Lecture 16

Image Reconstruction as Optimization: ℓ_1 Formulations and Compressed Sensing

16.1 Lecture Objectives

- Review image reconstruction formulations including ℓ_1-based regularization terms
- Connect this material to previous lectures describing convex optimization formulation and algorithms
- Recognize the existence of Compressed Sensing theory, which has fueled empirical research in ℓ_1-based medical image reconstruction techniques over the past 12 years

16.2 Refresher on ℓ_p Norms

As we have seen before in this course, norms are essentially metrics that describe the size of a vector. Note that when we say 'vector', this may actually refer to a 2D image, or a 3D image, etc (we can always reshape an image into a column vector in order to do algebraic operations on it). Throughout this course, we focus on a type of norms called ℓ_p norms. For a given value of $p \geq 1$, the ℓ_p norm of a vector is defined as:

$$\|\mathbf{x}\|_p = \left[\sum_n |x_n|^p\right]^{1/p} \tag{16.1}$$

In the previous lecture, as well as homework sets 2 and 3, we have focused on the ℓ_2 (aka Euclidean) norm:

$$\|\mathbf{x}\|_2 = \sqrt{\sum_n |x_n|^2} \tag{16.2}$$

This norm is often adequate for the data term of image reconstruction formulations, since the measurement noise can often be approximated as a Gaussian random variable (remember the maximum-likelihood derivation in homework 1).

However, in the regularization term, the ℓ_2 norm usually leads to excessive smoothing: if there are sharp features in the image and we use ℓ_2-based smoothness regularization, we will generally smooth over such features. In other words, the problem with the ℓ_2 norm is that it is 'too harsh' on large features: because of its quadratic penalty nature, it favors spreading the energy across multiple voxels rather than having a few voxels with high energy.

An exception to this limitation of the ℓ_2 norm is the case when we have a-priori knowledge of the location of the features (eg: where are the edges?). In this case, we can use weighting in the regularization term to avoid penalizing image features in those locations where we expect to encounter them. But in general, if we do not know the location of the image 'features', ℓ_2-based regularization will lead to smoothing.

Importantly, it is not always possible to know a priori where the features will be within our image. For such cases, the fundamental question that we are faced with is, *is there a way to simultaneously estimate the location of the features and reconstruct the image from the same data?*

This is where the ℓ_1 norm comes to the rescue. For a vector $\mathbf{x}$, we can define the ℓ_1 norm as follows:

$$\|\mathbf{x}\|_1 = \sum_n |x_n| \tag{16.3}$$

16.3 ℓ_1-Regularized Formulations

In this lecture, we will focus on the following formulation:

$$\min \|\mathbf{A}\mathbf{x} - \mathbf{d}\|_2^2 + \lambda \|\mathbf{C}\mathbf{x}\|_1 \tag{16.4}$$

where the data term is similar to the previous lecture. In contrast, the regularization term uses the ℓ_1 norm instead of the ℓ_2 norm - a somewhat subtle, yet profound, modification.

From a fundamental optimization perspective, this ℓ_1-regularized formulation has one key similarity and one key difference with respect to our previous ℓ_2-regularized formulation:

- Similarly to the ℓ_2 case, the ℓ_1 formulation leads to a convex optimization problem. In other words, regardless of our initialization, we should be able to converge to a globally optimal solution.

- In contrast to the ℓ_2 case, the ℓ_1 formulation does not lead to a closed-form or linear solution. In other words, in order to solve the ℓ_1 formulation we typically need to start iterating (eg: using steepest descent, conjugate gradients, or other descent-based methods) and see where we end up.

In other words, ℓ_1 formulations lead to *easy* optimization problems, but not as easy as ℓ_2 formulations.

16.4 ℓ_1 versus ℓ_2: 1D examples

In this section we will go over a 1D example where we try to recover a simple 1D signal from a subset of its DFT samples. This is a similar problem to the 2D image reconstruction from 2D DFT samples, but much easier to solve computationally. In this problem we will explore the different behavior of ℓ_2- versus ℓ_1 regularization. This behavior is driven by the way each of these approaches penalizes large values (see figure 16.1).

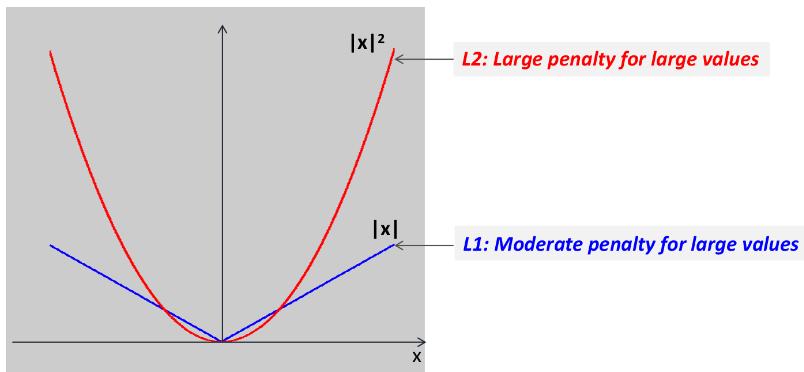

Figure 16.1: ℓ_2- versus ℓ_1 regularization: the ℓ_2 norm results in a large (quadratic) penalty for large values. In contrast, the ℓ_1 norm results in a moderate (linear) penalty for large values.

Specifically, let us consider the problem of recovering a 1D signal from half of its DFT samples. These samples are picked such that we select all the low-frequency samples (where most of the signal energy is contained), we select a subset of the rest of the (high-frequency) samples at random, and skip the rest. In other words, $mathbfd = \mathbf{A_S F x}_{\text{true}}$ where $\mathbf{F}$ is a DFT matrix, and $\mathbf{A_S}$ is an undersampling matrix that picks the selected 50% of the samples.

Next, we try to reconstruct our signal from these undersampled DFT samples using our usual regularized formulation:

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} \|\mathbf{A_S F x} - \mathbf{d}\|_2^2 + \lambda \|\mathbf{D x}\|_p^p \tag{16.5}$$

where $\mathbf{D}$ calculates finite differences, ie:

$$\mathbf{Dx} = \begin{bmatrix} x_2 - x_1 \\ x_3 - x_2 \\ \vdots \\ x_N - x_{N-1} \end{bmatrix} \tag{16.6}$$

Based on the formulation given in equation 16.5 above, we will implement it twice, using $p = 2$ (ℓ_2 regularization, which is solved directly as we have covered in this course) and $p = 1$ (ℓ_1 regularization, which is solved using a numerical descent algorithm in Matlab), respectively. Representative results from this exercise are shown in figure 16.2. As can be seen from these results, accurate recovery of this type of signal with ℓ_2 regularization is not possible. Indeed, the ℓ_2 norm places a large penalty on large features (large finite differences in this case), and therefore pushes the reconstruction to 'explain' the data using smooth features.

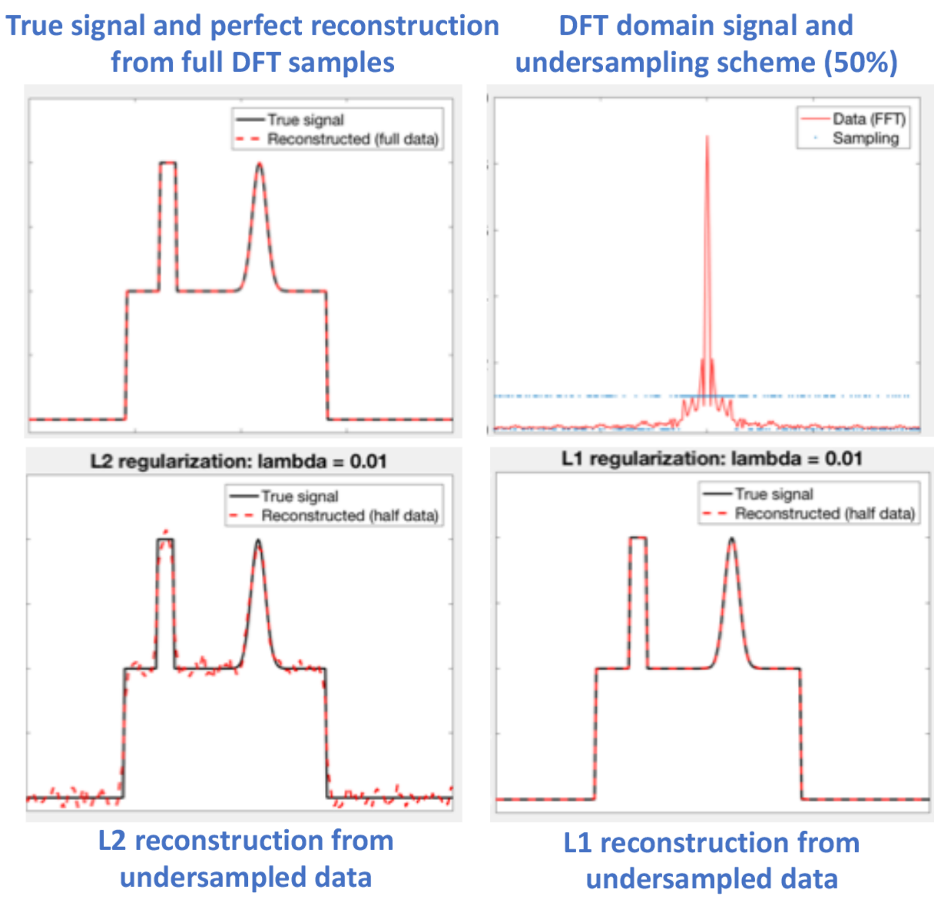

Figure 16.2: ℓ_2- versus ℓ_1-regularized reconstruction of a 1D signal from an undersampled set of DFT samples (half of the total samples were used). Note that , regardless of the choice of regularization parameter λ, accurate recovery of this type of signal with ℓ_2 regularization is not possible. This is due to the high penalty that the ℓ_2 norm places on large features: since we are penalizing finite differences, this translates into a high penalty for sharp edges. The ℓ_1 norm overcomes this challenge by enabling large features (tall edges) to 'explain' the data.

In contrast, ℓ_1 regularization is more 'tolerant' of large features and is able to accurately reconstruct this 1D signal using 50% of its DFT samples. In fact, it is a known empirical fact that ℓ_1 regularization enables perfect reconstruction (in the noiseless case) of many signals, including our 1D example as well as in a multitude of other cases.

Is this perfect reconstruction a coincidence, a fluke, or a feature? Does ℓ_1 regularization provide guarantees of accurate reconstruction for a certain class of signals?

16.5 Connection to Compressed Sensing Theory

Starting around the early 2000's, the theory of compressed sensing was developed. This theory guarantees that a certain class of signals (or images) can be recovered perfectly even from a number of measurements much less than the length of the signal, if sampled in specific ways (see below), and reconstructed using ℓ_1 regularization. These conditions are described in a bit more detail next:

- *What signals?* Compressed sensing theory deals mainly with sparse finite-length signals, ie: vectors where most of the entries are zero. These signals are important as sparsity occurs in a variety of applications. Note that we know a priori that our desired signal is sparse, ie: has few non-zero entries, but we do not know a priori which of its entries are non-zero. If our desired signal is length-N, then at most M of its entries are non-zero, with $M << N$.

- *How do we sample?* Here is the trickiest part (mathematically), so we will not describe this in detail. We sample (ie: obtain a data vector $\mathbf{d} = \mathbf{A}\mathbf{x}$) by multiplying our desired signal with a matrix of size $P \times N$. If $P \geq N$, then recovering $\mathbf{x}$ is generally easy (since we have as many equations as we have unknowns). However, in compressed sensing we are interested in the case when $P < N$ (more unknowns than equations, as in the example in the previous section, were we used only half of the DFT samples of our signal). The essential sampling condition for compressed sensing is based on a so-called 'restricted isometry property' of the matrix $\mathbf{A}$[1]. A central (and rather spectacular) result of Compressed Sensing theory is that, assuming that our sampling matrix satisfies this property, then we only need $M \log N$ samples (instead of the usual N samples) to recover our signal perfectly as described next.

- *How do we reconstruct?* Using ℓ_1 regularization by finding the solution with smallest ℓ_1 norm among those candidate solutions that match the data. This is similar (actually the equality constrained version) to our regularized formulation shown in the previous section.

- *What are our guarantees?* Perfect reconstruction of $\mathbf{x}$.

Figure 16.3 provides a graphical description of Compressed Sensing, depicting the analogies to Nyquist sampling. Note that Compressed Sensing is a rich mathematical area, with plenty of materials available for the interested reader[2].

[1]What this means is that, if $\mathbf{x_1}$ and $\mathbf{x_2}$ are two *sparse* vectors, and they are 'pretty different' from each other, then $\mathbf{Ax_1}$ and $\mathbf{Ax_2}$ will also be pretty different from each other. Note that this needs to be true for any pair of sufficiently sparse $\mathbf{x_1}$ and $\mathbf{x_2}$ in order for the matrix to have this restricted isometry property (in other words, this property is hard to check for a given matrix $\mathbf{A}$).

[2]For a relatively gentle introduction to the mathematics and potential applications of Compressed Sensing, here is a review article by one of the main contributors to the theory: https://authors.library.caltech.edu/10092/1/CANieeespm08.pdf

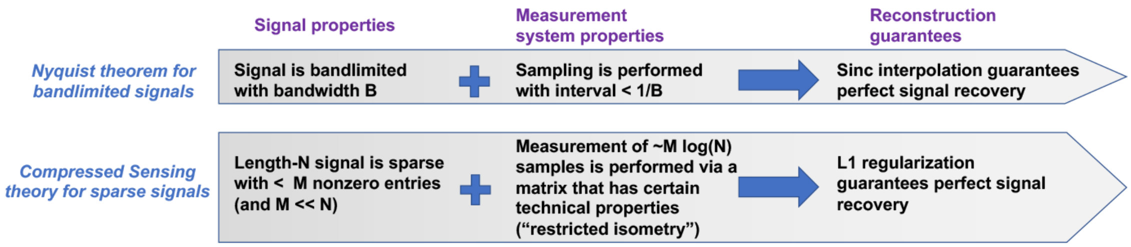

Figure 16.3: Analogy between Nyquist sampling/reconstruction of bandlimited signals, and Compressed Sensing sampling/reconstruction of sparse signals.

16.6 What does Compressed Sensing Theory have to do with Medical Imaging?

This connection has not been established: to the best of the instructor's knowledge, no one has sampled/reconstructed a medical image while strictly following the conditions of Compressed Sensing theory. This is partly because these conditions are rather difficult to satisfy for practical imaging systems, and partly because they are difficult to even check for a specific imaging system (ie: for a given 'sensing' matrix $\mathbf{A}$).

Despite this lack of direct connection to the theory, sparsity-seeking ℓ_1-based reconstructions have been empirically successful in various fields, including MRI and CT reconstruction[3], with various promising applications. Some of the most promising applications may lie in high-dimensional imaging problems, were we seek to reconstruct 3D images, or multiple images over times, etc: in these cases, there appears to be more 'structure' that can be exploited by the regularization to pick a reasonable (if not perfect) reconstructed image.

[3]Here is Michael Lustig's paper on regularized MRI reconstruction, which sparked a great deal of interest in the medical imaging community: https://www.ncbi.nlm.nih.gov/pubmed/17969013