Introduction to Compressive Sensing

Collection Editor:
Marco F. Duarte

Introduction to Compressive Sensing

Collection Editor:

Marco F. Duarte

Authors:

Mark A. Davenport
Ronald DeVore
Marco F. Duarte
Chinmay Hegde
Jason Laska
Michael A Lexa
Shriram Sarvotham
Mona Sheikh
Wotao Yin

Online:

< http://cnx.org/content/col11355/1.2/ >

C O N N E X I O N S

Rice University, Houston, Texas

Collection structure revised: August 29, 2011
PDF generated: February 19, 2013
For copyright and attribution information for the modules contained in this collection, see p. 106.

Table of Contents

1 Analog Sampling Theory

1.1 The Shannon-Whitaker Sampling Theorem 1
1.2 Stable Signal Representations 3
1.3 Optimal Encoding 5
1.4 Kolmogorov Entropy 7
1.5 Optimal Encoding of Bandlimited Signals 8

2 Sparsity and Compressibilty

2.1 Introduction to vector spaces 11
2.2 Bases and frames 13
2.3 Sparse representations 14
2.4 Compressible signals 17

3 Compressive Sensing

3.1 Sensing matrix design 21
3.2 Null space conditions 22
3.3 The restricted isometry property 24
3.4 The RIP and the NSP 28
3.5 Matrices that satisfy the RIP 30
3.6 Coherence 32
3.7 Sub-Gaussian random variables 33
3.8 Concentration of measure for sub-Gaussian random variables 35
3.9 Proof of the RIP for sub-Gaussian matrices 38

4 ℓ_1-norm minimization

4.1 Signal recovery via ℓ_1-norm minimization 43
4.2 Noise-free signal recovery 45
4.3 Signal recovery in noise 47
4.4 Instance-optimal guarantees revisited 51
4.5 The cross-polytope and phase transitions 53
4.6 Sparse recovery algorithms 54
4.7 Convex optimization-based methods 55
4.8 Greedy algorithms 58
4.9 Combinatorial algorithms 62
4.10 Bayesian methods 64

5 Applications of Compressive Sensing

5.1 Linear regression and model selection 67
5.2 Sparse error correction 67
5.3 Group testing and data stream algorithms 68
5.4 Compressive medical imaging 69
5.5 Analog-to-information conversion 69
5.6 Single-pixel camera 72
5.7 Hyperspectral imaging 76
5.8 Compressive processing of manifold-modeled data 81
5.9 Inference using compressive measurements 85
5.10 Compressive sensor networks 88
5.11 Genomic sensing 90

Glossary 92
Bibliography 93

Index 104
Attributions 106

Chapter 1

Analog Sampling Theory

1.1 The Shannon-Whitaker Sampling Theorem[1]

The classical theory behind the encoding analog signals into bit streams and decoding bit streams back into signals, rests on a famous sampling theorem which is typically refereed to as the Shannon-Whitaker Sampling Theorem. In this course, this sampling theory will serve as a benchmark to which we shall compare the new theory of compressed sensing.

To introduce the Shannon-Whitaker theory, we first define the class of bandlimited signals. A bandlimited signal is a signal whose Fourier transform only has finite support. We shall denote this class as B_A and define it in the following way:

$$B_A := \{f \in L_2(\mathbb{R}) : \hat{f}(\omega) = 0,\ |\omega| \geq A\pi\}. \tag{1.1}$$

Here, the Fourier transform of f is defined by

$$\hat{f}(\omega) := \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} f(t)\, e^{-i\omega t}\, dt. \tag{1.2}$$

This formula holds for any $f \in L_1$ and extends easily to $f \in L_2$ via limits. The inversion of the Fourier transform is given by

$$f(t) := \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} \hat{f}(\omega)\, e^{i\omega t}\, d\omega. \tag{1.3}$$

Theorem 1.1: Shannon-Whitaker Sampling Theorem

If $f \in B_A$, then f can be uniquely determined by the uniformly spaced samples $f\left(\frac{n}{A}\right)$ and in fact, is given by

$$f(t) = \sum_{n \in \mathbb{Z}} f\left(\frac{n}{A}\right) \text{sinc}\left(\pi\left(\text{At} - n\right)\right), \tag{1.4}$$

where $\text{sinc}(t) = \frac{sint}{t}$.

Proof:

It is enough to consider $A = 1$, since all other cases can be reduced to this through a simple change of variables. Because $f \in B_{A=1}$, the Fourier inversion formula takes the form

$$f(t) = \frac{1}{\sqrt{2\pi}} \int_{-\pi}^{\pi} \hat{f}(\omega)\, e^{i\omega t}\, d\omega. \tag{1.5}$$

[1] This content is available online at <http://cnx.org/content/m15146/1.2/>.

Define $F(\omega)$ as the 2π periodization of $\hat{f}$,

$$F(\omega) := \sum_{n\in\mathbb{Z}} \hat{f}(\omega - 2n\pi). \tag{1.6}$$

Because $F(\omega)$ is periodic, it admits a Fourier series representation

$$F(\omega) = \sum_{n\in\mathbb{Z}} c_n e^{-in\omega}, \tag{1.7}$$

where the Fourier coefficients c_n given by

$$\begin{aligned} c_n &= \tfrac{1}{2\pi}\int_{-\pi}^{\pi} F(\omega)\, e^{in\omega}\, d\omega \\ &= \tfrac{1}{2\pi}\int_{-\pi}^{\pi} \hat{f}(\omega)\, e^{in\omega}\, d\omega. \end{aligned} \tag{1.8}$$

By comparing ((1.8)) with ((1.5)), we conclude that

$$c_n = \frac{1}{\sqrt{2\pi}} f(n). \tag{1.9}$$

Therefore by plugging ((1.9)) back into ((1.6)), we have that

$$F(\omega) = \frac{1}{\sqrt{2\pi}} \sum_{n\in\mathbb{Z}} f(n)\, e^{-in\omega}. \tag{1.10}$$

Now, because

$$\hat{f}(\omega) = F(\omega)\,\chi_{[-\pi,\pi]} = \frac{1}{\sqrt{2\pi}} \sum_{n\in\mathbb{Z}} f(n)\, e^{-in\omega}\chi_{[-\pi,\pi]}, \tag{1.11}$$

and because of the facts that

$$\begin{aligned} \mathcal{F}\left(\chi_{[-\pi,\pi]}\right) &= \tfrac{1}{\sqrt{2\pi}}\text{sinc}\,(\pi\omega) \qquad \text{and} \\ \mathcal{F}(g(t-n)) &= e^{-in\omega}\mathcal{F}(g(t)), \end{aligned} \tag{1.12}$$

we conclude

$$f(t) = \sum_{n\in\mathbb{Z}} f(n)\ \text{sinc}\,(\pi(t-n)). \tag{1.13}$$

Comments:

1. (Good news) The set $\{\text{sinc}\,(\pi(t-n))\}_{n\in\mathbb{Z}}$ is an orthogonal system and therefore, has the property that the L_2 norm of the function and its Fourier coefficients are related by,

$$\| f \|_{L_2}^2 = 2\pi \sum_{n\in\mathbb{Z}} |f(n)|^2 \tag{1.14}$$

2. (Bad news) The representation of f in terms of sinc functions is not a stable representation, i.e.

$$\sum_{n\in\mathbb{Z}} |\text{sinc}\,(\pi(t-n))| \approx \sum_{n\in\mathbb{Z}} \frac{1}{|t-n|+1} \to \text{divergences} \tag{1.15}$$

1.2 Stable Signal Representations[2]

To fix the instability of the Shannon representation, we assume that the signal is slightly more bandlimited than before

$$\hat{f}(\omega) = 0 \quad \text{for} \quad |\omega| \geq \pi - \delta, \;\; \delta > 0, \tag{1.16}$$

and instead of using $\chi_{[-\pi,\pi]}$, we multiply by another function $\hat{g}(\omega)$ which is very similar in form to the characteristic function, but decays at its boundaries in a smoother fashion (i.e. it has more derivatives). A candidate function $\hat{g}$ is sketched in Figure 1.1.

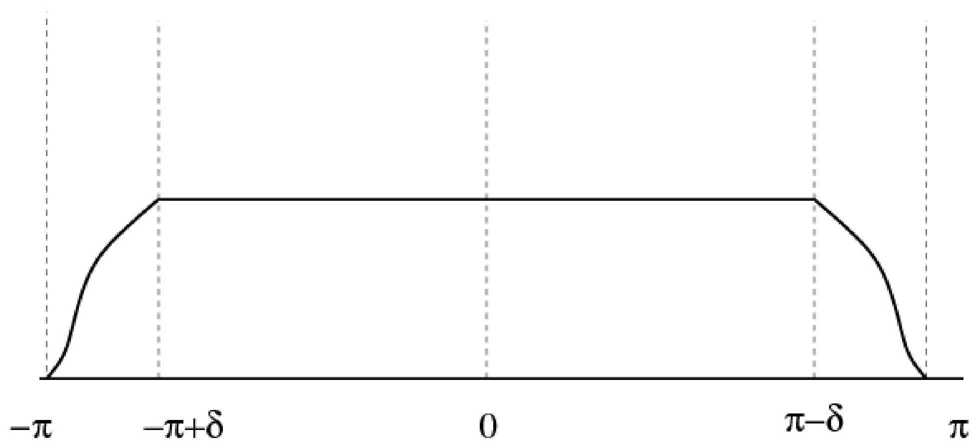

Figure 1.1: Sketch of $\hat{g}$.

Now, it is a property of the Fourier transform that an increased smoothness in one domain translates into a faster decay in the other. Thus, we can fix our instability problem, by choosing $\hat{g}$ so that $\hat{g}$ is smooth and $\hat{g}(\omega) = 1$, $|\omega| \leq \pi - \delta$ and $\hat{g} = 0$, $|\omega| > \pi$. By choosing the smoothness of g suitably large, we can, for any given $m \geq 1$, choose g to satisfy

$$|g(t)| \leq \frac{C}{(|t|+1)}^{m} \tag{1.17}$$

for some constant $C > 0$.

Using such a $\hat{g}$, we can rewrite () as

$$\hat{f}(\omega) = F(\omega)\,\hat{g}(\omega) = \frac{1}{\sqrt{2\pi}} \sum_{n\in\mathbb{Z}} f(n)\, e^{-in\omega}\,\hat{g}(\omega). \tag{1.18}$$

Thus, we have the new representation

$$f(t) = \sum_{n\in\mathbb{Z}} f(n)\;\, g(t-n), \tag{1.19}$$

where we gain stability from our additional assumption that the signal is bandlimited on $[-\pi - \delta, \pi - \delta]$.

[2]This content is available online at <http://cnx.org/content/m15144/1.1/>.

Does this assumption really hurt? No, not really because if our signal is really bandlimited to $[-\pi,\pi]$ and not $[-\pi-\delta,\pi-\delta]$, we can always take a slightly larger bandwidth, say $[-\lambda\pi,\lambda\pi]$ where λ is a little larger than one, and carry out the same analysis as above. Doing so, would only mean slightly oversampling the signal (small cost).

Recall that in the end we want to convert analog signals into bit streams. Thus far, we have the two representations

$$\begin{aligned} f(t) &= \sum_{n\in\mathbb{Z}} f(n)\ \mathrm{sinc}\,(\pi(t-n)), \\ f(t) &= \sum_{n\in\mathbb{Z}} f\left(\tfrac{n}{\lambda}\right)\ g(\lambda t-n). \end{aligned} \tag{1.20}$$

Shannon's Theorem tells us that if $f\in B_A$, we should sample f at the Nyquist rate A (which is twice the support of $\hat{f}$) and then take the binary representation of the samples. Our more stable representation says to slightly oversample f and then convert to a binary representation. Both representations offer perfect reconstruction, although in the more stable representation, one is straddled with the additional task of choosing an appropriate λ.

In practical situations, we shall be interested in approximating f on an interval $[-T,T]$ for some $T>0$ and not for all time. Questions we still want to answer include

1. How many bits do we need to represent f in $B_{A=1}$ on some interval $[-T,T]$ in the norm $L_\infty[-T,T]$?
2. Using this methodology, what is the optimal way of encoding?
3. How is the optimal encoding implemented?

Towards this end, we define

$$B_A := \{f\in L_2(\mathbb{R}) : |\hat{f}(\omega)| = 0, |\omega|\geq A\pi\}. \tag{1.21}$$

Then for any $f\in B_A$, we can write

$$f = \sum_n f\left(\frac{n}{A}\right)\cdot \mathrm{sinc}\ \pi(At-n). \tag{1.22}$$

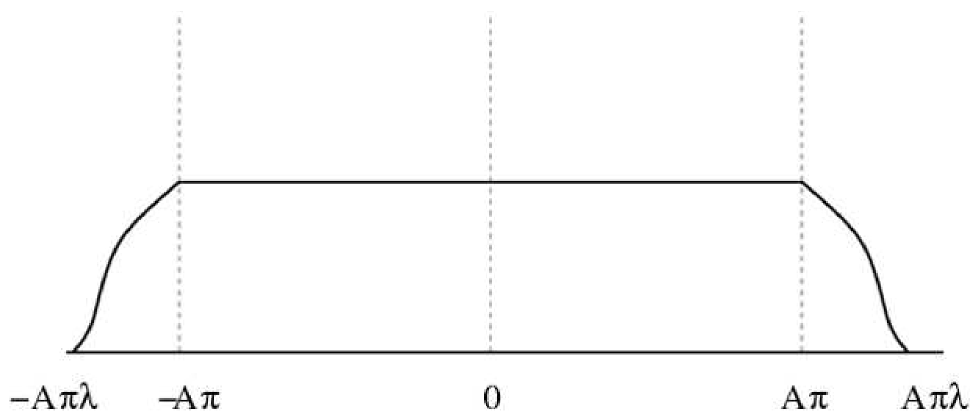

Figure 1.2: Fourier transform of $g_\lambda(\cdot)$.

In other words, samples at $0, \pm\frac{1}{A}, \pm\frac{2}{A}, \cdots$ are sufficient to reconstruct f. Recall also that $\mathrm{sinc}\,(x) = \frac{\sin(x)}{x}$ decays poorly (leading to numerical instability). We can overcome this problem by slight over-sampling. Say we over-sample by a factor $\lambda>1$. Then, we can write

$$f = \sum f\left(\frac{n}{\lambda A}\right) g_\lambda(\lambda At-n). \tag{1.23}$$

Hence we need samples at 0, $\pm\frac{1}{\lambda A}$, $\pm\frac{2}{\lambda A}$, etc. What is the advantage? Sampling more often than necessary buys us stability because we now have a choice for $g_\lambda(\cdot)$. If we choose $g_\lambda(\cdot)$ infinitely differentiable whose Fourier transform looks as shown in Figure 1.2 we can obtain

$$|g_\lambda(t)| \leq \frac{c_{\lambda,k}}{(1+|t|)}^{k}, \quad k = 1, 2, ... \tag{1.24}$$

and therefore $g_\lambda(\cdot)$ decays very fast. In other words, a sample's influence is felt only locally. Note however, that over-sampling generates basis functions that are redundant (linearly dependent), unlike the integer translates of the sinc $(\cdot)$ function.

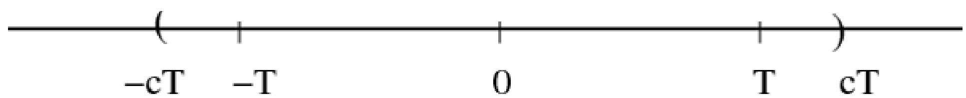

Figure 1.3: To reconstruct signals in $[-T, T]$, the sampling interval is $[-cT, cT]$.

If we restrict our reconstruction to t in the interval $[-T, T]$, we will only need samples only from $[-cT, cT]$, for $c > 1$ (see Figure 1.3), because the distant samples will have little effect on the reconstruction in $[-T, T]$.

1.3 Optimal Encoding[3]

We shall consider now the encoding of signals on $[-T, T]$ where $T > 0$ is fixed. Ultimately we shall be interested in encoding classes of bandlimited signals like the class B_A However, we begin the story by considering the more general setting of encoding the elements of any given compact subset K of a normed linear space X. One can determine the best encoding of K by what is known as the Kolmogorov entropy of K in X.

To begin, let us consider an encoder-decoder pair (E, D) E maps K to a finite stream of bits. D maps a stream of bits to a signal in X. This is illustrated in Figure 1.4. Note that many functions can be mapped onto the same bitstream.

[3] This content is available online at <http://cnx.org/content/m15139/1.1/>.

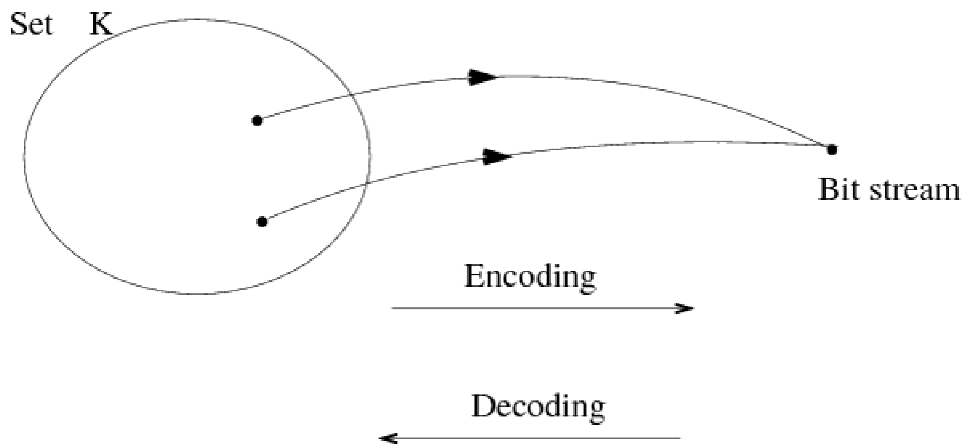

Figure 1.4: Illustration of encoding and decoding.

Define the distortion d for this encoder-decoder by

$$d(K,E,D,X) := \sup_{f\in K} \| f - D(Ef) \|_{\underline{X}}. \tag{1.25}$$

Let $n(K,E) = \sup_{f\in K} \# Ef$ where $\# Ef$ is the number of bits in the bitstream Ef. Thus n is the maximum length of the bitstreams for the various $f \in K$. There are two ways we can define optimal encoding:

1. Prescribe ε, the maximum distortion that we are willing to tolerate. For this ε, find the smallest $n_\varepsilon(K,X) := \inf_{(E,D)}\{n(K,E) : d(K,E,D,X) \leq \varepsilon\}$. This is the smallest bit budget under which we could encode all elements of K to distortion ε.
2. Prescribe N : find the smallest distortion $d(K,E,D,X)$ over all E,D with $n(K,E) \leq N$. This is the best encoding performance possible with a prescribed bit budget.

There is a simple mathematical solution to these two encoding problems based on the notion of Kolmogorov Entropy.

1.4 Kolmogorov Entropy[4]

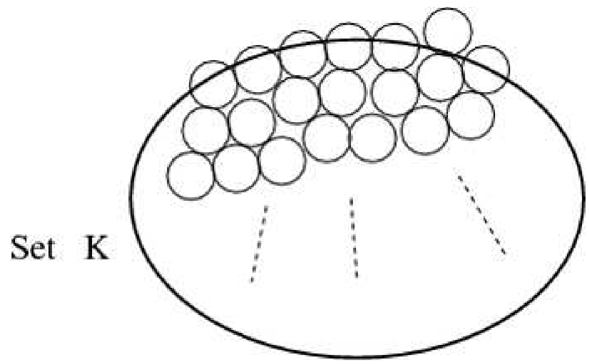

Figure 1.5: Coverings of K by balls of radius ε.

Given $\varepsilon > 0$, and the compact set K, consider all coverings of K by balls of radius ε, as shown in Figure 1.5. In other words,

$$K \subseteq U_{i=1}^{N} b\left(f_i, \varepsilon\right). \tag{1.26}$$

Let $N_\varepsilon := \inf \{N : \text{over all such covers}\}$. $N_\varepsilon(K)$ is called the covering number of K. Since it depends on X and K, we write it as $N_\varepsilon = N_\varepsilon(K, X)$.

Rule 1.1: Kolmogorov entropy
The Kolmogorov entropy, denoted by $H_\varepsilon(K, X)$, of the compact set K in X is defined as the logarithm of the covering number:

$$H_\varepsilon(K, X) = log N_\varepsilon(K, X). \tag{1.27}$$

The Kolmogorov entropy solves our problem of optimal encoding in the sense of the following theorem.

Theorem 1.2:
For any compact set $K \subset X$, we have $n_\varepsilon(K, X) = \lceil H_\varepsilon(K, X) \rceil$, where $\lceil \cdot \rceil$ is the ceiling function.

Proof:
Sketch: We can define an encoder-decoder as follows To encode: Say $f \in K$. Just specify which ball it is covered by. Because the number of balls is $N_\varepsilon\left(K, \underline{\overline{X}}\right)$, we need at most $\lceil log N_\varepsilon\left(K, \underline{\overline{X}}\right) \rceil$ bits to specify any such ball ball.

To decode: Just take the center of the ball specified by the bitstream.

It is now easy to see that this encoder-decoder pair is optimal in either of the senses given above.

□

The above encoder is not practical. However, the Kolmogorov entropy tells us the best performance we can expect from any encoder-decoder pair. Kolmogorov entropy is defined in the deterministic setting. It is the analogue of the Shannon entropy which is defined in a stochastic setting.

[4]This content is available online at <http://cnx.org/content/m15137/1.1/>.

1.5 Optimal Encoding of Bandlimited Signals[5]

We now turn back to the encoding of signals. We are interested in encoding the set

$$B_A(M) = \{f \in B_A : |f(t)| \leq M, t \in \mathbb{R}\} \tag{1.28}$$

where M is arbitrary but fixed. We shall restrict our discussion to the case where distortion is measured in $L_\infty[-T,T]$ where $T > 0$ is arbitrary but fixed. Then, $B_A(M)$ is a compact subset of L_∞: $B_A(M) \subseteq L_\infty[-T,T]$.

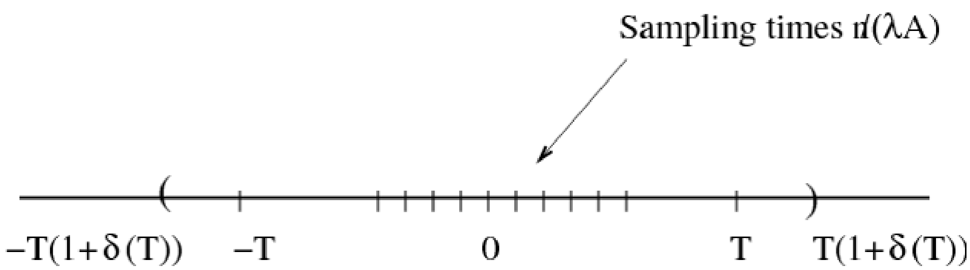

Figure 1.6: Sample points $\frac{n}{\lambda A}$ are chosen in the interval $[-T(1+\delta), T(1+\delta)]$.

We shall sketch how one can construct an asymptotically optimal encoder/decoder for B_A. The details for this construction can be found in .

We know $\hat{f}(\omega) = 0$ for $|\omega| \geq A\pi$, and $|f| \leq M$. How can we encode f in practice? We begin by chosing $\lambda = \lambda(T) > 1$ (see Figure 1.6) which will represent a slight oversampling factor we shall utilize. Given a target distortion $\varepsilon > 0$, we choose k so that $2^{-k-1} < \varepsilon \leq 2^{-k}$. Given f, we shall encode f by first taking samples $f\left(\frac{n}{\lambda A}\right)$ for $\frac{n}{\lambda A} \in [-T(1+\delta), T(1+\delta)]$ where $\delta(T) > 0$. In other words, we sample f on a slightly larger interval than $[-T,T]$. For each sample $f\left(\frac{n}{\lambda A}\right)$, we shall use the first $k + k_0(T)$ bits of its binary expansion. In other words, our encoder takes f and the samples $f\left(\frac{n}{\lambda A}\right)$ and then assigns to $f\left(\frac{n}{\lambda A}\right)$ the first $k + k_0(T)$ bits of this number.

To decode, the receiver would take the bits and construct the approximation $\overline{f}\left(\frac{n}{\lambda A}\right)$ to $f\left(\frac{n}{A\lambda}\right)$ from the bits provided. Notice that we have the accuracy

$$\left|f\left(\frac{n}{\lambda A}\right) - \overline{f}\left(\frac{n}{\lambda A}\right)\right| \leq 2^{-k-k_0} \cdot M. \tag{1.29}$$

We utilize the function g_λ satisfying () to define

$$\overline{f}(t) = \sum_{n \in N_T} \overline{f}\left(\frac{n}{\lambda A}\right) g_\lambda(\lambda A t - n), \tag{1.30}$$

where

$$N_T := \{n : \ -T(1+\delta) \leq \frac{n}{\lambda A} \leq T(1+\delta)\}. \tag{1.31}$$

We then have

$$\begin{aligned} |f(t) - \overline{f}(t)| \ &\leq \ \textstyle\sum_{n \in N_T} \left|f\left(\frac{n}{\lambda A}\right) - \overline{f}\left(\frac{n}{\lambda A}\right)\right| \cdot |g_\lambda(\lambda A t - n)| \\ &\quad + \textstyle\sum_{\left|\frac{n}{\lambda A}\right| > T(1+\delta)} \left|f\left(\frac{n}{\lambda A}\right)\right| \cdot |g_\lambda(\lambda A t - n)| \end{aligned} \tag{1.32}$$

[5] This content is available online at <http://cnx.org/content/m15140/1.2/>.

The term $\left|f\left(\frac{n}{\lambda A}\right)-\overline{f}\left(\frac{n}{\lambda A}\right)\right|$ that appears in the first summation in ((1.32)) is bounded by $M\cdot 2^{-k-k_0}$. The term $\left|f\left(\frac{n}{\lambda A}\right)\right|$ that appears in the second summation in the same equation is bounded by M. Therefore,

$$\begin{array}{rcl}\left|f(t)-\overline{f}(t)\right| & \leq & \sum_{n\in N_T} M\cdot 2^{-k-k_0}\cdot\left|g_\lambda(\lambda At-n)\right| \\ & & +\sum_{\left|\frac{n}{\lambda A}\right|>T(1+\delta)} M\cdot\left|g_\lambda(\lambda At-n)\right| =: S_1+S_2\end{array} \quad (1.33)$$

We can estimate S_1 by

$$\begin{array}{rcl}S_1 & = & \sum_{n\in N_T} M\cdot 2^{-k-k_0}\cdot\left|g_\lambda(\lambda At-n)\right| \\ & \leq & M\cdot 2^{-k-k_0}\cdot\sum_n\left|g_\lambda(\lambda At-n)\right| \\ & \leq & M\cdot C_0(\lambda)\cdot 2^{-k-k_0} \qquad \text{(because } g(\cdot) \text{ decays fast)}\end{array} \quad (1.34)$$

Therefore, if we choose k_0 sufficiently large, then $S_1\leq M\cdot C_0(\lambda)\cdot 2^{-k-k_0}\leq\frac{\varepsilon}{2}$. The second summation S_2 can also be bounded by $\varepsilon/2$ by using the fast decay of the function g_λ (see ()).

To make the encoder/decoder specific we need to precisely define δ and λ. It turns out that the best choices (in terms of bit rate performance on the class B_A) depend on T. But $\delta_T\to 0$ and $\lambda_T\to 1$ as $T\to\infty$. Recall that Shannon sampling requires $2T\lambda A$ samples. Since our encoder/decoder uses $k+k_0$ bits per sample, the total number of bits is $(k+k_0)\cdot 2\lambda AT(1+\delta)$, and so coding will require roughly k bits per Shannon sample.

This encoder/decoder can be proven to be optimal in the sense of averaged performance as we shall now describe. The average of performance of optimal encoding is defined by

$$\lim_{T\to\infty}\frac{n_\varepsilon\left(B_A(M),L_\infty\lfloor -T,T\rfloor\right)}{2T} \quad (1.35)$$

If we replace the optimal bit rate n_ε in ((1.35)) by the number of bits required by our encoder/decoder then the resulting limit will be the same as that in ((1.35)).

In summary, to encode band limited signals on an interval $[-T,T]$, an optimal strategy is to sample at a slightly higher rate than Nyquist and on a slightly large interval than $[-T,T]$. Each sample should then be quantized by using the binary expansion of the sample. In this way, for an investment of k bits per Nyquist rate sample, we get a distortion of 2^{-k}.

To get a feel for the number of bits required by such an encoder, let us say $A=10^6$ (signals band limited to 1Mhz). Say $T=24$ hours $\approx 10^5$ seconds, and $k=10$ bits. Then, $A\cdot k\cdot 2T=10^6\cdot 10\cdot 10^5=10^{12}$ bits. This is too BIG!

The above encoding is is known as Pulse Coded Modulation (PCM). In practice, people frequently use another encoder called Sigma-Delta Modulation. Instead of oversampling just slightly, Sigma Delta over samples a lot and then assign only one (or a few) bits per sample.

Why is Sigma-Delta preferred to PCM in practice? There are two reasons commonly given:

1. Getting accurate samples, quantization, etc. is not practical because of noise. For better accuracy, we need more expensive hardware.
2. Noise shaping. In Sigma-Delta, the distortion is higher but the distortion is spread over frequencies outside of the desired range.

In PCM, the distortion decays exponentially (like 2^{-k}), whereas for Sigma-Delta, the distortion decays like a polynomial (like $\frac{1}{k^m}$). Although the distortion decays faster in PCM, the distortion in Sigma-Delta is spread outside the desired frequency range.

Chapter 2

Sparsity and Compressibilty

2.1 Introduction to vector spaces[1]

For much of its history, signal processing has focused on signals produced by physical systems. Many natural and man-made systems can be modeled as linear. Thus, it is natural to consider signal models that complement this kind of linear structure. This notion has been incorporated into modern signal processing by modeling signals as *vectors* living in an appropriate *vector space*. This captures the linear structure that we often desire, namely that if we add two signals together then we obtain a new, physically meaningful signal. Moreover, vector spaces allow us to apply intuitions and tools from geometry in $\mathbb{R}^3$, such as lengths, distances, and angles, to describe and compare signals of interest. This is useful even when our signals live in high-dimensional or infinite-dimensional spaces.

Throughout this course[2], we will treat signals as real-valued functions having domains that are either continuous or discrete, and either infinite or finite. These assumptions will be made clear as necessary in each chapter. In this course, we will assume that the reader is relatively comfortable with the key concepts in vector spaces. We now provide only a brief review of some of the key concepts in vector spaces that will be required in developing the theory of compressive sensing[3] (CS). For a more thorough review of vector spaces see this introductory course in Digital Signal Processing[4].

We will typically be concerned with *normed vector spaces*, i.e., vector spaces endowed with a *norm*. In the case of a discrete, finite domain, we can view our signals as vectors in an N-dimensional Euclidean space, denoted by $\mathbb{R}^N$. When dealing with vectors in $\mathbb{R}^N$, we will make frequent use of the ℓ_p norms, which are defined for $p \in [1, \infty]$ as

$$\|x\|_p = \begin{cases} \left(\sum_{i=1}^{N} |x_i|^p\right)^{\frac{1}{p}}, & p \in [1, \infty); \\ \underset{i=1,2,\ldots,N}{max} |x_i|, & p = \infty. \end{cases} \tag{2.1}$$

In Euclidean space we can also consider the standard *inner product* in $\mathbb{R}^N$, which we denote

$$< x, z >= z^T x = \sum_{i=1}^{N} x_i z_i. \tag{2.2}$$

This inner product leads to the ℓ_2 norm: $\|x\|_2 = \sqrt{< x, x >}$.

In some contexts it is useful to extend the notion of ℓ_p norms to the case where $p < 1$. In this case, the "norm" defined in (2.1) fails to satisfy the triangle inequality, so it is actually a quasinorm. We will also

[1] This content is available online at <http://cnx.org/content/m37167/1.6/>.
[2] *An Introduction to Compressive Sensing* <http://cnx.org/content/col11133/latest/>
[3] "Introduction to compressive sensing" <http://cnx.org/content/m37172/latest/>
[4] *Digital Signal Processing* <http://cnx.org/content/col11172/latest/>