EECS 16A Designing Information Devices and Systems I
Spring 2018 Homework 13

This homework is due April 30, 2018, at 23:59.
Self-grades are due May 3, 2018, at 23:59.

Submission Format
Your homework submission should consist of **two** files.

- `hw13.pdf`: A single PDF file that contains all of your answers (any handwritten answers should be scanned) as well as your IPython notebook saved as a PDF.

  If you do not attach a PDF of your IPython notebook, you will not receive credit for problems that involve coding. Make sure that your results and your plots are visible.

- `hw13.ipynb`: A single IPython notebook with all of your code in it.

  In order to receive credit for your IPython notebook, you must submit both a "printout" and the code itself.

Submit each file to its respective assignment on Gradescope.

1. Constrained Least-Squares Optimization

In this problem, you'll go through a process of guided discovery to solve the following optimization problem: Consider a matrix $\mathbf{A} \in \mathbb{R}^{M \times N}$, of full column rank, where $M > N$. Determine a unit vector $\vec{x}$ that minimizes $\|\mathbf{A}\vec{x}\|$, where $\|\cdot\|$ denotes the 2-norm—that is,

$$\|\mathbf{A}\vec{x}\|^2 \triangleq \langle \mathbf{A}\vec{x}, \mathbf{A}\vec{x} \rangle = (\mathbf{A}\vec{x})^T \mathbf{A}\vec{x} = \vec{x}^T \mathbf{A}^T \mathbf{A}\vec{x}.$$

This is equivalent to solving the following optimization problem:

$$\text{Determine} \quad \vec{\hat{x}} = \text{argmin}_{\vec{x}} \|\mathbf{A}\vec{x}\|^2 \quad \text{subject to the constraint} \quad \|\vec{x}\|^2 = 1.$$

This task may *seem* like solving a standard least-squares problem $\mathbf{A}\vec{x} = \vec{b}$, where $\vec{b} = \vec{0}$, but it isn't. An important distinction is that in our problem, $\vec{x} = \vec{0}$ is *not* a valid solution, because the zero vector does not have unit length. Our optimization problem is a least squares problem with a constraint—hence the term *Constrained Least-Squares Optimization*. The constraint is that the vector $\vec{x}$ must lie on the unit sphere in $\mathbb{R}^N$. You'll tackle this problem in a methodical, step-by-step fashion.

Let $(\lambda_1, \vec{v}_1), \ldots, (\lambda_N, \vec{v}_N)$ denote the eigenpairs (i.e., eigenvalue/eigenvector pairs) of $\mathbf{A}^T\mathbf{A}$. Assume that the eigenvalues are all real and indexed in an ascending fashion—that is,

$$\lambda_1 \leq \cdots \leq \lambda_N.$$

Assume, too, that each eigenvector has been normalized to have unit length—that is, $\|\vec{v}_k\| = 1$ for all $k \in \{1, \ldots, N\}$.

(a) Show that $0 < \lambda_1$.

(b) Consider two eigenpairs $(\lambda_k, \vec{v}_k)$ and $(\lambda_\ell, \vec{v}_\ell)$ corresponding to distinct eigenvalues of $\mathbf{A}^T\mathbf{A}$—that is, $\lambda_k \neq \lambda_\ell$. Prove that the corresponding eigenvectors $\vec{v}_k$ and $\vec{v}_\ell$ are orthogonal: $\vec{v}_k \perp \vec{v}_\ell$.

To help you get started, consider the two equations

$$\mathbf{A}^T\mathbf{A}\,\vec{v}_k = \lambda_k\,\vec{v}_k \tag{1}$$

and

$$\vec{v}_\ell^{\,T}\mathbf{A}^T\mathbf{A} = \lambda_\ell\,\vec{v}_\ell^{\,T}. \tag{2}$$

Premultiply Equation 1 with $\vec{v}_\ell^{\,T}$, postmultiply Equation 2 with $\vec{v}_k$, compare the two, and explain how one may then infer that $\vec{v}_k$ and $\vec{v}_\ell$ are orthogonal.

(c) Since the N eigenvectors of $\mathbf{A}^T\mathbf{A}$ are mutually orthogonal—and each has unit length—they form an orthonormal basis in $\mathbb{R}^N$. This means that we can express an arbitrary vector $\vec{x} \in \mathbb{R}^N$ as a linear combination of the eigenvectors $\vec{v}_1, \ldots, \vec{v}_N$, as follows:

$$\vec{x} = \sum_{n=1}^{N} \alpha_n\,\vec{v}_n.$$

i. Determine the n^{th} coefficient α_n in terms of $\vec{x}$ and one or more of the eigenvectors $\vec{v}_1, \ldots, \vec{v}_N$.

ii. Suppose $\vec{x}$ is a unit-length vector (i.e., a unit vector) in $\mathbb{R}^N$. Show that

$$\sum_{n=1}^{N} \alpha_n^2 = 1.$$

(d) Now you're well-positioned to tackle the grand challenge of this problem—determine the unit vector $\vec{x}$ that minimizes $\|\mathbf{A}\vec{x}\|$.

Note that the task is the same as finding a unit vector $\vec{x}$ that minimizes $\|\mathbf{A}\vec{x}\|^2$.

Express $\|\mathbf{A}\vec{x}\|^2$ in terms of $\{\alpha_1, \alpha_2 \ldots \alpha_N\}$, $\{\lambda_1, \lambda_2 \ldots \lambda_N\}$, and $\{\vec{v}_1, \vec{v}_2 \ldots \vec{v}_N\}$, and find an expression for $\vec{x}$ such that $\|\mathbf{A}\vec{x}\|^2$ is minimized. You may *not* use any tool from calculus to solve this problem—so avoid differentiation of any flavor.

For the optimal vector $\vec{x}$ that you determine—that is, the vector

$$\vec{x} = \text{argmin}_{\vec{x}}\,\|\mathbf{A}\vec{x}\|^2 \quad \text{subject to the constraint} \quad \|\vec{x}\|^2 = 1,$$

determine a simple, closed-form expression for the minimum value

$$\min_{\|\vec{x}\|=1} \|\mathbf{A}\vec{x}\| = \left\|\mathbf{A}\vec{x}\right\|.$$

2. Labeling Patients Using Gene Expression Data

Least squares techniques are useful for many different kinds of prediction problems. The core ideas we learned in class have been extensively further developed. These ideas are commonly used in machine learning for finance, healthcare, advertising, image processing, and many other fields. Here, we'll explore how least squares can be used for classification of data in a medical context.

Gene expression data of patients, along with other factors such as height, weight, age, family history, is often used to understand the likelihood that a patient might develop certain common diseases such as diabetes. Gene expression profiles can be read using DNA microarray technology, which uses tissue samples from a

patient. This data, along with the patient specific characteristics above, can be combined into a vector to get a set of features that describe each patient.

Many scientific studies look at models in mice to understand how gene expression relates to diabetes. Previous studies have shown that the expression of the tomosin2 and ts1 genes are correlated to the onset of diabetes in mice. How can we predict whether or not a mouse will develop diabetes based on data about this expression as well as other factors of the mouse? We will use some (fake) data to explore this.

We are given information about the age and weight of the mouse and additionally have access to data about whether the genes tomosin2, ts1 and chn1 (a third gene) were expressed or not. The gene expression data is captured using vectors that are $+1$ if the gene is expressed and -1 if the gene is not expressed. Similarly, whether or not a mouse has diabetes is also captured using a $+1, -1$ vector, where $+1$ indicates that the mouse has diabetes. Using this data we would like to develop a linear model that predicts whether or not a mouse will have diabetes.

$$\alpha_1(\text{age}) + \alpha_2(\text{weight}) + \alpha_3(\text{tomosin2}) + \alpha_4(\text{ts1}) + \alpha_5(\text{chn1})$$

We would like the above expression to be positive if the mouse has diabetes and negative if the mouse does not have diabetes.

(a) In problems such as this, it is common to use some *training* data to generate a model. Turns out, a good heuristic for this can be developed using a least squares technique. Set up a linear model for the problem in a format we have used for least squares problems $\mathbf{A}\vec{x} = \vec{b}$. Here, $\vec{b}$ will be a vector with $+1, -1$ entries. The α_i's are your unknowns.

(b) Using the (fake) *training* data `diabetes_train.npy`, generate the linear model using the least squares technique, i.e. find the unknown model parameters for the given data set. Include the unknown parameter values in the writeup of your homework. Use the provided IPython notebook.

(c) Now it is time to use the model you have developed to make some predictions! It is interesting to note here that we are not looking for a real number to model whether each mouse has diabetes or not; we are looking for a binary label. Therefore, we will use the *sign* of the expression above to assign a ± 1 value to each mouse. Predict whether each mouse with the characteristics in the *test* data set `diabetes_test.npy` will get diabetes. There are four mice in the test data set. Include the ± 1 vector that indicates whether or not they have diabetes in your writeup.

3. How Much Is Too Much?

When discussing circuits in this course, we only talked about resistor *I-V* curves. There are many other two terminal devices that can be found in nature that do not have linear *I-V* relations. Instead, I is some general function of V, that is $I = f(V)$. Often times, the function describing the *I-V* relationship is not known beforehand. Furthermore, noise is present in every measurement. The function f is assumed to be a polynomial, and the parameters of f (the coefficients for every power of V) are computed using least squares.

Throughout this problem, we are provided with $\vec{x}$, a set of voltage measurements, and $\vec{y}$, a set of current measurements.

(a) Let's first try to model a resistor *I-V* curve. Run the code in the attached IPython notebook. Comment on the fit with different degree polynomials.

(b) In the attached IPython Notebook, fill out the code for the cost function. The cost function returns the mean squared error, or $\|\vec{I} - \mathbf{A}\vec{f}\|$, where $\mathbf{A}$ is the appropriately sized matrix containing powers of V. Then plot the cost of various degree polynomials fitting to the measured I-V points for a resistor. Comment on the shape of the Cost vs. Degree graph.

(c) One issue with testing on the same data that we perform least squares on is that we end up fitting to noise. For this reason, the data set is generally divided into two sets of data. One set is the training set, which is used to train the model. In this case, we will use least squares on the training set to calculate the parameters to our polynomial approximation. The other data set is the test set, where we test our model. This is the data set we will use to compute the cost for a given degree polynomial. In the attached IPython notebook, fill in the code for the `improvedCost` function. Then plot the Cost vs. Degree graph as before and comment on the results.

(d) So far, we've only used least squares to fit to a resistor. Let's repeat this procedure for a non-linear device. In the attached IPython notebook, plot the Cost vs. Degree curve for this new device. From this graph, what order polynomial should we use to approximate f?

(e) Let's repeat this procedure for one final device. Use the attached IPython notebook to plot the I-V data points for this device. From the I-V curve of the device, what type of function if f? Is it a polynomial? Plot the Cost vs. Degree graph. Explain the shape of the Cost vs. Degree graph given what you know about f.

4. OMP Practice

(a) Suppose we have a vector $\vec{x} \in \mathbb{R}^4$. We take 3 measurements of it, $b_1 = \vec{m}_1^T\vec{x} = 4$, $b_2 = \vec{m}_2^T\vec{x} = 6$, and $b_3 = \vec{m}_3^T\vec{x} = 3$, where $\vec{m}_1$, $\vec{m}_2$ and $\vec{m}_3$ are some measurement vectors. In the general case when there are 4 unknowns in $\vec{x}$ and we only have 3 measurements, it is not possible to solve for $\vec{x}$. Furthermore, there could be noise in the measurements. However in this case, we are given that $\vec{x}$ is sparse and only has 2 non-zero entries. In particular,

$$\mathbf{M}\vec{x} \approx \vec{b}$$

$$\begin{bmatrix} - & \vec{m}_1^T & - \\ - & \vec{m}_2^T & - \\ - & \vec{m}_3^T & - \end{bmatrix} \vec{x} \approx \vec{b}$$

$$\begin{bmatrix} 1 & 1 & 0 & -1 \\ 1 & 0 & -1 & 0 \\ 0 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} \approx \begin{bmatrix} 4 \\ 6 \\ 3 \end{bmatrix}$$

where exactly 2 of x_1 to x_4 are non-zero. Use Orthogonal Matching Pursuit to estimate x_1 to x_4.

(b) We know that OMP works only when the vector $\vec{x}$ is sparse, which means that it has very few non-zero entries. What if $\vec{x}$ is not sparse in the standard basis but is only sparse in a different basis? What we can do is to change to the basis where $\vec{x}$ is sparse, run OMP in that basis, and transform the result back into the standard basis. Suppose we have a $m \times n$ measurement matrix $\mathbf{M}$ and a vector of measurements $\vec{b} \in \mathbb{R}^m$ where $\mathbf{M}\vec{x} = \vec{b}$ and we want to find $\vec{x} \in \mathbb{R}^n$. The basis that $\vec{x}$ is sparse in is defined by basis vectors $\vec{a}_1 \cdots \vec{a}_n$, and we define:

$$\mathbf{A} = \begin{bmatrix} | & & | \\ \vec{a}_1 & \cdots & \vec{a}_n \\ | & & | \end{bmatrix}$$

such that $\vec{x} = \mathbf{A}\vec{x}'$ and that $\vec{x}'$ is sparse.

Suppose that we have an OMP function that will compute $\vec{x}'$ given $\mathbf{M}'$ and $\vec{b}'$ **only when** $\vec{x}'$ **is sparse**: $\vec{x}' = \text{OMP}(\mathbf{M}', \vec{b}')$. Assuming that the change of basis does not significantly affect the orthogonality of vectors, describe how you would compute $\vec{x}$ using the function OMP.

5. Sparse Imaging

Recall the imaging lab where we projected masks on an object to scan it to our computer using a single pixel measurement device, that is, a photoresistor. In that lab, we were scanning a 30×40 image having 1200 pixels. In order to recover the image, we took exactly 1200 measurements because we wanted our 'measurement matrix' to be invertible.

However, we saw that an iterative algorithm that does "matching and peeling" can enable reconstruction of a sparse vector while reducing the number of samples that need to be taken from it. In the case of imaging, the idea of sparsity corresponds to most parts of the image being black with only a small number of light pixels. In these cases, we can reduce the overal number of samples necessary. This would reduce the time required for scanning the image. (This is a real-world concern for things like MRI where people have to stay still while being imaged.)

In this problem, we have a 2D image I of size 91×120 pixels for a total of 10920 pixels. The image is made up of mostly black pixels except for 476 of them that are white.

Although the imaging illumination masks we used in the lab consisted of zeros and ones, in this question, we are going to have masks with real values — i.e. the light intensity is going to vary in a controlled way. Say that we have an imaging mask M_0 of size 91×120. The measurements using the solar cell using this imaging mask can be represented as follows.

First, let us vectorize our image to $\vec{i}$ which is a length 10920 column vector. Likewise, let us vectorize the mask M_0 to $\vec{m}_0$ which is a length 10920 column vector. Then the measurement using M_0 can be represented as

$$b_0 = \vec{m}_0^T \vec{i}.$$

Say we have a total of M measurements, each taken with a different illumination mask. Then, these measurements can collectively be represented as

$$\vec{b} = \mathbf{A}\vec{i},$$

where $\mathbf{A}$ is an $M \times 10920$ size matrix whose rows contain the vectorized forms of the illumination masks, that is

$$\mathbf{A} = \begin{bmatrix} \vec{m}_1^T \\ \vec{m}_2^T \\ \vdots \\ \vec{m}_M^T \end{bmatrix}.$$

To show that we can reduce the number of samples necessary to recover the sparse image I, we are going to only generate 6500 masks. The columns of $\mathbf{A}$ are going to be approximately uncorrelated with each other. The following question refers to the part of IPython notebook file accompanying this homework related to this question.

(a) In the IPython notebook, we call a function `simulate` that generates masks and the measurements. You can see the masks and the measurements in the IPython notebook file. Complete the function `OMP` that does the OMP algorithm described in lecture.

Remark: Note that this remark is not important for solving this problem; it is about how such measurements could be implemented in our lab setting. When you look at the vector `measurements`

you will see that it has zero average value. Likewise, the columns of the matrix containing the masks **A** also have zero average value. To satisfy these conditions, they need to have negative values. However, in an imaging system, we cannot project negative light. One way to get around this problem is to find the smallest value of the matrix **A** and subtract it from all entries of **A** to get the actual illumination masks. This will yield masks with positive values, hence we can project them using our real-world projector. After obtaining the readings using these masks, we can remove their average value from the readings to get measurements as if we had multiplied the image using the matrix **A**.

(b) Run the code `rec = OMP((height, width), sparsity, measurements, A)` and see the image being correctly reconstructed from a number of samples smaller than the number of pixels of your figure. What is the image?

(c) **PRACTICE:** We have supplied code that reads a PNG file containing a sparse image, takes measurements, and performs OMP to recover it. An example input file is also supplied together with the code. Generate an image of size 91×120 pixels of sparsity less than 400 and recover it using OMP with 6500 measurements.

You can answer the following parts of this question in very general terms. Try reducing the number of measurements. Does the algorithm start to fail in recovering your sparse image? Why do you think it fails? Make an image having fewer white pixels. By how much can you reduce the number of measurements that needs to be taken?

6. Mechanical Gram-Schmidt

Use Gram-Schmidt to find a matrix **U** whose columns form an orthonormal basis for the column space of **V**.

$$\mathbf{V} = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix}$$

Show that you get the same resulting vector when you project $\vec{w} = \begin{bmatrix} 1 & -1 & 0 & -1 & 0 \end{bmatrix}^T$ onto **V** and onto **U**, i.e. show that

$$\mathbf{V}(\mathbf{V}^T\mathbf{V})^{-1}\mathbf{V}^T\vec{w} = \mathbf{U}(\mathbf{U}^T\mathbf{U})^{-1}\mathbf{U}^T\vec{w}.$$

7. Homework Process and Study Group

Who else did you work with on this homework? List names and student ID's. (In case of homework party, you can also just describe the group.) How did you work on this homework?