Sampling and Quantization (and Reconstruction)

Güntürk, Spring 2010,
Scribe: Evan Chou

Table of contents

Week 1 (1/25/2010)

Overview of Sampling and Quantization

The Big Picture: The goal is to find/study/analyze methods to efficiently and effectively describe analog objects (continuous objects, e..g functions / signals) by digital (discrete, quantized) representations.

For example, we could be dealing with audio signals or visual signals like images or video. The context is data acquisition (analog-to-digital (A/D) conversion) and digital representation (storage, transmission).

More specifically, the fundamental question is the following:

Given a bit budget (storage space), how well can we represent objects in a given class of objects?

The class of objects is an important consideration. In the silliest case, if we only have to distinguish between two objects, then we can simply use 1 bit, representing one object with 0 and the other with 1.

Concerns:

- *Accuracy* - Given the representation we use, we should be able to recover the original object as "closely" as possible. This involves the notion of distance, or a metric.
- *Efficiency* - We want to be able to compute the representation quickly. So we will deal with algorithms and computation.
- *Robustness* - The representation should be resilient to noise, uncertainty.

Fundamental Notions:

- Rate (R) - What is the bit budget?
- Distortion (D) - How accurate is the representation?

R and D are inversely related: a larger bit budget allows for a smaller distortion.

Example 1. Suppose we are representing real numbers in $[0, 1]$ with a bit budget of r bits. What is the best distortion? Using the truncated binary representation, we can represent a number in $[0, 1]$ with distortion 2^{-r}.

A different example would be some metric space of objects (X, d). Suppose we want to represent some compact subset K with a bit budget of r bits. One way is to cover K with 2^r balls of some radius D, and represent each point of K by the center of one of the balls containing the point. Note we have thus represented K by 2^r points, which can be described by r bits.

Since the balls have fixed radius we will incur a distortion of D. In this case we want to minimize D such that we can still cover K with 2^r balls.

The tools we have are Sampling, Quantization, and Reconstruction. In a block diagram,

$$f \to \boxed{\text{Sampling}} \to (y_k)_{k\in\Lambda} \to \boxed{\text{Quantization}} \to (q_k)_{k\in\Lambda} \to \boxed{\text{Reconstruction}} \to \tilde{f}$$

f is the signal we want to represent. Then we sample in some way to reduce the complexity. In the common setting, $y_k = f(t_k)$ for $k \in \Lambda$ and some choice of t_k. Note that $f \mapsto y_k$ is linear. We will first study linear processes for sampling and reconstruction.

The quantization step is nonlinear. The sequence (y_k) is transformed to another sequence (q_k) where $q_k \in \mathcal{A}$ and $\mathcal{A}$ is the quantization alphabet, for instance $\mathbb{Z}$ or $\{-1, 1\}$. It will not necessarily be the case that each y_k is mapped to q_k directly, i.e. it is not necessarily true that $q_k = F(y_k)$ for some function F.

Then finally, given the quantized coefficients we then represent some approximation to f.

Quantization is in general very nonlinear. If we are working with a very fine resolution, for instance if $\mathcal{A} = \delta\mathbb{Z}$ for small δ, then it is possible for the quantization to be close to linear. This is not too interesting for us, as generally we have a fixed bit budget and therefore there will be some fixed resolution. We say that we work with "Coarse Quantization".

First let us consider various examples to fix ideas about sampling and reconstruction.

Example 2. (Lagrange Interpolation) Let us look at the space of polynomials of degree $\leq d$, i.e. $p \in \mathcal{P}_d$, so that $p(x) = \sum_{j=0}^{d} a_j t^j$. Then we sample at $d+1$ points $t_0 < t_1 < ... < t_d$, so $y_k = p(t_k)$, $k = 0, ..., d$. Then given these samples, we can reconstruct p exactly. We can either use a linear system of equations (which gives a Vandermonde matrix), or more directly, Lagrange interpolation.

Use $l_k(t) = \prod_{j=0, j\neq k}^{d} \frac{t - t_j}{t_k - t_j}$ so that $l_k(t_j) = \delta_{jk}$. Then $p(t) = \sum_{k=0}^{d} y_k\, l_k(t)$. So, we can see this as a sampling process $Sp = y$ where $y_k = p(t_k)$. The mapping $p \mapsto p(t_k)$ is a linear functional, so the process is linear. Then the reconstruction process gives $R(y) = \sum\ y_k\ l_k$. In this case we have reconstructed the polynomial exactly.

Note that if we used fewer than $d+1$ samples, we are *undersampling*, which means that there are many possible candidates for what p could have been, and so we may not be able to reconstruct p.

If we use more than $d+1$ samples, we are *oversampling*, which means that there are now multiple possible representations for p (can use Lagrange interpolation on any subset of the samples of size $d+1$).

Example 3. (Classical Shannon Sampling Theorem) As an aside, the sampling theorem is not due to Shannon. Shannon simply popularized the idea. It was known much earlier by Whittaker, Kotel'nikov, Nyquist, and even Cauchy had some form of sampling theorem.

We are interested in sampling band-limited functions

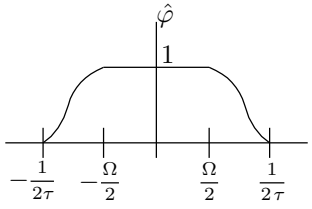
$$B_\Omega = \left\{ f : \mathbb{R} \to \mathbb{R},\, f \in L^2, \operatorname{supp}(\hat{f}) \subset \left[-\frac{\Omega}{2}, \frac{\Omega}{2} \right] \right\}$$

which is also known as the Paley-Wiener Space. Ω is called the bandwidth. Considering band-limited functions is a natural assumption for audio signals, since the audible range for the human ear is between 20 Hz and 20 kHz.

Above $\hat{f}$ is the Fourier transform, and we will be using the definition $\hat{f}(\xi) = \int f(t)\, e^{-2\pi i \xi t}\, dt$ with inverse $f(x) = \int \hat{f}(\xi) e^{-2\pi i \xi t}\, d\xi$. This gives the nicest form of Plancherel (without constants).

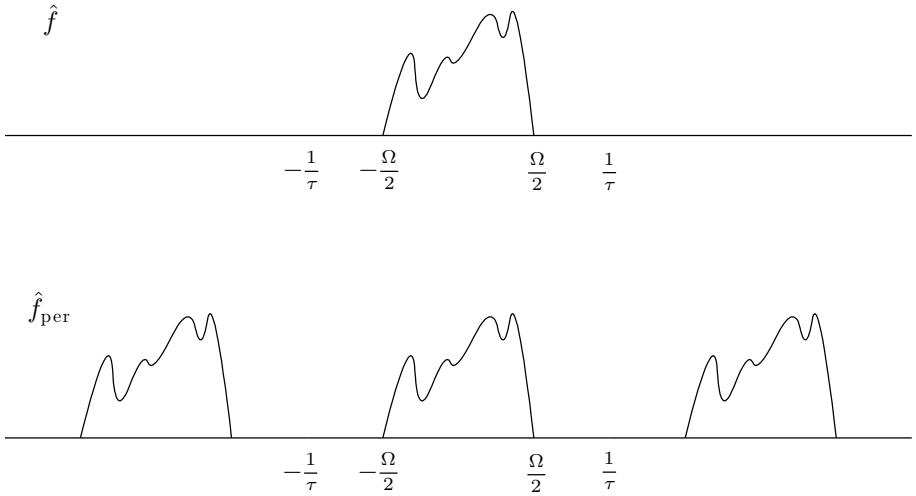
Consider taking samples $y_k = f(k\tau)$ where $\tau > 0, k \in \mathbb{Z}$. τ small corresponds to higher rate of sampling. The result of sampling theorem is that if $\tau \leq \frac{1}{\Omega}$, then f can be recovered exactly from its samples (y_k). The reconstruction procedure is the following:

Let φ satisfy $\hat{\varphi}(\xi) = \begin{cases} 1 & |\xi| \leq \frac{\Omega}{2} \\ 0 & |\xi| > \frac{1}{2\tau} \end{cases}$ where the values in the range $\frac{\Omega}{2} \leq |\xi| \leq \frac{1}{2\tau}$ can be arbitrary, though smoothness is preferred. This is a cutoff function in frequency.



Then $f(t) = \tau \sum_{k\in\mathbb{Z}} f(k\tau)\, \varphi(t - k\tau)$.

Proof. The first step is to periodize $\hat{f}$, with $\hat{f}_{\text{per}}(\xi) = \sum_{l\in\mathbb{Z}} \hat{f}(\xi + l\,\frac{1}{\tau})$. This makes copies of $\hat{f}$ so that $\hat{f}_{\text{per}}$ is $\frac{1}{\tau}$ -periodic:



Then using the Fourier series, we can write

$$\hat{f}_{\text{per}}(\xi) = \sum_{k\in\mathbb{Z}} a_k\, e^{-2\pi i k\tau\xi}$$

where

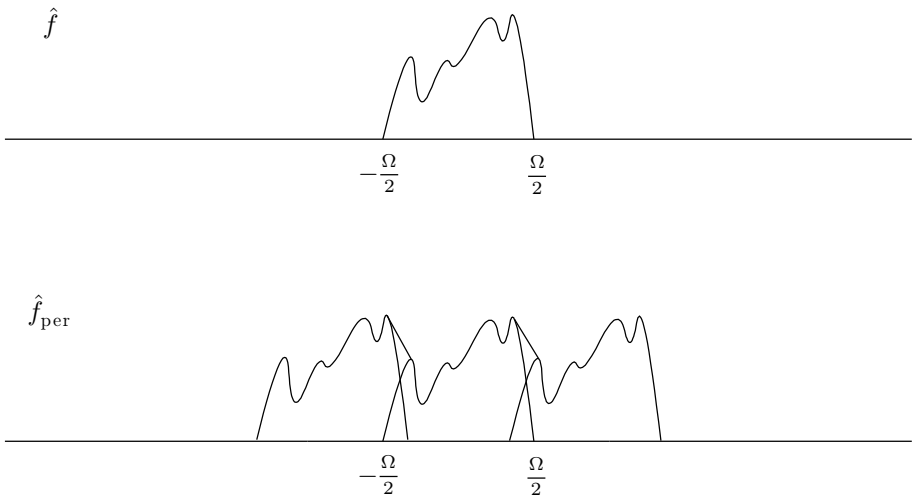
$$a_k = \tau \int_{-\frac{1}{2\tau}}^{\frac{1}{2\tau}} \hat{f}_{\text{per}}\, e^{2\pi i k\tau\xi}\, d\xi = \tau \int_{-\infty}^{\infty} \hat{f} e^{2\pi i k\tau\xi}\, d\xi = \tau f(k\tau)$$

noting that $\hat{f} = \hat{f}_{\text{per}}$ on $\left[-\frac{1}{2\tau}, \frac{1}{2\tau}\right]$ and $\hat{f} = 0$ outside $\left[-\frac{\Omega}{2}, \frac{\Omega}{2}\right]$. Continuing we have that

$$\begin{aligned} \hat{f}(\xi) &= \hat{f}_{\text{per}}(\xi)\hat{\varphi}(\xi) \\ &= \tau \sum_{k\in\mathbb{Z}} f(k\tau)\, e^{-2\pi i k\tau\xi}\, \hat{\varphi}(\xi) \\ f(t) &= \tau \sum_{k\in\mathbb{Z}} f(k\tau)\, \varphi(t - k\tau) \end{aligned}$$

as desired. □

We can also do this procedure for when $\frac{1}{\tau} < \Omega$ (undersampling), but will get "aliasing", the periodic copies will overlap::



In this case, it is not possible in general to recover the original function. There are multiple possibilities for f that may have caused the aliasing (for instance, using $\hat{f}$ with support in $\left[-\frac{1}{2\tau}, \frac{1}{2\tau}\right]$).

If $\frac{1}{\tau} > \Omega$, the oversampling case, we note from the constructions above that many possiblities for φ exist, even when $\frac{1}{\tau}$ is close to Ω (though some constants may blow up).

If $\frac{1}{\tau} = \Omega$, this is called critical sampling, and in this case $\hat{\varphi} = \mathbf{1}_{\left[-\frac{\Omega}{2}, \frac{\Omega}{2}\right]}$, or $\varphi = \Omega \operatorname{sinc}(\Omega t)$, where $\operatorname{sinc}(x) = \frac{\sin \pi t}{\pi t}$. In this case, the sinc kernel decays slowly $\sim \frac{1}{t}$, though it is in L^2. The representation $f = \sum \tau f(k\tau)\, \varphi(t - k\tau)$ converges in L^2, but it may not converge pointwise, and so the reconstruction is not local. Locality is necessary for robustness, and ideally the representation should be absolutely summable (uniform convergence in this case). In the oversampling case, we can choose φ so that we have very good decay, and then the kernel becomes more localized.

Example 4. (Relation between Shannon Sampling and Lagrange Interpolation) Use spacing $t_k = k$ (for simplicity let $\Omega = \tau = 1$). Recall the interpolant $l_k(t) = \prod_{j=1,\, j\neq k}^{d} \frac{t-j}{k-j}$. We take the limit as $d \to \infty$. Then

$$\begin{aligned}
\lim_{d\to\infty} l_k(t) &= \prod_{j\neq k} \frac{t-j}{k-j} \\
&= \prod_{j\neq k} \left(1 - \frac{t-k}{j-k}\right) \\
&= \prod_{n\neq 0} \left(1 - \frac{t-k}{n}\right) \\
&= \operatorname{sinc}(t-k)
\end{aligned}$$

So the sampling theorem in the critical sampling case is like the limit of polynomial interpolation.

General Setting. Suppose we have samples $y_k = \langle f, \varphi_n \rangle$ with linear functionals φ_n which are overcomplete, in the sense that

$$f = \sum_n \langle f, \varphi_n \rangle \psi_n$$

for many choices of ψ_n, called dual functionals. The dual picture in this case is that given ψ_n that span the space of interest, there are many functionals φ_n such that

$$f = \sum_n \langle f, \varphi_n \rangle \psi_n$$

This flexibility is great for quantization and noise. Let us now introduce quantization into the picture. The samples y_k are mapped to q_k, and we want the reconstruction

$$\tilde{f} = \sum q_k \psi_k$$

to be as close to $f = \sum y_k \psi_k$ as possible. One option is to simply "round" each y_k to the nearest $\mathcal{A}$. But this does not fully take advantage of oversampling.

If we consider the error $e = f - \tilde{f} = \sum (y_k - q_k)\psi_k = R(y-q)$, then we can rephrase this problem as finding q so that $y - q$ is as close to $\ker(R)$ as possible. In general, note it is not likely that we can find q so that $y - q \in \ker(R)$, as the possible choices for q are very restrictive!

Concrete Example. In the context of sampling theorem,

$$f(t) = \tau \sum y_k \varphi(t - k\tau) = [R(y)](t)$$

Then for some sequence $z = (z_k)$,

$$\widehat{Rz}(\xi) = \left[\tau \sum z_k e^{-2\pi i k \tau \xi}\right] \hat{\varphi}(\xi)$$

Thus, if $Rz = 0$, then $\widehat{Rz} = 0$, which occurs if $\sum_k z_k e^{-2\pi i k \tau \xi} = 0$ whenever $\xi \in \operatorname{supp}(\hat{\varphi}) \subset \left[-\frac{1}{2\tau}, \frac{1}{2\tau}\right]$, i.e. only high frequency components are present. Thus $\ker(R)$ has "high-pass" sequences, and we want to arrange q so that $y - q$ is a "high-pass" sequence, so that the low frequency components of y, q cancel. This may not even be possible, for instance in the situation of critical sampling:

Given any orthonormal system ψ_k, we note we have Parseval's

$$\left\| \sum (y_k - q_k)\psi_k \right\|_{L^2}^2 = \sum |y_k - q_k|^2$$

and thus to minimize LHS, we have no choice but to take $q_k \approx y_k$ (Greedy choice for q), so

$$q_k = \operatorname{argmin}_{r \in \mathcal{A}} |y_k - r|$$

How to design q_k? The idea is called $\Sigma\Delta$-modulation. Try picking $q_k \in \mathcal{A}$ so that

$$y_k - q_k = u_k - u_{k-1} = (\Delta u)_k$$

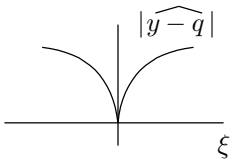
for some bounded sequence u_k. First, let's see why such a $y - q$ is a high-pass sequence:

$$\begin{aligned} \widehat{y-q} &= \sum (y_k - q_k) e^{ik\xi} \\ &= \sum (u_k - u_{k-1}) e^{ik\xi} \\ &= \left(\sum u_k e^{ik\xi}\right)(1 - e^{i\xi}) \end{aligned}$$

The last equality is from reindexing u_{k-1} to u_k after splitting the sum. Taking absolute values, we have that

$$|\widehat{y-q}|(\xi) = \left|\sum u_k e^{ik\xi}\right| 2\left|\sin\frac{\xi}{2}\right| = C\left|\sin\frac{\xi}{2}\right|$$

This shows that low frequencies (those close to 0) are vanishing:



Now to choose q so that $y - q = \Delta u$, we can use a Greedy algorithm. Suppose q_0, ..., q_{k-1} has been selected. We then set

$$q_k = \text{argmin}_{r\in\mathcal{A}}|u_{k-1} + y_k - r|$$

(this can be done backwards as well). This procedure works under certain conditions. One simple one is the following:

Exercise 1. Consider $y_k \in [-1, 1]$ and $\mathcal{A} = \{-1, 1\}$. Then with this procedure, $|u_k| \leq 1$ for all k.

Proof. This is done inductively. Note if u_{k-1} is bounded by 1, then $u_{k-1} + y_k \in [-2, 2]$. Then we choose $q_k = \text{sgn}(u_{k-1} + y_k)$ to push back $u_k = u_{k-1} + y_k + q_k \in [-1, 1]$. □

Overview of Compressive Sampling

Usually we deal with the situation where S, R are linear processes. However, recently there has been rapid developments with "undersampling" with S, R not necessarily linear. This goes under the many names: Compressed Sensing (Donoho), Compressive Sampling (Candés, Tao), Compressive Sensing (Rice U. Group).

We deal with the setting in which information is "sparse" with respect to some dictionary of functions, or basis, which is often the case. For instance, even though images contain many pixels, using Discrete Cosine Bases (JPEG), or Wavelet Bases (JPEG-2), we can drastically reduce the number of coefficients, since in general only a few coefficients with respect to these bases are large. Thus currently we collect all the pixels of the image, compute the representation with respect to these more efficient bases, and toss out all the small coefficients. This is a relatively wasteful procedure, and we ask whether we can sample less.

Given some basis, every signal corresponds to the coefficients with respect to the basis,

$$f = \sum x_k \psi_k$$

We say that f is sparse if there are only a few nonzero coefficients. More specifically, we say f is s-sparse if $\#\{i \leq j \leq N, x_j \neq 0\} \leq s$. Note that $\Sigma_s^N = \{x \in \mathbb{R}^N, x \text{ is } s-\text{sparse}\}$ is a nonlinear set. In fact, due to the fact that two signals may be sparse on two different sets, we have

$$\Sigma_s^N \subset \Sigma_s^N + \Sigma_s^N \subset \Sigma_{2s}^N$$

Now taking measurements $y_j = \varphi_j(x)$, for $j = 1, ..., m$. Note $m \geq s$ or we have no hope of recovering the function, and if $m = s$, we need φ_j to be exactly the entries for which the coefficients of f are nonzero. Unfortunately, we do not know which φ_j these are, and so we need $m > s$. For appropriately chosen φ_j, it turns out that we can use $m \ll N$ measurements so that exact recovery is possible. Even though this is "undersampling" with respect to the ambient dimension N, it is actually "oversampling" with respect to the sparsity level. We will discuss this is more detail next time.

Week 2 (2/1/2010)

Overview of Compressive Sampling (continued)

Essentially compressive sampling is a finite dimensional theory (though it can be extended to infinite dimensional settings), but deals with very high dimensional signals.

Setting: $f \in \mathbb{R}^N$ where N is very large. Given the signal f, we then take generalized "samples", not necessarily sampling the coordinates, but instead taking linear functionals

$$y_k = \langle \varphi_k, f \rangle,\ k = 1, ..., m$$

which we call "measurements" (that's what m stands for). We will be thinking of $m \ll N$. We can represent this in matrix form with

$$\Phi = \begin{pmatrix} - & \varphi_1 & - \\ & \vdots & \\ - & \varphi_m & - \end{pmatrix}$$

and so $y = \Phi f$.

To talk about sparsity, we consider a particular basis ψ_1, ..., ψ_N of $\mathbb{R}^N$, usually orthogonal, for instance "Fourier" bases (sines and cosines for real space theory), or Wavelets, or splines.

We will use the notation

$$\Psi = \begin{pmatrix} | & & | \\ \psi_1 & \dots & \psi_N \\ | & & | \end{pmatrix}$$

to describe both the matrix and the basis. We then assume f is S-sparse with respect to Ψ, so that

$$f = \sum_{k=1}^{N} x_k \psi_k$$

where most of the x_k are zero, i.e. $x \in \Sigma_S^N = \{x \in \mathbb{R}^N, \text{at most } S \text{ of the } x_k \text{ are nonzero}\}$. Or more generally, we consider f to be "compressible", meaning that if we reorder x in decreasing magnitude, x^*, there is some power law decay:

$$x_j^* \lesssim C j^{-\sigma},\ j = 1, ..., N$$

where the constant C is independent of N. This is analogous to "weak-L^p" spaces from real analysis, where

$$\mu\{|f|^p \geq M\} \leq \frac{C}{M^p}$$

Now we can represent the measurements on f in terms of the basis where f is sparse:

$$y = \Phi f = \Phi \Psi x$$

Note that we may just as well consider x to be the signal, and $\Phi\Psi$ as the measurement operator, which is still $m \times N$.

Thus, there is no loss in generality if we consider Ψ to be the identity, i.e. that f is sparse in the standard basis for $\mathbb{R}^N$. From now on we will assume that x is the signal and Φ is the measurement matrix where we take $y = \Phi x$.

First, we ask the question: Under what conditions for Φ is it possible to recover a sparse signal x from the measurements Φx?

This is answered by the following Proposition:

Proposition 5. *Let $N \geq 2S$. Then Φ is one to one on Σ_S^N if and only if $\ker \Phi \cap \Sigma_{2S}^N = \{0\}$.*

Proof. ($\Longrightarrow$) Suppose Φ is one to one on Σ_S^N, and let $x \in \ker \Phi \cap \Sigma_{2S}^N$. Since x is $2S$ sparse, we can write $x = \alpha + \beta$ where α, β are S sparse. Then since $x \in \ker \Phi$, $0 = \Phi x = \Phi\alpha + \Phi\beta$ so that $\Phi\alpha = \Phi(-\beta)$. However, since Φ is one to one on Σ_S^N and both $\alpha, -\beta$ are S-sparse, we have that $\alpha = -\beta$ so that $x = 0$.

($\Longleftarrow$) The proof is essentially reversible. Now suppose that $\ker \Phi \cap \Sigma_{2S}^N = \{0\}$, and suppose that $\Phi\alpha = \Phi\beta$ where $\alpha, \beta \in \Sigma_S^N$. We show that $\alpha = \beta$. Note $\Phi(\alpha - \beta) = 0$ and $\alpha - \beta$ is $2S$ sparse. Thus $\alpha - \beta = 0$ so that $\alpha = \beta$. □

The condition $\ker \Phi \cap \Sigma_{2S}^N$ is equivalent to saying that when x is $2S$ sparse, and $x \neq 0$, then $\Phi x \neq 0$. This implies that if we look at any $2S$ columns of Φ, they are linearly independent. If the support of x is $T \subset \{1, ..., N\}$ where $|T| \leq 2S$, then we note that Φx is a linear combination of the columns of Φ corresponding to T, which we will denote Φ_T.

There is a simple example of a matrix satisfying this condition.

Example 6. Consider $m = 2S$, and let $t_1 < t_2 < ... < t_N$ be fixed. Set $\Phi_{ij} = t_j^{i-1}$. Then any $2S$ columns of Φ form a Vandermonde matrix, which is invertible.

$$\Phi = \begin{pmatrix} 1 & 1 & \cdots & \cdots & 1 \\ t_1 & t_2 & \cdots & \cdots & t_N \\ \vdots & \vdots & & & \vdots \\ t_1^{2S-1} & t_2^{2S-1} & \cdots & \cdots & t_N^{2S-1} \end{pmatrix}$$

This is not practical due to numerical instability in inverting Vandermonde matrices. Even then, it is not clear how to recover a given sparse vector x from its measurements Φx, since we need do not know a priori what the support of x is.

To put things into context, in signal processing on images and audio, for instance, we want to find a compact representation, exploiting the fact that the signals are sparse with respect to a suitable basis. However, we do not know what the support of the signal is with respect to the basis, and so we end up computing all the coefficients, and then throwing out the coefficients that are near zero. This procedure is considered to be adaptive, since which samples we take depend on the signal.

Compressive Sampling bypasses this process of computing all the coefficients, with fewer nonadaptive samples.

Continuing on, given a measurement matrix Φ which satisfies injectivity on Σ_S^N, we need to know how to recover a sparse signal $x \in \Sigma_S^N$ from the measurements Φx. As mentioned earlier, the difficulty lies in determining the support. One possibility is to enumerate all $\binom{N}{2S}$ possibilities for supports $T \subset \{1, ..., N\}$ with $|T| \leq 2S$, and computing $\tilde{x} = \Phi_T^{-1} y$. If $\tilde{x}$ is S-sparse (out of $2S$ total coordinates), then we have found the solution (recall that there is a unique S-sparse solution since Φ is injective on Σ_S^N). Of course, this is very intractable, since for N large $\binom{N}{2S}$ blows up exponentially.

Under more restrictive assumptions, however, there does exist an efficient algorithm.

Restricted Isometry Property (RIP), introduced by Candés, Romberg, Tao [CRT]. Φ is said to have RIP of order (δ, s), denoted RIP(δ, s) with $0 \leq \delta < 1$, if

$$(1 - \delta)\|x\|_2 \leq \|\Phi x\|_2 \leq (1 + \delta)\|x\|_2 \text{ for all } x \in \Sigma_s^N$$

In other words, any s columns of Φ are nearly orthogonal. We will use the notation $\delta_s(\Phi)$ to denote the smallest δ for which RIP(δ, s) holds fixing Φ, s. Note that if Φ satisfies RIP$(\delta, 2s)$, then Φ is injective on Σ_s^N, so the RIP condition is stronger.

So far we have discussed how we sample data with Φ. Now we discuss how to reconstruct the data from the samples.

Recall that in inverting $\Phi x = y$, the difficulty is in finding the support of x, or in other words, which s columns of Φ were used to generate y. Then as a first idea, given y let's guess which s columns are used in a greedy fashion. First, pick the column which best approximates y, i.e. if we choose v_1, then there exists c_1 for which $\|c_1 v_1 - y\| \leq \|dv - y\|$ for any other column v and scalar d. Then looking at the residual $r_1 = y - c_1 v_1$, we repeat on the other columns, finding the best column which approximates r_1. This is known as "Greedy Pursuit" or "Matching Pursuit". Unfortunately, this doesn't always terminate in s steps, so it produces a solution that is more than s sparse.

A slight generalization that does better is called "Orthogonal Matching Pursuit". The idea is similar here, except having chosen j columns $v_1, ..., v_j$, we first project y onto the span of $\{v_1, ..., v_j\}$ and then compute the residual $r_j = y - P_j y$ and find the column v_{j+1} which best approximates r_j.

l^1 minimization

There is a different method which works well, and that is to use l^1 minimization.

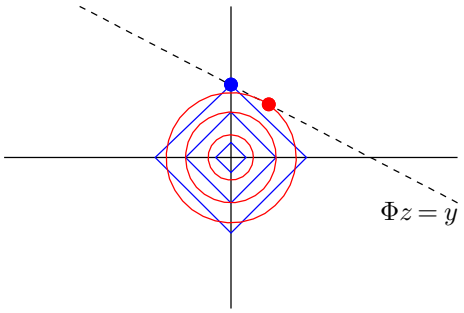
The initial approach is to seek the sparsest possible solution z to $\Phi z = y$, where $y = \Phi x$ and x is s-sparse. Note that such a z will also be s-sparse, and thus $z = x$ by injectivity of Φ on Σ_s^N. A quantity that meausres sparsity is called the l^0 "norm" (not a norm), $\|z\|_0 := |\text{supp}(z)|$. Though it is not a norm, it is the limit of l^p as $p \to 0$, which are "quasinorms". So we want

$$\mathcal{P}_0\text{: } \min \|z\|_0 \text{ s.t. } \Phi z = y$$

This is not easy, as $\{x\text{: } \|x\|_0 = c\}$ is non-convex. Instead, we consider a convex relaxation, but still close to $\|z\|_0$, and that is the l^1 relaxation

$$\mathcal{P}_1\text{: } \min \|z\|_1 \text{ s.t. } \Phi z = y$$

It is known in general that the l^1 minimization gives rise to sparse solutions, but the question is whether it is sparse enough? Also, note that l^2 minimization is much easier to compute, but does not give rise to sparse solutions. As an illustration of the difference between l^1 and l^2 minimization, consider $m = 1, N = 2$, so $\Phi z = y$ describes a line.



In the illustration above, the blue squares are level sets for the l^1 norm, i.e. $\{x\text{: } \|x\|_1 = r\}$, and to find the l^1 minimizer for $\Phi z = y$, we increase r until the first time we touch the solution space. Likewise, the same procedure is done in red with the l^2 norm. Here the l^2 minimizer is not sparse, but the l^1 minimizer is. Note Φ does not satisfy RIP or injectivity, since there are two one-sparse solutions, but this still illustrates the essential difference.

The problem $\mathcal{P}_1$ is a convex optimization problem, which is still not too bad. For one, we can recast the problem in terms of a linear program by introducing slack variables.

$$\mathcal{P}_1 \Longleftrightarrow \min \sum u_i \text{ s.t. } -u_i \leq z_i \leq u_i, \Phi z = y$$

where we note that $\|z\|_1 = \min_{-u_i \leq z_i \leq u_i} \sum u_i$. This increases the dimensionality of the problem to $2N$, but has efficient algorithms such as simplex and interior point methods.

Alternatively, we can use iterataively reweighted least squares (IRLS), which uses weighted l^2 norms with appropriately chosen weights to approximate behavior of the l^1 norm. Under a suitable RIP assumption, it can be shown that it converges to a $\mathcal{P}_0$ solution.

In general, l^1 minimization goes by the name "Basis Pursuit".

The natural question is then to ask when $\mathcal{P}_0 = \mathcal{P}_1$. For our purposes, we want the two to be true for all sparse vectors.

Theorem 7. *For all $x \in \Sigma_s^N$, x is the unique solution to $\mathcal{P}_1$ if and only if for all $\eta \in \ker \Phi$ and sparsity set $T \subset \{1, ..., N\}$ with $|T| \leq s$, that*

$$\|\eta_T\|_1 < \|\eta_{T^c}\|_1$$

($\eta_T := \eta \mathbf{1}_T$).

Remark 8. This property has been termed the "Null Space Property" (NSP), but says that we cannot have vectors in $\ker \Phi$ that are highly concentrated in s coordinates, and in particular, there cannot be vectors with sparsity $\geq 2s$ in $\ker \Phi$ or else we can take T to be the largest s components of such vectors, contradicting NSP.

Proof. ($\Longleftarrow$) Assume Φ has NSP. Let $x \in \Sigma_s^N$ and $y = \Phi x$. We know that all solutions z satisfy $z = x + \eta$ for $\eta \in \ker \Phi$. Then we show that $\|z\|_1 \geq \|x\|_1$. Let $T = \text{supp}(x)$. Then

$$\begin{aligned} \|z\|_1 &= \|(x+\eta)_T\|_1 + \|(x+\eta)_{T^c}\|_1 \\ &= \|x + \eta_T\|_1 + \|\eta_{T^c}\|_1 \\ &> \|x + \eta_T\|_1 + \|\eta_T\|_1 \\ &\geq \|x + \eta_T - \eta_T\|_1 \\ &= \|x\|_1 \end{aligned}$$

($\Longrightarrow$) Now pick any $\eta \in \ker \Phi$ and $T \subset \{1, ..., N\}$ with $|T| \leq s$, and suppose that x is the unique solution to $\mathcal{P}_1$ for all $x \in \Sigma_s^N$. We write $\eta = \eta_T + \eta_{T^c}$, so that $0 = \Phi(\eta) = \Phi(\eta_T) - \Phi(-\eta_{T^c})$. Let $y = \Phi(\eta_T) = \Phi(-\eta_{T^c})$, so both η_T and $-\eta_{T^c}$ solve $\Phi z = y$. Since η_T is s sparse, η_T is the unique l^1 minimizer for $\Phi z = y$, so

$$\|\eta_T\|_1 < \|-\eta_{T^c}\|_1 = \|\eta_{T^c}\|_1$$

□

So now we have a complete characterization for the equivalence of $\mathcal{P}_0$ and $\mathcal{P}_1$ for all $x \in \Sigma_s^N$. Note that the NSP is not easy to verify, since we have to consider all possible support sets. Currently there is work in trying to use optimization techniques to verify NSP.

It turns out that a sufficiently strong RIP condition implies NSP.

Generalized NSP: Given $0 < \gamma \leq 1$, we say that Φ has NSP(γ, s) if

$$\|\eta_T\|_1 \leq \gamma \|\eta_{T^c}\|_1$$

for all $\eta \in \ker \Phi$ and $T \subset \{1, ..., N\}$ with $|T| \leq s$. Smaller γ is desirable for robustness when handling compressible signals or noisy measurements / quantization.

Theorem 9. *If Φ has* RIP$(\delta, J + J')$*, then Φ has* NSP$\left(\frac{1+\delta}{1-\delta}\sqrt{\frac{J}{J'}}, J\right)$.

Remark 10. To interpret this result, we want δ to be near zero, and for a given sparsity level J, taking J' large enough gives NSP for smaller γ.

This theorem brings RIP into the picture, but how do we find Φ satisfying RIP? It turns out that choosing Φ randomly gives a high probability of satisfying RIP.

Theorem 11. *Let $\Phi_{i,j} \sim N(0, 1/m)$ for $1 \leq i \leq m$ and $1 \leq j \leq N$. If $m \geq c s \log \frac{N}{s}$, with c a constant depending only on δ, then Φ satisfies* RIP(δ, s) *with probability $1 - e^{-Cm}$, where C is an absolute constant.*

What is important about the theorem is how m scales with respect to s. Note that in particular $m \ll N$ and is essentially linear in s. Before this result, there were previous results which needed $m \sim s^2$, and these can be proven by studying diagonal dominance in $\Phi_T^* \Phi$. We can use Gershgorin circle theorem to place the spectrum so that $\sigma(\Phi_T^* \Phi) \subset [1 - \delta, 1 + \delta]$. This leads to explicit combinatoric constructions that achieve RIP, but m is stuck in the regime $m \sim s^2$. In fact, this barrier cannot be broken using diagonal dominance arguments, we need to understand the cancellations that can occur in the off diagonal terms. So it is still an open question to look at deterministic constructions that satisfy RIP with a good range of m.

For how to implement compressive sampling in practice, random constructions for Φ poses some problems. Since we have been assuming that the signal x is sparse with respect to the standard basis, we need to transfer the result back to f where the original signal is. In this setting $y = \Phi f = \Phi \Psi x$, and to obtain the samples for y, we select $(\Phi\Psi)$ randomly so that RIP is achieved, and then compute $(\Phi\Psi)\Psi^{-1}$ to obtain the measurement matrix Φ for f. Usually Ψ^{-1} can be implemented efficiently for certain bases, for instance FFT for Fourier.

- The first issue is how to even store $(\Phi\Psi)$ and communicate this to whoever is doing the sampling or reconstruction. This can be resolved by using pseudorandomly generated $(\Phi\Psi)$ based on a seed. We can store the seed used to generate the matrix so that it can be reproduced elsewhere.

- The second issue is the efficiency of computing $(\Phi\Psi)\Psi^{-1} f$, since $(\Phi\Psi)$ may be a very dense matrix. This is resolved by using *structured* random matrices:

    - Randomly selecting rows from the full FFT matrix (selecting frequency components at random) satisfies RIP, but the regime is for $m \geq c\, s\, (\log N/s)^4$, which is not too bad.

    - Other orthogonal systems can be used as well.

    - Perhaps we want a measurement matrix that satisfies some sort of causality, in that we can compute measurements on the fly, as the input arrives (streaming). These require banded matrices, and for instance random Toeplitz matrices have been used (behave like convolutions, same entries along diagonals)

The proofs of all these results use sophisticated machinery, probability, geometry of Banach spaces, and recently there have been many constructions based on graph theory or number theory. In the random construction for Φ satisfying RIP, it has also been shown that it works using $\Phi_{ij} \sim \text{Bern}(p, \{-1, 1\})$ with the same result (for m).

It also becomes an issue for what "deterministic" construction actually means. For instance, in the Bernoulli random matrix case above, there are 2^{nN} possible ± 1 matrices, and with high probability they satisfy RIP. If I enumerate all of these, at some point we will have an RIP matrix. This should not be considered a deterministic algorithm, however.

Extension to Compressible and Noisy Case

In reality,

- signals are not strictly sparse but compressible.
- measurements are noisy, or even quantized

So we will have $y = \Phi x + e$ for some error e. In this case, it turns out that l^1 minimization still works well!

Theorem 12. (Candés, Romberg, Tao) *Assume Φ satisfies RIP$(\delta, 2s)$ for δ sufficiently small (for instance, take $\delta < 1/4$). Then the following holds for all $x \in \mathbb{R}^N$. Given $y = \Phi x + e$, where e is unknown but $\|e\|_2 \leq \varepsilon$, consider the problem*

$$\min \|z\|_1 \text{ s.t. } \|\Phi z - y\|_2 \leq \varepsilon$$

(this is a second order cone program, still convex optimization). Let $x^\sharp$ be the solution. Then

$$\|x^\sharp - x\|_2 \leq C_1 \varepsilon + C_2 \frac{\sigma_s(x)_{l^1}}{\sqrt{s}}$$

where C_1, C_2 only depend on δ and $\sigma_s(x)_{l^1} = \min_{v \in \Sigma_s^N} \|x - v\|_{l^1}$, the best s-sparse approximant in l^1.

Here l^2 norm is used since for Gaussian noise typically we have good control over the l^2 norm. Note that this generalizes previous results. If x happens to be s-sparse, then the second term drops. If there is no noise, then the first term drops.

It is known that this result is near optimal for all perturbations $\|e\|_2 \leq \varepsilon$. However, it is not known if the result is optimal for quantized perturbations, which we have some degree of control over.

There is an interesting application of compressive sampling in image processing. A 1-pixel camera has been built, where light enters and hits an array of mirrors, which can be controlled to either reflect light or not. Whatever is reflected is then read by a single detector. There are m measurements each made with random setting of the mirrors, and then the image is reconstructed using l^1 minimization. It works okay, and similar apparatus may have applications for instance to the invisible spectrum (the reason digital cameras work well is because the silicon is sensitive to the visible light). Compressive sampling can also be applied for processing ultra wide-bandwidth signals, in which we cannot afford to oversample, and must undersample. However the signals do not take up the full bandwidth, so compressive sampling is well suited for this problem.

From a mathematical standpoint, compressive sampling has led to many new ideas, and is very interesting. From a practical standpoint, we do not know whether compressive sampling is the future of technology. That remains to be seen.

Week 3 (2/8/2010)

Introduction to Frames

We will be covering frames for finite dimensions, where every result is essentially linear algebra. The ideas extend to infinite dimensions, but require technical details such as convergence issues.

Let $\mathcal{H}$ be a finite dimensional inner product space, with $\dim \mathcal{H} = n$. Let $\{f_1, ..., f_m\} \subset \mathcal{H}$ be a collection of vectors that span $\mathcal{H}$. Of course, we must have $m \geq n$ for this to hold.

Recall that in the case where f_k form an orthonormal basis, we have a nice representation for any $f \in \mathcal{H}$:

$$f = \sum_{k=1}^{n} \langle f, f_k \rangle f_k$$

In general, we can take a similar approach, looking at the inner products with f_k and piecing them back together in some way. We define the "analysis operator" $T^*: \mathcal{H} \to \mathbb{C}^m$ by

$$T^*(f) = (\langle f, f_k \rangle)_{k=1}^m$$

The "synthesis operator" $T: \mathbb{C}^m \to \mathcal{H}$ is defined by

$$T(c) = \sum_{k=1}^{m} c_k f_k$$

Note that T, T^* are adjoints of each other:

$$\begin{aligned} \langle Tc, g \rangle_{\mathcal{H}} &= \sum_{k=1}^{m} c_k \langle f_k, g \rangle \\ &= \sum_{k=1}^{m} c_k \overline{\langle g, f_k \rangle} \\ &= \left\langle c, (\langle g, f_k \rangle_{\mathcal{H}})_{k=1}^m \right\rangle_{\mathbb{C}^m} \\ &= \langle c, T^* g \rangle_{\mathcal{H}} \end{aligned}$$

If we identify $\mathcal{H}$ with $\mathbb{C}^n$, the two operators can be written as matrices:

$$T = \begin{pmatrix} | & & | \\ f_1 & \cdots & f_m \\ | & & | \end{pmatrix}, \; T^* = \begin{pmatrix} — & f_1^* & — \\ & \vdots & \\ — & f_m^* & — \end{pmatrix}$$

Then we define

$$Sf = TT^* f = \sum_{k=1}^{m} \langle f, f_k \rangle f_k$$

(In the orthonormal case, $S = I$), and S is called the "frame operator" associated to $(f_1, ..., f_m)$. Note that

$$\langle Sf, f \rangle = \sum_{k=1}^{m} |\langle f, f_k \rangle|^2 = \|T^* f\|_2^2$$

Thus, S is a positive, self-adjoint ($S = TT^*$) operator. Let's find simple bounds on the operator norm of S. By Cauchy-Schwarz,

$$\sum_{k=1}^{m} |\langle f, f_k \rangle|^2 \leq \left(\sum_{k=1}^{m} \|f_k\|^2 \right) \|f\|^2$$

To get a lower bound, consider the map $\varphi(f) = \sum_{k=1}^{m} |\langle f, f_k \rangle|^2$ restricted to the set $S_1 = \{\|f\| = 1\}$. Note that φ is a continuous function of f, since it is a sum of squares of continuous functions of f. Then since S_1 is compact, f attains its minimum on S_1. Thus there is a unit vector u for which

$$\sum_{k=1}^{m} |\langle f, f_k \rangle|^2 \geq \varphi(u)$$

for all $\|f\| = 1$. By scaling, we have that

$$\sum_{k=1}^{m} |\langle f, f_k \rangle|^2 \geq \varphi(u) \, \|f\|^2$$

for arbitrary f. This lower bound will only be meaningful if $\varphi(u) > 0$ (gives invertibility, for one), and we note that this is indeed the case. For suppose $\varphi(u) = 0$. Then $\langle u, f_k \rangle = 0$ for all f_k, and hence $u = 0$. But this contradicts $\|u\| = 1$.

In summary, we can find two constants $0 < A \leq B < \infty$ for which

$$A \, \|f\|^2 \leq \sum_{k=1}^{m} |\langle f, f_k \rangle|^2 \leq B \, \|f\|^2$$

The optimal bounds A, B are called the lower and upper frame bounds, respectively. This condition turns out to be a good definition for frames in infinite dimensional spaces. In finite dimensions, the spanning condition gives this ineqeuality for free. We can think of this inequality as a Parseval-like inequality (Parseval is when $A = B = 1$, for orthonormal bases).

The frame lower bound shows that S is invertible, since

$$\langle Sf, f \rangle \geq A \, \|f\|^2$$

i.e. $Sf = 0$ implies $f = 0$, and in finite dimensions injectivity is equivalent to surjectivity.

Frame Decomposition

We can use the invertibility of S to get representations for f:

$$\begin{aligned} f &= SS^{-1}f \\ &= \sum_{k=1}^{m} \left\langle S^{-1}f, f_k \right\rangle f_k \\ &= \sum_{k=1}^{m} \left\langle f, S^{-1}f_k \right\rangle f_k \end{aligned}$$

Letting $g_k = S^{-1}f_k$, we have the representation $f = \sum_{k=1}^{m} \langle f, g_k \rangle f_k$. Note in general there are many other representations as well, and we will address this issue shortly. First, we note that we can get another reprsentation of f in terms of g_k:

$$\begin{aligned} f &= S^{-1}Sf \\ &= S^{-1} \sum_{k=1}^{m} \langle Sf, f_k \rangle f_k \\ &= \sum_{k=1}^{m} \langle Sf, f_k \rangle g_k \end{aligned}$$

which we consider as a "dual representation". We call $(g_k)_{k=1}^m$ the "canonical" dual frame to $(f_k)_{k=1}^m$.

Now if f_k are linearly dependent, then there are many representations for f of the form

$$f = \sum_{k=1}^{m} c_k f_k$$

In particular, if we take any $\eta \in \ker(T)$ (i.e. $\sum_{k=1}^m \eta_k f_k = 0$), we have that

$$f = \sum_{k=1}^{m} (\langle f, g_k \rangle + \eta_k) f_k$$

for any $\eta \in \ker(T)$. What makes $(\langle f, g_k \rangle)_{k=1}^m$ special is that it minimizes the l^2 norm of all such coefficients.

Proposition 13. *Among all $f = \sum_{k=1}^m c_k f_k$, $\sum_{k=1}^m |c_k|^2$ is minimized if and only if $c_k = \langle f, g_k \rangle$.*

Proof. Let $d_k = \langle f, g_k \rangle$. We will show that $c = (c-d) + d$ where $c - d \perp d$, from which it follows that

$$\|c\|^2 = \|c-d\|^2 + \|d\|^2 \geq \|d\|^2$$

with equality if and only if $d = c$. More specifically, $c - d \in \ker(T)$, and $d \in \operatorname{ran}(T^*)$, and we use the result from linear algebra that $\operatorname{ran}(T^*) = \ker(T)^\perp$. Since $\sum c_k f_k = \sum d_k f_k$, we have that $\sum_{k=1}^m (c_k - d_k) f_k = 0$ so that $c - d \in \ker(T)$. Also,

$$d_k = \langle f, g_k \rangle = \left\langle S^{-1} f, f_k \right\rangle$$

and thus $d \in \operatorname{ran}(T^*)$. □

Note that in the context of compressive sampling, we care more about minimizing the l^1 norm, but there is not a straightforward way to do this.

As a special case, if $m = n$, then $\{f_1, ..., f_m\}$ is a basis for $\mathcal{H}$, and if we consider

$$f_j = \sum_{k=1}^{m} \langle f_j, g_k \rangle f_k$$

and by the uniqueness of the coefficients with respect to a basis, we have $\langle f_j, g_k \rangle = \delta_{jk}$, and in this case we say that $\{g_1, ..., g_m\}$ is biorthogonal to $\{f_1, ..., f_m\}$.

Tight Frames

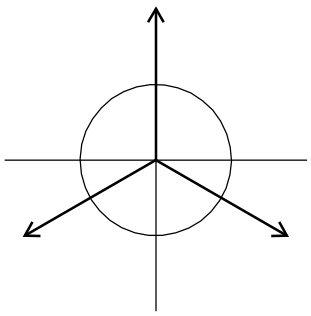
Consider the case where $S = TT^* = A\,\mathrm{Id}$ where A is a constant. Then we have that

$$f = \frac{1}{A} \sum_{k=1}^{m} \langle f, f_k \rangle f_k$$

which is the closest we can get to an orthogonal expansion when f_k is not an orthonormal basis. In this case $(f_k)_{k=1}^m$ is called a **tight frame**.

Trivial examples can be found in one dimensional spaces, and from orthonormal bases, or unions of orthonormal bases. First, a nontrivial example:

Example 14. For $\mathcal{H} = \mathbb{R}^2$, let $m = 3$, and consider the frame:



appropriately named, the Mercedes-Benz frame. In fact, any other collection of roots of unity in $\mathbb{R}^2$ will work as well. Later we will characterize tight frames and show how to construct general examples.

Terminology: If $\|f_k\| = 1$ for all k then we say the frame is *normalized*, or say that f_k forms a *unit norm tight frame*.

Pseudoinverse: If $TT^* = A\,\mathrm{Id}$, then $\frac{1}{A}T$ is a left inverse of T^*. In general, $S^{-1}T = (TT^*)^{-1}T$ is a left inverse of T^*, and is called the pseudoinverse of T^*. Also, T^*S^{-1} is the pseudoinverse of T. Recall that the pseudoinverse of an operator gives the element of the preimage that minimizes l^2 norm. We have already seen this in the frame decomposition, that out of all coefficients c such that $Tc = f$, the one that minimizes the l^2 norm is precisely $c = \langle S^{-1}f, f_k\rangle = T^*S^{-1}f$.

More Precise Frame Bounds

Returning to the frame operator S, we note that in

$$A\|f\|^2 \le \langle Sf, f\rangle \le B\|f\|^2$$

the best lower and upper frame bounds are given by $A = \lambda_{\min}(S)$ and $B = \lambda_{\max}(S)$. An especially easy way to see this is with Rayleigh quotients. Both bounds are attained when f is an eigenvector of S corresponding to the min or max eigenvalue. Note that S is a positive operator, so $\lambda_{\min}(S) > 0$.

Furthermore, note that

$$\sum_{k=1}^{m} \lambda_k(S) = \mathrm{tr}(S) = \mathrm{tr}(TT^*) = \|T\|_F^2 = \sum_{k=1}^{m} \|f_k\|^2$$

where $\|T\|_F$ is the Frobenius norm $\|T\|_F = \Big(\sum_{i,j} |T_{i,j}|^2\Big)^{1/2}$. For this we used the identification $\mathcal{H} = \mathbb{C}^n$ and the matrix form of T.

In the case of a unit norm frame, we have that

$$\sum_{k=1}^{m} \lambda_k(S) = m$$

and in the case of a unit norm tight frame, we have that $S = A\ \mathrm{Id}$ so that all eigenvalues are $\lambda_k(S) = A$, which implies that $A\ n = m$ and thus $\lambda_k(S) = A = \frac{m}{n}$. This describes the "redundancy ratio" of the unit norm tight frames. So frame bounds in some sense completely capture the "redundancy" of the frame in the case of a unit norm tight frame. Otherwise, can still get some feeling for how the redundancy of the frame after normalizing the frame...

Digression to Infinite Dimensions

Recall the condition for frames, which generalizes to infinite dimensions:

$$A\|f\|^2 \leq \sum_{k=1}^{\infty} |\langle f, f_k\rangle|^2 \leq B\|f\|^2$$

Note that just having a spanning/complete set $\{f_k\}_{k=1}^{\infty}$, i.e. $\mathcal{H} = \overline{\text{span}\{f_k\}_{k=1}^{\infty}}$ is not enough. We need some sort of stability, which is given by the frame bounds.

Example 15. There exists a set $\{f_k\}_{k=1}^{\infty}$ that is complete (closed linear span is all of $\mathcal{H}$), but where not every element of $\mathcal{H}$ can be represented as an infinite combination of f_k. Consider $\mathcal{H} = l^2(\mathbb{N})$ and the sequence

$$\begin{aligned} f_1 &= (1,1,0,0,0,...) \\ f_2 &= (0,1,1,0,0,...) \\ f_3 &= (0,0,1,1,0,...) \\ &\vdots \end{aligned}$$

Suppose that $c \in l^2$ such that $\langle c, f_k\rangle = 0$ for all k. Then we show that $c = 0$, in which case $(\overline{\text{span}\,\{f_k\}})^{\perp} = \{0\}$, or $\overline{\text{span}\{f_k\}} = \mathcal{H}$. Note that

$$0 = \langle c, f_k\rangle = c_k + c_{k+1}$$

this implies that $|c_k| = |c_{k+1}|$ for all k, but since $c \in l^2$, we must have $|c_k| = 0$ for all k (needs to decay at ∞). Thus $c = 0$ if $\langle c, f_k\rangle = 0$ for all k.

On the other hand, we note that e_1 cannot be expressed as an (infinite) linear combination of f_k, i.e. there does not exist a sequence d_k such that $e_1 = \sum_{k=1}^{\infty} d_k f_k$. Suppose there does exist such d_k. Then we have that $1 = \langle e_1, e_1\rangle = d_1$, and for $k > 1$,

$$0 = \langle e_1, e_k\rangle = d_k + d_{k-1}$$

so that $d_1 = 1$, $d_2 = -1$, $d_3 = 1$, and etc. However, $\sum_{k=1}^{\infty} (-1)^k f_k$ does not converge.

In infinite dimensions, the simplest example we have encountered that is a frame is given by the sampling theorem in the case where we oversample. Then $\{\varphi(t - n\tau)\}_{n\in\mathbb{Z}}$ form a frame.

Other examples are Gabor systems, which are time-frequency shifts of a fixed "window" $g(t)$ (smooth, with decay), and Wavelet systems.

Characterization of Tight Frames

Considering again the identification $\mathcal{H} = \mathbb{C}^n$, consider writing

$$T^* = \begin{pmatrix} — & f_1^* & — \\ & \vdots & \\ — & f_m^* & — \end{pmatrix}_{m\times n} = \begin{pmatrix} | & & | \\ h_1 & \dots & h_n \\ | & & | \end{pmatrix}_{m\times n}$$

Then the frame inequality

$$A\|f\|^2 \leq \sum_{k=1}^{m} |\langle f, f_k\rangle|^2 \leq B\|f\|^2$$

can be expressed as

$$A\|c\|_2^2 \leq \left\| \sum c_j h_j \right\|_2^2 \leq B\|c\|_2^2$$

noting that $\langle f, f_k \rangle_{k=1}^m = T^* f = \sum c_j h_j$, where c_j is the j-th coordinate of f. In the condition above, we call h_k a Riesz basis. Thus, note that

$$\begin{aligned} \{f_k\}_{k=1}^m \text{ is a frame} &\iff TT^* \text{ is invertible} \\ &\iff \{h_j\}_{j=1}^n \text{ is linearly independent} \end{aligned}$$

And moreover,

$$\begin{aligned} \{f_k\}_{k=1}^m \text{ is a tight frame} &\iff TT^* = A\,\mathrm{Id} \\ &\iff \{h_j\}_{j=1}^n \text{ is an orthogonal set} \end{aligned}$$

This implies that we can generate frames by taking m vectors from a basis in $\mathbb{C}^n$.

Example 16. Use the Fourier basis in $\mathbb{C}^n$, with

$$e_k(j) = \frac{1}{\sqrt{n}} e^{2\pi i kj/n} = \frac{1}{\sqrt{n}} \omega_n^{kj}$$

where k, j range from 0 to $m-1$. For any distinct set of frequencies $k_1, ..., k_m$ from 0 to $m-1$, define

$$f_j(l) = e_{k_l}(j)^*$$

for l from $1, ..., n$ (i.e. use e_{k_l} as the h_l in the discussion above). This gives a tight frame.

Application to Signal Processing

The redundancy in frames can be exploited to give us robustness against noise. Consider the problem of transmitting a signal $f \in \mathcal{H}$ across a noisy channel. Recall that given f, we have the representation

$$f = \sum_{k=1}^m \langle f, f_k \rangle S^{-1} f_k$$

and thus a plausible communication scheme is to transmit the coefficients $\langle f, f_k \rangle$ across the channel, so that the receiver just needs to compute the canonical dual frame to recover f in the ideal situation without noise. However, now we introduce noise:

$$f \to \boxed{\text{Transmitter}} \to \langle f, f_k \rangle_{k=1}^m \to \underset{\uparrow \xi_k}{\oplus} \to (\langle f, f_k \rangle + \xi_k)_{k=1}^m \to \boxed{\text{Receiver}} \to \tilde{f}$$

with the white noise assumption (not true when we introduce quantization, for instance) ξ_k is i.i.d. with zero mean and variance $\mathbb{E}[\xi_k^2] = \sigma^2$. The receiver, having received the noisy coefficients, chooses some dual basis h_k and reconstructs with

$$\tilde{f} = \sum_{k=1}^m (\langle f, f_k \rangle + \xi_k) h_k$$

We measure the error as the mean square error per dimension:

$$\varepsilon_n(f) = \frac{1}{n}\mathbb{E}[\|f - \tilde{f}\|^2]$$

We ask two questions:

Question 1: Given a frame f_k, which dual frame h_k will minimize the error measure?

Question 2: Having answered 1, which initial frame f_k will minimize the error measure?

First, let's compute the error measure explicitly:

$$\begin{aligned}
\varepsilon_n(f) &= \frac{1}{n}\mathbb{E}\left[\left\|\sum_{k=1}^{m} \xi_k h_k\right\|^2\right] \\
&= \frac{1}{n}\mathbb{E}\left[\sum_{k,l} \xi_k \xi_l \langle h_k, h_l\rangle\right] \\
&= \frac{\sigma^2}{n}\sum_{k=1}^{m} \|h_k\|^2
\end{aligned}$$

where we have used linearity of expectation and the fact that $\mathbb{E}[\xi_k \xi_l] = \sigma^2\delta_{k,l}$. Thus, we want to minimize $\sum_{k=1}^{m} \|h_k\|^2$.

Let $T = \begin{pmatrix} | & & | \\ f_1 & \cdots & f_m \\ | & & | \end{pmatrix}$ and $U = \begin{pmatrix} | & & | \\ h_1 & \cdots & h_m \\ | & & | \end{pmatrix}$. Then $UT^* = \mathrm{Id}$ if h_k is dual to f_k. Also, we can write

$$\varepsilon_n(f) = \frac{\sigma^2}{n}\mathrm{Tr}(UU^*)$$

and thus **Question 1** translates to the following optimization problem:

$$\min \mathrm{Tr}(UU^*) \text{ s.t. } UT^* = \mathrm{Id}$$

Solution: Let $V = (TT^*)^{-1}T$ be the pseudoinverse of T^* (i.e. corresponding the canonical dual). Clearly $VT^* = \mathrm{Id}$, and $(U - V)T^* = 0$. Denoting $R = U - V$, let $U = V + R$. Since $RT^* = 0$, note

$$RV^* = RT^*(TT^*)^{-1} = 0$$

thus

$$\begin{aligned}
\mathrm{Tr}(UU^*) &= \mathrm{Tr}((V+R)(V^*+R^*) \\
&= \mathrm{Tr}(VV^*) + \mathrm{Tr}(VR^*) + \mathrm{Tr}(RV^*) + \mathrm{Tr}(RR^*) \\
&= \mathrm{Tr}(VV^*) + \mathrm{Tr}(RR^*) \\
&\geq \mathrm{Tr}(VV^*)
\end{aligned}$$

since $\mathrm{Tr}(RR^*) \geq 0$ (eigenvalues of RR^* are nonnegative). This expression is minimized if and only if $R = 0$, i.e. $U = V$. Thus, the solution to question 1 is that the canonical dual is best.

Solution to Question 2. Now that we know that the canonical dual is best, the goal is to minimize

$$\varepsilon_n(f) = \frac{\sigma^2}{n}\mathrm{Tr}(VV^*) = \frac{\sigma^2}{n}\mathrm{Tr}(S^{-1})$$

noting $V = (T\,T^*)^{-1}T$ and $S = T\,T^*$. Here we note that that if we scale f_k large (so that noise is negligible), we can make the error artificially small. Hence this question makes more sense with an additional condition of normalizing the frame $\|f_k\| = 1$ for all k. With this condition, the question is equivalent to

$$\min \mathrm{Tr}(S^{-1}) \text{ s.t. } \|f_k\| = 1$$

Note that $\mathrm{Tr}(S^{-1}) = \sum_{k=1}^n \lambda_k(S^{-1}) = \sum_{k=1}^n \frac{1}{\lambda_k(S)}$. Also, $\sum_{k=1}^n \lambda_k(S) = \sum_{k=1}^m \|f_k\|^2 = m$ under the constraint. We will make use of the inequality between the arithmetic mean and harmonic mean ($\mathrm{AM} \geq \mathrm{HM}$):

$$\frac{1}{m}\sum_{j=1}^m x_j \geq \frac{m}{\sum_{j=1}^m \frac{1}{x_j}}$$

with equality if and only if $x_1 = x_2 = ... = x_m$. Then we have that

$$\mathrm{Tr}(S^{-1}) = \sum_{k=1}^n \frac{1}{\lambda_k(S)} \geq \frac{n^2}{\sum_{k=1}^n \lambda_k(S)} = \frac{n^2}{m}$$

and thus

$$\varepsilon_n(f) \geq \frac{\sigma^2}{n}\cdot\frac{n^2}{m} = \frac{\sigma^2}{(m/n)}$$

where the bound is obtained when all the λ_k are equal, $\lambda_k(S) = \frac{m}{n}$. This corresponds to a tight frame, i.e. $S = \frac{m}{n}\mathrm{Id}$. Thus, the solution is to take $\{f_k\}_{k=1}^m$ to be a unit norm tight frame with frame bound m/n, and in this case

$$\varepsilon_n(f) = \frac{\sigma^2}{(m/n)}$$

i.e. the larger the redundancy ratio, the better the error.

Recall that this is under the white noise assumption. When introducing quantization, a very simple scheme is to round each coefficient to the nearest quantization value. It can be shown that such a rounding procedure behaves like an i.i.d. random variable, satisfying the white noise assumption. Then we are in the setting above, and we have the given lower bound for the error. However, this is not optimal. There is no reason to choose the quantization values independently for each coefficient. If we quantize collectively, we may be able to reduce the approximation error further.

Week 4 (2/22/2010)

General ∞-dimensional Theory of Frames

We now turn to the general frame theory. With the addition of a little bit of analysis, we can generalize the previous results to this general setting. Let $\mathcal{H}$ a separable Hilbert space. Let $\{f_k\}_{k=1}^\infty$ be a sequence in $\mathcal{H}$.

Definition. We say that $\{f_k\}_{k=1}^\infty$ is a **frame** if there exists two constants $A, B > 0$ (finite) such that for all $f \in \mathcal{H}$,

$$A\,\|f\|^2 \leq \sum_{k=1}^\infty |\langle f, f_k\rangle|^2 \leq B\,\|f\|^2$$

A and B are called the **frame bounds**.

Observation: If $\{f_k\}_{k=1}^{\infty}$ is a frame, then $\{f_k\}_{k=1}^{\infty}$ is complete, i.e. $\overline{\text{span}\,\{f_k\}} = \mathcal{H}$. This is because if $\langle f, f_k\rangle = 0$ for all k, then the lower frame bound shows that $f = 0$.

If the frame inequality above holds for every $f \in \overline{\text{span}\{f_k\}_{k=1}^{\infty}}$, then $\{f_k\}_{k=1}^{\infty}$ is called a **frame sequence** (and $\{f_k\}_{k=1}^{\infty}$ is a frame for its closed span)

Example 17. *Complete, but not a frame.* We return to an example from last time. Given any orthonormal basis $\{e_k\}_{k=1}^{\infty}$, define $f_k = e_k + e_{k+1}$ for $k = 1, 2, ...$

1. For completeness, if $\langle f, f_k\rangle = 0$ for all k, then $\langle f, e_k\rangle = -\langle f, e_{k+1}\rangle$ for all k. Thus $\big|\langle f, e_k\rangle\big| = c$ for all k, and by Parseval, $\|f\|^2 = \sum_k |\langle f, e_k\rangle|^2$ and thus $\langle f, e_k\rangle = 0$ for all k. (Same proof as last time)

2. To show that it is not a frame, we'll exhibit a sequence $\{u_n\}_{n=1}^{\infty} \in \mathcal{H}$ such that $\|u_n\| = 1$ for all n but

$$\sum_{k=1}^{\infty} |\langle u_n, f_k\rangle|^2 \to 0$$

   This shows that there is no lower frame bound. Define $v_n = \sum_{j=1}^{n} (-1)^j e_j$, and let $u_n = \frac{v_n}{\|v_n\|} = \frac{v_n}{\sqrt{n}}$. Note that

$$\begin{aligned}
\langle v_n, f_k\rangle &= \left\langle \sum_{j=1}^{n} (-1)^j e_j, e_k + e_{k+1} \right\rangle \\
&= \sum_{j=1}^{n} (-1)^j \delta_{jk} + \sum_{j=0}^{n-1} (-1)^{j+1} \delta_{j+1,k+1} \\
&= \sum_{j=1}^{n-1} (-1)^j \underbrace{[\delta_{jk} - \delta_{j+1,k+1}]}_{=0} - \underbrace{\delta_{1,k+1}}_{=0} + (-1)^n \delta_{nk} \\
&= (-1)^n \delta_{n,k}
\end{aligned}$$

   This implies that

$$\sum_{k=1}^{\infty} |\langle u_n, f_k\rangle|^2 = \frac{1}{n} \sum_{k=1}^{n} |\delta_{n,k}|^2 = \frac{1}{n} \longrightarrow 0$$

Side note: Consider e_1. By completeness, there exists a sequence $\sum_{k=1}^{n} c_k^{(n)} f_k \to e_1$. But on the other hand, there does not exist a convergent representation of e_1 of the form

$$e_1 = \sum_{k=1}^{\infty} c_k f_k$$

Proof. We already proved there does not exist a convergent representation of e_1 in terms of the f_k in Example 15. To find a sequence of linear combinations $\sum_{k=1}^{n} c_k^{(n)} f_k \to e_1$, let's examine

$$e_1 - \sum_{k=1}^{n} c_k^{(n)} f_k = \left(1 - c_1^{(n)}\right) e_1 - \left(c_1^{(n)} + c_2^{(n)}\right) e_2 + ... - \left(c_{n-1}^{(n)} + c_n^{(n)}\right) e_n - c_n^{(n)} e_{n+1}$$

so intuitively we want $c_1 \approx 1$ and $c_j \approx -c_{j+1}$ but $c_n \approx 0$. Thus, let's set

$$c_k^{(n)} = (-1)^k \left(1 - \frac{k}{n}\right)$$

so that

$$\left\| e_1 - \sum_{k=1}^{n} c_k^{(n)} f_k \right\|_2^2 \leq \frac{n}{n^2} \to 0$$

There are many other possibilities of course.

□

Now let $\{f_k\}_{k=1}^{\infty}$ be a frame. As before, we consider the analysis/coefficient operator $C: \mathcal{H} \to l^2(\mathbb{N})$ given by

$$Cf = (\langle f, f_k \rangle)_{k=1}^{\infty}$$

By the frame property, this is a bounded oeprator with $\|C\| \leq \sqrt{B}$.

For defining the synthesis/reconstruction operator, we need to be a little careful. First, let c be a sequence with finitely many nonzero coefficients. Then we can define

$$Tc = \sum_{k=1}^{N} c_k f_k$$

This definition extends to $c \in l^2(\mathbb{N})$ in the following way. Let J be a finite subset of $\mathbb{N}$. Then

$$\begin{aligned}
\left\| \sum_{k \in J} c_k f_k \right\|_{\mathcal{H}} &= \sup_{\substack{\|g\|=1 \\ g \in \mathcal{H}}} \left| \left\langle \sum_{k \in J} c_k f_k, g \right\rangle_{\mathcal{H}} \right| \\
&= \sup_{\substack{\|g\|=1 \\ g \in \mathcal{H}}} \left| \sum_{k \in J} c_k \langle f_k, g \rangle_{\mathcal{H}} \right| \\
&= \sup_{\substack{\|g\|=1 \\ g \in \mathcal{H}}} \left(\sum_{k \in J} |c_k|^2 \right)^{1/2} \left(\sum_{k \in J} |\langle f_k, g \rangle_{\mathcal{H}}|^2 \right)^{1/2} \\
&\leq \sup_{\substack{\|g\|=1 \\ g \in \mathcal{H}}} \|c\|_{l^2(J)} \sqrt{B} \, \|g\|_{\mathcal{H}} \\
&= \sqrt{B} \, \|c\|_{l^2(J)}
\end{aligned}$$

This uniform bound for all J says that $\left(\sum_{k=1}^{N} c_k f_k \right)_{N=1}^{\infty}$ is a Cauchy sequence (take $J = \{M, ..., N\}$). Thus $\sum_{k=1}^{\infty} c_k f_k$ converges in $\mathcal{H}$ for all $c \in l^2(\mathbb{N})$, and thus defining Tc to be the limit (extension by continuity), T is defined on $l^2(\mathbb{N})$ with $\|T\| \leq \sqrt{B}$.

In fact, the convergence is stronger:

Proposition 18. *$\sum_{k=1}^{\infty} c_k f_k$ converges unconditionally, i.e. given any reordering $f_{\sigma(k)}$ where $\sigma: \mathbb{N} \to \mathbb{N}$ is a bijection, $\sum_{k=1}^{\infty} c_k f_{\sigma(k)}$ also converges to the same vector.*

Proof. The uniform bound in the proof of the extension above shows that for $J = \{\sigma(k),\, M \le k \le N\}$ $\sum_{k=1}^{N} c_{\sigma(k)}\, f_{\sigma(k)}$ is Cauchy and hence converges. (Also, can use the fact that $f_{\sigma(k)}$ is also a frame with the same frame bounds). To show that the convergence is to the same vector, the idea is that we can take N_1 so that $\sum_{k=N_1}^{\infty} c_k\, f_k$ is small, norm $< \varepsilon$, and N_2 so that $\sum_{k=N_2}^{\infty} c_{\sigma(k)} f_{\sigma(k)}$ is small, norm $< \varepsilon$, then

$$\left\| \sum_{k=1}^{\infty} c_k\, f_k - \sum_{k=1}^{\infty} c_{\sigma(k)} f_{\sigma(k)} \right\| \le \left\| \sum_{k=N_1}^{\infty} c_k\, f_k \right\| + \left\| \sum_{k=N_2}^{\infty} c_{\sigma(k)}\, f_{\sigma(k)} \right\| + \left\| \sum_{k=1}^{N_1} c_k\, f_k - \sum_{k=1}^{N_2} c_{\sigma(k)} f_{\sigma(k)} \right\|$$

The first two terms are bounded by ε by our choice of N_1, N_2, and the last term is bounded by 2ε, since terms cancel whenever indices overlap, i.e. $k \in \{1, ..., N_1\} \cap \{\sigma(1), ..., \sigma(N_2)\}$, and otherwise the leftover terms are contained tail of the other rearrangement, i.e.

$$\left\| \sum_{k=1}^{\infty} c_k\, f_k - \sum_{k=1}^{\infty} c_{\sigma(k)} f_{\sigma(k)} \right\| \le 2 \left\| \sum_{k=N_1}^{\infty} c_k\, f_k \right\| + 2 \left\| \sum_{k=N_2}^{\infty} c_{\sigma(k)}\, f_{\sigma(k)} \right\| \le 4\varepsilon$$

□

As with before, we have that the the two operators are adjoints:

Proposition 19. $C^* = T$

Proof.

$$\begin{aligned} \langle C^* c, f \rangle_{\mathcal{H}} &= \langle c, Cf \rangle_{l^2(\mathbb{N})} \\ &= \sum_{k=1}^{\infty} c_k \overline{\langle f\,,\, f_k \rangle}_{\mathcal{H}} \\ &= \left\langle \sum_{k=1}^{\infty} c_k f_k, f \right\rangle_{\mathcal{H}} \\ &= \langle Tc, f \rangle_{\mathcal{H}} \end{aligned}$$

□

Note that so far we have only used the upper bound of the frame property. The lower bound is used for reconstruction. Now that $T^* = C$, we define the frame operator as last time, $S: \mathcal{H} \to \mathcal{H}$ with

$$S(f) = TT^*(f) = \sum_{k=1}^{\infty} \langle f\,,\, f_k \rangle\, f_k$$

Note $\|S\| = \|T\|^2 \le B$, and $\langle Sf\,,\, f \rangle = \sum\, |\langle f\,,\, f_k \rangle|^2$. The frame property implies that

$$AI_{\mathcal{H}} \le S \le BI_{\mathcal{H}}$$

where the notation $A \le B$ means that $B - A$ is a nonnegative definite operator. Since $A > 0$, S is invertible and furthermore

$$B^{-1} I_{\mathcal{H}} \le S^{-1} \le A^{-1} I_{\mathcal{H}}$$

Note that the statement $AI_{\mathcal{H}} \le S \le BI_{\mathcal{H}}$ shows that the eigenvalues of S are between A and B, and since the eigenvalues of the inverse are the reciprocal of the eigenvalues, S^{-1} has eigenvalues between B^{-1} and A^{-1}.

Now we have the same reconstruction as before:

$$f = S^{-1}Sf = S^{-1}\left(\sum_{k=1}^{\infty} \langle f, f_k\rangle f_k\right) = \sum_{k=1}^{\infty} \langle f, f_k\rangle S^{-1}f_k$$

using continuity of S^{-1} for the last equality. Letting $g_k = S^{-1}f_k$, g_k is the canonical dual and

$$f = \sum_{k=1}^{\infty} \langle f, f_k\rangle g_k$$

Likewise,

$$f = SS^{-1}f = \sum_{k=1}^{\infty} \left\langle S^{-1}f, f_k\right\rangle = \sum_{k=1}^{\infty} \left\langle f, S^{-1}f_k\right\rangle$$

and thus we have the dual representation

$$f = \sum_{k=1}^{\infty} \langle f, g_k\rangle f_k$$

Note that $\left\{S^{-1}f_k\right\}_{k=1}^{\infty}$ is also a frame, because $\sum_{k=1}^{\infty} \left|\left\langle f, S^{-1}f_k\right\rangle\right| = \sum_{k=1}^{\infty} \left|\left\langle S^{-1}f, f_k\right\rangle\right|$ and that

$$\frac{A}{B^2}\|f\|^2 \le A\|S^{-1}f\|^2 \le \sum_{k=1}^{\infty} \left|\left\langle S^{-1}f, f_k\right\rangle\right| \le B\|S^{-1}f\|^2 \le \frac{B}{A^2}\|f\|^2$$

where we have used the bounds for the operator S^{-1}.

Note: These bounds are not optimal. The optimal frame bounds for $S^{-1}f_k = g_k$ are $\frac{1}{B}$ and $\frac{1}{A}$, as before.

To show this, noting that S depends on the frame $\{f_k\}_{k=1}^{\infty}$, we introduce the temporary notation $S_f = S_{\{f_k\}_{k=1}^{\infty}}$ and $T_f = T_{\{f_k\}_{k=1}^{\infty}}$. Then examining the frame operator corresponding to the dual frame g_k, we have that

$$T_g c = \sum_{k=1}^{\infty} c_k g_k = \sum_{k=1}^{\infty} c_k S_f^{-1} f_k = S_f^{-1} T_f c$$

Thus

$$S_g = T_g T_g^* = S_f^{-1} T_f T_f^* S_f^{-1} = S_f^{-1}$$

and thus $\frac{1}{B}I_{\mathcal{H}} \le S_g \le \frac{1}{A}I_{\mathcal{H}}$, so $\frac{1}{B}$ and $\frac{1}{A}$ are frame bounds for the frame $\{g_k\}_{k=1}^{\infty}$. From this observation, we also see that the canonical dual of the canonical dual gives back the original frame.

Tight Frames

We say a frame is **tight** if $A = B$ in the frame bounds.

Proposition 20. *A unit norm tight frame with $A = B = 1$ is an orthonormal basis.*

Proof. Have $\sum_{k=1}^{\infty} |\langle f, f_k\rangle|^2 = \|f\|^2$ for all f, and setting $f = f_j$ we have

$$\|f_j\|^4 + \sum_{k\neq j} |\langle f_j, f_k\rangle|^2 = \|f_j\|^2$$

and since $\|f_j\|^2 = \|f_j\|^4 = 1$, we must have that $\langle f_j, f_k \rangle = 0$ for all $k \neq j$. Hence f_k is an orthonormal basis.

□

Tight frames give a very nice representation, since $S = AI_{\mathcal{H}}$:

$$f = \frac{1}{A} \sum_{k=1}^{\infty} \langle f, f_k \rangle \, f_k$$

Note that we can turn a frame into a tight frame as follows:

Proposition 21. *If $\{f_k\}_{k=1}^{\infty}$ is a frame, then $\left\{S^{-1/2} f_k\right\}_{k=1}^{\infty}$ is a tight frame with constant 1 (not necessarily orthonormal, since f_k is not required to be unit).*

Proof. From functional calculus results we know that $S^{-1/2}$ exists and is bounded, positive, and commutes with S. One way to define this is as in finite dimensions. Since S is nonnegative definite there exists an orthonormal basis of eigenvectors with nonnegative eigenvalues, and if we decompose

$$f = \sum_{k=1}^{\infty} c_k \, e_k$$

then

$$Sf = \sum_{k=1}^{\infty} c_k \lambda_k e_k$$

and

$$S^{-1} f = \sum_{k=1}^{\infty} c_k \lambda_k^{-1} e_k$$

then we define

$$S^{-1/2} f = \sum_{k=1}^{\infty} c_k \lambda_k^{-1/2} e_k$$

In any case, we have that

$$\begin{aligned} f &= S^{-1/2} S (S^{-1/2} f) \\ &= S^{-1/2} \sum_{k=1}^{\infty} \left\langle S^{-1/2} f, f_k \right\rangle f_k \\ &= \sum_{k=1}^{\infty} \left\langle f, S^{-1/2} f_k \right\rangle S^{-1/2} f_k \end{aligned}$$

Also, we have that

$$\|f\|^2 = \langle f, f \rangle = \sum_{k=1}^{\infty} \left| \left\langle f, S^{-1/2} f_k \right\rangle \right|^2$$

so that $S^{-1/2} f_k$ is a tight frame.

□

Remark 22. We will have more explicit forms for $S^{-1/2} f_k$ when we consider specific examples involving translations and Fourier transforms.

Frame Algorithm

What is the heart of the matter here? Given a frame $\{f_k\}_{k=1}^{\infty}$, we want to be able to reconstruct f from its frame coefficients $\langle f, f_k\rangle_{k=1}^{\infty}$. If we can compute a dual frame (not necessarily the canonical dual), then we have a reconstruction formula. However, if this is too costly, or not possible, then we want a method that recovers f without explicitly computing a dual frame. As a first attempt, we have

$$Sf = \sum_{k=1}^{\infty} \langle f, f_k\rangle f_k$$

where the RHS uses just the data and the frame. In general we know that $Sf \neq f$ unless f_k is tight with constant 1. It turns out there is an iterative procedure, the so called "frame algorithm", involving only S and the initial data $\langle f, f_k\rangle$. It is nothing more than an iterative method for computing S^{-1}.

The algorithm is as follows:

1. Set $f^{(1)} = Sf = \sum_{k=1}^{\infty} \langle f, f_k\rangle f_k$
2. Given $f^{(n)}$, set $f^{(n+1)} = f^{(n)} + \lambda(f^{(1)} - Sf^{(n)})$ where λ is a parameter to be chosen. The intuition here is that if this iteration converges, then $f^{(1)} - Sf^{(n)} = S(f - f^{(n)}) \to 0$, from which it follows by the frame bounds that $f - f^{(n)} \to 0$.

Note that we can rewrite this iteration as

$$f^{(n+1)} = (I_{\mathcal{H}} - \lambda S) f^{(n)} + \lambda f^{(1)}$$

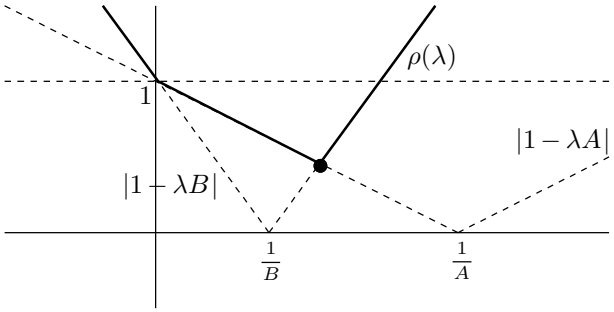
and convergence follows if λ is chosen so that $I_{\mathcal{H}} - \lambda S$ is a contraction. Since $AI_{\mathcal{H}} \leq S \leq BI_{\mathcal{H}}$, we have that

$$(1 - \lambda B) I_{\mathcal{H}} \leq I_{\mathcal{H}} - \lambda S \leq (1 - \lambda A) I_{\mathcal{H}}$$

So we want $\rho = \max(|1 - \lambda A|, |1 - \lambda B|) < 1$. Once we have this, then

$$\| f^{(n+1)} - f \| \leq \rho \| f^{(n)} - f \| \leq \rho^n \| f^{(1)} - f \|$$

and thus $f^{(n)} \to f$. To find the best possible ρ, we look at a graph of $\rho(\lambda)$:



The point of intersection is when $1 - \lambda A = \lambda B - 1$ or $\lambda = \frac{2}{A+B}$ which corresponds to $\rho = 1 - \frac{2A}{A+B} = \frac{B-A}{B+A}$.

Essentially, this is just a numerical analysis problem of recovering f from the equation $Sf = f^{(1)}$ using only applications of S. The algorithm presented here is not the fastest method, there are other methods based on conjugate gradient, for instance. Note here that the closer the frame is to being a tight frame, the better the convergence rate. This also shows the importance of having good frame bounds.

Note: Another way of seeing this iteration is through the identity

$$\lambda^{-1}S^{-1} = (I - (I - \lambda S))^{-1} = \sum_{n=0}^{\infty} (I - \lambda S)^n$$

which converges so long as $\|I - \lambda S\| < 1$ (i.e. a contraction). Note that compared to above, we note

$$f^{(n)} = \lambda \sum_{n=0}^{n} (I_{\mathcal{H}} - \lambda S)^n f^{(1)}$$

Important Examples of Frames

Frames in the Context of Sampling Theorem

Recall the sampling theorem on a uniform lattice, where we were working with the space of bandlimited functions

$$\mathcal{B}_\Omega = \left\{ f \in L^2(\mathbb{R}), \operatorname{supp} \hat{f}(\xi) \subset [-\Omega/2, \Omega/2] \right\}$$

The sampling theorem then shows that

$$f(t) = \tau \sum_{k \in \mathbb{Z}} f(k\tau)\, \varphi(t - k\tau)$$

where φ is any function such that

$$\hat{\varphi}(\xi) = \begin{cases} 1 & |\xi| \leq \Omega/2 \\ 0 & |\xi| > 1/2\tau \end{cases}$$

and $1/\tau \geq \Omega$. This gives quite a bit of flexibility when $1/\tau > \Omega$, and if $1/\tau = \Omega$, φ is the sinc function.

Setting $\varphi_k = \varphi(\cdot - k\tau)$, we note that

$$\langle f, \varphi_k \rangle = \left\langle \hat{f}, \widehat{\varphi_k} \right\rangle = \int_{-\Omega/2}^{\Omega/2} \hat{f}(\xi) e^{-2\pi i k\tau\xi}\, d\xi = \int_{-\infty}^{\infty} \hat{f}(\xi) e^{-2\pi i k\tau\xi}\, d\xi = f(k\tau)$$

noting that $\hat{f}$ is supported from $-\Omega/2$ to $\Omega/2$ so that extending the bounds of integration does not change the integral. Then we have that

$$f = \tau \sum_{k \in \mathbb{Z}} \langle f, \varphi_k \rangle\, \varphi_k$$

so that $\{\varphi_k\}_{k \in \mathbb{Z}}$ is a tight frame with frame bound $1/\tau$.

Also, $\|\varphi_k\| = \|\hat{\varphi}_k\| = \sqrt{\Omega}$, and if we renormalize $f_k := \frac{\varphi_k}{\sqrt{\Omega}}$, so that $\|f_k\| = 1$, we note that

$$f = \tau\Omega \sum_{k \in \mathbb{Z}} \langle f, f_k \rangle f_k$$

then $\{f_k\}_{k \in \mathbb{Z}}$ is a unit norm tight frame for $\mathcal{B}_\Omega$ with constant $\frac{1}{\tau\Omega} \geq 1$. Here $\frac{1}{\tau\Omega}$ describes the amount of redundancy, being precisely the ratio of the frequency sampling interval to the bandwidth $\frac{1/\tau}{\Omega}$. In general unit norm tight frames allow us to see the amount of redundancy there is.

Technical note: Note that technically $\{f_k\}_{k=1}^{\infty}$ is a frame for $\mathcal{B}_\Omega$ if $f_k \in \mathcal{B}_\Omega$ also, which is not the case for certain choices of φ_k. But this is an artifact from the definition of frames. Having the "tight frame" representation for f is very useful. We will revisit this point in the next lecture.

Week 5 (3/1/2010)

Frames of Translates

Generalizing the last example, we consider given $\varphi \in L^2(\mathbb{R})$ the following question:

Question 1: Is it possible to have $\{\varphi(\cdot - k\tau)\}_{k\in\mathbb{Z}}$ a frame for $L^2(\mathbb{R})$?

Answer: No. Note that the question asks whether we have a frame for *all* of $L^2(\mathbb{R})$, whereas the last example we considered a subspace $\mathcal{B}_\Omega$. We have the following theorem.

Theorem 23. *For all $\varphi \in L^2(\mathbb{R})$ and $\tau > 0$, $\{\varphi(\cdot - k\tau)\}_{k\in\mathbb{Z}}$ is not a frame.*

Proof. Given φ, τ, let $\varphi_k = \varphi(\cdot - k\tau)$. We will show that

$$\inf_{\|f\|_2=1} \sum_{k\in\mathbb{Z}} |\langle f, \varphi_k\rangle|^2 = 0$$

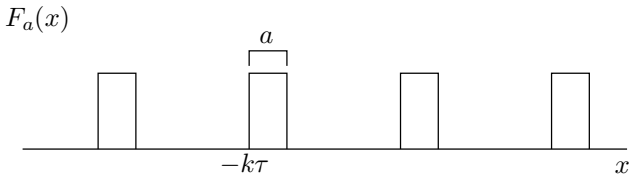
so that no lower frame bound exists, breaking the frame condition. The idea is simply to consider when f is an indicator:

$$f_a = \frac{1}{\sqrt{a}} \mathbf{1}_{[0,a)}$$

and we will vary a appropriately to get the infimum to be 0. Note

$$\begin{aligned}
\sum_{k\in\mathbb{Z}} |\langle f_a, \varphi_k\rangle|^2 &= \sum_{k\in\mathbb{Z}} \left|\left\langle f_a, \varphi_k \mathbf{1}_{[0,a]}\right\rangle\right|^2 \\
&\leq \sum_{k\in\mathbb{Z}} \|f_a\|_{L^2}^2 \|\varphi_k \mathbf{1}_{[0,a]}\|_{L^2}^2 \\
&= \sum_{k\in\mathbb{Z}} \int_0^a |\varphi(x - k\tau)|^2 \, dx \\
&= \sum_{k\in\mathbb{Z}} \int_{-k\tau}^{a-k\tau} |\varphi(x)|^2 \, dx \\
&= \int \left(\sum_{k\in\mathbb{Z}} \mathbf{1}_{[-k\tau, a-k\tau]}(x) \right) |\varphi(x)|^2 \, dx
\end{aligned}$$

The key is in the first line where we note that since f_a is supported in $[0, a]$, we can replace φ_k by $\varphi_k \mathbf{1}_{[0,a]}$ and the inner product does not change. If we denote $F_a(x) = \sum_{k\in\mathbb{Z}} \mathbf{1}_{[-k\tau, a-k\tau]}(x)$, then we note $\sum_{k\in\mathbb{Z}} \mathbf{1}_{[-k\tau, a-k\tau]}(x)$that for $a < \tau$ we have the following function:



and as $a \to 0$, $F_a \to 0$ pointwise. Then using dominated convergence, since

$$|F_a(x)|\,|\varphi(x)|^2 \leq |\varphi(x)|^2$$

which is integrable, we have that $\sum_{k\in\mathbb{Z}} |\langle f_a, \varphi_k\rangle|^2 \to 0$ as $a \to 0$, which proves that

$$\inf_{\|f\|_2=1} \sum_{k\in\mathbb{Z}} |\langle f, \varphi_k\rangle|^2 = 0$$

□

This leads us to a more relaxed question:

Question 2: When is $\{\varphi(\cdot - k\tau)\}_{k\in\mathbb{Z}}$ a frame sequence, i.e. a frame for its closed span.

Or an even more basic question is,

Question 3: When is $\{\varphi(\cdot - k\tau)\}_{k\in\mathbb{Z}}$ an orthonormal system?

First we address question 3.

Orthonormal System of Translates

Without loss of generality, note that we may assume $\tau = 1$ by simply scaling φ appropriately. Thus, let $\varphi_k = \varphi(\cdot - k)$ for $k \in \mathbb{Z}$. Note that since φ_k are translates,

$$\langle \varphi_k, \varphi_l\rangle = \delta_{kl} \iff \langle \varphi, \varphi_k\rangle = \delta_k$$

since $\langle \varphi_k, \varphi_l\rangle$ depends only on $k - l$. Then we compute:

$$\begin{aligned} \langle \varphi, \varphi_k\rangle &= \left\langle \hat{\varphi}, \hat{\varphi} e^{-2\pi i k\xi}\right\rangle \\ &= \int_{\mathbb{R}} |\hat{\varphi}|^2 e^{2\pi i k\xi} \end{aligned}$$

Right now this does not give a good characterization for when the result is δ_k. We will use a periodization trick, the same trick used for Poisson Summation Formula:

$$\begin{aligned} &= \sum_{k\in\mathbb{Z}} \int_l^{l+1} |\hat{\varphi}(\xi)|^2 e^{2\pi i k\xi}\, d\xi \\ &= \sum_{k\in\mathbb{Z}} \int_0^1 |\hat{\varphi}(\xi + l)|^2 e^{2\pi i k\xi}\, d\xi \\ &= \int_0^1 \left(\sum_{k\in\mathbb{Z}} |\hat{\varphi}(\xi + l)|^2 \right) e^{2\pi i k\xi}\, d\xi \end{aligned}$$

Now that the integral is on [0, 1], it is easy to see that the result is δ_k if and only if the periodization, which we denote by $\Phi(\xi) = \sum_{k\in\mathbb{Z}} |\hat{\varphi}(\xi + l)|^2$ is identically 1 ($e^{2\pi i k\xi}$ form an orthonormal basis for L^2[0, 1]). Thus,

$$\{\varphi_k\}_{k\in\mathbb{Z}} \text{ is an ONS} \iff \Phi = 1 \text{ a.e.}$$

Side remark: Note that if $F \in L^1(\mathbb{R})$, then the expression $F_{\text{per}}(x) = \sum_{k\in\mathbb{Z}} F(x + k)$ is meaningful. Can check that $F_{\text{per}} \in L^1(\mathbb{T})$:

$$\int_0^1 \left| \sum_{k\in\mathbb{Z}} F(x + k) \right| dx \le \sum_{k\in\mathbb{Z}} \int_0^1 |F(x + k)|\, dx = \int_{\mathbb{R}} |F(x)|\, dx$$

using Tonelli. In particular, the periodization Φ above makes sense since $|\hat{\varphi}|^2 \in L^1$.

Example 24. There are many examples that satisfy $\Phi = 1$ a.e.:

- An easy special case of this is when $\hat{\varphi} = \mathbf{1}_{[-1/2,1/2]}$, which corresponds to the sinc $\varphi(x) = \frac{\sin \pi x}{\pi x}$.
- The "dual" consideration to consider when $\varphi = \mathbf{1}_{[-1/2,1/2]}$, for which φ_k is an orthonormal system. The characterization above thus tells us that

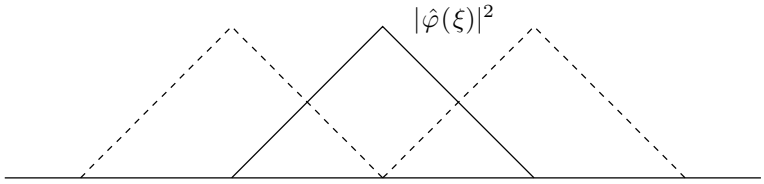
$$1 = \sum_{l \in \mathbb{Z}} |\hat{\varphi}(\xi + l)|^2 = \sum_{l \in \mathbb{Z}} \left(\frac{\sin \pi(\xi + l)}{\pi(\xi + l)} \right)^2 = \frac{\sin^2(\pi \xi)}{\pi^2} \sum_{l \in \mathbb{Z}} \frac{1}{(\xi + l)^2}$$

  so that

$$\sum_{l \in \mathbb{Z}} \frac{1}{(\xi + l)^2} = \frac{\pi^2}{\sin^2(\pi \xi)} \text{ for } \xi \notin \mathbb{Z}$$

  which is an interesting identity.

- Can get other identities as well, for instance we can also let $|\hat{\varphi}(\xi)|^2$ be



  (the dotted lines show the translates). Note that summing the translates gives 1 identically. The corresponding φ allows us to obtain more identities. There are many other examples as well.

Riesz Bases and Translates

Now we return to address question 2, about when translates are a frame sequence. Recall that we have a frame sequence when we can find two constants A, B positive and finite such that

$$A \| f \|^2 \leq \sum_k |\langle f, \varphi_k \rangle|^2 \leq B \| f \|^2$$

for all $f \in \overline{\text{span}\{\varphi_k\}_{k \in \mathbb{Z}}}$. In terms of the frame operator, note that the middle term is equal to

$$\langle Sf, f \rangle = \| T^* f \|_{l^2}^2$$

where S, T correspond to $\{\varphi_k\}_{k=1}^{\infty}$ ($S = TT^*$).

Before turning to frames, we consider a dual notion, which turns out to be a special case.

Definition: Given a sequence $\{f_k\}_{k=1}^{\infty} \subset \mathcal{H}$, we say that $\{f_k\}_{k=1}^{\infty}$ is a **Riesz sequence** if for all $c \in l^2(\mathbb{N})$, there exists A, B positive and finite such that

$$A \| c \|_{l^2}^2 \leq \left\| \sum_{k=1}^{\infty} c_k f_k \right\|_{\mathcal{H}}^2 \leq B \| c \|_{l^2}^2$$

Note here that in terms of the frame operator, the middle term is equal to $\| Tc \|_{\mathcal{H}}^2$, and we have required this property to hold for all $c \in l^2(\mathbb{N})$, in contrast to the definition for a frame sequence.

Remark 25. Note that under these conditions, $\{f_k\}_{k=1}^{\infty}$ are linearly independent, since if $\sum c_k f_k = 0$, then the lower bound shows that $c = 0$. In other words, there can only be one sequence $c \in l^2$ for which we have a representation of the form $f = \sum_k c_k f_k$ (i.e. if a representation of f in terms of f_k exists, it is unique).

Remark 26. If a frame $\{f_k\}_{k=1}^{\infty}$ is a Riesz sequence, then it is also a basis, since the frame property gives us an expansion for any f in terms of the frame f_k, and the Riesz property gives linear independence. In this case we say that $\{f_k\}_{k=1}^{\infty}$ a **Riesz basis**.

In fact, we will show that a Riesz sequence is also a frame sequence, so according to our definition every Riesz sequence is a basis for its closed span.

Remark 27. Recalling that $T^* f = \{\langle f, f_k\rangle\}_{k\in\mathbb{Z}}$ and $Tc = \sum_k c_k f_k$,

- If $\{f_k\}_{k=1}^{\infty}$ is a frame, then TT^* is invertible
- If $\{f_k\}_{k=1}^{\infty}$ is a Riesz sequence, then T^*T is invertible.

First we prove the connection between these two dual notions.

Proposition 28. *Given a sequence $\{f_k\}_{k=1}^{\infty} \subset \mathcal{H}$, then*

$$\begin{array}{ll} (\#) & A\|f\|_{\mathcal{H}}^2 \le \|T^* f\|_{l^2}^2 \le B\|f\|_{\mathcal{H}}^2 \text{ for all } f \in \overline{\text{ran}(T)} \\ & \Updownarrow \\ (*) & A\|c\|_{l^2}^2 \le \|Tc\|_{\mathcal{H}}^2 \le B\|c\|_{l^2}^2 \quad \text{for all } c \in \overline{\text{ran}(T^*)} \end{array}$$

Note that $(\#)$ says that $\{f_k\}_{k=1}^{\infty}$ is a frame sequence.

Also, this implies that if $\{f_k\}_{k=1}^{\infty}$ is a Riesz sequence (condition $(*)$ except for all $c \in l^2$), then it is also a frame sequence.

Proof. We will denote $(\#)_1$ to mean the left inequality of $(\#)$, i.e. $A\|f\|_{\mathcal{H}}^2 \le \|T^* f\|_{l^2}^2$ and $(\#)_2$ to mean the right inequality of $(\#)$ and likewise for $(*)_1$ and $(*)_2$.

Let us show that $(\#)_1 \Longrightarrow (*)_1$. First let $c \in \text{ran}(T^*)$ (will later approximate the closure). This means that $T^* f$ for some $f \in \mathcal{H}$. Then let $f = f_1 + f_2$ where $f_1 \in \overline{\text{ran}(T)}$ and $f_2 \in \ker T^* = \text{ran}(T)^{\perp}$. Study

$$\begin{aligned} \langle c, c\rangle^2 &= |\langle T^* f, T^* f\rangle|^2 \\ &= |\langle Sf, f\rangle|^2 \\ &\le \|Sf\|_{\mathcal{H}}^2 \|f\|_{\mathcal{H}}^2 \end{aligned}$$

Note that since $T^* f = T^* f_1$ and $Sf = Sf_1$, so that by $(\#)_1$ (since $f_1 \in \overline{\text{ran}(T)}$),

$$\begin{aligned} &\le \|Sf\|_{\mathcal{H}}^2 \frac{1}{A} \|T^* f\|_{l^2}^2 \\ &= \|Tc\|_{\mathcal{H}}^2 \frac{1}{A} \langle c, c\rangle \end{aligned}$$

where we note that

$$\|Sf\|_{\mathcal{H}}^2 = \langle TT^* f, TT^* f\rangle_{\mathcal{H}} = \langle Tc, Tc\rangle_{\mathcal{H}} = \|Tc\|_{\mathcal{H}}^2$$

Thus, rearranging we have that

$$\|Tc\|_{\mathcal{H}}^2 \geq A\,\|c\|_{l^2}^2$$

which shows $(*)_1$. Continuity shows that this holds for $c \in \overline{\mathrm{ran}(T^*)}$ also. Note that swapping c with f and T with T^* shows $(*)_1 \Longrightarrow (\#)_1$.

To show the equivalence $(*)_2 \Longleftrightarrow (\#)_2$, the same trick applies, using the inequality $\|f\|_{\mathcal{H}}^2 \leq \frac{1}{B}\|T^* f\|_{l^2}^2$ instead. □

Note that we have used nothing specific to frames to prove this result; it is just a general result concerning Hilbert space operators.

Now we address the following question:

Question 4: When is $\{\varphi_k = \varphi(\cdot - k)\}_{k\in\mathbb{Z}}$ a Riesz sequence?

Answer: We have the following theorem:

Theorem 29. *$\{\varphi_k = \varphi(\cdot - k)\}_{k\in\mathbb{Z}}$ is a Riesz sequence with constants A, B if and only if*

$$A \leq \Phi(\xi) \leq B$$

where $\Phi(\xi) = \sum_{l\in\mathbb{Z}} |\hat{\varphi}(\xi - l)|^2$.

Proof. Here we inspect $\|T_\varphi c\|_{\mathcal{H}}^2$. Note that

$$\begin{aligned}
\widehat{T_\varphi c}(\xi) &= \sum_{k\in\mathbb{Z}} c_k\,\hat{\varphi}(\xi)\,e^{2\pi i k\xi} \\
&= \hat{\varphi}(\xi)\left(\sum_{k\in\mathbb{Z}} c_k\,e^{-2\pi i k\xi}\right) \\
&= \hat{\varphi}(\xi)\,\hat{c}(\xi)
\end{aligned}$$

Note that $\hat{c}$ denotes Fourier transform of a sequence (in particular note $\hat{c}$ is 1-periodic), and $\hat{\varphi}$ is the Fourier transform on $\mathbb{R}$. Then computation using the 1-periodization trick shows that

$$\begin{aligned}
\|T_\varphi\,c\|_{\mathcal{H}}^2 &= \|\widehat{T_\varphi c}\,\|_{\mathcal{H}}^2 \\
&= \int_{\mathbb{R}} |\hat{\varphi}(\xi)|^2\,|\hat{c}(\xi)|^2\,d\xi \\
&= \int_0^1 |\hat{c}(\xi)|^2\left(\sum_{l\in\mathbb{Z}} |\hat{\varphi}(\xi + l)|^2\right) d\xi \\
&= \int_0^1 |\hat{c}(\xi)|^2\,\Phi(\xi)\,d\xi
\end{aligned}$$

Note that if $0 < A \leq \Phi(\xi) \leq B$, then

$$A\,\|c\|_{l^2}^2 \leq \|T_\varphi\,c\|_{\mathcal{H}}^2 = \int_0^1 |\hat{c}(\xi)|^2\,\Phi(\xi)\,d\xi \leq B\,\|c\|_{l^2}^2$$

which implies that $\{\varphi_k\}_{k=1}^{\infty}$ form a Riesz sequence with constants A, B. Note that in the two computations above we have used Parseval's equality (Fourier transform is an isometry).

Conversely, suppose that

$$A\,\|c\|_{l^2}^2 \le \|T_\varphi\, c\|_{\mathcal{H}}^2 = \int_0^1 |\hat{c}(\xi)|^2\,\Phi(\xi)\,d\xi \le B\,\|c\|_{l^2}^2$$

Then for any interval $I \subset [0,1]$, choose $c \in l^2$ so that $|\hat{c}(\xi)|^2 = \frac{1}{|I|}\mathbf{1}_I$. This implies that

$$A \le \frac{1}{|I|}\int_I \Phi(\xi)\,d\xi \le B$$

Taking $I = (x-\delta,\, x+\delta)$ and taking $\delta \to 0$, we have a result about Lebesgue points which tells us that almost all points satisfy $\frac{1}{2\delta}\int_{x-\delta}^{x+\delta} \Phi(\xi)\,d\xi \to \Phi(x)$ as $\delta \to 0$, and thus we have that $A \le \Phi(x) \le B$ a.e.

□

As before, this allows us to find many examples of translates that are Riesz sequences (and hence frame sequences as well). As long as $\hat{\Phi}(\xi)$ is bounded away from 0 and ∞, we have a corresponding Riesz sequence $\{\varphi(\,\cdot - k)\}_{k\in\mathbb{Z}}$. This generalizes the characterization we found earlier for when $\{\varphi(\,\cdot - k)\}_{k\in\mathbb{Z}}$ form an orthonormal basis.

Frame Sequences of Translates

Finally, we consider question 2, which asks when $\{\varphi(\,\cdot - k)\}_{k\in\mathbb{Z}}$ will form a frame sequence. Proposition 28 will help give the characterization we need. Recall that the frame sequence condition is equivalent to verifying that

$$A\,\|c\|_{l^2}^2 \le \|Tc\|_{\mathcal{H}}^2 \le B\,\|c\|_{l^2}^2 \quad \text{for all } c \in \overline{\operatorname{ran}(T^*)}$$

Remember that the difference between a frame sequence and a Riesz sequence is precisely the range of c we need to check. Also, recall that $\overline{\operatorname{ran}(T^*)} = (\ker T)^\perp$.

Thus, in our case, we want that

$$\|T_\varphi\, c\| = \int_0^1 |\hat{c}(\xi)|^2\,\Phi(\xi)\,d\xi$$

is controlled by constants times $\int_0^1 |\hat{c}(\xi)|^2\,d\xi$ for $c \in (\ker T)^\perp$. First let us compute exactly what $\ker T_\varphi$ is:

$$\begin{aligned} T_\varphi c = 0 &\iff \int_0^1 |\hat{c}(\xi)|^2\,\Phi(\xi)\,d\xi = 0 \\ &\iff \operatorname{supp}(\hat{c}) \cap \operatorname{supp}(\Phi) = \varnothing \end{aligned}$$

Thus

$$\ker T_\varphi = \{c \in l^2 \colon \operatorname{supp}(\hat{c}) \subset \operatorname{supp}(\Phi)^c\}$$

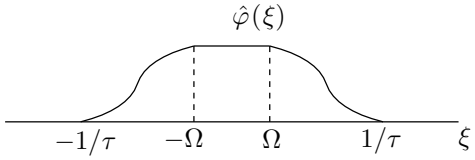
and

$$(\ker T_\varphi)^\perp = \{c \in l^2 \colon \operatorname{supp}(\hat{c}) \subset \operatorname{supp}(\Phi)\}$$

and thus with the same computation as before, we have that $\{\varphi(\,\cdot - k)\}_{k\in\mathbb{Z}}$ if and only if

$$A \le \Phi(\xi) \le B \text{ a.e. on } \operatorname{supp}(\Phi)$$

This is a bit of a strange condition, but note that for instance, if we have



the corresponding translates $\{\varphi(\,\cdot - k\tau)\}_{k\in\mathbb{Z}}$ do *not* form a frame sequence. Nevertheless, for $f \in \mathcal{B}_\Omega$ we still have an expansion of the form

$$f = \tau \sum_k f(k\tau)\,\varphi(x - k\tau)$$

This is not contradictory since $\varphi_k = \varphi(\,\cdot - k\tau) \notin \mathcal{B}_\Omega$ in the first place, and it is true that $\{\varphi_k\}_{k=1}^\infty$ does not form a frame for its span (by the equivalence we just proved). This is simply a technicality from the definition of frame. In this case the frame inequality holds for $f \in \mathcal{B}_\Omega \subset \overline{\text{span}\{\varphi_k,\, k \in \mathbb{Z}\}}$, and in this case we say that $\{\varphi_k\}_{k=1}^\infty$ is a **pseudoframe** (though really, it enjoys the same properties as a frame, especially the nice form for the expansion of f and robustness to noise, etc).

Week 6 (3/8/2010)

Quantization

We return to the topic of quantization, and here the setting will be for bandlimited functions. The main tool here is the sampling theorem. Recall for $f \in \mathcal{B}_\Omega$,

$$f(t) = \tau \sum_n f(n\tau)\,\varphi(t - n\tau)$$

where $\tau \le \frac{1}{\Omega}$ and $\hat{\varphi}(\xi) = \begin{cases} 1 & |\xi| \le \frac{\Omega}{2} \\ 0 & |\xi| > \frac{1}{2\tau} \end{cases}$.

Notation: We define the **sampling frequency** $\omega := \frac{1}{\tau}$, and thus we have $\omega \ge \Omega$.

Also recall we have the relations

$$f(a) = \langle f, \varphi_a \rangle$$

where $\varphi_a(x) = \varphi(x - a)$, noting that

$$\langle f, \varphi_a \rangle = \left\langle \hat{f}, \hat{\varphi}(\,\cdot\,) e^{2\pi i a \cdot} \right\rangle = \int_{-\Omega/2}^{\Omega/2} \hat{f}(\xi)\, e^{2\pi i a \xi}\, d\xi = f(a)$$

and in particular, $f(n\tau) = \langle f, \varphi_{n\tau} \rangle$ so that we have the expansion

$$f = \tau \sum_n \langle f, \varphi_{n\tau} \rangle\, \varphi_{n\tau}$$

which looks like a tight frame expansion for f. Again, recall that $\varphi_{n\tau}$ is not a frame for a trivial reason, that $\varphi_{n\tau} \notin \mathcal{B}_\Omega$ unless $\tau = \frac{1}{\Omega}$, in which case we have the sinc functions. Furthermore, $\{\varphi_{n\tau}\}_{n\in\mathbb{Z}}$ is not a frame sequence if $\hat{\varphi}$ is continuous (from last time, this implies that it decays to zero, so $\Phi = \sum_l |\varphi(\xi + l)|^2$ is not bounded away from 0 on its support). Nevertheless, we will make use of the *pseudoframe* expansion as a main tool for quantization.

Quantization Problem

Let $y_n := f(n\tau)$. We would like to replace the sequence $\{y_n\}_{n=1}^{\infty}$ with a quantized sequence $\{q_n\}_{n=1}^{\infty}$ where each $q_n \in \mathcal{A}$, some discrete alphabet (and preferrably finite). For instance,

$$\mathcal{A} = \text{arithmetic progression of step size } d$$

A natural (but naive) approach to quantization is to simply round each y_n to the nearest element in $\mathcal{A}$. In the case where $\mathcal{A}$ is an artihmetic progression of step size d, we will have that

$$\|y - q\|_\infty \leq \frac{d}{2}$$

Let's call $e := y - q$ the **quantization error/noise**. Then from the quantized sequence, we reconstruct

$$\tilde{f} := \tilde{f}_q = \tau \sum_{n\in\mathbb{Z}} q_n \, \varphi(\cdot - n\tau)$$

Formally, we have

$$f - \tilde{f} = \tau \sum_{n\in\mathbb{Z}} e_n \, \varphi(\cdot - n\tau)$$

There are a few questions to address before we continue (i.e. "formally")

- Do we even have $\tilde{f} \in L^2$?

  This depends on $\mathcal{A}$. Note that if $0 \in \mathcal{A}$, then there is no concern since $y \in l^2$ so that $y_n \to 0$ as $n \to \infty$. This implies that $\tilde{f}$ will be a *finite* sum.

  However, what if $0 \notin \mathcal{A}$, for instance $\mathcal{A} = \{\pm 1\}$? This may seem bad, especially with the current naive approach, since there is no resolution (since $q_n = \operatorname{sgn}(y_n)$, we see that every positive function f gives the same $\tilde{f} = \tau \sum_n \varphi(\cdot - n\tau)$). But we will see soon that we will that we can still work with this alphabet using a different approach.

- How do we measure the error?

  There are a few options here:

  - Consider the "time-averaged" L^2-norm:

    $$\limsup_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} |f(t) - \tilde{f}(t)|^2 \, dt$$

    Note that if $\tilde{f} \in L^2$, then this limit is 0 as then

    $$\frac{1}{2T} \int_{-T}^{T} |f(t) - \tilde{f}(t)|^2 \leq \frac{1}{2T} \|f - \tilde{f}\|_2^2 \leq \frac{C}{2T} \to 0$$

    Thus, this error only makes sense if $\tilde{f}$ is not in L^2 and $\int_{-T}^{T} |\tilde{f}(t)|^2 \, dt$ does not grow too fast as $T \to \infty$. For example, this error measure makes sense if $\tilde{f}$ is bounded or $\tilde{f}$ has sustained oscillations.

    We will discuss this error measure later on.

  - We can also consider the L^∞ metric $\|f - \tilde{f}\|_\infty$.

Note that $f \in L^\infty$ since $\hat{f} \in L^2[-\frac{\Omega}{2}, \frac{\Omega}{2}]$, so $\hat{f} \in L^1$ and $f \in C_0 \subset L^\infty$ by Riemann Lebesgue Lemma. Also, $\tilde{f} \in L^\infty$ if φ is sufficiently localized. Specifically, consider

$$\tilde{f}(t) = \tau \sum_{n \in \mathbb{Z}} q_n \, \varphi(t - n\tau)$$

Since $f \in L^\infty$, the sample sequence $\{y_n\}_{n=1}^\infty \in l^\infty$ and no matter what the alphabet is, we can get q_n to within a finite distance from y_n, i.e. $|y_n - q_n| \leq C$. This implies that $\{q_n\}_{n=1}^\infty \in l^\infty$ as well. For simplicity, suppose $\mathcal{A}$ is an arithmetic progression with step size d, and hence $\|q_n - y_n\|_\infty \leq \frac{d}{2}$. In this case, we have

$$\|\tilde{f}\|_\infty \;\leq\; \|q_n\|_{l^\infty} \underbrace{\sup_t \tau \sum_{n \in \mathbb{Z}} |\varphi(t - n\tau)|}_{\text{finite if } \varphi \in L^1} \tag{1}$$

noting that the RHS is like a Riemann sum, and φ is smooth since $\hat{\varphi}$ is compactly supported.

Also, by this same argument, note that

$$\|f - \tilde{f}\|_{L^\infty} \leq C_\varphi \, \|y - q\|_{l^\infty}$$

where C_φ depends on $\|\varphi\|_{L^1}$.

For now, we will be using this error measure.

- Here we note that the most relevant error depends on context. For instance, for compression and reconstruction of audio signals, the relevant error measure would account for the presence of undesirable high pitched tones that may not be accounted for if we were to use the L^∞ metric instead, for instance.

Remark 30. In the estimate $\|f - \tilde{f}\|_{L^\infty} \leq C_\varphi \, \|y - q\|_{l^\infty}$ above, we can choose φ so that C_φ is independent of Ω.

Let $\varphi = \varphi_\Omega$ with $\varphi_\Omega(t) = \Omega \, \varphi_0(\Omega t)$ and $\hat{\varphi}_0(\xi) = 1$ for $|\xi| \leq \frac{1}{2}$. Then $\hat{\varphi}_\Omega(\xi) = \hat{\varphi}_0\Big(\frac{\xi}{\Omega}\Big)$ and

$$C_\varphi \sim \|\varphi\|_1 = \|\varphi_0\|_1$$

Note that we do not have this stability if $\hat{\varphi} = \chi_{[-\Omega/2, \Omega/2]}$ (i.e. $\varphi = \text{sinc}$) since $\varphi \notin L^1$, and hence we must oversample if we want such an error estimate. In general, the more we oversample, the more possibilities there are for quantization.

Recall that $T_\varphi(c) = \sum_{n \in \mathbb{Z}} c_n \, \varphi(\cdot - n)$. Define

$$T_{\varphi,\tau}(c) = \tau \sum_{n \in \mathbb{Z}} c_n \, \varphi(\cdot - n\tau)$$

The same method used to obtain the estimate (1) for $\|\tilde{f}\|_\infty$ above can be used to show that $T_{\varphi,\tau}$ is bounded from $l^\infty \to L^\infty$ if $\varphi \in L^1 \cap \mathcal{B}_\omega$ (recall $\omega = \frac{1}{\tau}$). Since

$$f - \tilde{f} = T_{\varphi,\tau}(y - q) = T_{\varphi,\tau}(e)$$

we can do better than the naive approach if we can choose q so that e is "close" to $\ker T_{\varphi,\tau}$. This is the idea of **noise-shaping**.

Recall from last lecture that

$$\ker T_\varphi = \{c \in l^2 \text{ s.t. } \operatorname{supp}(\hat{c}) \cap \operatorname{supp}(\Phi) = \varnothing\}$$

where $\Phi(\xi) = \sum_{l\in\mathbb{Z}} |\hat{\varphi}(\xi + l)|^2$. Similarly, we have that

$$\ker T_{\varphi,\tau} = \{c \in l^2 \text{ s.t. } \operatorname{supp}(\hat{c}) \cap \operatorname{supp}(\Phi_\tau) = \varnothing\}$$

where $\Phi_\tau(\xi) = \sum_{l\in\mathbb{Z}} \left|\hat{\varphi}\Big(\frac{\xi+l}{\tau}\Big)\right|^2$.

To translate, note that

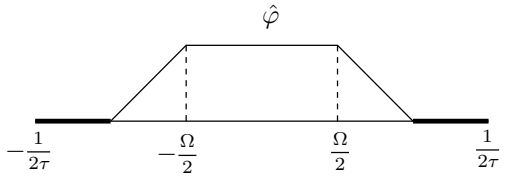
$$\begin{aligned}
(T_{\varphi,\tau}c)(x) &= \tau \sum_{n\in\mathbb{Z}} c_n\, \varphi(x - n\tau) \\
&= \sum_{n\in\mathbb{Z}} c_n\, \tau\varphi\Big(\tau\Big(\frac{x}{\tau} - n\Big)\Big) \\
&= \sum_{n\in\mathbb{Z}} c_n\, \psi\Big(\frac{x}{\tau} - n\Big) \\
&= (T_\psi c)\Big(\frac{x}{\tau}\Big)
\end{aligned}$$

where $\psi(x) = \tau\varphi(\tau x)$. Thus, $\ker T_{\varphi,\tau} = \ker T_\psi$ and

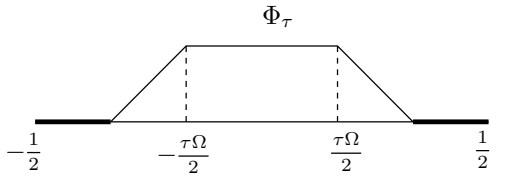
$$\Psi(\xi) = \sum_{l\in\mathbb{Z}} |\hat{\psi}(\xi + l)|^2 = \sum_{l\in\mathbb{Z}} \left|\hat{\varphi}\Big(\frac{\xi+l}{\tau}\Big)\right|^2 = \Phi_\tau(\xi)$$

which gives us the characterization of $\ker T_{\varphi,\tau}$ above.

To illustrate, suppose we have the following $\hat{\varphi}$:



where we have thickened where $\hat{\varphi} = 0$. Note that this corresponds to the following Φ_τ:



Then $c \in \ker T_{\varphi,\tau}$ if and only if $\hat{c}$ is supported in the thickened strip. In particular, note that as $\tau \to 0$ (oversampling), the region where $\Phi_\tau = 0$ gets larger, and hence there are more opportunities to make the noise smaller through redundancy.

Thus, we seek q such that $y - q$ is a high-pass sequence, and then we reduce the error $T_{\varphi,\tau}(y - q)$. This is the idea behind $\Sigma\Delta$ modulation.

ΣΔ Modulation

A typical high pass sequence has 0 mean. Thus, we'll try to satisfy the following difference equation

$$y_k - q_k = u_k - u_{k-1}$$

for some auxiliary sequence $\{u_k\}_{k=1}^{\infty}$ and choice of $\{q_k\}_{k=1}^{\infty}$. Note that if there are no conditions on u, there are plenty of pairs (u, q) that will work. To be useful, however, we will see that it suffices to restrict u to be a bounded sequence.

Assuming we have found a pair (u, q) satisfying the difference equation and with u a bounded sequence, we have that

$$\begin{aligned} f(t) - \tilde{f}(t) &= \tau \sum_{k\in\mathbb{Z}} (y_k - q_k)\, \varphi(t - k\tau) \\ &= \tau \sum_{k\in\mathbb{Z}} (u_k - u_{k-1})\, \varphi(t - k\tau) \\ &= \tau \sum_{k\in\mathbb{Z}} u_k[\varphi(t - k\tau) - \varphi(t - (k+1)\tau)] \end{aligned}$$

Above we have used summation by parts, which is justified since $\varphi(\,\cdot\,)$ has sufficient decay so that the series can be split into two parts and recombined after shifting indices.

This implies that

$$\|f - \tilde{f}\,\|_\infty \le \tau\, \|u\|_\infty \sup_t \sum_{k\in\mathbb{Z}} |\varphi(t - k\tau) - \varphi(t - (k+1)\tau)|$$

Now note that $t_k = t - k\tau$ form a partition of $\mathbb{R}$, so that

$$\sup_t \sum_{k\in\mathbb{Z}} |\varphi(t - k\tau) - \varphi(t - (k+1)\tau)| \le \|\varphi\|_{\mathrm{TV}}$$

Since φ is smooth, it follows that $\|\varphi\|_{\mathrm{TV}} = \|\varphi'\|_{L^1}$, and thus

$$\|f - \tilde{f}\,\|_\infty \le \tau \|u\|_\infty \|\varphi'\|_{L^1}$$

Note that u is independent of τ (the differnec eequation has no τ dependence). The claim is that φ can be chosen so that $\|\varphi'\|_{L^1}$ is independent of τ, from which it follows that as $\tau \to 0$, $\tilde{f} \to f$.

First, note that with $\varphi(t) = \varphi_\Omega(t) = \Omega\varphi_0(\Omega t)$ as before, we have that

$$\varphi'(t) = \Omega(\Omega\, \varphi_0'(\Omega t))$$

so that $\|\varphi'\|_1 = \Omega \|\varphi_0'\|_1$. Now if we introduce the notation $\lambda = \frac{\omega}{\Omega} = \frac{1}{\tau\Omega}$, we have that

$$\|f - \tilde{f}\,\|_\infty \le \Omega\tau\, \|u\|_\infty\, \|\varphi_0'\|_1 = \frac{1}{\lambda}\, \|u\|_\infty\, \|\varphi_0'\|_1$$

and so as $\lambda \to \infty$, $\tilde{f} \to f$.

Solving the Difference Equation

Now we show that the difference equation can be solved with u bounded. We want to find $\{q_k\}_{k=1}^{\infty}$ so that

$$y_k - q_k = u_k - u_{k-1}$$

with u bounded. We will use the assumption on the alphabet that $\mathcal{A}$ is an infinite (temporary assumption) progression of size d. We will solve it recursively. As an initial condition, we can set u_0 to be arbitrary, say $u_0 = 0$ (this will not matter) and define

$$q_k = \text{Argmin}_{p \in \mathcal{A}} |u_{k-1} + y_k - p|$$

This guarantees that

$$|u_{k-1} + y_k - q_k| = |u_k| \leq \frac{d}{2}$$

and hence $|u_k| \leq \frac{d}{2}$ for all $k > 0$. Since $|y_k| \leq \|f\|_\infty$, $|u_{k-1}| \leq \frac{d}{2}$, the largest value of p in magnitude chosen in the Argmin above will be $|p| \leq \|f\|_\infty + d$. In other words, if we assume an apriori bound on $\|f\|_\infty \leq 1$, for instance, $\mathcal{A}$ can be chosen to be a finite arithmetic progression. We need $\mathcal{A}$ to be finite so that implementation becomes practical.

In the extreme case, we have $|\mathcal{A}| = 2$, or 1-bit quantization, taking $\mathcal{A} = \{-1, 1\}$, so $d = 2$ and $|u_k| \leq 1$ for all $k > 0$. In this case, the quantization rule will be

$$q_k = \text{sign}(u_{k-1} + y_k)$$

and so

$$u_k = u_{k-1} + y_k - \text{sign}(u_{k-1} + y_k)$$

In this case we can verify that $|u_k| \leq 1$ for all $k > 0$. Then, given that $|u_{k-1}| \leq 1$ and $|y_k| \leq \|f\|_\infty \leq 1$, we see that $u_{k-1} + y_k \in [-2, 2]$. Then $q_k = \text{sign}(u_{k-1} + y_k)$ gives that $u_k = u_{k-1} + y_k - q_k \in [-1, 1]$ as desired. Thus $|u_k| \leq 1$ for all $k > 0$.

To solve the difference equation for $k < 0$, the situation is quite symmetric. If we write the equation as

$$u_{k-1} = u_k + (-y_k) - (-q_k)$$

we can think of $-y_k$ as the input and u_k given, and the goal is to choose $-q_k$, in the context of the previous discussion. Then we set

$$-q_k = \text{sign}(u_k - y_k)$$

or

$$q_k = -\text{sign}(u_k - y_k)$$

and we have solved the difference equation for all $\mathbb{Z}$ with $\|u\|_\infty \leq 1$. If we did the same computations in general for an alphabet that is an arithmetic progression with step d, then we would get $\|u\|_\infty \leq \frac{d}{2}$ if $\|f\|_\infty \leq \max_{p \in \mathcal{A}} |p|$.

Summarizing the results, after we have solved the difference equation we will have

$$\|f - \tilde{f}\|_\infty \leq \frac{d}{2} \cdot \frac{1}{\lambda} \cdot \|\varphi_0'\|_{L^1}$$

This reflects two facts:

- As $d \to 0$ (the resolution in the alphabet becomes finer), the error goes to 0.
- As $\lambda \to \infty$ (the oversampling ratio becomes larger), the error goes to 0.

Notice that the error decays in λ at the rate $\frac{1}{\lambda}$. We say that this scheme is a **first-order** $\Sigma\Delta$ scheme. We will discuss higher order schemes shortly.

Remark 31. This scheme is already popular in practice. We have already talked about applications for A/D conversion (analog-to-digital) for storage and reconstruction, but in practice a particular application is in D/A conversion in mp3 players. The setting is as follows:

An audio signal has been sampled at 44 kHz (i.e. oversampling at a rate much higher than the highest audible frequency) to get y_k and each sample y_k has been truncated to a specified number of bits for storage (Pulse Code Modulation (PCM)), and we denote the rounded samples by $\tilde{y}_k$. The goal is to reconstruct an analog signal from these digital samples to be played back to the human ear.

Note that we already have the sampling theorem which tells us how to reconstruct the signal approximately. We add up the pulses $\tilde{y}_k \varphi(\cdot - k\tau)$, and furthermore since φ is local, at a given time we only need to add finitely many of these at a given time (i.e. $\tilde{f}(t) = \sum_{|k| \le K} \tilde{y}_k \varphi(t - k\tau)$). However, in the interpolation formula it is difficult in hardware to reproduce a pulse $\tilde{y}_k \varphi(t - k\tau)$ for a large range of $\tilde{y}_k$.

Thus, what is done in practice is to replace $\tilde{y}_k$ even further by coefficients in a coarser alphabet $q_k \in \mathcal{A}$, for instance $\mathcal{A} = \{-1, 1\}$ using $\Sigma\Delta$ modulation (usually higher than first order). Then we just have to reproduce

$$\tilde{f}(t) = \sum_{|k| \le K} q_k \varphi(t - k\tau)$$

and since $q_k \in \{-1, 1\}$ it is easier for hardware to reproduce.

Now consider the worst case error

$$\varepsilon(\lambda) = \sup_{\substack{f \in \mathcal{B}_\Omega \\ \|f\| \le \mu}} \inf_{q \in \mathcal{A}_d^{\mathbb{Z}}} \|f - \tilde{f}\|_\infty$$

where $\tilde{f} = T_{\varphi,\tau} q$ and $\max \mathcal{A}_d > \mu$. As before, setting $\lambda = \frac{1}{\tau\Omega} = \frac{\omega}{\Omega}$ the oversampling ratio, we want to consider what is the worst case error as we take $\lambda \to \infty$. It turns out that the best we can do is

$$\varepsilon(\lambda) \gtrsim 2^{-\lambda}$$

where the proof uses the metrical entropy of $\mathcal{B}_\Omega$. Recall that the first order $\Sigma\Delta$ scheme above gave $\varepsilon(\lambda) \sim \frac{C}{\lambda}$. As a first step, we can consider higher order $\Sigma\Delta$ schemes:

Higher Order $\Sigma\Delta$ Schemes

Define $(\Delta u)_k = u_k - u_{k-1}$. The first order scheme solves the difference equation

$$\Delta u = y - q$$

with $u \in l^\infty$. A natural generalization is to consider the difference equation

$$\Delta^r u = y - q$$

with $u \in l^\infty$. The corresponding error bound is

$$\begin{aligned} \|f - \tilde{f}\|_\infty &= \|T_{\varphi,\tau}(\Delta^r u)\|_\infty \\ &\le \tau^r \|u\|_\infty \|\varphi^{(r)}\|_1 \end{aligned}$$

and we will prove the specifics next time. We can control $\|\varphi^{(r)}\|_1$ using Bernstein's inequality:

Proposition 32. (Bernstein's Inequality) *Let $f \in \mathcal{B}_\Omega \cap L^p$ for $1 \le p \le \infty$. Then*

$$\|f'\|_p \le (\pi\Omega)\,\|f\|_p$$

and consequently

$$\|f^{(r)}\|_p \le (\pi\Omega)^r\,\|f\|_p$$

We will also prove this next time.

Thus if $\varphi(t) = \Omega\,\varphi_0(\Omega t)$ where supp $\hat{\varphi}_0 \subset \left[-\frac{1}{2}(1+\varepsilon_0), \frac{1}{2}(1+\varepsilon_0)\right]$ (the ε_0 is from oversampling), then we have that

$$\|\varphi^{(r)}\|_1 \le [\pi(1+\varepsilon_0)\Omega]^r\,\|\varphi_0\|_1$$

and above we have that

$$\|f - \tilde{f}\,\|_\infty \le C^r\,\frac{\|u\|_\infty\,\|\varphi_0\|_1}{\lambda^r}$$

where $C = \pi(1+\varepsilon_0)$ (not dependent on r). There are two questions here:

1. How do we solve the difference equation with $u \in l^\infty$?
2. What is the size of $\|u\|_\infty$?

It turns out that we need to make a slight adjustment to the difference equation, and even then, the corresponding $\|u\|_\infty$, grows very fast in r, at a rate $\sim r!$.

Week 7 (3/22/2010)

Higher Order ΣΔ (continued)

We continue with a discussion of the higher order ΣΔ schemes. Recall in the greedy solution to the difference equation $y_k - q_k = u_k - u_{k-1}$, we had that we recursively define

$$q_k := \text{argmin}_{p\in\mathcal{A}}\,|y_k + u_{k-1} - p|$$

To simplify notation, given the alphabet $\mathcal{A}$ we define the rounding function $Q_\mathcal{A}\colon \mathbb{R} \to \mathcal{A}$ by

$$Q_\mathcal{A}(w) = \text{argmin}_{p\in\mathcal{A}}|w - p|$$

and so the difference equation becomes

$$u_k = u_{k-1} + y_k - Q_\mathcal{A}(u_{k-1} + y_k)$$

For higher order ΣΔ modulation, we set up the equation

$$y - q = \Delta^r u$$

where $(\Delta u)_k = u_k - u_{k-1}$ is the difference operator. We want a solution with u bounded. Given that we have found q, u which solve the difference equation with u bounded, we now compute more systematically estimates for the corresponding error

$$f(t) - \tilde{f}(t) = \tau \sum_{k\in\mathbb{Z}} (y_k - q_k)\varphi(t-k\tau) = \tau \sum_{k\in\mathbb{Z}} (\Delta^r u)_k \varphi(t-k\tau)$$

Define an operator Δ_τ on functions by

$$(\Delta_\tau g)(t) = g(t) - g(t-\tau)$$

Then

$$f(t) - \tilde{f}(t) = \tau \sum_{k\in\mathbb{Z}} (\Delta^r u)_k \varphi(t-k\tau) = \tau \sum_{k\in\mathbb{Z}} u_k (\Delta_\tau^r \varphi)(t-k\tau)$$

and

$$\begin{aligned} \|f - \tilde{f}\|_\infty &\le \tau \|u\|_\infty \|\Delta_\tau^{r-1}\varphi\|_{\mathrm{TV}} \\ &\le \tau \|u\|_\infty \|(\Delta_\tau^{r-1}\varphi)'\|_1 \\ &\le \tau \|u\|_\infty \|\Delta_\tau^{r-1}\varphi'\|_1 \end{aligned}$$

where the last line follows from translation invariance of the operator Δ (alternatively, can expand all the terms and differentiate each term). Now we claim that

$$\|\Delta_\tau g\|_1 = \tau \|g'\|_1$$

This is just a computation: First note that $g(t) - g(t-\tau) = \int_{t-\tau}^{t} g'(x)\,dx$ by Fundamental Theorem of Calculus. Then

$$\begin{aligned} \|\Delta_\tau g\|_{L^1} &= \int_{-\infty}^{\infty} |g(t) - g(t-\tau)|\,dt \\ &\le \int_{-\infty}^{\infty} \int_{t-\tau}^{t} |g'(x)|\,dx\,dt \qquad (t-\tau \le x \le t) \\ &= \int_{-\infty}^{\infty} |g'(x)| \Big(\int_{x}^{x+\tau} dt \Big) dx \\ &= \tau \|g'\|_1 \end{aligned}$$

using Fubini to swap the order of integration. Then by induction, we have that

$$\|\Delta_\tau^r g\|_1 = \|\Delta_\tau(\Delta_\tau^{r-1} g)\|_1 \le \tau \|\Delta_\tau^{r-1} g'\|_1 \le \tau^r \|g^{(r)}\|_1$$

Thus the bound above becomes

$$\|f - \tilde{f}\|_\infty \le \tau^r \|u\|_\infty \|\varphi^{(r)}\|_1$$

and using the fact that as before, $\hat{\varphi}(\xi) = \hat{\varphi}_0\Big(\frac{\xi}{\Omega}\Big)$ where $\hat{\varphi}_0(\xi) = \begin{cases} 1 & |\xi| \le \frac{1}{2} \\ 0 & |\xi| > \frac{1}{2}(1+\varepsilon_0) \end{cases}$ (i.e. independent of Ω), we have that $\varphi(x) = \Omega\,\varphi(\Omega x)$, and $\|\varphi^{(r)}\|_1 = \Omega^r \|\varphi_0\|_1$, so we can write the estimate as

$$\|f - \tilde{f}\|_\infty \le \frac{\|u\|_\infty \|\varphi_0^{(r)}\|_1}{\lambda^r}$$

Now we will make use of Bernstein's inequality, Proposition 32, and sketch a proof of the inequality. Restating the Bernstein inequality, if $f \in \mathcal{B}_\Omega \cap L^p$ for $1 \le p \le \infty$, then

$$\|f'\|_p \le \pi\,\Omega \|f\|_p$$

Proof. (Sketch) Note that if we take any ψ with $\hat{\psi}(\xi) = 1$ for $|\xi| \le \Omega$, then $\hat{f} = \hat{f}\,\hat{\psi}$. Taking the inverse Fourier transform, we have that

$$f = f * \psi$$

and so $f' = f * \psi'$ and $\|f'\|_p \le \|f\|_p \, \|\psi'\|_1$ by Young's inequality. Note that

$$\|\psi'\|_1 \ge \|(\varphi')^\wedge\|_\infty \ge |2\pi i\, \hat{\varphi}(\Omega/2)| = \pi\Omega$$

so that $\pi\Omega$ is the best constant we can hope for. To achieve the constant, we optimize over potential choices for ψ. □

Applying Bernstein's, we have that since $\varphi_0 \in \mathcal{B}_{1+\varepsilon_0} \cap L^1$, we have $\|\varphi_0^{(r)}\|_1 \le (\pi(1+\varepsilon_0))^r \, \|\varphi_0\|_1$ and hence

$$\|f - \tilde{f}\,\|_\infty \lesssim \|u\|_\infty \left(\frac{\pi(1+\varepsilon_0)}{\lambda} \right)^r$$

the $\lesssim$ denotes up to an absolute constant factor, which in this case is $\|\varphi_0\|_1$. This is of course assuming that we can solve $\Delta^r u = y - q$ for u bounded.

Greedy Quantization for r-th order $\Sigma\Delta$

If we expand the equation $\Delta^r u = y - q$, we get

$$u_k = \left(\sum_{j=1}^{r} (-1)^{j-1} \binom{r}{j} u_{k-j} \right) + y_k - q_k$$

For instance, for $r = 1$, $(\Delta u)_k = u_k - u_{k-1}$ and for $r = 2$, $(\Delta^2 u)_k = u_k - 2u_{k-1} - u_{k-2}$, and we see binomial coefficients with alternating signs.

Let us consider the more general recurrence relation (since we will be considering other recurrences later)

$$v_k = \left(\sum_{j=1}^{\infty} h_j\, v_{k-j} \right) + y_k - q_k = (h * v)_k + y_k - q_k$$

where in the right hand side $h_k = 0$ for $k \le 0$. This is the requirement so that the right hand side is only in terms of the previous coefficients $v_{k-1}, v_{k-2}, \ldots$. As before, the greedy rule is simply

$$q_k = Q_{\mathcal{A}}[(h * v)_k + y_k]$$

We have the following lemma, which gives a condition for when the resulting v_k is bounded:

Lemma 33. *If*

$$\frac{d}{2}\|h\|_1 + \|y\|_\infty \le \frac{d}{2}|\mathcal{A}|$$

then v_k remains bounded in $\left[-\frac{d}{2}, \frac{d}{2}\right]$, assuming that the initial condition $|v_0| \le \frac{d}{2}$.

As a special case, note that if $h_1 = 1, h_j = 0, j > 1$, then $\|h\|_1 = 1$, and we are in the 1-bit case. In the 1-bit case we noted the condition

$$\|y\|_\infty \leq \max_{p\in\mathcal{A}} |p| = \frac{d}{2}|\mathcal{A}| - \frac{d}{2}$$

(recall $\mathcal{A}$ is an symmetric arithmetic progression with step size d. The number of intervals is $|\mathcal{A}| - 1$, and thus $\max - \min = d(|\mathcal{A}| - 1)$ and the largest value is $\frac{\max - \min}{2} = \frac{d}{2}(|\mathcal{A}| - 1)$).

Proof. The proof of the lemma is identical to the 1-bit case.

$$|(h * v)_k + y_k| \leq \frac{d}{2}\|h\|_1 + \|y\|_\infty \leq \frac{d}{2} + \max_{p\in\mathcal{A}} p$$

where we have used the inequality $|(h * v)_k| \leq \max_{1\leq j\leq k-1} |v_j| \, \|h\|_1$. Thus with $q_k = Q_{\mathcal{A}}((h * v)_k + y_k)$,

$$|v_k| = |(h * v)_k + y_k - q_k| \leq \frac{d}{2}$$

□

As a special case of this lemma, we can study the difference equation $\Delta^r u = y - q$. In this case we have

$$h_j = (-1)^{j-1}\binom{r}{j}$$

for $1 \leq j \leq r$, and hence $\delta^0 - h = \Delta^r$ if we identify the operator Δ^r with the corresponding convolution vector. Then we have that $\|h\|_1 = \sum_{j=1}^r \binom{r}{j} = 2^r - 1$. The condition in the lemma becomes

$$|\mathcal{A}| \geq \|h\|_1 + \frac{2}{d}\|y\|_\infty = 2^r - 1 + \frac{2}{d}\|y\|_\infty$$

and if this is satisfied, then $\|u\|_\infty \leq \frac{d}{2}$. Noting that if we fix a resolution $d = 2$, and as before take $\|y\|_\infty \leq 1$, then the condition is that $|\mathcal{A}| \geq 2^r$ (i.e. we need at least r-bits in the quantization alphabet to solve the given recurrence).

Another interesting observation we can make here is that if $|\mathcal{A}| = \infty$, i.e. $\mathcal{A} = d\mathbb{Z}$, then the conditions of the lemma above are always satisfied, and we don't even need y_k to be bounded (though this will be the case since we are considering $f \in \mathcal{B}_\Omega$). Then the r-th order $\Sigma\Delta$ error bound from above applies with $\|u\|_\infty \leq \frac{d}{2}$, and we have

$$\|f - \tilde{f}\|_\infty \lesssim \frac{d}{2}\left(\frac{\pi(1+\varepsilon_0)}{\lambda}\right)^r$$

and note that no matter how coarse our quantization is, i.e. no matter what d is, we can make this error arbitrarily small if we take $\lambda > \pi(1 + \varepsilon_0)$ and take $r \to \infty$. Of course having an infinite quantization alphabet is not realistic, and we observe that as $r \to \infty$, we require larger quantization alphabets to maintain the conditions of the lemma.

Studying the Optimal Error

Consider the estimate

$$\|f - \tilde{f}\|_\infty \lesssim \|u\|_\infty \left[\frac{\pi(1+\varepsilon_0)}{\lambda}\right]^r$$

which holds if $\Delta^r u = y - q$. If we are in the conditions of the previous lemma, then $\|u\|_\infty \leq \frac{d}{2}$. In general, we want to consider other potential solutions, with bounds $\|u\|_\infty$ that may depend on r. We wish to ask what is the best we can expect?

We want to study the worst case error

$$E_{\text{opt}}(\mu, \mathcal{A}, \lambda) = \sup_{f \in \mathcal{B}_\Omega, \|f\|_\infty \leq \mu} \|f - \tilde{f}\|_\infty$$

(λ is the oversampling ratio and appears in the formula for $f, \tilde{f}$)

Define

$$\mathcal{U}(\mu, \mathcal{A}, r) := \sup_{\|y\|_\infty \leq \mu} \inf_{q \in \mathcal{A}^{\mathbb{Z}}} \|u\|_\infty$$

where u satisfies $\Delta^r u = y - q$ with $\|y\|_\infty \leq \mu$ and $q \in \mathcal{A}^{\mathbb{Z}}$. Here the sup over $\|y\|_\infty \leq \mu$ is describing the worst case input, and the inf over $q \in \mathcal{A}^{\mathbb{Z}}$ solving the recurrence describes the best possible error given y, and hence $\mathcal{U}$ describes the best possible error for the worst case input.

A fundamental question is then understanding the behavior of $\mathcal{U}(\mu, \mathcal{A}, r)$ as a function of r (for now, later we can also consider what happens when we change μ or $\mathcal{A}$).

The reason is that if we know $\mathcal{U}(\mu, \mathcal{A}, r)$, we have that

$$E_{\text{opt}}(\mu, \mathcal{A}, \lambda) = \sup_{f \in \mathcal{B}_\Omega, \|f\|_\infty \leq \mu} \|f - \tilde{f}\|_\infty \leq \inf_{r > 0} \left(\mathcal{U}(\mu, \mathcal{A}, r) \left[\frac{\pi(1 + \varepsilon_0)}{\lambda} \right]^r \right)$$

The strategy is that given the oversampling ratio λ, we may choose the optimal value of r adaptively.

What is interesting is that we can study $E_{\text{opt}}(\mu, \mathcal{A}, \lambda)$ through other means, and a lower bound for this error can be obtained through the study of covering numbers.

Note that if we consider all possible quantizations

$$\{T_{\varphi,\tau} q,\, q \in \mathcal{A}^{\mathbb{Z}}\}$$

then if we look at the ε-balls $\{B_\varepsilon(T_{\varphi,\tau} q),\, q \in \mathcal{A}^{\mathbb{Z}}\}$, this forms an ε-cover for $\{f \in \mathcal{B}_\Omega, \|f\|_\infty \leq \mu\}$ with $\varepsilon = E_{\text{opt}}(\mu, \mathcal{A}, \lambda)$. Note that this cover is not a finite cover, and in general we cannot expect a finite cover since $\{f \in \mathcal{B}_\Omega, \|f\|_\infty \leq \mu\}$ is not a compact space.

To study ε-covering numbers (the size of the minimal ε-cover) for the space $\mathcal{B}_\Omega$, we need to consider compact subsets of $\mathcal{B}_\Omega$. For an interval I, consider

$$\mathcal{B}_{\Omega,I} = \{f \chi_I \colon f \in \mathcal{B}_\Omega, \|f\|_\infty \leq \mu\}$$

i.e. bounded bandlimited functions restricted to the interval I. This is a compact subspace of $C(I)$ by Arzela-Ascoli, noting the Bernstein inequality $\|f'\|_\infty \leq \pi\Omega \|f\|_\infty$. Then it makes sense to talk about the minimal ε-cover for $\mathcal{B}_{\Omega,I}$, and we can consider the behavior as we take the interval $|I| \to \infty$. This was studied (along with covering numbers for other spaces) by Kolmogorov. If we define $N_\varepsilon(\mathcal{B}_{\Omega,I})$ to be the size of the minimal ε-cover for $\mathcal{B}_{\Omega,I}$, then Kolmogorov showed that

$$\frac{1}{|I|} \log(N_\varepsilon(\mathcal{B}_{\Omega,I})) \longrightarrow \Omega \log\left(\frac{1}{\varepsilon}\right)$$

as $|I| \to \infty$.

From this we can show that a lower bound for $E_{\text{opt}}(\mu, \mathcal{A}, \lambda)$ behaves like $e^{-c\lambda}$ for some c, which we will do next time.

Week 8 (3/29/2010)

Studying the Optimal Error (continued)

We turn to studying the behavior of

$$\mathcal{U}(\mu,\mathcal{A},r) := \sup_{\|y\|_\infty \le \mu} \inf_{q\in\mathcal{A}^{\mathbb{Z}}} \|u\|_\infty$$

which as we described last time tells us about

$$E_{\text{opt}}(\mu,\mathcal{A},\lambda) = \sup_{f\in\mathcal{B}_\Omega,\|f\|_\infty\le\mu} \|f-\tilde{f}\,\|_\infty \le \inf_{r>0}\left(\mathcal{U}(\mu,\mathcal{A},r)\left[\frac{\pi(1+\varepsilon_0)}{\lambda}\right]^r\right)$$

First we consider a calculus lemma

Lemma 34. *Let $\alpha,\beta>0$. Then*

$$\exp\Big\{-\beta e^{-1}\alpha^{-1/\beta}\Big\} \le \inf_{r\in\mathbb{Z}_+}\alpha^r r^{\beta r} \le \exp\Big\{\beta-\beta e^{-1}\alpha^{-1/\beta}\Big\}$$

Remark 35. This lemma will then be used to transfer bounds on $\mathcal{U}(\mu,\mathcal{A},r)$ to bounds on $E_{\text{opt}}(\mu,\mathcal{A},\lambda)$ by substituting $\alpha=\frac{\pi(1+\varepsilon_0)}{\lambda}$ and $\mathcal{U}(\mu,\mathcal{A},r)\le r^{\beta r}$ (almost).

Proof. Let $F(t)=\alpha^t t^{\beta t}=\exp\{t\log\alpha+\beta t\log t\}$ for $t\in\mathbb{R}^+$. Then

$$F'(t)=(\log\alpha+\beta\log t+\beta)\exp\{t\log\alpha+\beta t\log t\}$$

and

$$F''(t)=\frac{\beta}{t}\exp\{t\log\alpha+\beta t\log t\}+(\log\alpha+\beta\log t+\beta)^2\exp\{t\log\alpha+\beta t\log t\}>0,\text{ for } t>0$$

so that F is convex for $t>0$. We have a critical point where $F'(t_*)=0$ or

$$\begin{aligned}\log\alpha+\beta\log t_*+\beta &= 0\\ \log t_* &= -1-\frac{1}{\beta}\log\alpha\\ t_* &= e^{-1}\alpha^{-1/\beta}\end{aligned}$$

and since $\log\alpha+\beta\log t_*=-\beta$, we have that $F(t_*)=\exp\{t_*(\log\alpha+\beta\log t_*)\}=\exp\{-\beta e^{-1}\alpha^{-1/\beta}\}$.

By convexity, t_* is at the global minimum of $F(t)$, and thus this is a lower bound for $\inf_{k\in\mathbb{Z}_+}F(k)$. To get an upper bound, we just need to find a $k\in\mathbb{Z}_+$ close to the minimum, and it suffices to use $k=\lfloor t_*\rfloor$. If we write $k=\lfloor t_*\rfloor=t_*-\theta$ for $\theta=\langle t_*\rangle\in[0,1)$, then we have

$$\begin{aligned}F(k) &= \alpha^k k^{\beta k}\\ &\le \alpha^k (t_*)^{\beta k}\\ &= (\alpha^{t_*}t_*^{\beta t_*})\alpha^{-\theta}t_*^{-\beta\theta}\\ &= F(t_*)\alpha^{-\theta}(e^{-1}\alpha^{-1/\beta})^{-\beta\theta}\\ &= e^\beta F(t_*)\end{aligned}$$

Thus

$$F(t_*) \le \inf_{r\in\mathbb{Z}_+} \alpha^r r^{\beta r} \le e^{\beta} F(t_*)$$

which proves the lemma. □

Corollary 36. *Assume $\mathcal{U}(r,\mathcal{A},\mu) \le A^r r^{Br}$, with $B>0$. Then*

$$E_{\text{opt}} \lesssim_B \exp\left\{ -Be^{-1}\left[\frac{\lambda}{A\pi(1+\varepsilon_0)} \right]^{1/B} \right\} = \exp\left\{ -C(A,B)\lambda^{1/B} \right\}$$

setting $\alpha = A\left[\frac{\pi(1+\varepsilon_0)}{\lambda}\right]$ and $\beta = B$. In particular, if $B<1$, then $E_{\text{opt}}(\lambda)$ decays faster than exponentially.

As mentioned last time, we will show that $E_{\text{opt}}(\lambda) \le e^{-c\lambda}$ for some c through other means, and this will show that in the corollary above, we must have $B \ge 1$ (otherwise it violates the lower bound).

Kolmogorov Entropy Based Lower Bound

Let I be an interval, and define

$$\mathcal{B}(\Omega, I, \mu) = \{ f\chi_I,\, f \in \mathcal{B}_\Omega, \|f\|_\infty \le \mu \}$$

the restrictions of bandlimited functions to the interval I. Note that by Bernstein's inequality,

$$\|f'\|_\infty \le \pi\Omega \|f\|_\infty \le \pi\Omega\,\mu$$

and so we have a family of functions with uniformly bounded derivatives on a compact interval I, and thus by Arzela Ascoli $\mathcal{B}(\Omega, I, \mu)$ is a compact subset of $C(I)$ with respect to $\|\cdot\|_\infty$.

Let ε be given. Let $N = N_\varepsilon(\Omega, I, \mu)$ be the ε-covering number for $\mathcal{B}(\Omega, I, \mu)$. Then define

$$H_\varepsilon := \log_2 N$$

is called the **metric (or Kolmogorov) entropy** of $\mathcal{B}(\Omega, I, \mu)$.

What we observe is that $H_\varepsilon = H_\varepsilon(\Omega, I, \mu)$ only depends on the interval length $|I|$ and not the location of the interval by the shift-invariance of $\mathcal{B}_\Omega$ (translation in time domain corresponds to modulation in frequency, so the support of the Fourier transform is preserved). Furthermore, as $|I| \to \infty$, H_ε increases linearly in $|I|$. These properties are specific to the space of bandlimited functions; for instance this is not the case for Lipschitz or C^2.

Then we define the average ε-entropy per unit interval

$$\overline{H_\varepsilon} = \lim_{|I|\to\infty} \frac{1}{|I|} H_\varepsilon(\Omega, I, \mu)$$

(apriori we do not know the limit exists, so we should define quantities with limsup and liminf, but it turns out to exist).

Theorem 37. (Kolmogorov)

$$\overline{H_\varepsilon} = [1+o(1)]\,\Omega \log_2\left(\frac{1}{\varepsilon}\right)$$

as $\varepsilon \to 0$.

This is a fairly technical result proved in a paper by Kolmogorov and Tihomirov entitled *ε-Entropy and ε-Capacity of Sets in Functional Spaces* (American Mathematical Society Translations, Series 2 Volume 17).

We will not prove the result here, but the intuition in the context of sampling theorem is that if we consider the sampling at the Nyquist rate (critical sampling $\frac{1}{\tau} = \Omega$), then we need to encode one sample $f(k\tau)$ per Nyquist interval $[k\tau,\ (k\ +\ 1)\tau]$, and to achieve a resolution of ε we need at least $\log_2\Big(\frac{2\mu}{\varepsilon}\Big)$ bits to encode $f(k\tau)$. Thus a lower bound on the average ε-entropy per unit interval should behave on the order of $\frac{1}{\tau}\log_2\Big(\frac{1}{\varepsilon}\Big) = \Omega\log_2\Big(\frac{1}{\varepsilon}\Big)$.

The proof of the Theorem is similar in spirit; however, if we want to localize the sampling theorem

$$f(t) = \tau\sum_{k\in\mathbb{Z}} f(k\tau)\,\varphi(t-k\tau) \approx \tau\sum_{\frac{k}{\Omega}\in I} f(k\tau)\,\varphi(t-k\tau)$$

to samples where $\frac{k}{\Omega}\in I$, we need to oversample so that φ is more localized, and so we cannot just consider sampling at the Nyquist rate.

Using this result we can prove a lower bound for $E_{\mathrm{opt}}(\lambda)$.

Corollary 38.

$$E_{\mathrm{opt}}(\lambda) \gtrsim_\mu |\mathcal{A}|^{-\lambda}$$

where $\gtrsim_\mu$ denotes that the inequality is up to a constant factor $C(\mu)$ which depends on μ.

Proof. (Sketch) Consider

$$\mathcal{F} = \left\{ (T_{\varphi,\tau}\,q)(t) = \tau\sum_{k\tau\in I} q_k\,\varphi(t-k\tau),\, q\in\mathcal{A}^{\mathbb{Z}}, t\in I \right\}$$

a collection of points in $C(I)$, which do not necessarily fall in $\mathcal{B}(\Omega,\, I)$. Note that for computing covering numbers, the cover does not necessarily have to consist of balls with centers in $\mathcal{B}(\Omega,\, I)$. Consider the ε-cover generated by $\mathcal{F}$:

$$\mathcal{C}_\varepsilon = \{B_\varepsilon(f),\, f\in\mathcal{F}\}$$

Note that if $E_{\mathrm{opt}}(\lambda) = \varepsilon$, then $\mathcal{C}_\varepsilon$ forms an ε-covering for $\mathcal{B}(\Omega, I)$. The size of $\mathcal{C}_\varepsilon$ is counted roughly by

$$|\mathcal{C}_\varepsilon| \approx |\mathcal{A}|^{|I|/\tau}$$

($|\mathcal{A}|$ choices for each q_k, and there are $|I|/\tau$ integers with $k\tau\in I$). We bound $|\mathcal{C}_\varepsilon|$ below by the optimal size of an ε-cover (covering number) to get

$$\begin{aligned}
|\mathcal{C}_\varepsilon| \approx |\mathcal{A}|^{|I|/\tau} &\geq N_\varepsilon(\Omega, I, \mu) \\
&= 2^{H_\varepsilon(\Omega,I,\mu)} \\
(\text{as } |I|\to\infty) &\approx 2^{|I|\,\overline{H_\varepsilon}} \\
&\approx 2^{|I|\,\Omega\log_2(1/\varepsilon)} \\
&= \left(\frac{1}{\varepsilon}\right)^{|I|\,\Omega}
\end{aligned}$$

Solving for ε gives

$$\varepsilon \geq |\mathcal{A}|^{-1/(\tau\Omega)} = |\mathcal{A}|^{-\lambda}$$

and therefore $E_{\text{opt}}(\lambda) \geq |\mathcal{A}|^{-\lambda}$. The μ dependence is hidden from approximation details. $\square$

Using this result, by Corollary 36 we have that $\mathcal{U}(r, \mathcal{A}, \mu) \geq c^r r^r$ for $c = c(\mathcal{A}, \mu)$. This result is quite indirect and does not examine the difference equation corresponding to $\mathcal{U}(r, \mathcal{A}, \mu)$, and for this reason it is a little unsatisfactory. With a slight modification to the problem, we will find a more direct way to bound $\mathcal{U}(r, \mathcal{A}, \mu)$.

Direct Lower Bounds for the Difference Equation

Let's modify the problem slightly and consider solving the difference equation

$$\Delta^r u = y - q$$

with $\|y\|_\infty \leq \mu$ and $q_k \in \mathcal{A}$ only for positive indices $\mathbb{Z}^+$. Thus, we set $u_k = 0$ for $k < 0$ and solve for positive indices. This will allow us to work with the generating function of u_k.

Note that the error signal is then

$$\begin{aligned} f(t) - \tilde{f}(t) &= \tau \sum_{k<0} y_k \varphi(t - k\tau) + \tau \sum_{k \geq 0} (y_k - q_k) \varphi(t - k\tau) \\ &= E_1(t) + E_2(t) \end{aligned}$$

The first error $E_1(t)$ is unavoidable, as we have set $u_k = 0$ for $k < 0$ which means that $q_k = 0$ for $k < 0$. Furthermore, since φ is localized (oversampling case), for t large $E_1(t)$ vanishes very quickly. This means that in practice we can work with one-sided sequences without much loss, which is a practical necessity so that the resulting quantization system is causal (depending only on present and past samples, if we consider the index to be time).

Thus we are interested in the second component

$$E_2(t) = \tau \sum_{k \geq 0} (y_k - q_k) \varphi(t - k\tau)$$

and define as in the double-sided case

$$\mathcal{U}_+(r, \mathcal{A}, \mu) := \sup_{\|y\|_\infty \leq \mu} \inf_{q : \Delta^r u = y - q} \|u\|_\infty$$

As mentioned above, in studying one-sided sequences we can use generating functions, which are very effective with handling difference equations.

For an arbitrary bounded sequence $a \in l^\infty$, we define the generating function

$$F_a(z) = \sum_{k \geq 0} a_k z^k$$

Note that $F_a(z)$ is defined for $|z| < 1$, the radius of convergence for the power series $F_a(z)$. Note that with a simple reindexing,

$$F_{\Delta a}(z) = \sum_{k \geq 0} (a_k - a_{k-1}) z^k = (1 - z) \sum_{k \geq 0} a_k z^k = (1 - z) F_a(z)$$

If we consider the inverse operation $(Sa)_k = \sum_{j=0}^k a_j$, so that $\Delta \circ S = S \circ \Delta = \text{Id}$, we note that

$$F_{Sa}(z) = \frac{F_a(z)}{1 - z}$$

Also note that $|F_a(z)| \le \|a\|_\infty \sum_{k\ge 0} |z|^k = \frac{\|a\|_\infty}{1-|z|}$.

Now if we consider the difference equation

$$w := \Delta^r u = y - q$$

we have that $F_w(z) = F_{\Delta^r u}(z) = (1-z)^r F_u(z)$, and we have the bound

$$|F_w(z)| \le \|u\|_\infty \frac{|1-z|^r}{1-|z|}$$

for $|z|<1$. This allows us to transfer between bounded sequences and analytic functions on the unit disk.

Theorem 39. (Borwein-Erdelyi-Kós) *Let $f(z)$ be an analytic function on $|z|<1$ such that*

$$|f(z)| \le \frac{1}{1-|z|},\ |z|<1$$

Then there exist absolute constants $c_1, c_2 > 0$ such that

$$\sup_{x\in[1-\varepsilon,1]} |f(x)| \ge |f(0)|^{c_1/\varepsilon} e^{-c_2/\varepsilon} \text{ for all } \varepsilon \in (0,1]$$

If $|f(0)|<1$ then we have an exponentially small lower bound as $\varepsilon \to 0$.

This is a nontrivial result, proved in a paper *Littlewood-type problems on* [0, 1] (Theorem 5.1), using the Hadamard Three Circles Theorem. We will use this to obtain a bound on $\mathcal{U}(r, \mathcal{A}, \mu)$.

If we consider

$$f(z) = \frac{F_w(z)}{\|w\|_\infty} = \sum_{k\ge 0} \frac{w_k}{\|w\|_\infty} z^k$$

so that f satisfies the hypotheses of this theorem [BEK]. Thus,

$$\sup_{x\in[1-\varepsilon,1]} \frac{|F_w(x)|}{\|w\|_\infty} \ge \left(\frac{F_w(0)}{\|w\|_\infty}\right)^{c_1/\varepsilon} e^{-c_2/\varepsilon}$$

Note that using the bound for $|F_w(z)|$ with $z = x \in [0,1]$ we have that

$$(1-x)^{r-1}\|u\|_\infty \ge |F_w(x)|$$

Then

$$\begin{aligned} \|u\|_\infty \sup_{x\in[1-\varepsilon,1]} (1-x)^{r-1} &\ge \sup_{x\in[1-\varepsilon,1]} |F_w(x)| \\ &\ge \|w\|_\infty \left(\frac{|w_0|}{\|w\|_\infty}\right)^{c_1/\varepsilon} e^{-c_2/\varepsilon} \end{aligned}$$

noting $w_0 = F_w(0)$. This implies that

$$\|u\|_\infty \ge \|w\|_\infty \sup_{0<\varepsilon\le 1} \left(\frac{1}{\varepsilon}\right)^{r-1} \left(\frac{|w_0|}{\|w\|_\infty}\right)^{c_1/\varepsilon} e^{-c_2/\varepsilon}$$

We can optimize over ε, though it turns out it suffices to let $\varepsilon = \frac{1}{r}$ so that

$$\|u\|_\infty \geq \|w\|_\infty r^{r-1} C^r$$

where $C = \left(\frac{|w_0|}{\|w\|_\infty}\right)^{c_1} e^{-c_2}$, which is the bound obtained for the version of the problem using double-sided sequences, so long as $C > 0$.

Recall that we are studying

$$\mathcal{U}_+(r, \mathcal{A}, \mu) = \sup_{\|y\|_\infty \leq \mu} \inf_{q_k : (\Delta^r u)_k = y_k - q_k, k \geq 0} \|u\|_\infty$$

and thus we are okay as long as there are sequences y for which there are solutions (u, q) with $w_0 = y_0 - q_0 \neq 0$, and there are indeed plenty by simply looking at sequences where $y_0 \notin \mathcal{A}$. In fact, we can easily choose y_0 so that $\|y - q\|_\infty \geq |y_0 - q_0| \geq \frac{d}{2}$. Thus, in the worst case input y, $\|w\|_\infty \geq |w_0| \geq \frac{d}{2}$, and with the upper bound $\|w\|_\infty \leq \|y\|_\infty + \|q\|_\infty \leq \mu + \frac{d}{2}(|\mathcal{A}| - 1)$ we have that

$$\mathcal{U}(r, \mathcal{A}, \mu) = \sup_{\|y\|_\infty \leq \mu} \inf_{q : \Delta^r u = y - q} \|u\|_\infty \geq \frac{d}{2} r^{r-1} \left(\left[\frac{\frac{d}{2}}{\mu + \frac{d}{2}(|\mathcal{A}| - 1)} \right]^{c_1} e^{-c_2} \right)^r$$

In other words,

$$\mathcal{U}_+(r, \mathcal{A}, \mu) \geq \frac{d}{2} C^r r^{r-1}$$

for some $C(\mathcal{A}) > 0$, which is the desired lower bound. This bound involves only analysis, and is quite direct. The generating function approach cannot be directly applied to the double-sided sequences since $\sum_{k \in \mathbb{Z}} u_k z^k$ is not guaranteed to converge for any z, but perhaps there is some other way to find a more direct proof for the double-sided sequences (food for thought).

Week 9 (4/5/2010)

Upper Bounds for the Difference Equation

Recall that given the sequence y_n coming from sampling $y_n = f(k\tau)$, we want to find $q_n \in \mathcal{A}$ and a sequence $u \in l^\infty$ with $\Delta^r u = y - q$ and $\|u\|_\infty$ as small as possible, solving for any $\|y\|_\infty \leq \mu$. From last time we have the lower bound $\sup_{\|y\|_\infty \leq \mu} \inf_{q \in \mathcal{A}^{\mathbb{N}}} \|u\|_\infty \geq (cr)^r$ for some constant $c = c(\mu, \mathcal{A})$.

An initial attempt is to consider probabilistic choices for q, but it turns out that such probabilistic arguments do not yield a generic solution to the problem. We'll stick to $\Sigma\Delta$ constructions.

Recall also that using the greedy quantization rule for $\Delta^r u = y - q$ requires $|\mathcal{A}| \geq 2^r$, and hence for a fixed alphabet size the greedy rule cannot be used for large enough r.

The first ∞-family of $\Sigma\Delta$ schemes of arbitrary order (for a fixed alphabet) is due to Daubechies and Devore around 1998. With a specially designed sequence

$$q_n = \mathcal{Q}(u_{n-1}, u_{n-2}, ..., u_{n-r})$$

they found that the difference equation can be satisfied with $\|u\|_\infty \leq c^{r^2}$, and note $c^{r^2} \gg (cr)^r$, and hence is highly suboptimal. This bound translates to having $\|f - \tilde{f}\|_\infty \lesssim \lambda^{-c \log \lambda}$, and $\lambda^{-c \log \lambda} \gg e^{-\lambda}$, which is the corresponding lower bound for $\|f - \tilde{f}\|_\infty$. It had been an open question for awhile whether it was possible to achieve exponential accuracy in λ. This was resolved by Güntürk recently around 2002.

Infinite-Order ΣΔ Schemes with Exponential Accuracy

One ingredient is a switch in the difference equation from $\Delta^r u = y - q$ to a general difference equation $(\delta^0 - h) * v = y - q$. The goal is to find a difference equation which behaves "like" an r-th order difference equation but still allows us to use the greedy algorithm.

An initial observation is that if $h = (h_n)_{n>0}$ (causal) is such that $\delta^0 - h = \Delta^r g$ for some $g \in l^1(\mathbb{N})$, then any bounded solution v to $(\delta^0 - h) * v = y - q$ yields a bounded solution u to $\Delta^r u = y - q$ via $u = g * v$ and $\|u\|_\infty \leq \|g\|_1 \|v\|_\infty$. This is by simply plugging into the formula:

$$\Delta^r u = \Delta^r (g * v) = (\Delta^r g) * v = (\delta^0 - h) * v = y - q$$

Note that we can use the greedy algorithm to find a bounded solution for v so long as

$$\|h\|_1 + \frac{\|y\|_\infty}{d/2} \leq |\mathcal{A}|$$

Then the goal is to design h such that $\|h\|_1$ is well controlled (so that the greedy algorithm can be used) and $\delta^0 - h = \Delta^r g$ with $g \in l^1$. In particular, we would like to minimize $\|g\|_1$.

What can we expect? From last time, we have that if $\Delta^r u = w$ then

$$\|u\|_\infty \geq \|w\|_\infty \left(c_1 \frac{|w_0|}{\|w\|_\infty} \right)^{c_2 r} r^r$$

Let $u = g$ so that $w = \delta^0 - h$, and $w_0 = 1$. Then

$$\|w\|_\infty \leq \|w\|_1 = 1 + \|h\|_1$$

which shows that

$$\|g\|_1 \geq \|g\|_\infty \geq \left(\frac{c_1}{1 + \|h\|_1} \right)^{c_2 r} r^r$$

Now for the greedy rule to be applicable, we need that

$$\|h\|_1 \leq \|h\|_1 + \frac{\|y\|}{d/2} \leq |\mathcal{A}|$$

and thus we have

$$\|g\|_1 \geq \left(\frac{c_1}{1 + |\mathcal{A}|} \right)^{c_2 r} r^r \tag{2}$$

Main Optimization Problem:

$$\text{Minimize } \|g\|_1 \text{ subject to } \begin{cases} \Delta^r g = \delta^0 - h \\ \|h\|_1 \leq \gamma \\ h_j = 0, j \leq 0 \end{cases}$$

We have here l^1 norm objective function and in the constraint as well, and by introducing the appropriate slack variables we can turn this into a linear program (increasing the dimensionality of the problem)

We will be analyzing special solutions to the optimization problem.

First Reduction: Consider (h, g) pairs that have finite support. Note that given h, if $\delta^0 - h = \Delta^r g$ then $g = S^r(\delta^0 - h)$ where S as before is defined by $(S(w))_k := \sum_{j=0}^{k} w_j$.

Question: Suppose supp $h \subset \{1, ..., L(r)\}$ and $\delta^0 - h = \Delta^r g$, so that g is also finitely support. With the constraint that $\|h\|_1 \leq \gamma$, how small can we take $L(r)$ to be?

Proposition 40.

$$(S^r w)_k = \sum_{j=0}^{k} \binom{k-j+r-1}{r-1} w_j$$

Proof. Note the generating function of $S^r w$ is $\frac{w(z)}{(1-z)^r}$, which we rewrite as

$$\left(\sum_{k=0}^{\infty} z^k \right)^r \sum_{k=0}^{\infty} w_k z^k$$

If we look at the coefficient of z^k after expanding the product, contributions come from the product of $w_j z^j$ and a z^{k-j} term from $\left(\sum_{k=0}^{\infty} z^k \right)^r$. The number of such terms in the product is precisely the number of ways to split $k - j = a_1 + ... + a_r$ where $a_i \geq 0$, which is $\binom{k-j+r-1}{r-1}$, and thus the coefficient of z^k is

$$\sum_{j=0}^{k} \binom{k-j+r-1}{r-1} w_j$$

□

Corollary 41.

$$\|S^r w\|_{l_1^N} \leq \binom{N+r}{r} \|w\|_{l_1^N}$$

Proof. We have that

$$\begin{aligned} \|S^r w\|_{l_1^N} &= \sum_{k=0}^{N} \left| \sum_{j=0}^{k} \binom{k-j+r-1}{r-1} w_j \right| \\ &\leq \sum_{j=0}^{N} |w_j| \sum_{k=j}^{N} \binom{k-j+r-1}{r-1} \end{aligned}$$

Bounding the inner sum by when $j = 0$, we have

$$\begin{aligned} &\leq \|w\|_{l_1^N} \sum_{k=0}^{N} \binom{k+r-1}{r-1} \\ &= \binom{N+r}{r} \|w\|_{l_1^N} \end{aligned}$$

where the last equality follows using Pascal triangle identities:

$$\begin{aligned} \binom{N+r}{r} &= \binom{N+r-1}{r-1} + \binom{N+r-1}{r} \\ &= \binom{N+r-1}{r-1} + \binom{N+r-2}{r-1} + \binom{N+r-2}{r} \\ &\vdots \\ &= \sum_{k=0}^{N} \binom{k+r-1}{r-1} \end{aligned}$$

□

Using these observations, we have that if $\delta^0 - h = \Delta^r g$, then

$$\underbrace{1 - \sum_{n\geq 0} h_n z^n}_{\deg \leq L} = (1-z)^r \underbrace{\sum_{n\geq 0} g_n z^n}_{\deg \leq L-r}$$

Using $N = L - r$ in the previous Corollary 41, we have that

$$\begin{aligned} \|g\|_1 &= \|S^r(\delta^0 - h)\|_1 \\ &\leq \binom{L}{r} \|\delta^0 - h\|_1 \\ &\leq \binom{L}{r}(1+\gamma) \end{aligned}$$

From earlier (2) we also have the lower bound $\|g\|_1 \geq (cr)^r$. Matching the two bounds, we then have that a necessary condition for admissibility for L is

$$\binom{L}{r} \gtrsim (cr)^r$$

Using the simple bound $\binom{L}{r} \leq \frac{L^r}{r!} \lesssim \frac{L^r}{r^r e^{-r}\sqrt{2\pi r}}$ we have that

$$L^r \gtrsim (\tilde{c}\, r^2)^r$$

and hence

$$L \gtrsim \tilde{c}\, r^2$$

This addresses the question of how low the sparsity of h can be.

Let us also make another observation:

Proposition 42. *If H, G are two finite sequences such that $\Delta^r G = H$, then H has r vanishing moments, i.e.*

$$\sum_{n\geq 0} H_n\, n^j = 0,\ j = 0, 1, ..., r-1$$

Proof. The proof is by generating functions. If we consider $H(z) = \sum_{n\geq 0} H_n\, z^n$ and $G(z) = \sum_{n\geq 0} G_n\, z^n$, then we have that

$$\begin{aligned} \Delta^r G = H &\implies (1-z)^r G(z) = H(z) \\ &\implies H(z) \text{ has a zero of order 1 at } z = 1 \\ &\implies H^{(j)}(1) = 0,\ j = 0, 1, ..., r-1 \end{aligned}$$

Now we note

$$H^{(j)}(z) = \sum_{n\geq 0} H_n\, n(n-1)\cdots(n-j+1)\, z^{n-j}$$

and

$$0 = H^{(j)}(1) = \sum_{n\geq 0} H_n (n^j + P_{j-1}(n))$$

where $P_{j-1}(n)$ is a polynomial in n of degree $\leq j-1$. If we use induction, supposing that $\sum_{n\geq 0} H_n\, z^k = 0$ for $k \leq j-1$, then $\sum_{n\geq 0} H_n\, P_{j-1}(n) = 0$, then we see that from the equality above,

$$\sum_{n\geq 0} H_n n^j = 0$$

The base case is trivial, since $\sum_{n\geq 0} H_n = H^{(0)}(1) = 0$. □

So, if we have $\Delta^r g = \delta^0 - h$, then $\delta^0 - h$ has r vanishing moments.

Let us write out these conditions, which form linear constraints on $h = (0, h_1, ..., h_L)$. Note that

$$0 = \sum_{n\geq 0} (\delta^0 - h)_n\, n^l \Longleftrightarrow \sum_{n\geq 0} h_n\, n^l = \begin{cases} 1 & l = 0 \\ 0 & l \neq 0 \end{cases}$$

In matrix form,

$$\begin{pmatrix} 1 & 1 & ... & 1 \\ 1 & 2 & ... & L \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 2^{r-1} & ... & L^{r-1} \end{pmatrix} \begin{pmatrix} h_1 \\ h_2 \\ \vdots \\ h_L \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

Note that we have an $r \times L$ matrix, and we make the following observation:

Remark 43. Under the conditions above, h must have at least r nonzero coordinates. Hence, the *minimally* supported choices of h will have exactly r nonzero coefficients.

Proof. Note that the above equation reduces to taking only the columns of the matrix corresponding to where h is nonzero. Then we show that if we take fewer than r columns of the matrix, $\begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$ will not be in its span, and hence there is no solution h satisfying the equation.

To show this, the idea is that if we take columns $n_1, n_2, ..., n_{r-1}$, we have

$$\begin{pmatrix} 1 & 1 & ... & 1 \\ n_1 & n_2 & ... & n_{r-1} \\ \vdots & \vdots & \ddots & \vdots \\ n_1^{r-1} & n_2^{r-1} & ... & n_{r-1}^{r-1} \end{pmatrix} \begin{pmatrix} h_{n_1} \\ h_{n_2} \\ \vdots \\ h_{n_{r-1}} \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

This is an $r \times r-1$ matrix, and if we look at the $(r-1) \times (r-1)$ submatrix

$$\begin{pmatrix} n_1 & ... & n_{r-1} \\ \vdots & \ddots & \vdots \\ n_1^{r-1} & ... & n_{r-1}^{r-1} \end{pmatrix} \begin{pmatrix} h_{n_1} \\ \vdots \\ h_{n_{r-1}} \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}$$

we note that the matrix is invertible since it is the product of a Vandermonde matrix and a diagonal matrix

$$\begin{pmatrix} n_1 & ... & n_{r-1} \\ \vdots & \ddots & \vdots \\ n_1^{r-1} & ... & n_{r-1}^{r-1} \end{pmatrix} = \begin{pmatrix} 1 & ... & 1 \\ \vdots & \ddots & \vdots \\ n_1^{r-2} & ... & n_{r-1}^{r-2} \end{pmatrix} \begin{pmatrix} n_1 & & \\ & \ddots & \\ & & n_{r-1} \end{pmatrix}$$

and thus $h_{n_1} = ... = h_{n_{r-1}} = 0$. But this contradicts the first equation $\sum h_{n_k} = 1$. Thus h must have at least r nonzero coordinates. $\square$

Let us consider minimally supported h on r nonzero coordinates, so that

$$h = \sum_{j=1}^{r} c_j \delta^{n_j}$$

i.e. $h_{n_j} = c_j$ and is zero elsewhere. By the matrix formulation, we see that as soon as we specify $n_1, ..., n_r$, then the c_j are uniquely determined. There is a nice explicit solution for the c_j which can be found through ideas of Lagrange interpolation.

Let

$$V = \begin{pmatrix} 1 & 1 & ... & 1 \\ n_1 & n_2 & ... & n_r \\ \vdots & \vdots & \ddots & \vdots \\ n_1^{r-1} & n_2^{r-1} & ... & n_r^{r-1} \end{pmatrix}$$

and $c = \begin{pmatrix} c_1 \\ \vdots \\ c_r \end{pmatrix}$ be the coefficients, so that we wish to solve $Vc = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$.

We will seek rows $d^{(j)} := (d_1^{(j)}, ..., d_r^{(j)})$ so that $d^{(j)}V = e_j$ (1 at j, 0 elsewhere). this implies that if we take

$$D = \begin{pmatrix} — & d^{(1)} & — \\ & \vdots & \\ — & d^{(r)} & — \end{pmatrix}$$

then $DV = I$, so $D = V^{-1}$. This implies that $c_j = DVc = De_1 = d_1^{(j)}$. Now, the condition $d^{(j)}V = e_j$ means that

$$\sum_{i=1}^{r} d_i^{(j)} n_k^{j-1} = e_j(k) = \delta_{j,k}$$

If we let $P_j(t) := \sum_{i=1}^{r} d_i^{(j)} t^{i-1}$, then this implies that $P_j(n_k) = \delta_{j,k}$ for $1 \le k \le r$. Then P_j has an explicit form as a Lagrange polynomial:

$$P_j(t) = \prod_{i \neq j} \frac{t - n_i}{n_j - n_i}$$

Now we are only interested in $d_1^{(j)}$, which is the constant term of P_j, which can be extracted by $P_j(0)$, so

$$c_j = d_1^{(j)} = P_j(0) = \prod_{i \neq j} \frac{-n_i}{n_j - n_i} = \prod_{i \neq j} \frac{n_i}{n_i - n_j}$$

So what we have shown is that if we are seeking a minimally sparse h for this $\Sigma\Delta$ scheme, then we have an explicit form for h. Thus, we have reduced the problem to selecting $n_1 < n_2 < ... < n_r$ such that

$$\|h\|_1 = \sum_{j=1}^{r} |c_j| = \sum_{j=1}^{r} \prod_{i \neq j} \frac{n_i}{|n_i - n_j|} \le \gamma$$

so that $\|g\|_1 = \|S^r(\delta^0 - h)\|_1$ is minimized.

We are now confronted with a highly nonlinear problem. Noting the upper bound

$$\|g\|_1 \leq \binom{n_r}{r}(\gamma + 1)$$

it seems that a first step would be to choose n_r as small as possible. It turns out that

Proposition 44.

$$\|g\|_1 = \frac{1}{r!} \prod_{i=1}^{r} n_i$$

Proof. This is a long exercise. A hint is to show that $g_n \geq 0$ and then compute $\sum g_n$. □

Notice that

$$\frac{n_1 \cdots n_r}{r!} \leq \frac{n_r(n_r - 1)\cdots(n_r - r + 1)}{r!} = \binom{n_r}{r}$$

so that the upper bound is not very good. Thus we have

The Full Minimization Problem:

$$\text{Minimize } \prod_{i=1}^{r} n_i \text{ such that } \sum_{j=1}^{r} \prod_{i \neq j} \frac{n_i}{|n_i - n_j|} \leq \gamma$$

We know that $n_r \geq r^2$ from the bounds on $\|g\|_1$. Perhaps we can try $n_j = j^2$ as a natural boundary case. It turns out that

$$\|h\|_1 = \sum_{j=1}^{r} \prod_{k \neq j} \frac{k^2}{|k^2 - j^2|} \approx \sqrt{\pi r} \to \infty$$

(left as an exercise) which is no good for large r. This is slightly discouraging, but experimentally there are integer programming solutions to be found. After analyzing the pattern of these solutions, a special sequence that turns out to work is given in the following theorem.

Theorem 45. *Let $\sigma \geq 1$ be an integer, and set $n_j = 1 + \sigma(j-1)^2$ for $j = 1, 2, \ldots$ Then*

$$\|h\|_1 \leq \cosh\left(\frac{\pi}{\sqrt{\sigma}}\right) \quad \textit{for all } r > 0$$

This bound is tight as $r \to \infty$.

Remark 46. A few notes:

- In our problem the condition that $\sum_n h_n = 1$ shows that $\sum |h_n| \geq 1$, and hence $\gamma \geq 1$
- For any $\gamma > 1$, we can find σ so that $1 \leq \cosh\left(\frac{\pi}{\sqrt{\sigma}}\right) < \gamma$, noting $\cosh(0) = 1$ and is continuous.
- $\|g\|_1 \leq \frac{1}{r!} \prod_{j=1}^{r} (1 + \sigma(j-1)^2) \leq \left(\frac{\sigma}{e} r\right)^r$, which matches the asymptotic order of the lower bound (up to the constant in the exponent). This means that exponential accuracy is achievable for $\Sigma\Delta$!

Special Case: (arguably the most important case) For 1-bit quantization, $\mathcal{A} = \{\pm 1\}$, we have that

$$2^{-\lambda} = |\mathcal{A}|^{-\lambda} \leq \|f - \tilde{f}\|_\infty \leq 2^{-c\lambda}$$

where $c = 0.078$. This is an order of magnitude off from optimal.

Later, from the PhD thesis of F. Krahmer, it was shown that in the optimization problem

$$\min \prod_{i=1}^{r} n_i \text{ s.t. } \sum_{j=1}^{r} \prod_{i \neq j} \frac{n_i}{|n_i - n_j|} \leq \gamma$$

the optimal $(n_j)_1^r$ are distributed asymptotically according to the zeros of the (Type II) Chebyshev polynomials. Using this result gives the upper bound $\|f - \tilde{f}\|_\infty \leq 2^{-c\lambda}$ with constant $c \sim 0.102$. This means that if we restrict our attention to minimally supported h, then $c \sim 0.102$ is the best possible constant.

Intuitively, we expect this to be the best we can do, since the structure of the l^1 norm gives rise to sparse solutions when minimizing the l^1 norm. There is some ongoing work to show that the constant $c \sim 0.102$ is near optimal for all schemes based on the difference equation $(\delta^0 - h) * v = y - q$.

Week 10 (4/12/2010)

Compressed Sensing

In the second week there was an overview of compressed sensing. Here we quickly recall the setting and notation. We are interested in recovering sparse vectors $x \in \Sigma_s^N = \{x \in \mathbb{R}^N, |\text{supp}(x)| \leq s\}$ from m linear measurements $\{l_i(x)\}_{i=1}^m$, where $s < m \ll N$. In matrix notation,

$$y = \Phi x$$

and Φ is an $m \times N$ measurement matrix (linear measurements can be expressed as row vectors).

In general, we call $f \in \mathbb{R}^N$ sparse with respect to a given [orthonormal] basis ψ_i (which we place in the columns of Ψ) if $f = \sum x_i \psi_i = \Psi x$ with $x \in \Sigma_s^N$. Note that we can figure out which s coefficients are nonzero by computing all the coefficients: $x = \Psi^* f$, but this requires N measurements. We want to recover with fewer measurements, using some measurement matrix $y = \mathbb{M} f$. Thus the problem reduces to recovering x from $y = \mathbb{M} f = \mathbb{M}\Psi x$, and we can then consider $\Phi = \mathbb{M}\Psi$ as our measurement matrix on Σ_s^N.

Recall that if there is any hope for recovery, a necessary condition is that Φ should be injective on Σ_s^N, so that if $\Phi x_1 = \Phi x_2$ for $x_1, x_2 \in \Sigma_s^N$, then $x_1 = x_2$. An equivalent condition examines the kernel of Φ, that

$$\ker \Phi \cap \Sigma_{2s}^N = \{0\}$$

or that every $m \times 2s$ submatrix of Φ is full rank ($2s$ linearly independent columns). Note that this means $m \geq 2s$.

Also, the main consequence is that if there is an s-sparse solution $x \in \Sigma_s^N$ to the equation $y = \Phi z$ (solving for z), then x is the sparsest such solution. This leads to the problem

$$(P_0) \quad \min \|z\|_0 \quad \text{s.t.} \quad \Phi z = y$$

which is a combinatorially difficult problem (naively can enumerate all $\binom{N}{s}$ potential supports of sparse solutions z and solve). The convex relaxation of this problem is

$$(P_1) \quad \min \|z\|_1 \quad \text{s.t.} \quad \Phi z = y$$

which will be useful only if it yields the same solution. Under stronger conditions on Φ this is the case, and in fact we saw an equivalent condition for $P_0 = P_1$.

Proposition 47. *For all $x \in \Sigma_s^N$, x is the unique solution to $y = \Phi z$ (solving for z) if and only if for all $\eta \in \ker \Phi$, and for all $T \subset [N]$ with $|T| \leq s$,*

$$\|\eta_T\|_1 < ||\eta_{T^c}\|_1$$

($\eta_T = \eta\chi_T$). This is called the "Null Space Property" (NSP) of order s.

We also proved this in the overview. Verifying the null space property is also a combinatorially hard property to verify deterministically, and for this reason we will turn to random constructions.

We also turn to a stronger, more accessible property, called the Restricted Isometry Property (RIP).

Definition: We say that an $m \times N$ matrix Φ with $s < m$ satisfies RIP with constant $\delta < 1$ and order k which we denote by $\mathrm{RIP}(k, \delta)$ if

$$(1-\delta)\|x\|_2^2 \leq \|\Phi x\|_2^2 \leq (1+\delta)\|x\|_2^2, \quad \text{for all } x \in \Sigma_k^N$$

(every k columns of Φ is a near isometry on $\mathbb{R}^k$). Another way to write this is to extract the columns of T corresponding to an index set T with $|T| \leq k$, and writing

$$(1-\delta)\|x\|_2^2 \leq \|\Phi_T x\|^2 \leq (1+\delta)\|x\|_2^2 \text{ for all } x \in \mathbb{R}^{|T|}$$

we will be using both notations interchangeably.

Fixing k, we denote $\delta_k(\Phi)$ by the smallest such δ satisfying this property.

Remark 48. A sufficiently strong RIP condition will imply NSP of order s, but the benefit of this property is that we have $\|\cdot\|_2$ norms now, with more tools for dealing with these, such as studying eigenvalues. In fact, $\mathrm{RIP}(k, \delta)$ is equivalent to the condition that for all index sets $|T| \leq k$,

$$1 - \delta \leq \lambda_i(\Phi_T^t \Phi_T) \leq 1 + \delta \quad \text{for } i = 1, ..., |T|$$

or in other words

$$\sqrt{1-\delta} \leq \sigma_i(\Phi_T) \leq \sqrt{1+\delta} \text{ for } i = 1, ..., |T|$$

where λ_i denote eigenvalues and σ_i denote singular values.

Also, we have the following properties:

- $\delta_1(\Phi) \leq \delta_2(\Phi) \leq ...$ noting that if $\Phi \in \mathrm{RIP}(k, \delta)$, then $\Phi \in \mathrm{RIP}(k', \delta)$ for $k' < k$, i.e. the same δ will work for smaller orders. This is just by definition. In fact, for the above remark this shows that it suffices to consider $|T| = k$.

- We have the following characterization of $\delta_k(\Phi)$:

$$\delta_k(\Phi) = \max_{|T|=k} \|\Phi_T^t \Phi_T - \mathrm{Id}_T\|_{\mathrm{op}}$$

Proof. Rearranging the definition of RIP, the RIP condition says that

$$\max_{|T|=k} \sup_{x\in\mathbb{R}^{|T|},\|x\|=1} \left|\left\langle \Phi_T^t\Phi_T x, x\right\rangle - \langle x, x\rangle\right| \le \delta_k(\Phi)$$

From linear algebra, we know that for a symmetric matrix A, $\|A\|_{\text{op}} = \sup_{\|x\|_2=1} |\langle Ax, x\rangle|$. Thus, the above becomes

$$\max_{|T|=k} \|\Phi_T^t\Phi_T - \text{Id}_{|T|}\|_{\text{op}} \le \delta_k(\Phi)$$

since $\delta_k(\Phi)$ is the smallest such δ, we have equality here. □

- Here is a technical lemma that will help in verifying that RIP implies NSP. Let $x \in \Sigma_k^N$ and $x' \in \Sigma_{k'}^N$ such that $\text{supp}(x)\cap\text{supp}(x') = \varnothing$. Then

$$|\langle \Phi x, \Phi x'\rangle| \le \delta_{k+k'}(\Phi)\,\|x\|_2\,\|x'\|_2 \tag{3}$$

Proof. This essentially follows by Cauchy Schwarz, and the computation uses the fact that $\langle x, x'\rangle = 0$ since the supports are disjoint:

$$\begin{aligned} |\langle \Phi x, \Phi x'\rangle| &= |\left\langle \Phi^t\Phi x, x'\right\rangle| \\ &= |\left\langle (\Phi^t\Phi - \text{Id})x, x'\right\rangle| \\ &\le \|\Phi_{T\cup T'}^t\Phi_{T\cup T'} - \text{Id}_{T\cup T'}\|_{\text{op}}\,\|x\|_2\,\|x'\|_2 \\ &\le \delta_{k+k'}(\Phi)\,\|x\|_2\,\|x'\|_2 \end{aligned}$$

The last inequalities follow by noting that x, x' are supported on $T\cup T'$ and hence we can consider only those columns of Φ. □

These observations lead us to the following implication:

Theorem 49. (Candés, Romberg Tao) *Let $\delta_{2s}(\Phi) < \frac{1}{3}$. Then Φ satisfies NSP of order s. Consequently, every s-sparse vector $x\in\Sigma_s^N$ solves (P_1).*

Proof. We want to show that

$$\|\eta_T\|_1 < \|\eta_{T^c}\|_1 \quad \text{for all } |T|\le s, \eta\in\ker\Phi$$

Given $\eta\in\ker\Phi$, it suffices to show this result for the s largest entries of η (in absolute value). Let T be the indices corresponding to the s largest entries of η. Now we split T^c into blocks of size s as well, in *decreasing order*: so let S_1 be the next s largest entries of η after T, S_2 be the following s largest entries, etc. Thus we have $T^c = \bigcup_{i=1}^K S_k$ where

$$\min_{i\in S_k} |\eta_i| \ge \max_{j\in S_{k+1}} |\eta_j|$$

with $|S_k| = s$ except possibly the left-over last term $|S_K| \le s$.

Since $\eta\in\ker\Phi$, we have that $\Phi\eta = 0$, $\Phi\eta_T = \Phi(-\eta_{T^c})$ so that

$$\Phi\,\eta_T = \Phi(-\eta_{T^c}) = \sum_{k=1}^K \Phi(-\eta_{S_k})$$

Now note that by Cauchy Schwarz with η and χ_T, and applying RIP,

$$\|\eta_T\|_1 \leq \sqrt{s}\,\|\eta_T\|_2 \leq \frac{\sqrt{s}}{1-\delta_s}\frac{\|\Phi\,\eta_T\|_2^2}{\|\eta_T\|_2}$$

Studying $\|\Phi\eta_T\|^2$ now, we use observation (3) above so that

$$\begin{aligned}\|\Phi\,\eta_T\|_2^2 &= \left\langle \Phi\,\eta_T,\ \sum_{k=1}^{K}\Phi(-\eta_{S_k})\right\rangle \\ &= \sum_{k=1}^{K}\langle \Phi\,\eta_T, \Phi(-\eta_{S_k})\rangle \\ &\leq \delta_{2s}(\Phi)\,\|\eta_T\|_2\left(\sum_{k=1}^{K}\|\eta_{S_k}\|_2\right)\end{aligned}$$

We would like to control the 2-norms here in terms of 1-norms, and for this we can make use of the property of S_k being ordered. Since all the values of S_{k+1} are less than those of S_k, we note that they are also less than the average:

$$\|\eta_{S_{k+1}}\|_\infty \leq \frac{1}{s}\sum_{i\in S_k}|\eta_i| = \frac{1}{s}\|\eta_{S_k}\|_1$$

Consequently we have

$$\|\eta_{S_{k+1}}\|_2 \leq \sqrt{s}\,\|\eta_{S_{k+1}}\|_\infty \leq \frac{1}{\sqrt{s}}\|\eta_{S_k}\|_1$$

and also $\|\eta_{S_1}\|_2 \leq \frac{1}{\sqrt{s}}\|\eta_T\|_1$

This implies that above we have

$$\begin{aligned}\|\eta_T\|_1 &\leq \frac{\sqrt{s}}{1-\delta_s}\frac{\|\Phi\,\eta_T\|_2^2}{\|\eta_T\|_2} \\ &\leq \frac{\delta_{2\,s}\sqrt{s}}{1-\delta_s}\left(\sum_{k=1}^{K}\|\eta_{S_k}\|_2\right) \\ &\leq \frac{\delta_{2s}}{1-\delta_s}\left(\|\eta_T\|_1 + \sum_{k=1}^{K-1}\|\eta_{S_k}\|_1\right) \\ &\leq \frac{\delta_{2s}}{1-\delta_s}(\|\eta_T\|_1 + \|\eta_{T^c}\|)\end{aligned}$$

and since $\delta_s \leq \delta_{2s} < \frac{1}{3}$, this means $\frac{\delta_{2s}}{1-\delta_s} < \frac{1/3}{2/3} = \frac{1}{2}$ and rearranging above, we have that

$$\|\eta_T\|_1 < \|\eta_{T^c}\|_1$$

as desired. □

This theorem is not completely optimized, can improve constants by examining different size blocks, etc. In fact work has been done to push for $\delta_{2s} < 0.47$ by tweaking the proof.

Coherence and RIP

Another way to view the RIP condition is through the idea of coherence. Recall

$$\delta_k(\Phi) = \max_{|T|=k}\|\Phi_T^t\Phi_T - \mathrm{Id}\|_{\mathrm{op}}$$

and that if δ_k is small, then the columns of every $m \times k$ submatrix Φ_T with $|T| = k$ are "almost orthonormal" (Note that we cannot expect all submatrices Φ_T to be orthonormal without violating dimensionality constraints. For instance, $m < N$ means that the columns of Φ are necessarily linearly dependent, so they cannot all be orthonormal to each other, etc).

Denote the columns of Φ by

$$\Phi = \begin{pmatrix} | & & | \\ \varphi_1 & \cdots & \varphi_N \\ | & & | \end{pmatrix}, \|\varphi_i\| = 1$$

We will assume the columns to be normalized as notated above. We define the **coherence** $\mu(\Phi)$ to be

$$\mu(\Phi) = \max_{i \neq j} |\langle \varphi_i, \varphi_j \rangle|$$

Note that $(\Phi^t \Phi)_{ij} = \langle \varphi_i, \varphi_j \rangle$. We have the following facts from linear algebra:

Proposition 50. *If A is symmetric, then*

$$\|A\|_{2,2} \leq \|A\|_{1,1}$$

(In general, $\|A\|_{2,2} \leq \max\{\|A\|_{1,1}, \|A\|_{\infty,\infty}\}$. Also,

$$\|A\|_{1,1} = \max_i \|Ae_i\|_1$$

(i.e. the max column sum)

As a consequence, we see that for any $|T| \leq k$,

$$\|\Phi_T^t \Phi_T - \mathrm{Id}\|_{1,1} \leq (k-1)\,\mu(\Phi)$$

(summing $k-1$ inner products, the Id cancels out the diagonal of $\Phi_T^t \Phi_T$) This implies that

$$\delta_k(\Phi) \leq (k-1)\mu(\Phi)$$

Also, as an exercise, can check that $\delta_2(\Phi) = \mu(\Phi)$.

We want to see what sort of bounds on δ_k we can obtain. The Welch bound, which we will not prove here, says the following:

Proposition 51. (Welch Bound)

$$\mu(\Phi) \geq \frac{1}{\sqrt{m}} \left(\frac{N-m}{N-1} \right)^{1/2} \sim \frac{1}{\sqrt{m}} \quad \text{if } N \gg m$$

This implies that if we want to bound $\delta_k(\Phi) < 1$ via studying coherence, then we need $k \lesssim \frac{1}{\sqrt{m}}$ if we are using the bound

$$\delta_k(\Phi) \lesssim k\mu(\Phi)$$

(note that this may not even be sufficient).

For the Welch Bound, we can provide a heuristic argument for why we should expect such a bound via a simple example. Suppose that $\varphi_1, ..., \varphi_m$ is an orthornomal basis for $\mathbb{R}^m$. Let ψ be any other unit vector and consider

$$\Phi = \begin{pmatrix} | & & | & | \\ \varphi_1 & ... & \varphi_m & \psi \\ | & & | & | \end{pmatrix}$$

Then we have that $\psi = \sum_{i=1}^m \langle \psi, \varphi_i \rangle \varphi_i$ and $1 = \|\psi\|_2 = \sum_{i=1}^m |\langle \psi, \varphi_i \rangle|^2$. This implies that

$$m \max_i |\langle \psi, \varphi_i \rangle|^2 \geq \sum_{i=1}^m |\langle \psi, \varphi_i \rangle|^2 = 1$$

and thus $\mu(\Phi) = \max_i |\langle \psi, \varphi_i \rangle| \geq \frac{1}{\sqrt{m}}$

Now we ask whether the Welch Bound is sharp. Can we attain the lower bound?

The bound $\mu(\Phi) \leq \frac{c}{\sqrt{m}}$ can be easily achieved for $N = 2m$.

Example 52. Take

$$\Phi = \begin{pmatrix} I & | & D \end{pmatrix}$$

where D is the discrete cosine transform, a basis $d_1, ..., d_m$ with $\|d_i\|_\infty \leq \frac{c}{\sqrt{m}}$. Then

$$\mu(\Phi) \leq \frac{c}{\sqrt{m}}$$

Other examples with $N = m^2$ also exist (polynomials, chirps)

In any case, with just bounds using coherence and the Welch Bound (and the operator norm bound), the barrier $m \sim k^2$ cannot be broken.

Probabilistic Methods

Let

$$\Phi = \begin{pmatrix} \varphi_{ij} \end{pmatrix}_{\substack{1 \leq i \leq m \\ 1 \leq j \leq N}}$$

with φ_{ij} i.i.d. Recall that Φ satisfies RIP(k, δ) if (Remark 48)

$$\sqrt{1-\delta} \leq \sigma_i(\Phi_T) \leq \sqrt{1+\delta} \quad \text{for all } i = 1, ..., k$$

for all $|T| = k$. In other words, every $m \times k$ submatrix Φ_T is well conditioned. It is well known that rectangular random matrices (with i.i.d Gaussian entries, for instance) are well conditioned. In fact, we have the following theorem.

Theorem 53. *Let A be a random $m \times k$ matrix, with A_{ij} i.i.d $N\left(0, \frac{1}{m}\right)$. Then*

$$\Pr\left(\left\{ \sqrt{1-\delta} \leq \sigma_i(A) \leq \sqrt{1+\delta} \text{ for all } i \right\} \right) \geq 1 - e^{-c(\delta)m}$$

First, why $N\left(0, \frac{1}{m}\right)$? Note that

$$\mathrm{E}(\|Ae_j\|^2) = \mathrm{E}\left(\sum_{i=1}^{m} |A_{ij}|^2\right) = 1 \text{ for all } j$$

This also implies $\mathrm{E}(\|Ax\|^2) = \|x\|^2$ by linearity of expectation, and so the expected norm is preserved. We will prove this theorem next time, but first let's see how it is used.

Implication for RIP: Take Φ to be a random $m \times N$ matrix with $\Phi_{ij} \sim N\left(0, \frac{1}{m}\right)$ i.i.d. Then for any $|T| = k$, consider $A = \Phi_T$. In the theorem above let

$$\varepsilon(A) = \left\{\sqrt{1-\delta} \le \sigma_i(A) \le \sqrt{1-\delta}, \text{ for all } i\right\}$$

and so $\Pr(\varepsilon(A)) \ge 1 - e^{-c(\delta)m}$. Then we note that

$$\Pr(\delta_k(\Phi) \le \delta) = \Pr\left(\bigcap_{|T|=k} \varepsilon(\Phi_T)\right)$$

and using the "union bound" (probability of a union of events bounded by the sum of the probabilities), we have that

$$\begin{aligned}\Pr\left[\left(\bigcap_{|T|=k} \varepsilon(\Phi_T)\right)^c\right] &\le \sum_{|T|=k} \Pr[\varepsilon(\Phi_T)^c] \\ &\le \binom{N}{k} e^{-c(\delta)m}\end{aligned}$$

This shows that if $\ln\binom{N}{k} \le \frac{1}{2}c(\delta)m$, then this is bounded by $e^{-\frac{1}{2}c(\delta)m}$. Since

$$\binom{N}{k} \le \frac{N^k}{k!} \lesssim \frac{N^k}{k^k e^{-k}}$$

we have that

$$\ln\binom{N}{k} \lesssim k \ln\left(\frac{eN}{k}\right)$$

Therefore if $k \ln\left(\frac{eN}{k}\right) \le \frac{1}{2}c(\delta)m$, then $\delta_k(\Phi) \le \delta$ with probability $\ge 1 - e^{-\frac{1}{2}c(\delta)m}$. We can write this as

$$\frac{m}{k} \gtrsim \ln\left(\frac{N}{k}\right)$$

in terms of $\frac{m}{k}$, the ratio of measurements to sparsity, a measurement of redundancy. There are many ways to interpret this inequality (depending on which variable is under study). In the paper by Donoho and Tanner, they investigate phase transitions when taking $m, N \to \infty$ while fixing the ratios $\frac{m}{k}$ and $\frac{N}{k}$. In the end, everything boils down to bounds for singular values of random matrices.

Remark 54. We used the Gaussian here, but we can use other matrices such as Bernoulli random variables. The importance here is that "most" matrices will satisfy RIP with high probability (and with this exponential bound on the failure probability, the appropriate term is *overwhelming* probability).

Other considerations are practical, for instance, if we want to be able to store the matrices, then we should use Bernoulli if possible, $(0,1)$-valued, which will be relatively sparse. Or even better, we can look at structured random matrices such as randomly sampling rows of the discrete Fourier transform matrix (which gives a range like $\frac{m}{k} \gtrsim \left(\ln \frac{N}{k}\right)^4$.

Week 11 (4/19/2010)

Quickly reviewing what we did last time, given an $m \times N$ measurement matrix Φ, we defined $\delta_k(\Phi)$ to be the smallest number δ such that for all $x \in \Sigma_k^N$ such that

$$(1-\delta)\|x\|_2^2 \leq \|\Phi x\|_2^2 \leq (1+\delta)\|x\|_2^2$$

Also, $\delta_k(\Phi) = \max_{|T|=k} \|\Phi_T^t \Phi_T - \mathrm{Id}_T\|_{\mathrm{op}}$. Also, if $\Phi = \begin{pmatrix} | & & | \\ \varphi_1 & \dots & \varphi_N \\ | & & | \end{pmatrix}$, with $\|\varphi_i\| = 1$, then we considered the coherence $\mu(\Phi) = \max_{i \neq j} |\langle \varphi_i, \varphi_j \rangle|$, which led us to the simple bound

$$\delta_k(\Phi) \leq (k-1)\mu(\Phi)$$

Since it suffices for $\delta_k(\Phi) < \frac{1}{3}$ $(k = 2s)$, we would like $\mu(\Phi) \sim \frac{1}{k}$. The Welch bound tells us that

$$\mu(\Phi) \geq \frac{1}{\sqrt{m}} \sqrt{\frac{N-m}{N-1}} \sim \frac{1}{\sqrt{m}}$$

for $N \gg m$. Thus, through coherence methods, the best we can hope for in order to attain $\delta_k < 1$ is $k \leq \sqrt{m}$.

Last time we did not prove the Welch bound, but in fact it is not difficult. Welch showed that

$$\sum_{i,j} |\langle \varphi_i, \varphi_j \rangle|^2 \geq \frac{N^2}{m}$$

which implies the result since the LHS is just

$$\sum_{i=1}^N \|\varphi_i\|_2^4 + \sum_{i \neq j} |\langle \varphi_i, \varphi_j \rangle|^2 \leq N + N(N-1)\mu(\Phi)^2$$

Thus

$$\mu(\Phi)^2 \geq \frac{\frac{N^2}{m} - N}{N(N-1)} = \frac{\frac{N}{m} - 1}{N-1} = \frac{1}{m} \cdot \frac{N-m}{N-1}$$

which gives the Welch bound. We will show a stronger assumption, lifting the norm assumption on $\|\varphi_i\|$:

Proposition 55. (Waldon) *For all $\varphi_1, \dots, \varphi_N$,*

$$\sum_{i=1}^N \sum_{j=1}^N |\langle \varphi_i, \varphi_j \rangle|^2 \geq \frac{1}{m} \left(\sum_{i=1}^N \|\varphi_i\|_2^2 \right)^2$$

Proof. Note that $(\Phi^t \Phi)_{ij} = \langle \varphi_i, \varphi_j \rangle$, and that the LHS is just

$$\|\Phi_t \Phi\|_F^2 = \mathrm{tr}((\Phi^t \Phi)^2) = \|\Phi \Phi^t\|_F^2$$

($\|\cdot\|_F$ denotes the Frobenius norm, the l^2 norm on the entries of the matrix). Note $\Phi\Phi^t$ is an $m \times m$ matrix, and we can diagonalize to obtain $\Phi\Phi^t = U\Lambda U^t$ for U orthogonal and $\Lambda = \text{diag}(\lambda_1, ..., \lambda_m)$. Then we note that $(\Phi\Phi^t)^2 = U\Lambda^2 U^t$ so that $\text{tr}((\Phi\Phi^t)^2) = \text{tr}(\Lambda^2) = \sum_{i=1}^m \lambda_i^2$. Note that by Cauchy Schwarz,

$$\text{tr}(\Phi\Phi^t)^2 = \left(\sum_{i=1}^m \lambda_i\right)^2 \leq m \sum_{i=1}^m \lambda_i^2 = m\,\text{tr}((\Phi\Phi^t)^2)$$

and since $\text{tr}(\Phi\Phi^t) = \|\Phi\|_F^2 = \sum_{i=1}^m \|\varphi_i\|_2^2$, we have that

$$\sum_{i=1}^N \sum_{j=1}^N |\langle \varphi_i, \varphi_j\rangle|^2 = \text{tr}((\Phi\Phi^t)^2) \geq \frac{1}{m}\text{tr}(\Phi\Phi^t)^2 = \frac{1}{m}\left(\sum_{i=1}^N \|\varphi_i\|_2^2\right)^2$$

which is the desired result. The only inequality occurs in the application of Cauchy Schwarz, and we can then study when equality holds. Equality holds if and only if all the λ_i are identical, which implies that $\varphi_1, ..., \varphi_m$ form a tight frame.

□

Probabilistic Methods (continued)

With probabilistic methods, we can break the barrier imposed by Welch's bound and coherence methods, and we attain RIP with $\delta_k(\Phi) < 1$ with $m \approx k$ (up to a log N term). This is a nonconstructive method, which states that our random construction attains RIP with very high probability.

Theorem 56. *Let Φ be an $m \times N$ random matrix with $\Phi_{ij} \sim N(0, 1/m)$ i.i.d. Then for any $\delta > 0$,*

$$\Pr[\delta_k(\Phi) \geq \delta] \leq \exp(-c(\delta)m)$$

for $\frac{m}{k} \gtrsim c_1(\delta) \ln \frac{c_2(\delta)N}{k}$.

We will start working with individual $m \times k$ submatrices $A = \Phi_T$, where $|T| = k$. We need that for each such index set T,

$$(1-\delta)\|x\|_2^2 \leq \|Ax\|_2^2 \leq (1+\delta)\|x\|_2^2$$

i.e. that $\sqrt{1-\delta} \leq \sigma_i(A) \leq \sqrt{1+\delta}$ for $i = 1, ..., k$.

Strategy:

1. First we will show that for a *fixed* u, that $\|Au\|_2^2$ is concentrated around $\|u\|_2^2$. Without loss of generality we will assume $\|u\|_2 = 1$. Taking squares here will be convenient for examining sums of random variables, but we will be using the fact that

$$\{c_1\|u\|_2^2 \leq \|Au\|_2^2 \leq c_2\|u\|_2^2\} = \{\sqrt{c_1}\|u\|_2 \leq \|Au\|_2 \leq \sqrt{c_2}\|u\|_2\}$$

2. (ε-net argument) Next, we obtain an ε-net $X_\varepsilon \subset S^{k-1}$ for S^{k-1}, where $S^{k-1} = \{\|x\| = 1\}$. Having obtained bounds $C_1 \leq \|Au\|_2 \leq C_2$ for all $u \in X_\varepsilon$, then we will obtain bounds for $\|Ax\|$ for all $x \in S^{k-1}$. Given $x \in S^{k-1}$ we can find $u \in X_\varepsilon$ for which $\|x - u\|_2 \leq \varepsilon$ (from the ε-net), and then

$$\|Ax\|_2 \leq \|Au\|_2 + \|A(x-u)\|_2 \leq C_2 + \|A\|_{\text{op}}\,\varepsilon$$

Taking supremum over all $x \in S^{k-1}$ we thus have that $\|A\|_{\text{op}} \leq C_2 + \|A\|_{\text{op}}\varepsilon$, and hence

$$\|A\|_{\text{op}} \leq \frac{C_2}{1-\varepsilon}$$

We can also find a lower bound using reverse triangle:

$$\|Ax\|_2 \geq \|Au\|_2 - \|A(x-u)\|_2 \geq C_1 - \frac{C_2\,\varepsilon}{1-\varepsilon}$$

Thus, having shown $C_1 \leq \|Au\|_2 \leq C_2$ for all $u \in X_\varepsilon$, we have

$$C_1 - \frac{C_2\,\varepsilon}{1-\varepsilon} \leq \|Ax\|_2 \leq \frac{C_2}{1-\varepsilon}$$

and of course it will be in our interest to find C_1, C_2 as close to 1 as desired. Therefore, this shows that it is enough to obtain concentration results for $\|Au\|_2$ for $u \in X_\varepsilon$.

As an aside, above we showed that $\|A\|_{\text{op}} \leq \frac{1}{1-\varepsilon}\sup_{u\in X_\varepsilon}\|Au\|_2$. More generally, if Y_ρ is an ρ-net for S^{m-1}, then since $\|Au\|_2 = \sup_{\|y\|=1}|\langle Au, y\rangle|$, we see that

$$\|A\|_{\text{op}} \leq \frac{1}{(1-\varepsilon)(1-\rho)}\sup_{u\in X_\varepsilon}\sup_{y\in Y_\rho}|\langle Au, y\rangle|$$

This leads us to the following question concerning step (2) above:

Question: What is the smallest cardinality of an ε-net X_ε for S^{k-1}?

The answer is, consider a *maximal* ε-separated subset X_ε of S^{k-1}. This is an ε-net (if some point is not covered, we can throw it into our set and it will be a larger ε-separated set). Now we just bound the cardinality with a volume argument. Since $\{B_{\varepsilon/2}(x) : x \in X_\varepsilon\}$ are disjoint, we have that

$$\bigcup_{x\in X_\varepsilon} B_{\varepsilon/2}(x) \subset B_{1+\varepsilon/2}(0) \setminus B_{1-\varepsilon/2}(0)$$

and thus

$$|X_\varepsilon|\,\text{vol}(B_{\varepsilon/2}(0)) \leq \text{vol}(B_{1+\varepsilon/2}(0))$$

so

$$|X_\varepsilon| \leq \left(1 + \frac{2}{\varepsilon}\right)^k$$

and we have shown:

Proposition 57. *Given $\varepsilon > 0$, there exists an ε-net X_ε for S^{k-1} of size*

$$|X_\varepsilon| \leq \left(1 + \frac{2}{\varepsilon}\right)^k$$

This tells us the number of u that we need to obtain a concentration result for. By symmetry, the probability that $u \in X_\varepsilon$ satisfies some concentration result is the same for all u, and in fact, we have the following result:

Proposition 58. *Given an ε-net X_ε for S^{k-1}, and fixing some $\|u\|_2 = 1$, we have that*

$$\Pr\left(\left\{ \|Ax\|_2 \geq \frac{C_2}{1-\varepsilon} \text{ for all } \|x\|_2 = 1 \right\} \right) \leq |X_\varepsilon| \Pr(\{\|Au\|_2 \geq C_2\})$$

and

$$\Pr\left(\left\{ \|Ax\|_2 \leq C_1 - \frac{C_2\,\varepsilon}{1-\varepsilon} \text{ for all } \|x\|_2 = 1 \right\} \right) \leq |X_\varepsilon| \Pr(\{\|Au\|_2 \geq C_2\} \cup \{\|Au\|_2 \leq C_1\})$$

Proof. If $\|A\|_{\text{op}} \geq \frac{\gamma}{1-\varepsilon}$, then from above arguments there is some element $u \in X_\varepsilon$ in our ε-net such that $\|Au\|_2 \geq \gamma$. Thus, given our X_ε, we have that

$$\begin{aligned} \Pr\left(\left\{ \|Ax\|_2 \geq \frac{C_2}{1-\varepsilon} \text{ for all } \|x\|_2 = 1 \right\} \right) &= \Pr\left(\bigcup_{u \in X_\varepsilon} \{\|Au\|_2 \geq C_2\} \right) \\ &\leq \sum_{u \in X_\varepsilon} \Pr\{\|Au\|_2 \geq C_2\} \\ &= |X_\varepsilon| \Pr\{\|Au\|_2 \geq C_2\} \end{aligned}$$

The inequality is referred to as the union bound. The second statement holds in the same fashion, but we note that we need both bounds to hold.

Summarizing, here we are bounding the probability that the concentration result fails by the sum of the probabilities that the concentration result fails for some $u \in X_\varepsilon$. □

This type of argument is standard for proving results with random matrices, by first showing the result for fixed points and using an ε covering.

Concentration Bounds

Now let's derive the concentration result. Let $\|u\|_2 = 1$ be fixed, and let A be an $m \times k$ random matrix with $A_{ij} \sim N(0, 1/m)$ i.i.d. Let $\xi = Au$. We note that ξ_i are i.i.d, since if we denote the rows of A by

$$A = \begin{pmatrix} - & a_1 & - \\ & \vdots & \\ - & a_m & - \end{pmatrix}$$

then $\xi_i = \langle a_i, u \rangle$, and the rows of A are i.i.d. since the entries of A are i.i.d.

Furthermore,

$$\begin{aligned} E\big[\, \|\xi\|^2 \,\big] &= E\left[\sum_{i=1}^m \xi_i^2 \right] \\ &= \sum_{i=1}^m E[\xi_i^2] \\ &= \sum_{i=1}^m E\left[\left(\sum_j A_{ij} u_j \right)^2 \right] \\ &= \sum_{i=1}^m \sum_{1 \leq j,l \leq m} E[A_{ij} A_{il}]\, u_j u_l \qquad\qquad E[A_{ij}A_{il}] = \frac{1}{m}\delta_{jl} \\ &= \sum_{i=1}^m \sum_{j=1}^m \frac{1}{m} u_j^2 \\ &= 1 \end{aligned}$$

Here we have used linearity of expectation and independence. Note that we have not use the assumption that A_{ij} are normally distributed, just the fact that $E[A_{ij}]=0$ and $\mathrm{Var}[A_{ij}^2]=\frac{1}{m}$.

We are interested in controlling $\Pr[\|\xi\|_2>\gamma]$, for $\gamma>1$ which is close to 1. Instead let's work with

$$Y=m\,\|\xi\|_2^2=\sum_{i=1}^{m}\,(\sqrt{m}\xi_i)^2$$

so that $\sqrt{m}\,\xi_i\sim N(0,1)$. We will use Laplace's method:

$$\Pr[Y>\alpha]=\Pr[e^{tY}>e^{t\alpha}]\le\inf_{t>0}e^{-t\alpha}\,E[e^{tY}]$$

where we need $t>0$ so that $e^{t(\cdot)}$ is monotone (so the first equality holds). The second inequality is Markov. Now we have

$$\begin{aligned}
E[e^{tY}] &= \prod_{i=1}^{m}E[e^{t(\sqrt{m}\xi_i)^2}]\\
&= \left[\int_{-\infty}^{\infty}e^{ty^2}e^{-y^2/2}\frac{dy}{\sqrt{2\pi}}\right]^m\\
&= \left[\int_{-\infty}^{\infty}e^{-(1-2t)\,y^2/2}\frac{dy}{\sqrt{2\pi}}\right]^m,\ 0<t<\frac{1}{2}\\
&= \left[\underbrace{\left(\int_{-\infty}^{\infty}e^{-u^2/2}\frac{du}{\sqrt{2\pi}}\right)}_{=1}\frac{1}{\sqrt{1-2t}}\right]^m,\ u=(\sqrt{1-2t})y\\
&= (1-2t)^{-m/2}
\end{aligned}$$

Then we have that $\Pr[Y>\alpha]\le\inf_{0<t<1/2}\ e^{-t\alpha}(1-2t)^{-m/2}$. Placing everything in the exponent, we are trying to minimize

$$\exp\Big\{-t\alpha-\frac{m}{2}\log(1-2t)\Big\}$$

Since $\exp(\,\cdot\,)$ is monotonic, it suffices to minimize the exponent $-t\alpha-\frac{m}{2}\log(1-2t)$. The derivative is $-\alpha+\frac{m}{1-2t}$ and second derivative is $\frac{2m}{(1-2t)^2}>0$, and hence the critical point will be the minimum. Setting the derivative to zero yields $1-2t^*=\frac{m}{\alpha}$, and $t^*=\frac{1-\frac{m}{\alpha}}{2}$. Plugging this value in we have that

$$\Pr[Y>\alpha]\le e^{-\frac{1}{2}(\alpha-m)}\left(\frac{m}{\alpha}\right)^{-m/2}$$

Note we will be making $\alpha>m$ since we are deriving concentration around the mean of Y which is

$$E[Y]=E[m\,\|\xi\|_2^2]=m$$

Parametrizing $\alpha=me^{\beta}$ to simplify computation, for $\beta>0$, we have

$$\Pr\big[\|Au\|_2^2>e^{\beta}\big]=\Pr[Y>me^{\beta}]\le\exp\Big\{-\frac{e^{\beta}-1}{2}m+\frac{\beta m}{2}\Big\}=\exp\Big\{-\frac{m}{2}\big(-\beta-1+e^{\beta}\big)\Big\}$$

Since $e^{\beta}>1+\beta+\frac{\beta^2}{2}$ (Taylor expansion), we have that

$$\Pr[\|Au\|_2^2>e^{\beta}]\le\exp\Big\{-\frac{m\beta^2}{4}\Big\}$$

Since we are interested in $e^{\beta} \sim 1$, we will further set $e^{\beta} = 1+\rho$. Then $\beta = \ln(1+\rho)$, and we have that

$$\Pr[\|Au\|_2^2 > 1+\rho] \leq \exp\Big\{-\frac{m}{4}(\ln(1+\rho))^2\Big\}$$

By a very similar calculation, we can obtain lower bounds with

$$\Pr[Y < \alpha] = \Pr[e^{-tY} > e^{-t\alpha}] \leq \inf_{t>0} e^{t\alpha} E[e^{-tY}]$$

to obtain the bound

$$\Pr[\|Au\|_2^2 < e^{-\beta}] \leq \exp\Big\{-\frac{m}{2}(\beta + e^{-\beta} - 1)\Big\}$$

Now setting $e^{-\beta} = 1-\rho$, so that $\beta = \ln\Big(\frac{1}{1-\rho}\Big)$, we have that

$$\beta + e^{-\beta} - 1 = \ln\Big(\frac{1}{1-\rho}\Big) - \rho = \sum_{k=2}^{\infty} \frac{\rho^k}{k} \geq \frac{\rho^2}{2}$$

and thus

$$\Pr[\|Au\|_2^2 < 1-\rho] \leq \exp\Big\{-\frac{m\rho^2}{4}\Big\}$$

Summarizing the computation, we have:

Proposition 59. *If A is a random $m \times k$ matrix with $A_{ij} \sim N(0, \frac{1}{m})$ i.i.d, and fixing $\|u\|_2 = 1$, we have that*

$$\begin{aligned} \Pr[\|Au\|_2^2 > 1+\rho] &\leq \exp\Big\{-\frac{m}{4}(\ln(1+\rho))^2\Big\} \\ \Pr[\|Au\|_2^2 < 1-\rho] &\leq \exp\Big\{-\frac{m\rho^2}{4}\Big\} \end{aligned}$$

Now we bring in the ε-net into the picture. Taking an ε-net X_ε of $S^{k-1} = \{|x| = 1\}$, and making use of Proposition 58, we have that

$$\Pr\Big[\Big\{\sqrt{1-\rho} - \frac{\varepsilon\sqrt{1+\rho}}{1-\varepsilon} \leq \|Ax\|_2 \leq \frac{\sqrt{1+\rho}}{1-\varepsilon}\Big\}^c\Big] \leq |X_\varepsilon|\left(e^{-\frac{m}{4}(\ln(1+\rho))^2} + e^{-\frac{m\rho^2}{4}}\right)$$

(note the complement in the previous expression)

By Proposition 57, we can find an ε-net X_ε for S^{k-1} with size $|X_\varepsilon| \leq \Big(1+\frac{2}{\varepsilon}\Big)^k$.

Given any $\delta > 0$, we will be setting ρ, ε so that $\sqrt{1-\rho} - \frac{\varepsilon\sqrt{1+\rho}}{1-\varepsilon} > \sqrt{1-\delta}$ and $\frac{\sqrt{1+\rho}}{1-\varepsilon} < \sqrt{1+\delta}$, which is possible since we can just choose $\rho < \delta$ to give some wiggle room, and choose ε sufficiently small to close the gap and obtain the desired bounds.

Then what we have gained is the following:

Proposition 60. *Let $\delta > 0$, and A be a random $m \times k$ matrix where $A_{ij} \sim N\Big(0, \frac{1}{m}\Big)$ i.i.d. Then there exist constants $c_1(\delta)$ and $c_2(\delta)$ such that*

$$\Pr\Big[\big\{(1-\delta)\|x\|_2^2 \leq \|Ax\|_2^2 \leq (1+\delta)\|x\|_2^2 \ \textit{ for all } x\big\}^c\Big] \leq \exp\{c_1(\delta)k - c_2(\delta)m\}$$

Above we have converted $\left(1+\frac{2}{\varepsilon}\right)^k = \exp\left\{k\log\left(1+\frac{2}{\varepsilon(\delta)}\right)\right\}$, so $c_1(\delta) = \log\left(1+\frac{2}{\varepsilon(\delta)}\right)$ and from above we have $c_2(\delta) = \max\left\{\frac{1}{4}(\ln(1+\rho(\delta)))^2, \frac{\rho(\delta)^2}{4}\right\} \sim \frac{\rho(\delta)^2}{4}$.

Now returning to prove Theorem 56, we have just shown that for a particular $m \times k$ submatrix Φ_T, we have the concentration result above. We want this concentration to hold for all $m \times k$ submatrices Φ_T, and there are $\binom{N}{k}$ of these. Thus the probability that RIP fails for our $m \times N$ random matrix Φ is the probability that the concentration result fails for some submatrix Φ_T, and we can bound this with the union bound (sum of the failure probabilities):

$$\Pr[\Phi \notin \mathrm{RIP}(k,\delta)] = \Pr[\delta_k(\Phi) > \delta] \leq \binom{N}{k} e^{c_1(\delta)k - c_2(\delta)m} = \exp\left\{\ln\binom{N}{k} + c_1(\delta)k - c_2(\delta)m\right\}$$

Now as in the previous lecture, we can bound $\binom{N}{k} \leq \left(\frac{Ne}{k}\right)^k$ and so we need

$$k\ln\left(\frac{N}{k}\right) + c_1(\delta)\,k - c_2(\delta)\,m \leq -c_3(\delta)\,m$$

for some $c_3(\delta)$ (want the exponent to be negative). This means that

$$k\left[c_1(\delta) + \ln\left(\frac{Ne}{k}\right)\right] \leq [c_2(\delta) - c_3(\delta)]\,m$$

Choose $c_3(\delta) = \frac{1}{2}c_2(\delta) \sim \frac{\rho(\delta)^2}{8}$, for instance, and we have that a sufficient condition for $\Phi \in \mathrm{RIP}(k,\delta)$ with probability at least $1 - e^{-c_3(\delta)m}$ is

$$\frac{m}{k} \geq \frac{2}{c_2(\delta)}\ln\left(e^{1+c_1(\delta)}\frac{N}{k}\right) = c_1'(\delta)\ln\left(c_2'(\delta)\frac{N}{k}\right)$$

as desired.

What is important is that we can achieve RIP for any $\delta > 0$ with an appropriate choice of ρ, the parameter in the concentration result, and then a sufficiently small ε, the size of the ε-net in the proof. The probability is exponentially decaying in m, so having fixed ρ, we should not have to make m too large to make this probability overwhelmingly small.

Week 12 (4/26/2010)

Compressible Signals and Noise

Today we discuss near recovery of **compressible** (i.e. not necessarily sparse, but well-approximable by sparse vectors) vectors from noisy (arbitrarily small perturbations) measurements. We won't be looking at a stochastic model for noise here.

Let Φ be our $m \times N$ measurement matrix, and $x \in \mathbb{R}^N$, to be either sparse or compressible (definition to come in a moment). We then have a noisy measurement

$$y = \Phi x + e$$

with $\|e\|_2 \leq \varepsilon$ for some known quantity ε which is small. The goal is to recover a good approximation to x.

We already know from previously discussed results that if we have a s-sparse vector $x \in \Sigma_s^N$, and $\delta_{2s}(\Phi) < \frac{1}{3}$, then if we just solve $x^* = \text{argmin} \|z\|_1$ subject to $\Phi z = y$, then $x = x^*$, i.e. we have recovered the sparse vector exactly. This is the no noise case, with $\varepsilon = 0$.

Now in the presence of noise, we'll set up a similar problem:

$$\min \|z\|_1 \text{ s.t. } \|\Phi z - y\|_2 < \varepsilon$$

We remark quickly that we cannot use $\Phi z = y$ since there may not be a solution, and since x itself satisfies this constraint, it is the logical choice for the constraint. This is a convex minimization problem, convex constraints and convex objective function, and there are numerical solvers that can handle this type of problem.

Let us also define

$$\sigma_s(x)_X := \inf_{u \in \Sigma_s^N} \|x - u\|_X$$

where X is a normed space (we will be using l^1, l^2 on $\mathbb{R}^N$ here). This represents the best approximation error when representing x with an s-sparse signal u with respect to the norm in X.

Then we have the result:

Theorem 61. (Candés, Romberg, Tao; Candés 2008) *Assume $\delta_{2s}(\Phi) < \sqrt{2} - 1$, $\varepsilon > 0$ given, and $y = \Phi x + e$ with $\|e\|_2 \leq \varepsilon$. Also let*

$$x^* := \text{argmin} \|z\|_1 \text{ s.t. } \|\Phi z - y\|_2 \leq \varepsilon$$

Then

$$\|x - x^*\|_2 \leq C_0 \varepsilon + \frac{C_1}{\sqrt{s}} \sigma_s(x)_{l_1}$$

where C_0, C_1 are absolute constants depending only on $\delta_{2s}(\Phi)$.

Remark 62. A few remarks:

- Note that this is actually an improvement from the RIP result before, as $\sqrt{2} - 1 > \frac{1}{3}$, and in fact can be further improved to around 0.46.
- If $x \in \Sigma_s^N$, then $\sigma_s(x)_{l_1} = 0$ and $\|x - x^*\|_2 \leq C_0 \varepsilon$.

  Can we expect to do better? It turns out that this result is in some sense optimal. For instance, suppose an oracle tells us that $T = \text{supp}(x)$, $|T| \leq s$. Then if we minimize

  $$\|\Phi_T z - y\|_2 \text{ s.t. } z \in \mathbb{R}^{|T|}$$

  a least squares fit, then we have solution given by the psuedoinverse:

  $$z^* = (\Phi_T)^\dagger y = (\Phi_T^* \Phi_T)^{-1} \Phi_T^* y$$

  Let us now denote $x|_T$ by $x_0 \in \mathbb{R}^{|T|}$. Then $y = \Phi_T x_0 + e$, and

  $$z^* - x_0 = (\Phi_T^* \Phi_T)^{-1} \Phi_T^* \Phi_T x_0 + (\Phi_T^* \Phi_T)^{-1} \Phi_T^* e - x_0 = (\Phi_T^* \Phi_T)^{-1} \Phi_T^* e$$

Since by RIP, we have that $(1-\delta_s)\mathrm{Id} \leq \Phi_T^*\Phi_T \leq (1+\delta_s)\mathrm{Id}$, we know that

$$\|z^* - x_0\|_2 \geq \lambda_{\min}((\Phi_T^*\Phi_T)^{-1})\,\|\Phi_T^* e\|_2 = \frac{1}{1+\delta_s}\|\Phi_T^* e\|_2$$

and so long as we choose e with $\|\Phi_T^* e\|_2 \sim \|e\|_2 = \varepsilon$, which is possible since the nonzero singular values of Φ_T^* are near 1, we have that

$$\|z^* - x_0\|_2 \geq \frac{\varepsilon}{1+\delta_s}$$

In this sense, we cannot expect a better result.

- In the noiseless case, $\varepsilon = 0$, we have that

$$\|x - x^*\|_2 \leq \frac{C_1}{\sqrt{s}}\sigma_s(x)_{l_1}$$

What can we expect in this case? Is this optimal as well?

For this, we can consider **compressible** signals. Given any $(x_n)_{n\geq 1}$ (this does not depend on dimension, and we allow dimension to be infinite), let us denote the decreasing rearrangement of (x_n) by $|x|_{(n)}$, where

$$|x|_{(1)} \geq |x|_{(2)} \geq \ldots$$

This implies $\|x\|_p = \| \, |x|_{(\cdot)} \, \|_p$. We say that (x_n) is **compressible** iff

$$|x|_{(n)} \leq \frac{C}{n^\alpha} \text{ for all } n \geq 0$$

Note that for such compressible signals, the best s-sparse approximant in l^r is obtained by simply using the largest s entries. In other words,

$$\sigma_s(x)_{l^r} = \left(\sum_{n>s} |x|_{(n)}^r \right)^{1/r}$$

We will be using this with $r = 1, 2$.

Compressibility with power α is related to the weak l^p spaces (with $\alpha = 1/p$), defined like so:

$$\text{weak-}l^p := \left\{ x : |x|_{(n)} \leq \frac{C}{n^{1/p}} \text{ for all } n \geq 0 \right\}$$

and we denote the smallest such C to be the weak-l^p norm $\|x\|_{\text{w-}l^p}$.

Note that $l^p \subset \text{weak-}l^p$, since

$$\|x\|_p^p = \sum_{j\geq 1} |x_j|^p = \sum_{j\geq 1} |x|_{(j)}^p \geq \sum_{j=1}^{n} |x|_{(j)}^p \geq n\,|x|_{(n)}^p$$

and thus $|x|_{(n)} \leq \frac{\|x\|_p}{n^{1/p}}$.

Note that $l^p \subsetneq \text{weak-}l^p$ since $x_n = \frac{1}{n^{1/p}}$ is in weak-l^p but not l^p.

Now if $x \in$ weak-l^p, then we have that

$$\begin{aligned}\sigma_s(x)_{l_1} &= \sum_{n>s} |x|_{(n)} \\ &\leq \sum_{n>s} \frac{\|x\|_{\text{w-}l^p}}{n^{1/p}} \\ &\leq \|x\|_{\text{w-}l^p} s^{-\frac{1}{p}+1}\end{aligned}$$

(recall x is finite dimensional, and this estimate is obtained by approximating with an integral) This implies that

$$\frac{1}{\sqrt{s}} \sigma_s(x)_{l_1} \leq \|x\|_{\text{w-}l^p} s^{-\frac{1}{p}+\frac{1}{2}}$$

On the other hand, $\|x - x^*\|_2 \geq \sigma_s(x)_{l^2}$, and we have that

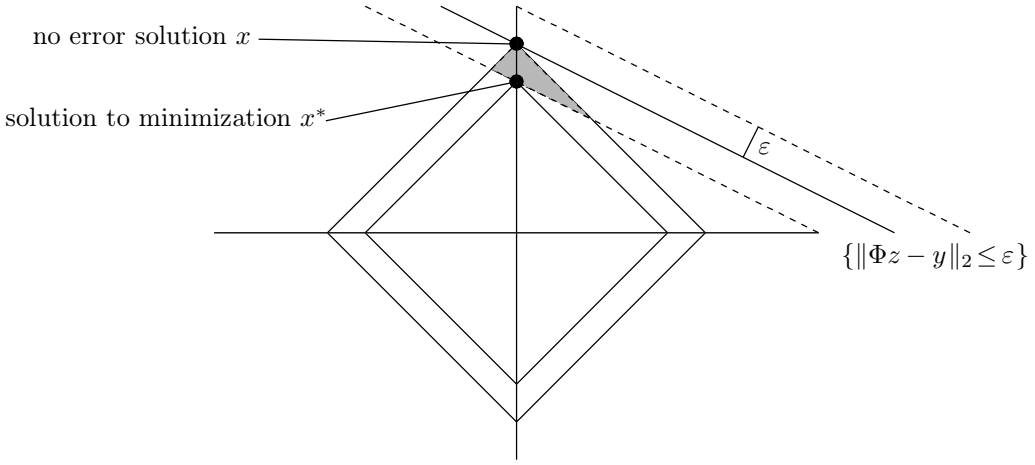
$$\sigma_s(x)_{l^2} = \left(\sum_{n>s} |x|^2_{(n)} \right)^{1/2} \leq \|x\|_{\text{w-}l^p} \left(s^{-\frac{2}{p}+1} \right)^{1/2} = \|x\|_{\text{w-}l^p} s^{-\frac{1}{p}+\frac{1}{2}}$$

which is a sharp inequality if we pick $x_n = \frac{1}{n^{1/p}}$ for instance. Thus, we cannot expect better results in the inequality

$$\|x - x^*\|_2 \leq \frac{C_1}{\sqrt{s}} \sigma_s(x)_{l^1}$$

given above.

Now we turn to the proof of Theorem 61. Here is a simple picture for ths situation where x is s-sparse.



The shaded region describes the potential locations for the solution to the minimization problem x^* (think higher dimensions). We expect the shaded region $B_{l_1, \|x\|_1} \cap \{\|\Phi z - y\|_2 \leq \varepsilon\}$ to be small (i.e. points are close to x). This isn't exactly close to the method of proof, but gives some intuition for what is happening. Contrast this for instance with the l^2 ball $B_{l_2, \|x\|_1} \cap \{\|\Phi z - y\|_2 \leq \varepsilon\}$, corresponding to replacing l^1 minimization with l^2. Even in 2 dimensions we can see that the region is significantly larger if we use the l^2 ball instead.

Proof. (of Theorem 61) This follows a 4 page paper by Candés in 2008. The proof is similar in spirit to the previous RIP result in Theorem 49.

A first observation is that

$$\|\Phi(x-x^*)\|_2 \leq \|\Phi x - y\|_2 + \|y - \Phi x^*\|_2 \leq 2\varepsilon$$

where $\|\Phi x - y\|_2 = \|e\|_2 \leq \varepsilon$ by assumption and $\|y - \Phi x^*\|_2$ by the constraints to the minimization problem (feasibility). This inequality tells us that $x - x^*$ is near $\ker(\Phi)$. Recall that in the previous proof we were processing $x - x^* \in \ker(\Phi)$ to obtain the null space property. We can then expect to do something similar here.

Let $h := x^* - x$, and we decompose as follows:

$$h = \sum_{j \geq 0} h_{T_j}\ , \quad |T_j| \leq s, \quad T_0, T_1, \ldots \text{ disjoint}$$

where

$$\begin{aligned} T_0 &:= \text{ locations of the } s \text{ largest entries of } x \\ T_1 &:= \text{ locations of the } s \text{ largest entries of } h_{(T_0)^c} \\ T_2 &:= \text{ locations of the } s \text{ largest entries of } h_{(T_0 \cup T_1)^c} \\ &\vdots \end{aligned}$$

note that the first index set T_0 is special here. On one hand, we want to capture the support of the best s-sparse approximant to x (and this relates to the $\sigma_s(x)_{l^1}$ term), and on the other we want to bound h to begin with.

Note that for $j \geq 2$, the entries in h_{T_j} are all less than entries in $h_{T_{j-1}}$ (in absolute value), and thus

$$\|h_{T_j}\|_2 \leq \sqrt{s}\, \|h_{T_j}\|_\infty \leq \frac{1}{\sqrt{s}}\, \|h_{T_{j-1}}\|_1$$

as in the proof of Theorem 49. This implies that

$$\sum_{j \geq 2} \|h_{T_j}\|_2 \leq \frac{1}{\sqrt{s}} \sum_{j \geq 2} \|h_{T_{j-1}}\|_1 \leq \frac{1}{\sqrt{s}} \|h_{(T_0)^c}\|_1$$

Furthermore, applying the triangle with the previous inequality, we have

$$(*)_1 \qquad \|h_{(T_0 \cup T_1)^c}\|_2 = \left\| \sum_{j \geq 2} h_{T_j} \right\|_2 \leq \sum_{j \geq 2} \|h_{T_j}\|_2 \leq \frac{1}{\sqrt{s}} \|h_{(T_0)^c}\|_1$$

(labeling the inequality as $(*)_1$).

Now we make use of the inequality $\|x^*\|_1 \leq \|x\|_1$ to relate $h = x^* - x$ with $\|x_{(T_0)^c}\|_1 = \sigma_s(x)_{l_1}$:

$$\begin{aligned} \|x_{T_0}\|_1 + \|x_{(T_0)^c}\|_1 = \|x\|_1 &\geq \|x^*\|_1 \\ &= \|x + h\|_1 \\ &= \|(x+h)_{T_0}\|_1 + \|(x+h)_{(T_0)^c}\|_1 \\ &\geq \|x_{T_0}\|_1 - ||h_{T_0}\|_1 + \|h_{(T_0)^c}\|_1 - \|x_{(T_0)^c}\|_1 \end{aligned}$$

where we have used reverse triangle inequality in the last line. Rearranging and using Cauchy Schwarz we have

$$(*)_2 \qquad \|h_{(T_0)^c}\|_1 \leq \|h_{T_0}\|_1 + 2\|x_{(T_0)^c}\|_1 \leq \sqrt{s}\|h_{T_0}\|_2 + 2\sigma_s(x)_{l_1}$$

and combining with $(*)_1$, we have

$$\|h_{(T_0\cup T_1)^c}\|_2 \le \|h_{T_0}\|_2 + \frac{2}{\sqrt{s}}\sigma_s(x)_{l^1} \le \|h_{(T_0\cup T_1)}\|_2 + \frac{2}{\sqrt{s}}\sigma_s(x)_{l_1}$$

Now we see that what remains is to bound $\|h_{(T_0\cup T_1)}\|_2$ in terms of ε and $\sigma_s(x)_{l_1}$ and we will be finished since then

$$(*)_3 \qquad \|h\|_2 \le \|h_{(T_0\cup T_1)^c}\|_2 + \|h_{(T_0\cup T_1)}\|_2 \le 2\|h_{(T_0\cup T_1)}\|_2 + \frac{2}{\sqrt{s}}\sigma_s(x)_{l_1}$$

Now by RIP, we have that

$$(1-\delta_{2s})\,\|h_{(T_0\cup T_1)}\|_2^2 \le \|\Phi(h_{(T_0\cup T_1)})\|_2^2$$

Since $h \in \ker(\Phi)$, we have that

$$\Phi(h_{T_0\cup T_1}) = \Phi h - \sum_{j\ge 2} \Phi h_{T_j}$$

and taking inner products with $\Phi(h_{T_0\cup T_1})$, we have

$$\|\Phi h_{T_0\cup T_1}\|_2^2 = \langle \Phi h_{T_0\cup T_1}, \Phi h\rangle - \sum_{j\ge 2} \left\langle \Phi h_{T_0\cup T_1}, \Phi h_{T_j}\right\rangle$$

We bound each term:

- Recall that as the first observation we had $\|\Phi h\|_2 = \|\Phi(x^* - x)\|_2 \le 2\varepsilon$. Then Cauchy Schwarz and RIP gives

$$|\langle \Phi h_{T_0\cup T_1}, \Phi h\rangle| \le \|\Phi h_{T_0\cup T_1}\|_2\,\|\Phi h\|_2 \le \sqrt{1+\delta_{2s}}\,\|h_{T_0\cup T_1}\|_2\,(2\varepsilon)$$

- For the second term, we recall a previous observation that if $u \in \Sigma_k^N$ and $v \in \Sigma_{k'}^N$ with $\operatorname{supp}(u)\cap \operatorname{supp}(v) = \varnothing$, then

$$|\langle \Phi u, \Phi v\rangle| \le \delta_{s+s'}(\Phi)\,\|u\|_2\,\|v\|_2$$

  To put everything in terms of δ_{2s}, we will break $\Phi h_{T_0\cup T_1}$ into $\Phi h_{T_0} + \Phi h_{T_1}$ above (otherwise $T_0 \cup T_1 \cup T_j$ is $3s$ sparse). Then,

$$\begin{aligned}
\left|\sum_{j\ge 2} \left\langle \Phi h_{T_0\cup T_1}, \Phi h_{T_j}\right\rangle\right| &\le \sum_{j\ge 2} \left|\left\langle \Phi h_{T_0}, \Phi h_{T_j}\right\rangle\right| + \sum_{j\ge 2} \left|\left\langle \Phi h_{T_1}, \Phi h_{T_j}\right\rangle\right| \\
&\le \sum_{j\ge 2} \delta_{2s}\,\|h_{T_0}\|_2\,\|h_{T_j}\|_2 + \sum_{j\ge 2} \delta_{2s}\,\|h_{T_1}\|_2\,\|h_{T_j}\|_2 \\
&\le \delta_{2s}\,(\|h_{T_0}\|_2 + \|h_{T_1}\|_2) \sum_{j\ge 2} \|h_{T_j}\|_2 \\
\text{Applying } (*)_1 \quad &\le \frac{\sqrt{2}}{\sqrt{s}}\delta_{2s}\,\|h_{T_0\cup T_1}\|_2\,\|h_{(T_0)^c}\|_1
\end{aligned}$$

  also noting a small application to Cauchy-Schwarz in

$$1\cdot\|h_{T_0}\|_2 + 1\cdot\|h_{T_1}\|_2 \le \sqrt{2}\cdot\sqrt{\|h_{T_0}\|_2^2 + \|h_{T_1}\|_2^2} = \sqrt{2}\|h_{T_0\cup T_1}\|_2$$

Now we combine both these estimates, and finally we have

$$\begin{aligned}(1-\delta_{2s})\,\|h_{(T_0\cup T_1)}\|_2^2 &\leq \|h_{T_0\cup T_1}\|_2\left(\sqrt{1+\delta_{2s}}\,(2\varepsilon)+\frac{\sqrt{2}}{\sqrt{s}}\delta_{2s}\,\|h_{(T_0)^c}\|_1\right)\\ \|h_{T_0\cup T_1}\|_2 &\leq \frac{2\varepsilon\,\sqrt{1+\delta_{2s}}}{1-\delta_{2s}}+\frac{\sqrt{2}\delta_{2s}}{\sqrt{s}(1-\delta_{2s})}\|h_{(T_0)^c}\|_1\\ &= \alpha\varepsilon+\frac{\beta}{\sqrt{s}}\,\|h_{(T_0)^c}\|_1\end{aligned}$$

where $\alpha=\frac{2\sqrt{1+\delta_{2s}}}{1-\delta_{2s}}$ and $\beta=\frac{\sqrt{2}\,\delta_{2s}}{1-\delta_{2s}}$. Applying $(*)_2$, we note that

$$\|h_{T_0\cup T_1}\|_2\leq\alpha\,\varepsilon+\frac{\beta}{\sqrt{s}}(\sqrt{s}\|h_{T_0}\|_2+2\sigma_s(x)_{l_1})\leq\alpha\varepsilon+\frac{2\beta}{\sqrt{s}}\sigma_s(x)_{l_1}+\beta\,\|h_{T_0\cup T_1}\|_2$$

and so long as $\beta<1$ (i.e. $\sqrt{2}\,\delta_{2s}<1-\delta_{2s}$, $\delta_{2s}<\frac{1}{1+\sqrt{2}}=\sqrt{2-1}$), we can rearrange to get

$$\|h_{T_0\cup T_1}\|_2\leq\frac{\alpha}{1-\beta}\varepsilon+\frac{2\beta}{1-\beta}\,\frac{1}{\sqrt{s}}\,\sigma_s(x)_{l_1}$$

and combining with $(*)_3$, we have

$$\|h\|_2\leq\frac{2\alpha}{1-\beta}\varepsilon+\left(\frac{2\beta}{1-\beta}+2\right)\frac{1}{\sqrt{s}}\sigma_s(x)_{l_1}=\frac{2\alpha}{1-\beta}\varepsilon+\frac{2(1+\beta)}{1-\beta}\,\frac{1}{\sqrt{s}}\,\sigma_s(x)_{l_1}$$

Thus

$$\|h\|_2\leq C_0\,\varepsilon+\frac{C_1}{\sqrt{s}}\sigma_s(x)_{l_1}$$

with $C_0=\frac{2\alpha}{1-\beta}$ and $C_1=\frac{2(1+\beta)}{1-\beta}$.

How good are these constants? We recall that with our earlier RIP result from random matrices in Theorem 56, given $\delta>0$, if $\Phi_{ij}\sim N(0,1/m)$ i.i.d, and $\frac{m}{2s}\geq c_1\ln\left(\frac{c_2\,N}{2s}\right)$, $(k=2s)$

$$\Pr[\delta_{2s}(\Phi)>\delta]\leq\exp(-c_3\,m)$$

where c_1,c_2,c_3 depend on δ, and we can push $\delta_{2s}\to 0$ for a suitable choice of parameters.

Now as $\delta_{2s}\to 0$, we see that above $\alpha\to 2$ and $\beta\to 0$ so that $C_0\to 4$ and $C_1\to 2$, so the constants are manageable.

□

We remark that in the result above we have mixed l_1, l_2 norms in the result. There is a corresponding result with just l^1 norms (similar proof, though not a corollary):

$$\|x-x^*\|_1\leq C_1\,\sigma_s(x)_{l_1}$$

for the case $\varepsilon=0$. There is not a corresponding bound for $\varepsilon>0$.

Also, we can ask whether it is possible to obtain a result of the form

$$\|x-x^*\|_2\leq C_2\,\sigma_s(x)_{l_2}$$

But it turns out this is not possible, and there is something special about l^1 (These results are considered in the paper by Cohen, Dahmen, Devore: *Compressed sensing and best k-term approximation*, which can be found at http://dsp.rice.edu/cs)