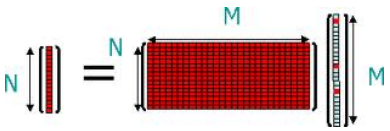
EE 381V: Large Scale Learning **Spring 2013**

Lecture 18 — March 21

Lecturer: Caramanis & Sanghavi *Scribe: Yicong Wang*

18.1 Sparsity

From this class we begin to study the problem of sparsity. In general, sparsity problem is that we want to find a data 'x' of high dimensions, e.g., 1000 dimensions, from a few measurements given that only 10/100 entries of x are non-zero. If the measurements we have are linear measurements, then the problem can be expressed as: $y = A \cdot x$, where x is the data we want to recover from the linear measurements y and matrix A is known.

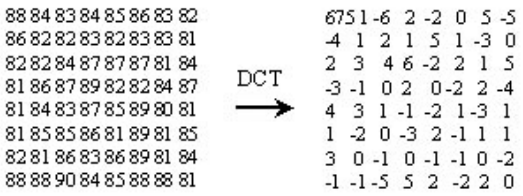


18.2 Applications of Sparsity

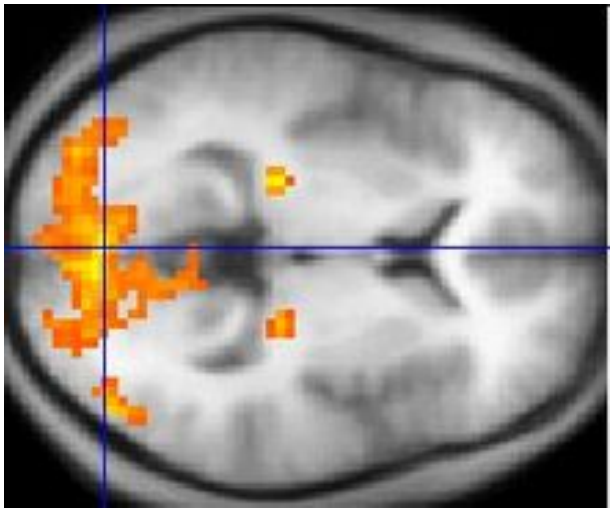
18.2.1 Image Processing

Nature and medical images are sparse in appropriate bases:

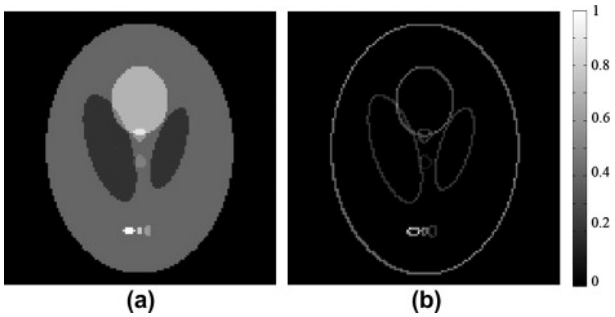
Example 1. JPEG uses Discrete Cosine Transform(DCT) and weight frequency based on perception to store image. After 2-D DCT, most energy is concentrated at the top left corner, which stands for low frequency and the entries on the rest part can be viewed as sparse data, with only a few entries having non-zero(or relatively large) values. The following picture gives an example of the 2-D DCT result of an image.



Example 2. MRI image processing. A typical method of getting the image is to select and measure a random line. Our measurement is the Fourier Transform result of the signal get from that line and the corresponding problem is to recover the image using the few measurement we have.

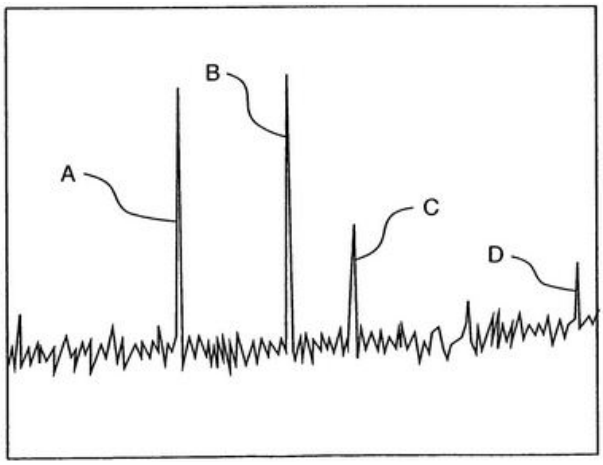


Example 3. Reconstruct image using sparsity total variation norm. For a simple image, we may focus on the contour of the image thus only the contour is useful to us and most points of the image can be viewed as zero. Then we can recover the image with a few measurements.



18.2.2 Radio Frequency / wide band monitoring

In cognitive radio, people need to monitor a wide band, e.g., 2-3 GHz band for idle spectrum bands. The band we want to monitor is very large thus we can only take a few measurements of the total band. Besides, only a few bands are occupied and each occupied band is narrow compared with the whole band so we can consider this problem as recovering the wide spectrum band with a few measurements.



18.2.3 Sparse Predictors in Statistics

In some statistics problems, we want to find the factors that determine the response of certain experiments. Let y be the response, $x_1, x_2, \ldots, x_n$ be the variables that determine the response. There are many variables while only a small number of them are significant. An example is the study in genetics, that is, decide the genome that results in certain response and the dependence of these response on the genome is sparse. This problem can be expressed as $y = x^T \cdot \beta$, where β is the sparse causative relation. Each pair of (y_i, x_i) is the test results of a person, with y be the response and x be the genome.

18.3 Number of Measurements Needed

A natural question that comes up is how many measurements we need to recover the data we want. Here we state the sparsity problem again: use the fact that $\underline{y} = X \cdot \underline{\beta}$, recover β given $\underline{y}$ and X. Whether the number of dependents is known or unknown may lead to different questions. Now we go back to the radio frequency monitoring problem discussed earlier and the question is how many measurements we need given that the signal is a simple sin signal.

From Nyquist theorem, we need 2 points per cycle to fully recover the signal. However, in our problem, the signal is a simple sin signal thus we can expect that fewer measurements are needed. A simple sin signal is: $x(t) = a \cdot \sin \omega(t - \tau)$, with a,ω and τ unknown.

The signal is sparse in frequency and our observation is in the time domain. From the functions listed below, we can see that in most cases only three measures are needed to get the three unknown variables.

$$y_1 = a \cdot \sin(\omega(t_1 - \tau)$$
$$y_2 = a \cdot \sin(\omega(t_2 - \tau)$$
$$y_3 = a \cdot \sin(\omega(t_3 - \tau)$$

Let $x = D_{IDFT} \cdot \tilde{x}$ be the IDFT transform of $\tilde{x}$. Here x is the signal in time domain and $\tilde{x}$ is the signal in frequency domain. Then clearly $\tilde{x}$ is sparse with only one entry being non-zero. We can randomly choose $x(t_1)$, $x(t_2)$ and $x(t_3)$ as y, the corresponding rows in IDFT matrix as A, and $\tilde{x}$ as the sparse unknown data, then the problem is expressed as $y = A \cdot \tilde{x}$, which is the general form of sparsity problem. With random t_1, t_2 and t_3, we can recover the sin signal with high probability.

We will study the more general form of similar problems later.

18.4 Methods for Sparse Recovery

In this section we begin to study the methods for sparse recovery.

18.4.1 Exhaustive Search

For a sparsity problem, $y = A \cdot x$, we can find x using exhaustive search of the columns in A given the sparsity of x.

For example, given x is 1-sparse, we can search the columns of A which best matches our observation y. For 2-sparse case, we need to check every pair of columns of A and for k-sparse case, search every groups of k columns.

The advantage of exhaustive method is that we can always get the correct answer, but search cost is high, especially when n and k are large.

18.4.2 Greedy Methods

Another group of methods are Greedy Methods and Orthogonal Matching Pursuit is an example of Greedy Methods.

Let $r = y$, then for each column a_i of A, we first find

$$f_i = \frac{< a_i, r >}{\|a_i\|}.$$

Pick $i^* \arg\max_i f_i$ as the first non-zero entry of x and update r as

$$r = r - \frac{< r, a_{i^*} >}{\|a_{i^*}\|} \frac{a_{i^*}}{\|a_{i^*}\|}.$$

Continue this process until we find all k non-zero entries of x. The intuition behind the method is that we greedily find the a_{i^*} which best matches y, update y and find a_i which best matches the residual of y. It is possible that the result of greedy methods is not correct.

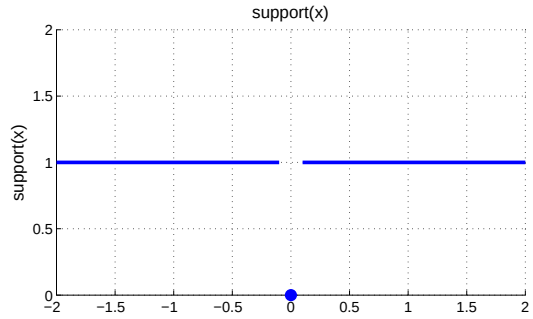
Another example of Greed Methods is that, let $\hat{x}$ be some solution to $y = A \cdot x$ and pick the two largest coordinates of $\hat{x}$ to solve the 2-sparse problem.

18.4.3 Optimization-based Methods

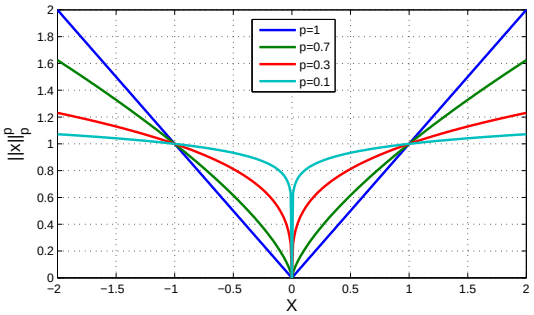
We can solve the Sparsity Problem based on optimization and the optimization problem is as follows:

$$\min f(x)$$
$$s.t. y = A \cdot x$$

$f(x) = |\text{support}(x)|$ is the most easy and straight forward objective. The solution to minimize such $f(x)$ gives us x that gives us observation y with fewest non-zero entries. For each x_i, the support function is shown as below:



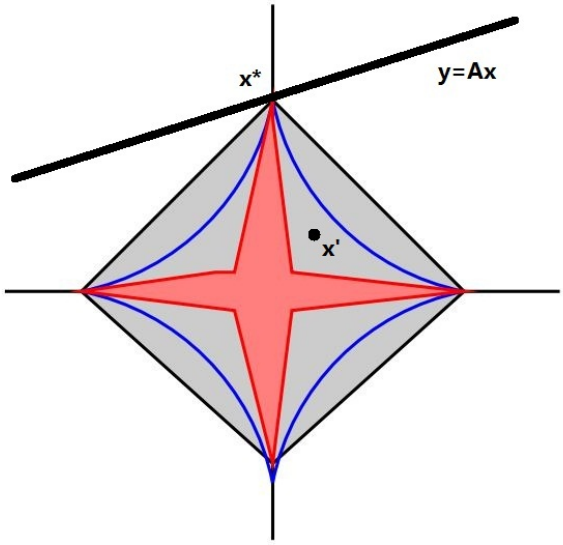
If $x_i = 0$, then x_i is zero and $\text{support}(x_i) = 0$, otherwise x_i is non-zero thus $\text{support}(x_i) = 1$. The problem with such $f(x)$ is that the function is not continuous and it is not feasible to solve the optimization problem. Instead, we use $\|x\|_p = (\sum |x_i|^p)^{1/p}, p < 1$ to approximate $|\text{support}(x)|$. As p becomes smaller, the norm becomes sharper and approximate *support* function better. Notice that $\|x\|_p, p < 1$ is not a "norm" and the function is not convex.



When $p = 1$, 1-norm is convex and sometimes we can get the sparse solution by solving the optimization problem:

$$\min \|x\|_1$$
$$s.t. y = A \cdot x$$

Now let us see when the solution of the optimization problem is also the solution to the sparse problem. Let x^* be 1-sparse data and the total number of dimensions is 2. Then the constraints $y = A \cdot x$ is a line crossing x^* as shown in the following figure.



As we can see from the figure, if $y = Ax$ is the black line in the figure, then the solution to the optimization problem is x^*, thus the global optimal solution can be achieved using $f(x) = \|x\|_1$. If $y = Ax$ passes x', then $\|x'\|_1 < \|x^*\|_1$ and the solution of the optimization problem is not the solution. However, as we makes p smaller, e.g., $\|x\|_p$ becomes the red function shown in figure, then x' is excluded. In general, error probability becomes smaller as p becomes smaller. This example intuitively tells us why small p-norm is better.

When $p > 1$, the optimization problem becomes a convex problem and we can get the sparse solution x^* only when the hyperplane of $y = A \cdot x$ is tangent to the norm ball at x^* which is generally not the case.

We will study more about optimization methods and other methods on Sparse problem in later classes.