MLISP: Machine Learning in Signal Processing

Lecture 12

Prof. V. I. Morgenshtern

Scribe: M. Elminshawi

Illustrations: Compressed sensing course by E. Candes

Agenda:

1. Discrete wavelet transform algorithm
2. Wavelet compression
3. Traditional information transmission pipeline
4. Magnetic resonance imaging (MRI)
5. Introduction to compressed sensing
6. Basics of convex optimization

1 Discrete wavelet transform algorithm

In the previous lecture we have seen that to compute the wavelet transform of a signal we need to multiply that signal by the orthogonal matrix

$$\mathbf{W}^{\mathsf{T}} = \begin{bmatrix} \phi(t_1) & \phi(t_2) & \dots & \phi(t_N) \\ \psi(t_1) & \psi(t_2) & \dots & \psi(t_N) \\ \sqrt{2}\psi(2t_1) & \sqrt{2}\psi(2t_2) & \dots & \sqrt{2}\psi(2t_N) \\ \sqrt{2}\psi(2t_1-1) & \sqrt{2}\psi(2t_2-1) & \dots & \sqrt{2}\psi(2t_N-1) \\ 2\psi(2^2t_1) & 2\psi(2^2t_2) & \dots & 2\psi(2^2t_N) \\ 2\psi(2^2t_1-1) & 2\psi(2^2t_2-1) & \dots & 2\psi(2^2t_N-1) \\ 2\psi(2^2t_1-2) & 2\psi(2^2t_2-2) & \dots & 2\psi(2^2t_N-2) \\ \vdots & & & \\ 2^{J/2}\psi(2^Jt_1-2^J+1) & 2^{J/2}\psi(2^Jt_2-2^J+1) & \dots & 2^{J/2}\psi(2^Jt_N-2^J+1) \end{bmatrix}$$

If the matrix is $N \times N$, the complexity of this multiplication is $\mathcal{O}(N^2)$. It turns out that there is a vastly more efficient algorithm to compute the wavelet transform of a signal that only takes $\mathcal{O}(N)$ steps. This is faster then the $\mathcal{O}(N\log(N))$ of Fast Fourier Transform. We derive the fast algorithm for wavelet transform next.

Since $2^{1/2}\phi(t/2) \in V_{-1} \subset V_0$ and the shifts of the pixel function $\{\phi(t-k)\}_{k\in\mathbb{Z}}$ form an orthonormal basis for V_0, we can decompose

$$\frac{1}{\sqrt{2}}\phi\left(\frac{t}{2}\right) = \sum_{n=-\infty}^{\infty} h[n]\phi(t-n) \tag{1}$$

where

$$h[n] = \left\langle \frac{1}{\sqrt{2}}\phi\left(\frac{t}{2}\right), \phi(t-n) \right\rangle. \tag{2}$$

The sequence $h[n]_{n\in\mathbb{Z}}$ is called the *conjugate mirror filter*. Since the function $\phi(\cdot)$ is usually chosen to be well concentrated, $h[n]_{n\in\mathbb{Z}}$ has very few nonzero elements.

Exercise: Compute $h[n]$ when $\phi(\cdot)$ is the pixel function of Haar wavelets. How many nonzero elements does $h[n]$ have?

Similarly, $2^{1/2}\psi(t/2) \in W_{-1} \subset V_0$ and, therefore we can write

$$\frac{1}{\sqrt{2}}\psi\left(\frac{t}{2}\right) = \sum_{n=-\infty}^{\infty} g[n]\phi(t-n) \tag{3}$$

where

$$g[n] = \left\langle \frac{1}{\sqrt{2}}\psi\left(\frac{t}{2}\right), \phi(t-n) \right\rangle. \tag{4}$$

The sequence $g[n]_{n\in\mathbb{Z}}$ is called the *mirror filter*. Again, since the function $\psi(\cdot)$ is usually chosen to be well concentrated, $g[n]_{n\in\mathbb{Z}}$ has very few nonzero elements.

Exercise: Compute $g[n]$ for Haar wavelets. How many nonzero elements does $g[n]$ have?

By performing the change of variable $t \leftarrow 2(2^j t - k)$, we get the following refinement relationship:

$$\frac{1}{\sqrt{2}}\phi(2^j t - k) = \sum_{n=-\infty}^{\infty} h[n]\phi(2^{j+1}t - (n+2k)) \tag{5}$$

$$\frac{1}{\sqrt{2}}\psi(2^j t - k) = \sum_{n=-\infty}^{\infty} g[n]\phi(2^{j+1}t - (n+2k)). \tag{6}$$

By a second change of variable $n \leftarrow n - 2k$ and a multiplication by $2^{j+1/2}$, we finally obtain

$$\phi_{j,k} = \sum_{n=-\infty}^{\infty} h[n-2k]\phi_{j+1,n} \tag{7}$$

$$\psi_{j,k} = \sum_{n=-\infty}^{\infty} g[n-2k]\phi_{j+1,n} \tag{8}$$

where $\phi_{j,k} = 2^{j/2}\phi(2^j t - k)$ and $\psi_{j,k} = 2^{j/2}\psi(2^j t - k)$ for all $t \in \mathbb{R}$. Therefore, for every signal $f(\cdot)$ we have

$$\langle f, \phi_{j,k} \rangle = \sum_{n=-\infty}^{\infty} h[n-2k] \langle f, \phi_{j+1,n} \rangle \tag{9}$$

$$\langle f, \psi_{j,k} \rangle = \sum_{n=-\infty}^{\infty} g[n-2k] \langle f, \phi_{j+1,n} \rangle. \tag{10}$$

Now denote $a_j[k] = \langle f, \phi_{j+1,n} \rangle$ and $w_j[k] = \langle f, \psi_{j+1,n} \rangle$ and rewrite the two equations above as

$$a_j[k] = \sum_{n=-\infty}^{\infty} h[n-2k] a_{j+1}[n] \tag{11}$$

$$w_j[k] = \sum_{n=-\infty}^{\infty} g[n-2k] a_{j+1}[n]. \tag{12}$$

If our signal consists of $N = 2^J$ samples as follows

$$f(t) = \sum_{k=0}^{2^J-1} x[k] 2^{J/2} \phi(2^J t - k) \in V_J \tag{13}$$

then $a_J[k] = x[k]$ and $w_j[k]$ for $j < J$ are the wavelet coefficients we are trying to compute. The recursive formulas (11) and (12) allow us to compute these coefficient starting from $a_J[n]$ by incrementally decreasing j.

At each step in the recursion, we convolve the filters $h[\cdot]$ and $g[\cdot]$ with the sequence $a_j[\cdot]$ and *subsample the result by a factor of* 2. (Subsampling corresponds to the term $-2k$ above.) Suppose that the width of the filters $h[\cdot]$ and $g[\cdot]$ is no larger than C. Then, the computation of a_{J-1} and w_{J-1} requires $C \cdot 2^{J+1}$ multiplications, the computation of a_{J-2} and w_{J-2} requires $C \cdot 2^J$ multiplications, etc. At each level, the number of necessary multiplications is reduced by the factor of 2 due to subsampling. Therefore,

$$\text{the total number of multiplications} = C \sum_{j=0}^{J+1} 2^j \approx C 2^{J+2} = 4CN. \tag{14}$$

We conclude that the algorithm is linear in N.

Using similar ideas one can derive recursive formulas for the inverse wavelet transform and show that the complexity of the inverse transform is also linear in N.

2 Wavelet compression

We have seen that in the case of wavelet transform orthogonal matrix $\mathbf{W}$, the optimization problem

$$\min_{\boldsymbol{\theta}} \|\mathbf{y} - \mathbf{W}\boldsymbol{\theta}\|_2^2 + 2\lambda \|\boldsymbol{\theta}\|_1$$

with $\lambda = \sigma\sqrt{2 \log N}$ recovers a *sparse* solution $\boldsymbol{\theta}$ due to soft-thresholding of small wavelet coefficients. We have seen also that natural images are nearly sparse in the wavelet domain. This means that most of the wavelet coefficients of a photograph are small and only few are large. Consider a 1-megapixel image:

1 megapixel image

To compress this image, do:

- compute 1,000,000 wavelet coefficients
- set to zero all but the 25000 largest coefficients
- invert the wavelet transform.

After doing this, we obtain the following result:

1 megapixel image

25k term approximation

We can see that the quality of the image is almost the same as before, but the image has been compressed 40 times. This is the principle that underlines modern lossy coders.

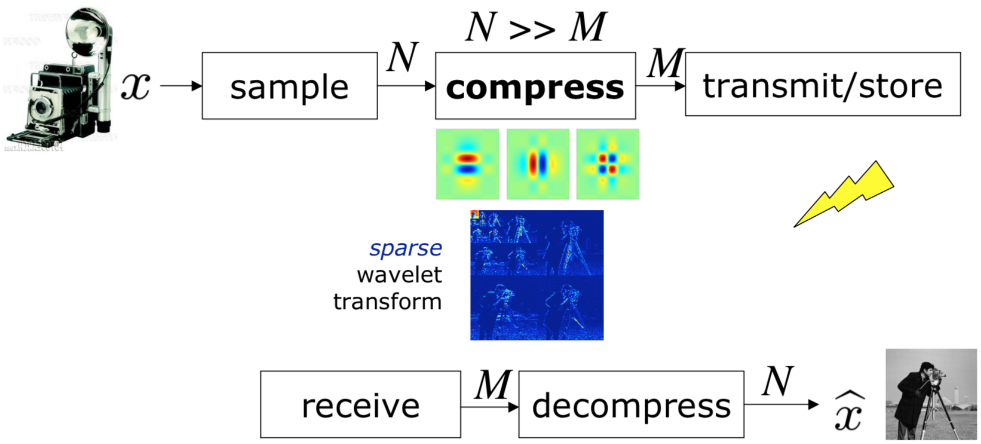
3 Traditional information transmission pipeline

If we want to transmit an image, the process would typically consist of the following steps:

- sample the image at high sampling rate, generate N samples

- compress the image using a nonlinear signal dependent algorithm, generate $M \ll N$ numbers that describe the image
- transmit the M numbers
- receive the M numbers
- decompress the image generating N pixels that approximate the image well.

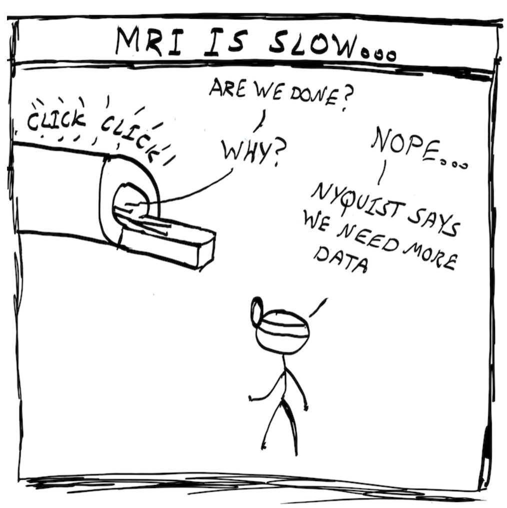
The process is illustrated in the following figure:



Do we really need to acquire $N \gg M$ pieces of information simply to throw most of that information away immediately afterwards? This is ok in photography, because the image acquisition costs is small. However, this is highly suboptimal in many other applications.

4 Magnetic resonance imaging (MRI)

MRI is the prominent example in which acquiring many samples is very expensive.



Here are the principles of MRI:

- Powerful magnetic field aligns nuclear magnetization of hydrogen atoms in water in the body.
- RF fields systematically alter the alignment of this magnetization, Hydrogen nuclei produce a rotating magnetic field detectable by the scanner.
- Make excitation strength space dependent.
- Goal is to recover proton density.

The data the MRI scanner captures can be modeled by the 2D Fourier transform of the image that we want to recover, $f[t_1, t_2], 0 \le t_1, t_2 \le N-1$:

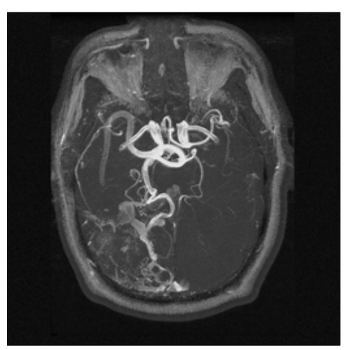
$$\hat{f}(\omega_1, \omega_2) = \sum_{t_1, t_2} f[t_1, t_2] e^{-i2\pi(\omega_1 t_1 + \omega_2 t_2)}.$$

Concretely, the MR scan records:

$$y(\omega_1, \omega_2) = \hat{f}(\omega_1, \omega_2) + \text{noise}.$$

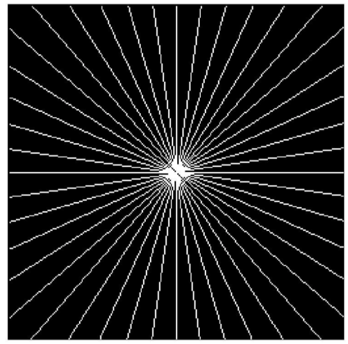
The set of frequencies $\{(\omega_1, \omega_2)\}$ where the data is recorded can be controlled. Recording one coefficient $y(\omega_1, \omega_2)$ costs certain amount of time. The more coefficients we record, the longer the scan takes. We wish to speed up MRI by sampling less. This is very important to widen applicability.

For example, consider magnetic resonance angiography of a 500×500 pixel image:



space

Instead of sampling 500×500 Fourier coefficients as Shannon-Nyquist theorem dictates, we will sample along about 20 radial lines in Fourier space and take 500 samples along each line:



frequency

Therefore, about 96% of Fourier coefficients are missing. Can we still recover the image from 4% of the samples that are normally considered necessary?

5 Introduction to compressed sensing

Compressed sensing theory tells us that this is possible, because the images are highly compressible in the wavelet domain.

Let $\boldsymbol{\theta}$ be a $N = (500)^2$-dimensional vector that contains the wavelet coefficients of our image:

$$\text{image} = \mathbf{W}\boldsymbol{\theta}.$$

We know that most of elements of $\boldsymbol{\theta}$ are *approximately* zero. Assume for simplicity that they are *exactly* zero, and there is only $s \ll N$ nonzero elements.

Our measurements are $M = 20 \cdot 500$ Fourier transform coefficients:

$$\mathbf{y} = \mathbf{F}^{\mathsf{H}} \cdot \text{image} = \mathbf{F}^{\mathsf{H}}\mathbf{W}\boldsymbol{\theta}$$

where $\mathbf{F}^{\mathsf{H}}$ is the $M \times N$ matrix whose M rows contain a subset of the N 2D Fourier transform vectors, corresponding to the M measurements.

We have $20 \cdot 500$ linear equations for $500 \cdot 500$ unknowns. Therefor, we have 25 times less equations than unknowns. In general, there are infinitely many solutions to this system of equations: if $\boldsymbol{\theta}_0$ is one solution, i.e.

$$\mathbf{y} = \mathbf{F}^{\mathsf{H}}\mathbf{W}\boldsymbol{\theta}_0$$

and $\mathbf{z}$ is any vector from the null space of $\mathbf{F}^{\mathsf{H}}\mathbf{W}$, i.e.

$$\mathbf{F}^{\mathsf{H}}\mathbf{W}\mathbf{z} = 0.$$

Then, $\boldsymbol{\theta}_0 + \mathbf{z}$ is another solution:

$$\mathbf{F}^{\mathsf{H}}\mathbf{W}(\boldsymbol{\theta}_0 + \mathbf{z}) = \mathbf{F}^{\mathsf{H}}\mathbf{W}\boldsymbol{\theta}_0 = \mathbf{y}.$$

However, we are searching for a very special solution, the one that is sparse. The following problem has a unique solution:

$$\text{Find the sparsest} \quad \boldsymbol{\theta} \quad \text{such that} \quad \mathbf{F}^{\mathsf{H}}\mathbf{W}\boldsymbol{\theta} = \mathbf{y}. \tag{NON-CVX}$$

How can we solve this? Recall that the $\ell 1$-norm promotes sparsity. Hence, we could attempt to instead solve:

$$\begin{aligned} \min_{\boldsymbol{\theta}} \quad & \|\boldsymbol{\theta}\|_1 \\ \text{subject to} \quad & \mathbf{F}^{\mathsf{H}}\mathbf{W}\boldsymbol{\theta} = \mathbf{y}. \end{aligned} \tag{CVX}$$

As we will see, if $M > \mathit{const} \cdot s \log N$ (s is the number of nonzero wavelet coefficients in $\mathbf{x}$), the solution of (CVX) is exactly equal to the solution of (NON-CVX). This is the magic of compressed sensing. We will prove a basic result of this form in 2 weeks. To do this we first need to study basics of convex optimisation.