

Summer Project Report

Tom Kealy

October 4, 2013

1 Introduction

Despite the ubiquity, capacity and apparent efficacy, modern communication systems are wasteful, inefficient and in need of reform. Most of the bits of data collected by our sensing systems are unessential, and only serve to necessitate data compression wasting computation time before transmission. For example, people regularly use a camera with a resolution of several megapixels to only to upload a file of a few kilobytes to Facebook. Devices are unable to make dynamic decisions about how to transmit this data, leading to both spectral exhaustion on some frequencies whilst much of the available radio spectrum lies fallow.

This project addresses these issues, by reviewing a novel acquisition and decompression framework for data: a way in which we need only sense the most informative bits of data. This framework is then applied to the problem of sensing spectral opportunities dynamically, to make better use of available spectrum.

The key uniting both these applications is that data and spectra are *sparse*: that is they have a representations which are 'smaller' than their respective dimension. For example, images and audio can be compressed into file formats much smaller than when initially recorded (compare the relative sizes of bitmap and JPEG images).

The sole focus of this research is to use the sparsity of the spectrum to uncover transmission opportunities, allowing better use of spectrum more generally.

We are motivated by the need to send more data over wireless networks, whilst at the same time having a constrained frequency set over which to transmit this information. This issue could be alleviated by users dynamically allocating spectrum on a per-transmission basis: without the ability to gain knowledge of spectral conditions this can never become a reality however.

The requirement for increasing bandwidth isn't just a pressing issue for today: in the next decade it is forecast that network operators will need to provide for

three-orders of magnitude (1000 times) more capacity. Demand is continually outstripping supply - motivated by the ubiquity of smart-phones, and the consumers appetites for media.

At the same time as this demand for ever more data, there is an increasing scarcity of radio spectrum over which to transmit. New frequencies are rarely cleared for commercial purposes, and when they are they go for high prices. A decade ago the UK auction for 3g spectrum licenses raised an estimated £22.4 billion [?] for the UK treasury, indicating the seriousness of the market players requirements for new spectrum. The recent 4g spectrum auction raised £2.3 billion [?]- with initial networks being rolled out by the end of 2013.

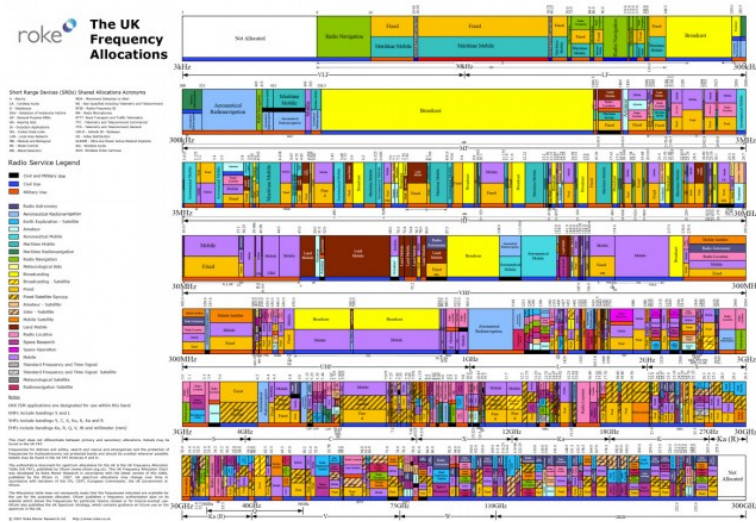


Figure 1: A digram of current Spectral allocation [?]

However, a closer inspection of the frequency allocation suggests this scarcity is artificial, it's more a product of regulatory oversight over time. As the constraints on spectrum requirement became more complex, so did the solutions to that problem - at the cost of leaving much of the spectrum idle for most of the time.

For example: much of the spectrum is allocated to TV broadcast, radio broadcast and mobile. However, if we look closer, the allocations aren't even for specific companies - they're simply categories. Within these, OFCOM may have many licensees within each category.

Also interesting to note is how much frequency the Government allocates to itself (the red bar underneath the blocks indicates Government use). Compare this to the actual utilisation of spectrum: much of it is not used at all. Figure 2 shows a snapshot of frequency utilisation in three diverse locations in the UK over te radio spectrum, note that many frequencies are not utilised (coloured blue)

whilst others have significant activity (coloured yellow). Note that the plot for Southwark (central London) is barely different from Braddock - a rural area.

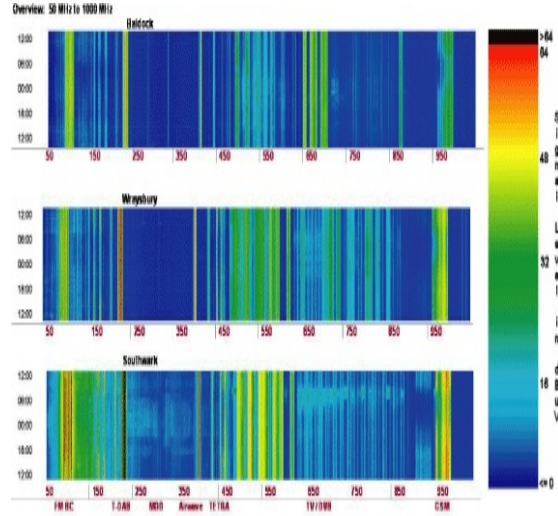


Figure 2: A snapshot of frequency utilisation in various areas: many frequencies are not used at all, whilst there is significant activity on others [?]

How do we then go about solving this issue - how can we obtain the most significant bits of information from our sensing mechanism, whilst obviating the need to compress the data once we are done? How do we dynamically assign spectrum? The work of Candes, Tao [?] and Donoho [?], has shown that instead of measuring the information we require directly (and then compressing it), we can measure 'holographic' and non-linear random projections between our measurement space and the space where our data is sparse. This requires only the knowledge that the signal is compressible via some transform - both the acquisition protocol and the reconstruction algorithm are agnostic to the type of signal. What is surprising is that the sampling kernels are fixed independently of the signal, are non-adaptive and these projections are sufficient to reconstruct the signal - as if we had an Oracle to tell us where the non-zero components of our signal are.

This work has had a large impact in medical imaging since it's inception: for example, it's now possible to take an image of a patient's heart within a single breath, as well as dynamic imaging of the heart ([?] figures 7 and 9).

Modern digital signal processing techniques (such as modulation techniques) are far more spectrally efficient than their historic analogue counterparts, which has in part contributed to the spectrum crisis. All this is changing though: from the beginning of 2013 all TV in the UK will be transmitted digitally. Historically, television in the UK was broadcast using analogue signals requiring 32 multiplexes. Digital TV requires 6 multiplexes, on the other hand.

This freeing up of TV frequencies represents an opportunity: these frequencies have good propagation characteristics (they suffer less with free space path loss relative to higher frequencies), whilst still providing good bandwidth for data transmission. These TV frequencies are being opened up to civilian and commercial users: spectral holes will be able to be exploited opportunistically by devices, so long as they don't interfere with the reception of TV. Historically, this is the single largest gift of new spectrum, and because there is no requirement for licensing this spectrum is free.

As with all technological innovations, this will not only improve existing infrastructure but also new classes of devices to transmit, for instance; applications such as passive sensor networks) which only need spectrum intermittently to transmit monitoring results), inter-vehicle communication for real time traffic monitoring and wireless internet at broadband data rates have all been proposed.

Despite all of this hype, dynamic spectrum access won't become a reality unless spectral holes can be robustly detected. The requirement that secondary users exploit the new spectrum politely, without interference to

primary user makes spectrum sensing essential to TV white-space (TVWS) technologies. The realisation of any Cognitive Radio standard (such as IEEE 802.22), requires the co-existence of primary (TV users) and secondary (everybody else who wants to use TVWS spectrum) users of the frequency spectrum to ensure proper interference mitigation and appropriate network behaviour.

Users of TVWS (Cognitive Radios) must sense whether spectrum is available, and must be able to detect very weak primary user signals. Furthermore they must sense over a wide bandwidth (due to the amount of TVWS spectrum proposed), which challenges traditional Nyquist sampling techniques, because the sampling rates required are not technically feasible with current RF or Analogue-to-Digital conversion technology.

Sensing should enable devices to detect the presence of TV signals in a band and provide smart and adaptive (and possibly distributed) solutions to band iden-

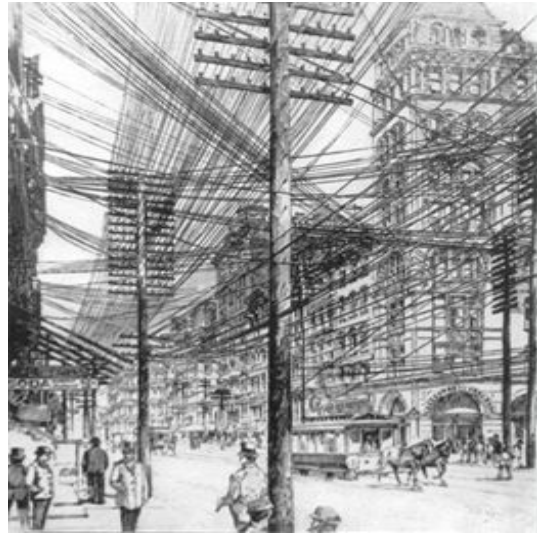


Figure 3: A picture of early 20th century New York: Bandwidth has always been an issue

tification.

Spectrum sensing should involve:

1. Sensing to detect white spaces.
2. Co-existence with similar devices.
3. Frequency monitoring of other devices.
4. Interference management.
5. Spectrum mobility and transmission power control when needed.

As described earlier, the available spectrum is highly underutilised, and can be thought of as a collection of narrowband transmissions over a wideband channel. As such, the spectrum we're sensing is sparse. This makes it an ideal candidate for sparse recovery techniques such as Compressive Sensing.

The structure of the report is as follows: first a few methods for sensing narrowband signals (i.e. channels where the frequency response is approximately flat, and where the bandwidth is smaller than the coherence bandwidth of the channel) are discussed and the limitations of these are highlighted for the problem of sensing spectrum for Cognitive Radios. Then (more appropriate) wideband sensing techniques are discussed.

Compressive Sensing, and Group Testing are then introduced and it is argued that they offer appropriate solutions to the sensing problems for cognitive radios. Sampling schemes married to these reconstruction algorithms are discussed.

Finally the results and simulations performed over the time-frame of the project are discussed, with reference to further work.

2 Narrowband Spectrum Sensing

The problem of spectrum sensing is to decide whether a particular band is available, or not [?]. That is, we wish to discriminate between the following two hypotheses:

$$H_0 : y[n] = w[n], n = 1 \dots N \quad (2.0.1)$$

$$H_1 : y[n] = x[n] + w[n], n = 1 \dots N \quad (2.0.2)$$

Where x is the primary users signal, having a specific structure which stems from modern coding and modulation techniques, w is additive white Gaussian noise and y is the received signal.

To decide whether the observations \mathbf{y} were generated under H_0 or H_1 is accomplished by forming a test statistic $\Gamma(\mathbf{y})$ and then comparing this statistic with a predefined threshold λ . Both classical methods, where the hypotheses are assumed to be deterministically true and the goal is to minimise the false detection probability, and Bayesian methods, where it is assumed that the source selects the true hypothesis at random according to some prior probabilities, agree that the test statistic should be likelihood ratio:

$$\Gamma(\mathbf{y}) = \frac{p(\mathbf{y} | H_0)}{p(\mathbf{y} | H_1)} \quad (2.0.3)$$

The performance of a detector is quantified in terms of the probability of detection

$$P_D = Pr(\Gamma(\mathbf{y}) > \lambda | H_1) \quad (2.0.4)$$

and the probability of false alarm

$$P_{FA} = Pr(\Gamma(\mathbf{y}) > \lambda | H_0) \quad (2.0.5)$$

By varying λ the operating point of a detector can be chosen anywhere along its receiver operating characteristics curve.

There are several proposed spectrum sensing methods that enable cognitive radios identify bands and perform dynamic frequency selection. Some of the common (narrowband) spectrum sensing techniques are described below.

2.0.1 Energy Detection

This is a common method for the detection of unknown signals in noise, due to low computational and implementation complexity. This method is quite generic as receivers need no knowledge of the primary users signal.

A typical method would be a bandpass filter with a centre frequency f_s and a bandwidth W . This is followed by a squaring device to measure the received energy and an integrator to determine the observation interval. Finally the output of the integrator is compared with a threshold to determine the presence of a signal. This threshold is determined based upon the noise variance of the channel. I.e. we have a decision metric of the following form:

$$M = \sum_{n=0}^N |y[n]|^2 \quad (2.0.6)$$

Assuming that the signal is a zero mean AWGN variable as well, we can derive expressions for the metric, the detection probability and the false alarm probability:

$$M = \begin{cases} \frac{\sigma_w^2}{2} & H_0 \\ \frac{\sigma_w^2 + \sigma_s^2}{2} & H_1 \end{cases} \quad (2.0.7)$$

$$P_D = 1 - \Gamma\left(1, \frac{\lambda}{1 + \frac{\sigma_s^2}{\sigma_w^2}}\right) \quad (2.0.8)$$

$$P_{FA} = 1 - \Gamma\left(1, \frac{\lambda}{\sigma_w^2}\right) \quad (2.0.9)$$

Where $\Gamma(1, x)$ is the incomplete gamma function. From these equations it's clear to see that the performance of energy detection based sensing faces challenges at low SNR values. See (REF) figure 3 for curves quantifying the performance. Also energy detectors perform poorly under extreme fading conditions as they are unable to distinguish primary users and noise. Further this type of detector is not efficient at detecting spread spectrum signals.

Similarly because the threshold used to make the decision is based on the noise variance σ_w^2 , any error in the noise power estimation can cause significant performance loss. Various algorithms have been proposed to estimate this variance adaptively. See [?] for more details.

2.0.2 Cyclostationary Feature Detection

Because the signals used in practical communication systems contain distinctive features that can be exploited for detection, it is possible to achieve a detection performance which substantially surpasses the energy detector. This is in contrast to the predictions of information theory where maximum entropy signals will be statistically white and Gaussian (if this were the case, then we could do no better than the energy detector). More importantly, known signal features can be exploited to estimate unknown parameters such as noise power.

Examples of well known patterns include pilot signals and spreading sequences. Other examples include preambles and midambles: known sequences transmitted before and in the middle of each slot, respectively. Others include redundancy added by coding, modulation and burst formatting used by the transmitter.

This method exploits cyclostationary features of received signals: man made periodicity in the signal (for example symbol rate, chip rate, cyclic prefix etc) or its statistics - mean, autocorrelation. A cyclic correlation function is used instead of PSD (or autocorrelation sequence) for detecting signals present in a given spectrum. This is able to differentiate noise from primary users signals since noise is wide-sense stationary with no correlation but modulated signals are cyclostationary due to the redundancy of signal correlations.

For clarity, the random processes encountered by a cognitive radio will have a period in both expectation and autocorrelation:

$$\mathbb{E}(t) = \mathbb{E}(t + mT) = \mathbb{E}[x(t)] \quad (2.0.10)$$

$$\mathbb{R}(t, \tau) = \mathbb{R}(t + mT, \tau) = \mathbb{E}[x(t)x(\bar{t} + \tau)] \quad (2.0.11)$$

where t is time, τ is the autocorrelation lag, $x(t)$ is the random process we are considering and m is an integer.

Due to the periodicity of the autocorrelation, it can be expressed as a Fourier series over integer multiples of the fundamental frequency in the signal as well as integer multiples of sums and differences of this frequency:

$$\mathbb{R}(t, \tau) = \sum_{\alpha} r(\alpha, \tau) e^{2\pi j \alpha t} \quad (2.0.12)$$

with Fourier coefficients:

$$r(\alpha, \tau) = \frac{1}{T} \int_T x\left(t + \frac{\tau}{2}\right) x\left(\bar{t} + \frac{\tau}{2}\right) e^{-2\pi j \alpha t} dt \quad (2.0.13)$$

where α is the cyclic frequency

From this we can define the Cyclic Power Spectrum of the signal:

$$S(f) = \int_{-\infty}^{\infty} r(\alpha, \tau) e^{-2\pi j f \tau} d\tau \quad (2.0.14)$$

For a fixed lag τ , 2.0.2 can be re-written as:

$$R_{xx}(t, \tau) = R_{xx}(\tau) + \sum_{\alpha} r(\alpha, \tau) e^{2\pi j \alpha t} \quad (2.0.15)$$

i.e. a part dependent on the lag only (the cyclic frequency is zero), and a part which is a periodic function of time.

Under both hypotheses, (2, 2), the continuous portion of the signal exists, but the cyclo-stationary portion only exists under 2 when $\alpha \neq 0$. Thus we only need to test for the presence of a cyclo-stationary component.

To this end re-write the hypotheses as:

$$H_0 : y[n] = S_w^\alpha[n], n = 1 \dots N \quad (2.0.16)$$

$$H_1 : y[n] = S_x^\alpha[n] + S_w^\alpha[n], n = 1 \dots N \quad (2.0.17)$$

where S_x^α is the CPS of white noise which is zero for $\alpha \neq 0$. Using the test statistic:

$$\chi = \sum_{\alpha \neq 0} \sum_n S_x^\alpha \bar{S}_x^\alpha \quad (2.0.18)$$

we can formulate the cyclo-stationary detector as:

$$d = \begin{cases} 0 & \chi < \lambda \\ 1 & \chi \geq \lambda \end{cases} \quad (2.0.19)$$

where λ is some pre-determined threshold [?].

2.0.3 Matched Filtering

If all the probability distributions and parameters - noise variance, signal variance, channel coefficients etc - are known under both hypotheses, and the signal to be detected is perfectly known then the optimal test statistic is a matched filter.

A matched filter is the convolution of a test signal with a template signal (or window) and detects the presence of the template in the unknown signal (as the convolution measures the overlap of two signals).

For example: for a given TV signal, $r(t)$ defined over $0 \leq t \leq T$ the corresponding matched filter is $h(t) = r(T - t)$.

A test statistic can be formed by sampling the output of the filter every nT seconds and choosing 1 if the statistic is below some threshold and 0 otherwise.

When compared to other methods, matched filtering takes a shorter time to achieve a threshold probability of false alarm. However, matched filtering requires that radios demodulate received signals, and so requires perfect knowledge of primary users signalling features. Matched filtering also requires a prohibitively large power consumption, as various algorithms need to be executed for detection.

2.0.4 Limitations

The methods described above, are appropriate for sensing whether a single channel is available for transmission, based upon the result of measurements of that channel. However, Cognitive Radios aim to exploit spectral holes in a wide band spectrum (i.e. a channel whose frequency response is not flat over the bandwidth) and will usually have to make a decision regarding transmission from measurements from this type of channel.

There are two proposed approaches to this: Multiband sensing and Compressive Sensing. Multiband sensing splits the wideband spectrum into a number of independent (not necessarily contiguous) sub-channels (whose frequency response is flat), and performs the hypothesis test for each sub-channel. However, in practice, there are correlations across sub-channels that this method fails to address. For example, digital TV signals are transmitted as spread spectrum signals so that primary user occupancy is correlated across channels. A related issue is that noise variance could be unknown but correlated across bands. Binary hypothesis testing then fails in this case, needing to be replaced by composite hypothesis tests which grow exponentially with the number of sub-channels. Such problems are typically non-convex and require prohibitively complex detectors.

3 Wideband Spectrum Sensing

This section presents a new method of sensing sparse signals, and its application to the problem of sensing over wideband spectra in Cognitive Radios. Initially we introduce Classical Sensing and then give an overview of both Compressive Sensing and Group Testing. Finally, we discuss some sub-Nyquist sampling techniques.

3.1 Classical Sensing

Classically, for perfect signal reconstruction, we must sample a signal such that the sampling rate must be at least twice the maximum frequency in the bandlimited signal. The continuous time signal can then be recovered using an appropriate reconstruction filter (e.g. a sinc filter). For example, we can represent a sampled continuous signal as a multiplication of the signal with a train of Dirac delta functions at multiples of the sampling period T .

$$x(nT) = \text{III}(t - nT) x(t) \quad (3.1.20)$$

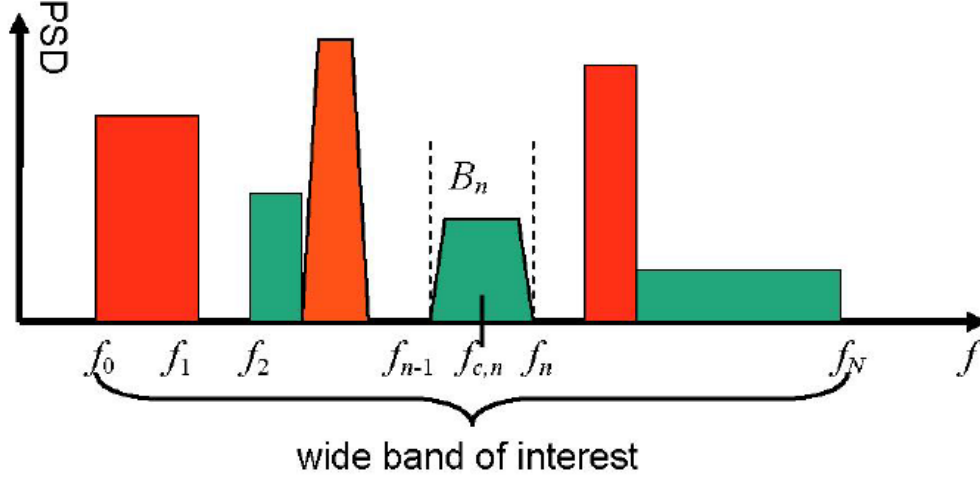


Figure 4: A diagram of the Spectrum Sensing model [?]

where

$$\text{III}(t - nT) = \sum_{k=-\infty}^{\infty} \delta(t - kT) \quad (3.1.21)$$

Working the frequency domain, this multiplication becomes convolution (which is equivalent to shifting):

$$\hat{X}_s(f) = \sum_{k=-\infty}^{\infty} x(t - kT) \quad (3.1.22)$$

Thus if the spectrum of the frequency is supported on the interval $(-B, B)$ then sampling at intervals $\frac{1}{2B}$ will contain enough information to reconstruct the signal $x(t)$. Multiplying the spectrum by a rectangle function (low-pass filtering), to remove any images caused by the periodicity of the function, and the signal $x(t)$ can be reconstructed from its samples:

$$x(t) = \sum_{n=-\infty}^{\infty} x(nT) \text{sinc}\left(\frac{t_n T}{T}\right) \quad (3.1.23)$$

3.2 Compressed Sensing

However, in practice many signals encountered 'in the wild' can be fully specified by much fewer bits than required by the sampling theorem above. For example, image compression algorithms can reduce the size of a stored image to about 1% of the size required by Nyquist sampling. If the reconstruction algorithm is able to reconstruct the image from this small amount of data, this raises the question: why collect all the data in the first place, when most of the information can be

thrown away? Why not directly measure the part that will not end up being thrown away?

Compressed Sensing considers situations where the signal is *undersampled* i.e. situations in which the number of samples is much smaller than the dimension of the signal (or the number of samples required by classical sampling theory). This is equivalent to a system of linear equations which is under-determined. That is, this is a method of measuring the informative parts of a signal directly without acquiring unessential information at the same time (i.e. the parts of the signal that would be discarded in traditional compression applications). The questions then are how can we acquire these measurements in the first place, and how to 'decompress' them once they are obtained [?].

To answer the first, note that 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

However, we may not be able to directly obtain those coefficients, as we may not possess an appropriate measuring device (or one may not exist). Yet we are able to measure correlations between the signal and the basis waveforms of the domain where the signal is sparse ϕ_k i.e.

$$y_k = \langle f, \phi_k \rangle \quad k = 1 \dots m \quad (3.2.24)$$

for $f \in \mathbb{R}^n$ expanded in an orthonormal basis ψ s.t.

$$f(t) = \sum_{i=1}^n x_i \psi_i(t) \quad (3.2.25)$$

where the x_i are the coefficient sequence of f .

An example of a practical Compressive Sensing system is the single-pixel camera at Rice University [?]. Typical camera devices obtain pixel samples by exposing a bank of photon detectors (one for each pixel) to the incident light field. This data is then processed into an image.

The single pixel camera takes pictures by first directing the incoming light field onto an array of tiny mirrors (one for each pixel). Each mirror can be either be oriented towards a single photon detector, or oriented away from the detector. In

this setup, a measurement is taken as the sum of all the incident light beams. Afterwards, the mirrors are flipped to a new random configuration, and another measurement is taken. This process is repeated, until enough information has been collected to reconstruct the image. Figure 5 shows the operation of the single pixel camera.

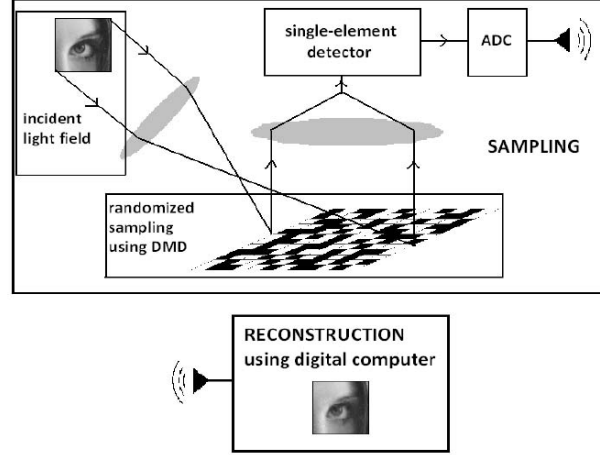


Figure 5: The operation of the single pixel camera at Rice University [?]

The question all this raises is where do we do our sensing? In other words, given that we know a basis in which our signal is sparse, ϕ , how do we choose ψ ? It's best to choose ψ so that the signal is 'spread out' relative to the signal's expansion in ϕ . Such pairs are said to be incoherent.

Definition 1 *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 pair of bases is denoted by μ .*

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

$$m \geq C\mu^2(\phi, \psi)S \log \left(\frac{n}{\delta} \right) \quad (3.2.26)$$

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.

Then f^* (the proposed reconstruction) is given by $f^* = \psi x^*$ where x^* is the solution to the convex optimisation program (n.b. $\|x\|_{l_1} := \sum_i |x_i|$):

$$\min \|\tilde{x}\|_{l_1} \text{ subject to } y_k = \langle \phi_k, \psi x^* \rangle \quad \forall k \in M \subset [1 \dots n] \quad (3.2.27)$$

i.e. CS: Sample non-adaptively in an incoherent domain and invoke linear programming after the acquisition step

Thus the role of l_1 minimisation is to decompress the data. There are many ways to perform this operation: some popular methods are basis pursuit [?] and Greedy approaches such as Orthogonal Matching Pursuit [?].

It may be remarked that the l_0 norm:

$$\|l_0\| = \{\#i, x_i \neq 0\} \quad (3.2.28)$$

is a more appropriate functional to minimise. In fact under this norm, $m \geq 2k$ measurements will suffice to uniquely determine all k -sparse signals. However, this norm is not convex and so minimising it is an NP-hard problem. As we are seeking sparse solutions the l_1 -norm will suffice [?]. This is because all vectors in a random k -dimensional subspace of an n -dimensional space are approximately Gaussian (in the sense that the components are distributed according to an approximate normal distribution). Such vectors have roughly equivalent norms, and so any solution to the l_1 minimisation problem will be the same solution to the l_0 minimisation problem for sufficiently sparse signals.

3.2.1 Incoherence, isometries and all that

To recover a sparse vector, we must make sure that the vectors are not in the null space of the sensing matrix (otherwise there would be no hope of recovery). We also require that any subset of S columns taken from the measurement matrix be nearly orthogonal w.r.t sparse vectors: i.e. all pairwise distances between S -sparse vectors be well preserved in the measurement space.

This can be summed up in the following inequality (Restricted Isometry Property) [?]:

$$(1 - \delta) \|x\|_{l_2}^2 \leq \|Ax\|_{l_2}^2 \leq (1 + \delta) \|x\|_{l_2}^2 \quad (3.2.29)$$

We are also in a position to evaluate the meaning of the constant μ in 3.2. We are considering sampling within orthonormal systems (for example, Time and Frequency):

$$A^* A = nI \quad (3.2.30)$$

so that each row or column has l_2 norm equal to \sqrt{n} . A is any matrix satisfying this property (examples include the Fourier matrix and the Dirac matrix). Thus μ must be in the set $[1, \sqrt{n}]$. μ then, is a measure of how concentrated the rows of our measurement matrix is - i.e. how much information is spread across each vector. If $\mu = 1$ then the rows are 'flat' - and we need relatively fewer samples to reconstruct an S -sparse signal (i.e. each sample provides the same amount of information). However, if the rows contain all non-zero entries except for a single component, then $\mu^2 = n$ and we will need to observe all components to determine the non-zero one (i.e. we have no guarantees of recovery from limited samples) [?].

Noting that the measurements we take are projections from our orthonormal system (from example time) onto a sparsifying basis (i.e. frequency) we can see that:

$$\mu = \max_{k,j} |\langle \phi_k, \psi_j \rangle| \quad (3.2.31)$$

So we need to choose a sensing basis, where the vectors will be 'spread out', and the degree of spreading is characterised by μ .

3.2.2 Short, Fat matrices

As remarked upon earlier: Compressive Sensing is equivalent to solving an under-determined linear system, with the constraint that we seek the sparsest solution. The content of the previous sections amounts to constraints on the number of rows of matrix of this linear system.

If we had an Oracle which could tell us where the non-zero components of our solution were, then we would need only as many rows of the matrix as there were non-zero components in the signal to fully specify the problem.

However, such an Oracle does not exist, and so we're left with the task of constructing a matrix to recover those components. Knowing that we're looking for k -sparse solutions, we need a matrix with at least $2k$ columns which are linearly independent. Equivalently, all images of k -sparse vectors under the operation of the sensing matrix Φ must be distinct. From this, any k -sparse signal can be reconstructed from Ax .

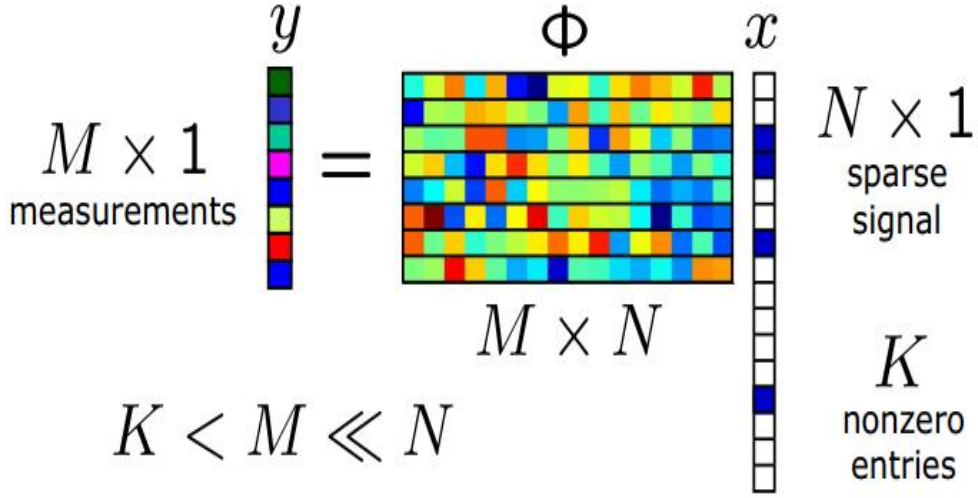


Figure 6: A visualisation of the Compressive Sensing problem as an under-determined system

To prove this assume the opposite - then there are two vectors $x, x' \in \mathbb{R}^n$ such that $Ax = Ax'$. I.e. $A(x - x') = 0$. However, $(x - x')$ is $2k$ -sparse and so there is a linear dependence between $2k$ columns of the sensing matrix A . We have a contradiction, and so $2k$ columns will suffice to reconstruct a k -sparse signal.

The problem with this is that we are trying to find the support of a k -sparse signal over a vector of length N , and so we would need to check all $\binom{N}{k}$ combinations of k -sparse signals which is prohibitively computationally expensive. Is there some way to gain the advantages of sparsity, without having to minimise a non-convex functional?

As it turns out, the answer is yes. If we take $m \geq C\mu^2(\phi, \psi)S \log(n)$ rows minimising the l_1 norm will find the sparsest solution. This is because the l_1 norm is an octahedron (in 3-dimensions, in higher dimensions it has an analogous spiky geometry), and solutions are more likely to intersect the norm at the points. Figure 7 shows this.

3.2.3 Bayesian Compressive Sensing

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

$$\mathbf{g} = \Phi \mathbf{w} \quad (3.2.32)$$

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

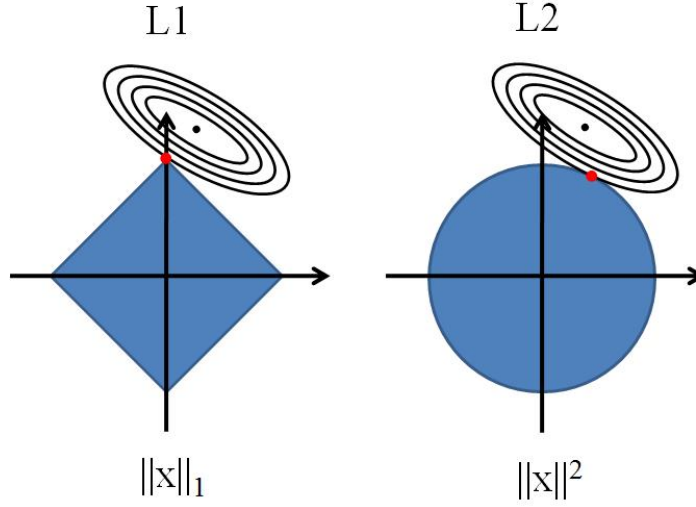


Figure 7: Solutions to the Compressive Sensing optimisation problem intersect the l_1 norm the points where all components (but one) of the vector are zero (i.e. it is sparsity promoting) [?]

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

$$\mathbf{g} = \Phi \mathbf{w} + \mathbf{n} \quad (3.2.33)$$

Where \mathbf{n} is the vector representing the vector of noise, and has the same support as the measurements.

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

From 3.2.3 we have a Gaussian likelihood model:

$$p(\mathbf{g} \mid \mathbf{w}, \sigma^2) = (2\pi\sigma^2)^{-K/2} \exp\left(-\frac{1}{2\sigma^2} \|\mathbf{g} - \Phi \mathbf{w}\|_2^2\right) \quad (3.2.34)$$

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

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

$$p(w \mid \lambda) = \left(\frac{\lambda}{2}\right)^N \exp\left(-\lambda \sum_{i=1}^N |w_i|\right) \quad (3.2.35)$$

Note that the solution the convex optimisation problem 3.2 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 8 [?]).

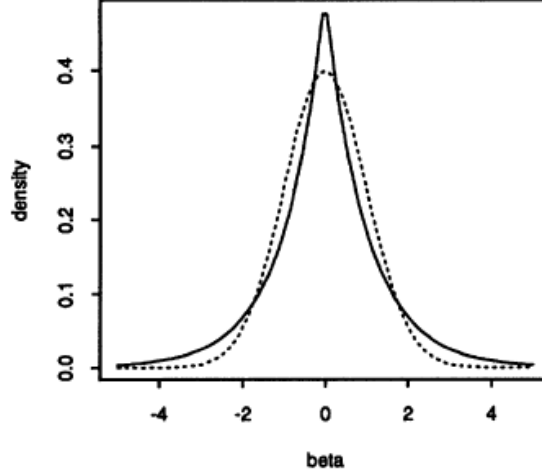


Figure 8: The Laplace (l_1 -norm, bold line) and Normal (l_2 -norm, dotted line) densities. Note that the Laplace density is sparsity promoting as it penalises solutions away from zero more than the Gaussian density. [?]

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 each element of w :

$$p(w | a) = \prod_{i=1}^N \mathbb{N}(w_i | 0, \alpha_i^{-1}) \quad (3.2.36)$$

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

$$p(\alpha | a, b) = \prod_{i=1}^N \Gamma(\alpha_i | a, b) \quad (3.2.37)$$

The overall prior is found by marginalising over the hyperparameters:

$$p(w | a, b) = \prod_{i=1}^N \int_0^\infty \mathbb{N}(w_i | 0, \alpha_i^{-1}) \Gamma(\alpha_i | a, b) \quad (3.2.38)$$

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 11. This model, and other CS models which not necessarily have closed form solutions, can be solved via belief-propagation [?]

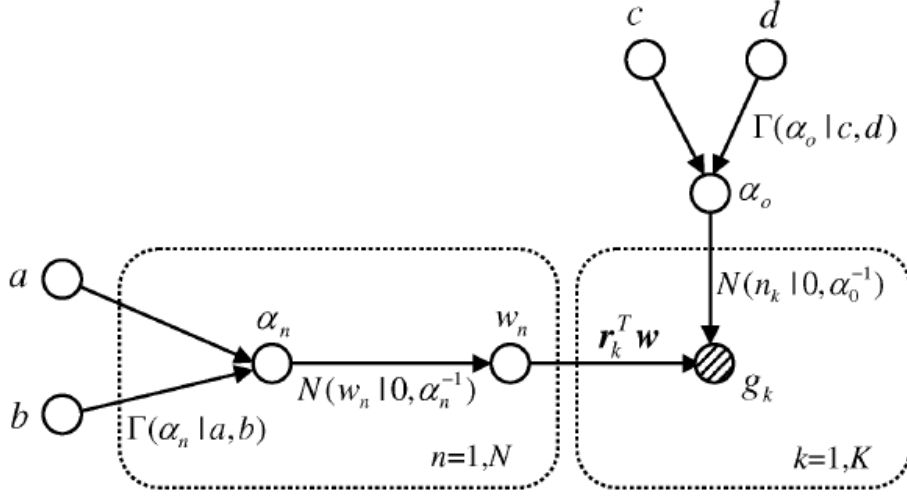


Figure 9: The hierarchical model for the Bayesian CS formulation [?]

3.3 Group Testing

Group Testing originated in the second world war because of the need to test all incoming conscriptees for syphilis. It would have been inefficient and expensive to test each soldier individually, as the rate for syphilis was only 10 per 100,000. Dorfman [?], considered the idea of pooling blood samples and testing the pooled samples for syphilis and only further testing the pools which come up positive.

A typical problem that can be solved by Group Testing is finding a counterfeit coin in a group of otherwise identical coins by weighing groups of coins on a pan balance. For example, given 80 coins known to contain a single counterfeit, which is lighter than the others, what is the minimum number of weighings needed to determine the counterfeit with certainty? You may get lucky and pick the counterfeit in the for the first go: but there's only a $\frac{1}{80}$ chance of that happening. There's also no need to check all $\binom{80}{1}$ combinations of pairs of coins. However, putting more than one coin on a pan reveals the same information - it's better to weigh groups of coins against each other.

Choose the groups so that each weighing can distinguish between the hypothesis that the pans will balance, or than there will be a heavier pan i.e. split the initial group into 3 (27, 27, 26). Continue this process recursively, splitting the remaining group into 3 each time, until you have found the counterfeit. If the two groups of 27 balance initially, take a coin from one of those groups and add it to the group of 26 to make a power of 3. This won't add any new information (you know this coin is not counterfeit) and so won't affect the inference.

The Group Testing problem can be formalised as follows: a set of items is given,

along with an upper bound on the number of defectives. The set is described as a vector, where if an item is 0 it is not defective and 1 if it is defective. Before the tests are run, the position of the 1's is unknown.

To find the defective items, a query is run against a subset of $[n]$, where the answer is defined as follows:

$$A(S) = 1 \sum_i x_i \geq 1 \quad (3.3.39)$$

Note that the addition is the binary-or in the above summation. The goal of Group Testing is to minimise the number of tests required to reconstruct the defective set.

Algorithms

An initial algorithm to consider is a simple binary search of the set to be tested. That is, given a set of size $N = 2^r$ i.e a powere of 2, we can recover a single defective in $\lceil \log_2 N \rceil$ tests.

To do this, create a new set of size $S = 2^{\lceil \log_2 n \rceil}$ which is guaranteed to contain a defective. Label the items with integers, and test the items in the sets $1, 2, \dots, S/2$ and $S/2 + 1, \dots, S/2$ separately. Then repeat the procedure on any groups which have a positive test.

To see why this testing procedure takes at most $\lceil \log_2 n \rceil$ tests, note that the procedure defines a binary tree over subsets of the N items, and so the depth of this tree is $\lceil \log_2 N \rceil$.

For input sets with more than a single defective (say K defectives) the binary search algorithm can be repeated, and each time a defective is found it is removed from the set. The binary search is then repeated, but on a set of size $N - 1$. Using this procedure we are guaranteed to find all the defectives in

$$K \lceil \log_2 N \rceil \leq K \log_2 N + K \quad (3.3.40)$$

tests. However, this is a very inefficient algorithm: early sets are large and so are likely to contain a defective.

The above algorithms return, with certainty, after at most $K \lceil \log_2 \binom{N}{K} \rceil$ tests, the defective set. Much work has gone into combinatorial search algorithms, often more complex than those described above.

This has been motivated by the analogy that the Group Testing problem can be considered a decoding problem where an experimenter receives a binary vector:

$$y = \mathbf{A}x \quad (3.3.41)$$

$y \in \{0,1\}^K, x \in \{0,1\}^N$, and wishes to decode the vector x to recover the defective set, subject to the constraints of the testing matrix \mathbf{A} . The matrix has to satisfy the property that the Boolean sum of any t columns was unique, and did not contain any other column in the matrix. These properties are known as separability and disjointness.

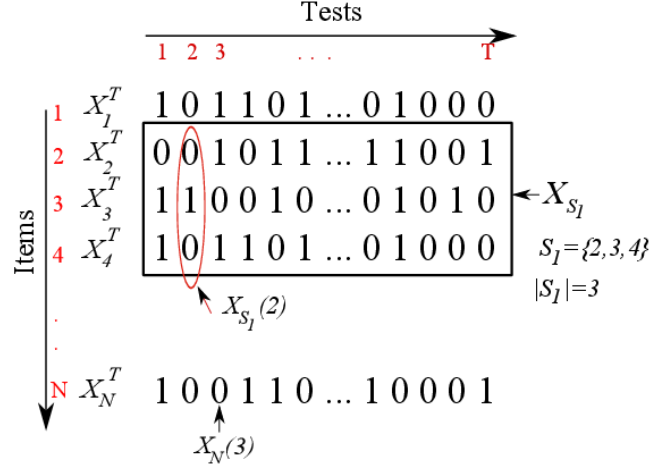


Figure 10: The Group Testing model: multiplication with a short, fat matrix [?]

See (REF HWANGDU) for more a detailed introduction and analysis of the requisite algorithms.

Hwang's Algorithm

In modern Coding Theory, there has been a move away from explicit combinatorial algorithms which return the codeword with certainty, towards probabilistic algorithms which return the correct codeword with an associated probability. The advantage of this has been the development of algorithms which can decode codes close to the Shannon capacity of the channel.

Similarly in Group Testing, the state of the art considers probabilistic algorithms instead of explicit combinatorial designs.

The problems with the binary search algorithm (that initial groups are very large and are highly likely to contain a defective) above can be overcome by instead considering groups whose size is chosen so that the probability that the group will have a positive test is close to half. Equivalently, given a set of size $N = 2^r$ which are known to contain K defectives, in expectation a group of size $\frac{K}{N}$ should contain a defective.

Thus we can use fewer tests than predicted by simple repeated binary search,

by testing 'pilot' groups of size roughly $\frac{K}{N}$. Hwang [?] gives such an algorithm, and provides an upper bound on the number of tests required to recover the defective set.

The steps for the algorithm are:

1. If $n \leq 2d - 2$ then test every item individually. Otherwise set $l = n - d + 1$ and define $\alpha := \log \lceil \frac{l}{d} \rceil$.
2. Test a group of size 2^α . If the outcome is negative, the group is good. Set $n := n - 2^\alpha$ and go to 1. If the outcome is positive, then use binary splitting on the group to identify a defective and x good items. Set $n := n - 1 - x$ and $d := d - 1$ and go to 1.

The upper bound on the number of tests is given by the following argument: as there are $\binom{n}{k}$ possible sets of defectives, and in t tests at most 2^t cases can be differentiated, $\lceil \log_2 \binom{n}{k} \rceil$ tests are needed.

Bounds

It has been previously believed that the success probability to recover the defective set given T tests was:

$$P(\text{Success}) \leq \frac{T}{\log_2 \binom{N}{K}} \quad (3.3.42)$$

However, a tighter upper bound has recently been found [?] at:

$$P(\text{Success}) \leq \frac{2^T}{\binom{N}{K}} \quad (3.3.43)$$

i.e. the probability of success increases exponentially with the number tests, opposed to linearly.

In the Group Testing literature there exists an 'adaptivity gap' - it seems that adaptive algorithms *do* give a performance improvement over non-adaptive algorithms, in terms of the number of tests required to recover the defective set. This is discussed here, using Hwang's algorithm as a test bed.

Hwang's algorithm is guaranteed to succeed in:

$$T = \log_2 \binom{N}{K} + K \quad (3.3.44)$$

tests. The Combinatorial Orthogonal Matching Pursuit algorithm, considered in [?], is guaranteed to recover the defective set with probability $N^{-\delta}$ in

$$T = ((1 + \delta) e) K \ln N \quad (3.3.45)$$

tests. For all N and K) we have:

$$K \log_2 \frac{N}{K} \leq \log_2 \binom{N}{K} \leq K \log_2 \frac{Ne}{K} \quad (3.3.46)$$

which follows from well-know bounds on binomial coefficients. This allows a contrast between the asymptotic bounds of previous algorithms to be considered in this section. We see that, the regime where $K = N^{1-\beta}$, Hwang's algorithm succeeds with:

$$T = \beta K \log_2 N + K (\log_2 e + 1) \quad (3.3.47)$$

tests, whilst the COMP algorithm succeeds with:

$$1.88(1 + \delta) K \log_2 N \quad (3.3.48)$$

tests. It's worthwhile to contrast these results, to gain some insight into the problem. 3.3 suggests that for very sparse problems (β tending towards 1) that Hwang's adaptive algorithm will outperform a simmlar non-adaptive algorithm. Even though the two procedures have the same complexity, they have different constants (1 v,s 1.88 in the sparse case). Thus, there are asymptotic gains (in terms of the number of tests required to recover the defective set) which are offered by adaptive algorithms, and not by non-adaptive algorithms.

These ideas can be summarised in the idea of a *capacity* for Group Testing [?]. That is, there is a constant C such that a sequence of Group Testing algorithms with $K = \omega(N)$ will succeed with probability tending to 1. This allows different noise, and dilution models to be considered so that a more complete characterisation of the structural properties of Group Testing is revealed.

3.3.1 Comparison to Compressive Sensing

The goal of Coding Theory is given a vector $x \in \mathbb{F}^m$, where \mathbb{F} is some finite-field, is to construct a 'code-book' C which produces a vector $y \in \mathbb{F}^n$, $n > m$, so that the original vector may be transmitted over a noisy-channel with vanishing error probability. This problem is structurally similar to the Compressive Sensing and Group Testing problems, but in reverse. In CS and GT we're given 'short' vector, and we wish to infer the 'longer' one satisfying the constraint that we seek the sparsest vector, under some conditions on the matrix Φ . This suggests that there may be some Information-theoretic framework uniting all three disciplines.

In [?] Tao and Candes consider the CS problem as one of error correction of a linear code: however in this case the codewords are drawn from \mathbb{R}^m as opposed to a finite alphabet more common in Coding Theory. This is done by considering Φ

as the parity check matrix of a linear code and the signal x as the error pattern. Linear programming can then be viewed as a method for decoding.

Group testing is a combinatorial variant of Compressive Sensing, where the sensing matrix is a binary matrix. The matrix represents combinations (or pools) of items, such that a 1 in the i^{th} row and j^{th} column means that the i^{th} item is tested in the j^{th} pool. The goal of Group testing can then be seen as designing testing pools so to accurately reconstruct the sparse set of interesting items.

In Group Testing, instead of the sensing matrix being subject to coherence constraints such as those above, the sensing matrices have the property that the support of any column is not contained in the union of the supports of any t other columns. Thus a t -disjunct matrix defines a group testing scheme which can identify any defective set up to size t .

The analogue between Group Testing and Coding is even closer, as GT explicitly considers signals and matrices from Binary alphabets. That is, Group Testing is a closer cousin of Coding Theory than Compressive Sensing, in a sense the inverse problem as in both Coding and Group Testing we are working over a finite field. This is encouraging, as it could allow the reconstruction of the defective set via methods developed in Coding Theory. There has been some work done on this, [?] considers the noisy Group Testing problem and the reconstruction of the defective set via belief propagation whilst [?] gives explicit theorems on conditions for the recovery of the defective set for the case of the binary symmetric channel. [?] takes this further and finds the capacity of Group Testing for a number of cases.

3.4 Sub-Nyquist Sampling techniques

This section presents some work on sampling methods for wide-band spectrum sensing

3.4.1 Wideband Modulated Converter

The sampling scheme proposed in [?] is capable of sampling wideband signals at rates below those predicted by Shannon-Nyquist sampling theory.

It works 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 N_{sig} uncorrelated transmissions (i.e. possible narrowband channels).

This process is repeated in parallel over M channels (unrelated to N_{sig} so that each band in x appears in baseband. The mixing functions are required to be periodic, with period T_p . Since p_i is periodic it has Fourier expansion:

$$p_i(t) = \sum_{l=-\infty}^{\infty} c_{il} \exp j l t \frac{2\pi}{T_p} \quad (3.4.49)$$

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 $x p_i$, with Fourier transform:

$$X_i(f) = \int_{-\infty}^{\infty} x(t) p_i(t) dt \quad (3.4.50)$$

$$= \sum_{l=-\infty}^{\infty} c_{il} X(f - l f_p) \quad (3.4.51)$$

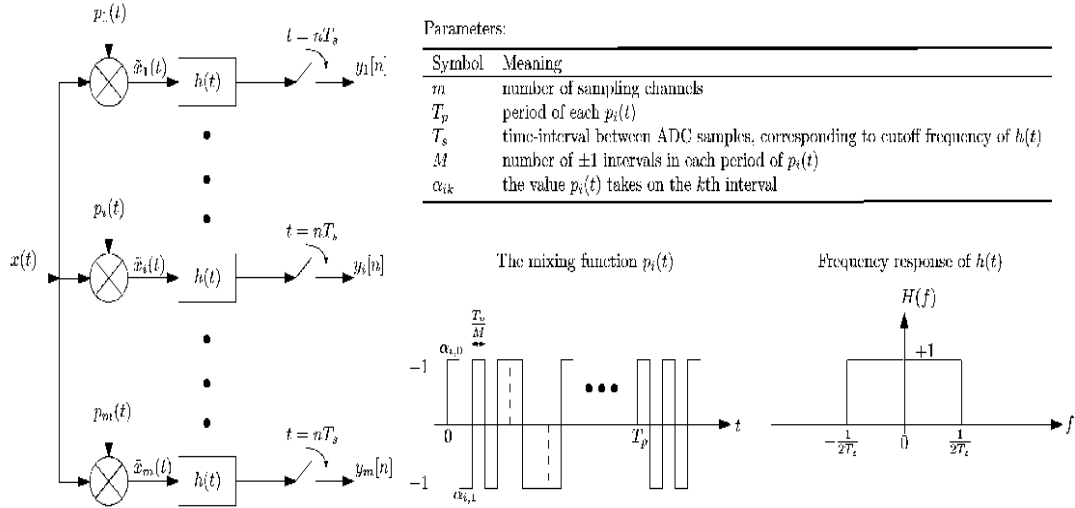


Figure 11: The operation of the Modulated Wideband Converter [?]

(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 it's support (we have assumed this Nyquist frequency exists, even though we never sample at that rate).

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:

$$Y_i(e^{j2\pi f T_s}) = \sum_{l=-L_0}^{+L_0} \quad (3.4.52)$$

since frequencies outside of $[-f_2/2, f_s/2]$ will be filtered out. L_0 is the smallest integer number of non-zero contributions in $X(f)$ over $[-f_2/2, f_s/2]$ - at most $\lceil f_N Y Q / f_p \rceil$ if we choose $f_s = f_p$. These relations can be written in matrix form as:

$$\mathbf{y} = \mathbf{A} \mathbf{x} \quad (3.4.53)$$

where \mathbf{y} contains the output of the WMC process, \mathbf{A} contains the Fourier coefficients of the mixing functions, and \mathbf{x} is the vector of unknown samples of $x(t)$.

4 Results and Simulations

To compare the efficacy of Group Testing and Compressive Sensing, Hwang's algorithm and the algorithm presented in [?] were simulated for a problem size of $N=1024$ and $K=10$. The problem was simulated 100 times and the cumulative distribution found - i.e. after how many tests or measurements were the respective problems solved? This allows the number of tests required by Group Testing to be compared to the number of measurements in Compressive Sensing. Figure 12 shows the results:

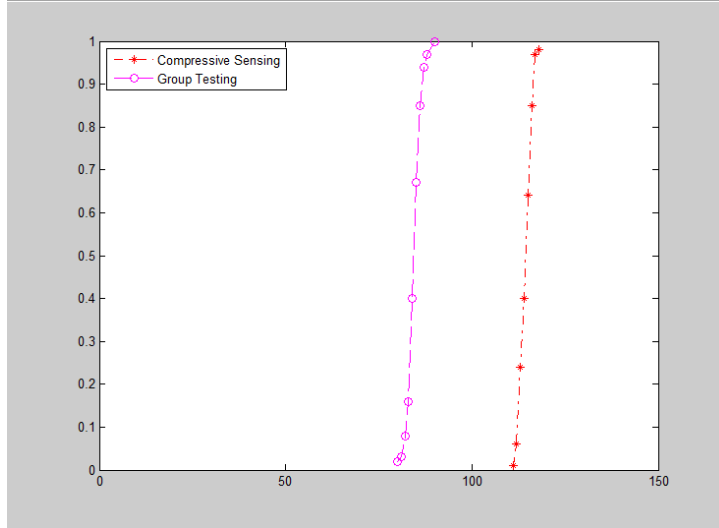


Figure 12: Group Testing vs Compressive Sensing

Note that both the algorithms meet their respective asymptotic bounds ($\log_2 \binom{N}{K}$ in the case of GT and $k \log N$ for CS). The main point of interest is that Group Testing requires roughly $\frac{2}{3}$ of the tests required by Compressive Sensing. This is encouraging: despite there being 'less' information - in the sense that the result

is a binary number as opposed to a real one - GT outperforms CS. Intuitively, one would conjecture the opposite - more information should allow you to locate the non-zero components faster. This justifies our interest in the problem, as the performance increase is substantial.

5 Further Work

Both the Compressive Sensing and Group Testing problems have been well studied for the case where there is no prior belief over where the non-zero components of the signal are (or equivalently, which items constitute the defective set). Despite some work being done solving the CS problem in the Bayesian paradigm, the prior distributions used are sparsifying (for example a Laplace prior), and not over the placement of the non-zero components.

As such, there is some 'low hanging fruit' to be picked: would using extra information about the likelihood of an item being defective improve the performance of recovery suitably defined? For example, would a Group Testing procedure require fewer tests if there was reason to believe that some subset of items was more likely to be defective. In the coin weighing example discussed above, the input set may contain a coin which is a different colour than the others - by testing this 'suspicious' coin in the first weighing we could find the counterfeit in a single test (otherwise we would find the counterfeit in at most five tests - one for the suspicious coin, and four following the procedure outlined in the Group Testing section).

This suggests that using prior information in could significantly reduce the number of tests required to recover the defective set, however if initial tests come back negative the performance of these algorithms could be worsened by using a non-iid distribution.

To try and characterise this numerical simulations and probabilistic analyses are required. The goal is to design algorithms which have better average case performance than current GT algorithms by incorporating prior beliefs, whilst at the same time do no worse than current performance guarantees. In the example above we'd seek an algorithm that could find the counterfeit coin in a single weighing (by testing the suspicious coin initially), yet could guarantee to find the counterfeit in at most four weighings of the pan balance.

For the 80 coins, 1 counterfeit discussed above this is probably not possible: adding an extra test which fails before proceeding with the standard approach will always be worse than simply assuming all coins are equally likely to be fake and proceeding accordingly. Do many cases like this exist, and what characterises

them? Will testing suspicious items initially always give a hit in performance should those tests fail?

A few directions into answering these questions are sketched below:

1. Implicit in the discussion of Hwang's algorithm above is that each item is defective with the same probability ($\frac{K}{N}$). Using prior knowledge over the input set we could test groups of size 2^β , where β is the empirical expectation of the input set. I.e. modify Hwang's algorithm to test groups in size roughly equal to the (empirical) expectation of the distribution over the input. This is a low risk activity and should take at most a fortnight - the Author has already implemented Hwang's algorithm, so all that needs to be done is to calculate the empirical mean of the input distribution, and characterise the algorithm in a few cases.
2. Re-formulate the problem in terms of source coding as per [?] and represent the items in the input set as a Huffman tree. Then the Group Testing algorithm will take place over the tree with the groups to test being the branches of the Huffman tree. This is a medium risk activity and should take a few months to solve.
3. It has usually been assumed that defective items are independent. Is it possible to relax this assumption, and if so what are the consequences of this for Group Testing algorithms? For example, what happens if defectives occur in clumps? This is a high risk activity, and should take over 6 months to begin to characterise.
4. Some work has been done considering Group Testing as the reverse of Coding [?], [?], but there is still much to be done. How close is the analogy? Coding theory has a wealth of noise cases - could these be imported into the Group Testing literature, and how useful will they be?
5. To make these models suitable for engineering applications, noisy cases must be considered. Again this is a high risk activity, but it can be considered in parallel to the 1st two cases, and so there should be some indication of a solution within a few months.
6. There currently exist no two-dimensional models of Group Testing in the literature. This is particularly relevant for the problem faced by Cognitive Radios: the radios could represent the measurement point on a geographic grid, and the 'groups' they need to infer are sets of available frequencies.

6 Project Plan

Item	Timeline
Modify Hwang's Algorithm for non-iid input distributions	The rest of October 2013
Reformulate as a source coding problem	October-December 2013
Correlations between defectives	October 2013-March 2014
Explore the connection (if any) between Coding and Group Testing	October 2013 - September 2014
Extend the Group Testing problem to 2-dimensional input sets	January-April 2014
Group Testing for Spectrum Sensing	January - September 2013