonina C Eldar. From theory to practice: Sub-nyquist sampling of sparse wideband analog signals. *Selected Topics in Signal Processing, IEEE Journal of*, 4(2):375–391, 2010.

[15] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge University Press, 2014.

[16] R Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (...*, 58(1):267–288, 1996.

[17] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[18] Joel A Tropp and Anna C Gilbert. Signal recovery from random measurements via orthogonal matching pursuit. *Information Theory, IEEE Transactions on*, 53(12):4655–4666, 2007.

[19] Jinming Wen, Xinen Zhu, and Di-Jie Li. Improved bounds on restricted isometry constant for orthogonal matching pursuit. *Electronics Letters*, 49(23):1487–1489, 2013.

[20] Jonathan S. Yedidia. Message-Passing Algorithms for Inference and Optimization. *Journal of Statistical Physics*, 145(4):860–890, October 2011.

[21] Huazi Zhang, Zhaoyang Zhang, and Yuen Chau. Distributed compressed wideband sensing in Cognitive Radio Sensor Networks. In *2011 IEEE Conference on Computer Communications Workshops, INFOCOM WKSHPS 2011*, pages 13–17, 2011.