Assignment 5: CS 754, Advanced Image Processing

Due: 19th April before 11:55 pm

Remember the honor code while submitting this (and every other) assignment. All members of the group should work on and understand all parts of the assignment. We will adopt a zero-tolerance policy against any violation.

Submission instructions: You should ideally type out all the answers in Word (with the equation editor) or using Latex. In either case, prepare a pdf file. Create a single zip or rar file containing the report, code and sample outputs and name it as follows: A5-IdNumberOfFirstStudent-IdNumberOfSecondStudent.zip. (If you are doing the assignment alone, the name of the zip file is A5-IdNumber.zip). Upload the file on moodle BEFORE 11:55 pm on 19th April, and the cutoff time is 20th April at 10 am. **There will be no late extensions for this assignments as the cutof date is the last day of classes**. Note that only one student per group should upload their work on moodle. Please preserve a copy of all your work until the end of the semester. If you have difficulties, please do not hesitate to seek help from me.

1. Consider an inverse problem of the form $\boldsymbol{y} = \mathcal{H}(\boldsymbol{x}) + \boldsymbol{\eta}$ where $\boldsymbol{y}$ is the observed degraded and noisy image, $\boldsymbol{x}$ is the underlying image to be estimated, $\boldsymbol{\eta}$ is a noise vector, and $\mathcal{H}$ represents a transformation operator. In case of denoising, this operator is represented by the identity matrix. In case of compressed sensing, it is the sensing matrix, and in case of deblurring, it represents a convolution. The aim is to estimate $\boldsymbol{x}$ given $\boldsymbol{y}$ and $\mathcal{H}$ as well as the noise model. This is often framed as a Bayesian problem to maximize $p(\boldsymbol{x}|\boldsymbol{y}, \mathcal{H}) \propto p(\boldsymbol{y}|\boldsymbol{x}, \mathcal{H})p(\boldsymbol{x})$. In this relation, the first term in the product on the right hand side is the likelihood term, and the second term represents a prior probability imposed on $\boldsymbol{x}$.
   With this in mind, we refer to the paper 'User assisted separation of reflections from a single image using a sparsity prior' by Anat Levin, IEEE Transactions on Pattern Analysis and Machine Intelligence. Answer the following questions:

   - In Eqn. (7), explain what $A_{j\rightarrow}$ and b_j represent, for each of the four terms in Eqn. (6).
     Answer: There are actually four terms in equation (7) or equation (6) per filter - see also slides for statistics of natural images. If there are K filters, the total number of terms would be $4K$. Following the order in the paper, the answer is as follows (for each term, the vector v is the vectorized form of the image I_1 and has n elements):
     (a) $A_{1\rightarrow}$ is the $n \times n$ block circulant matrix corresponding to each of the different filters indexed by k, and b_1 is the zero vector with n elements.
     (b) $A_{2\rightarrow}$ is the $n \times n$ block circulant matrix corresponding to each of the different filters indexed by k, and b_2 is the vector obtained by convolving the same k^{th} filter with the observed image I followed by vectorizing the output.
     (c) The third term is similar to the second one but the summation is only over the pixels in set S_1. Hence $A_{3\rightarrow}$ is constructed from the $n \times n$ block circulant matrix corresponding to each of the different filters indexed by k followed by deleting all rows corresponding to indices not in S_1, and b_3 is the vector obtained by convolving the same k^{th} filter with the observed image I followed by vectorizing the output and deleting elements corresponding to indices not in S_1.
     (d) The fourth term is similar to the second one but the summation is only over the pixels in set S_2. Hence $A_{4\rightarrow}$ is constructed from the $n \times n$ block circulant matrix corresponding to each of the different filters indexed by k followed by deleting all rows corresponding to indices not in S_2, and b_4 is a zero vector with $|S_2|$ elements.

Marking scheme: 7 points for a reasonably accurate answer - 3 points for part (a), 1 point for part (b), and 1.5 points each for the last two parts. 3 points are to be deducted if it is not clear from the answer that there is a different term effectively for each of the K filters. The word 'block circulant' in the solutions is not necessary, and not points are to be deducted for missing out on it.

- In Eqn. (6), which terms are obtained from the prior and which terms are obtained from the likelihood? What is the prior used in the paper? What is the likelihood used in the paper?
  Answer: The first two terms are from the prior which imposes sparsity on the gradient filter output in the images I_1 and $I - I_1$. The paper uses a mixture of Laplacians prior for the output of filtering images I_1 or $I_2 = I - I_1$ with gradient filters of some K different types. The outputs of all filters and at all pixel locations are assumed to be statistically independent. The likelihood terms are the last two terms of equation (6) which utilize the user-specified sets S_1 and S_2. The same mixture of Laplacians model is used as the likelihood as well.
  Marking scheme: 4 points for correctly identifying which terms were for the prior and which terms were for the likelihood. 4 points for explaining the prior - deduct 2 points if there is no mention of independence across filter outputs and locations. 4 points for mentioning the likelihood.

- Why does the paper use a likelihood term that is different from the Gaussian likelihood ~~prior~~? [7+12+6=25 points]
  Answer: The job of the likelihood term is to ensure that the gradients of the estimated image I_1 agree with the gradients of image I at all points in set S_1, and also to to ensure that the gradients of the estimated image $I_2 = I - I_1$ agree with the gradients of image I at all points in set S_2. However, one can expect that the obtained solution may have some outliers, i.e. at a small fraction of points, this rule may be violated by a large margin. If a Gaussian likelihood is used, this small fraction of points will create severe problems due to the squared distance in the exponential term of the Gaussian. Instead, a sparsity-promoting prior such as the Laplacian or the mixture of Laplacians is heavy tailed, and hence is gentler on outliers. In particular, for a single Laplacian, the term in the exponent is an absolute value term, which grows much more slowly than a squared term. This will lead to greater robustness to outliers.
  Marking scheme: 6 points for a sensible explanation. Deduct 3 points if the word 'heavy-tailed' or its equivalent ('robust') is missing.

2. Consider compressive measurements of the form $\boldsymbol{y} = \boldsymbol{\Phi x} + \boldsymbol{\eta}$ under the usual notations with $\boldsymbol{y} \in \mathbb{R}^m, \boldsymbol{\Phi} \in \mathbb{R}^{m \times n}, m \ll n, \boldsymbol{x} \in \mathbb{R}^n, \boldsymbol{\eta} \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I}_{m \times m})$. Instead of the usual model of assuming signal sparsity in an orthonormal basis, consider that $\boldsymbol{x}$ is a random draw from a zero-mean Gaussian distribution with known covariance matrix $\boldsymbol{\Sigma_x}$ (of size $n \times n$). Derive an expression for the maximum a posteriori (MAP) estimate of $\boldsymbol{x}$ given $\boldsymbol{y}, \boldsymbol{\Phi}, \boldsymbol{\Sigma_x}$. Also, run the following simulation: Generate $\boldsymbol{\Sigma_x} = \boldsymbol{U \Lambda U}^T$ of size 128×128 where $\boldsymbol{U}$ is a random orthonormal matrix, and $\boldsymbol{\Lambda}$ is a diagonal matrix of eigenvalues of the form $ci^{-\alpha}$ where $c = 1$ is a constant, i is an index for the eigenvalues with $1 \leq i \leq n$ and α is a decay factor for the eigenvalues. Generate 10 signals from $\mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma_x})$. For $m \in \{40, 50, 64, 80, 100, 120\}$, generate compressive measurements of the form $\boldsymbol{y} = \boldsymbol{\Phi x} + \boldsymbol{\eta}$ for each signal $\boldsymbol{x}$. In each case, $\boldsymbol{\Phi}$ should be a matrix of iid Gaussian entries with mean 0 and variance $1/m$, and $\sigma = 0.01\times$ the average absolute value in $\boldsymbol{\Phi x}$. Reconstruct $\boldsymbol{x}$ using the MAP formula, and plot the average RMSE versus m for the case $\alpha = 3$ and $\alpha = 0$. Comment on the results - is there any difference in the reconstruction performance when α is varied? If so, what could be the reason for the difference? [25 points]
   Solution and marking scheme: Check slides for 'Statistics of natural images' for the MAP estimate and for Statistical Compressed Sensing. Either of the two formulae is allowed. The statement of the formula along with a brief derivation (even if it is there in the slides) carries 7 points. 12 points for the simulations for both the cases. The results with $\alpha = 3$ will be better than with $\alpha = 0$ as there is a decay of eigenvalues in the former case due to which effectively the datapoints to be reconstructed lie close to a lower-dimensional subspace. 6 points for the comments on the results. If there are no plots in the report, a total of 6 points are to be deducted.

3. Read through the proof of Theorem 3.3 from the paper 'Guaranteed Minimum-Rank Solutions of Linear Matrix Equations via Nuclear Norm Minimization' from the homework folder. This theorem refers to the

optimization problem in Eqn. 3.1 of the same paper. Answer all the questions highlighted within the proof. You may directly use linear algebra results quoted in the paper without proving them from scratch, but mention very clearly which result you used and where. [12 × 2 +1 = 25 points]

Solution:

(a) $\|X_0\|_* \geq \|X\|_*$ follows because X is the matrix which minimizes the nuclear norm.

(b) The middle assertion follows from the reverse triangle inequality. Since $X_0 R_c^t = X_0^t R_c = 0$, we have $\|X_0 + R_c\|_* = \|X_0\|_* + \|R_c\|_*$ and hence the last equality follows.

(c) In equation 3.5 of the paper, we see that $k \in I_{i+1}$. As the singular values are in descending order, we clearly see that $\sigma_k \leq \sigma_j$ for all $j \in I_i$. As $|I_i| = 3r$, we must have $\sigma_k \leq \frac{1}{3r}\sum_{j \in I_i} \sigma_j$.

(d) Keeping the previous result (equation 3.5 of the paper) in mind, we note that $\sum_{j \in I_i} \sigma_j = \|R_i\|_*$. Hence we have $\|R_{i+1}\|_F^2 = \sum_{k \in I_{i+1}} \sigma_k^2 \leq |I_{i+1}| \times \frac{1}{9r^2}\|R_i\|_*^2 = \frac{1}{3r}\|R_i\|_*^2$. The last equality follows because $|I_{i+1}| = 3r$.

(e) This follows by considering the previous result (i.e. $\|R_{i+1}\|_F^2 \leq \frac{1}{3r}\|R_i\|_*^2$), taking square-root on either side and adding up on both sides across j.

(f) This follows because $\|R_0\|_* \geq \|R_c\|_*$ from equation 3.4 of the paper.

(g) This follows from equation 2.1 of the paper which gives a relation between the nuclear norm and the Frobenius norm. Also note that the rank of X_0 is upper bounded by $2r$.

(h) The rank of R_0 is at most $2r$, and the rank of R_1 is at most $3r$, and hence the rank of $R_0 + R_1$ is at most $5r$.

(i) We have $R = R_0 + R_1 + \sum_{j\geq 2} R_j$. Hence we have $\mathcal{A}(R) = \mathcal{A}(R_0) + \mathcal{A}(R_1) + \sum_{j \geq 2}\mathcal{A}(R_j)$. Using the triangle equality on the vector ℓ_2 norm, we have $\|\mathcal{A}(R)\|_2 \geq \|\mathcal{A}(R_0 + R_1)\|_2 - \sum_{j\geq 2}\|\mathcal{A}(R_j)\|_2$.

(j) Consider the previous result and also the fact that $(1-\delta_{5r})\|R_0 + R_1\|_F \leq \|\mathcal{A}(R_0 + R_1)\|_2$ which follows from the order-$5r$ RIP of $\mathcal{A}$. Note that we consider the fact that $R_0 + R_1$ has rank at most $5r$. Also, consider the fact that $(1-\delta_{3r})\|R_j\|_F \leq \|\mathcal{A}(R_j)\|_2$ from the order-$3r$ RIP of R_j. Again note that the rank of R_j is at the most $3r$. Combining all these facts, we obtain the result in question.

(k) We have $\mathcal{A}(R) = 0$, because we are considering noiseless measurements (see eqn. 3.1 of the paper), and hence $\mathcal{A}(X) = \mathcal{A}(X_0) = b$ where X_0 is the original matrix.

(l) This one is quite simple :-)

Marking scheme: 2 points for each question, and 1 bonus point for any attempt which involves at least one correct answer.

4. Read section 1 of the paper 'Exact Matrix Completion via Convex Optimization' from the homework folder. Answer the following questions: (1) Why do the theorems on low rank matrix completion require that the singular vectors be incoherent with the canonical basis (i.e. columns of the identity matrix)? (2) How would this coherence condition change if the sampling operator were changed to the one in Eqn. 1.13 of the paper? [10 points]
Solution: For the first part, we see that the theorems for matrix completion require a lower bound on the number of measurements (i.e. observed, valid entries of the matrix) for successful recovery. This lower bound is directly proportional to the coherence between the singular vectors of the matrix and columns of the identity matrix. This requirment essentially prevents a matrix which would lie in the nullspace of the sampling operator. Recall that the sampling operator in the matrix completion problem is a row-subsampled identity matrix. One example of such an adversarial matrix is one with all zeros except a one along two locations on the diagonal. Such a matrix will have singular vectors $\boldsymbol{u_1} = (\boldsymbol{e_1} + \boldsymbol{e_2})/\sqrt{2}$ and $\boldsymbol{u_2} = (\boldsymbol{e_1} - \boldsymbol{e_2})/\sqrt{2}$, which are clearly quite coherent with the canonical basis vectors. It is impossible to reconstruct such a matrix from undersampled entries. This is mentioned in section 1.2 of the paper.
For the second part, let the sampling happen using orthonormal bases $\boldsymbol{f_1}, \boldsymbol{f_2}, ..., \boldsymbol{f_n}$ and $\boldsymbol{g_1}, \boldsymbol{g_2}, ..., \boldsymbol{g_n}$ in the form $\boldsymbol{f_i}^* \boldsymbol{X} \boldsymbol{g_j}$ with $(i,j) \in \Omega$ where Ω is the sample-set. In such a case, we require the left singular vectors of the matrix $\boldsymbol{X}$ (i.e. its column space) to be incoherent with the vectors $\{\boldsymbol{f_i}\}_{i=1}^n$, and the right singular vectors (row space) of $\boldsymbol{X}$ to be incoherent with the vectors $\{\boldsymbol{g_i}\}_{i=1}^n$.

Marking scheme: 5 points for each part. In the first part, the point about the nullspace of the sampling operator needs to be mentioned. Otherwise, you lose 2.5 points. In the second part, the part about left and right singular vectors being incoherent with f and g (correctly) is required, else you lose 2.5 points.

5. Read section 5.9 of the paper 'Low-Rank Modeling and Its Applications in Image Analysis' from the homework folder. You will find numerous image analysis or computer vision applications of low rank matrix modelling and/or RPCA, which we did not cover in class. Your task is to glance through any one of the papers cited in this section and answer the following: (1) State the title and venue of the paper; (2) Briefly explain the problem being solved in the paper; (3) Explain how low rank matrix recovery/completion or RPCA is being used to solve that problem. Write down the objective function being optimized in the paper with meaning of all symbols clearly explained. [5 + 10 = 15 points]
   Solution: One example is the paper by Shen and Wu on saliency detection ('A Unified Approach to Salient Object Detection via Low Rank Matrix Recovery'). The aim is to detect salient regions from the image - see the paper for examples. The authors perform feature extraction, followed by image segmentation based on the extracted features by a clustering procedure called mean shift. The authors actually over-segment the image so that the background also contains multiple segments even if it is visually homogenous. A matrix is assembled containing the features of the various segments. The aim is that the segments belonging to the background forms the low rank part of the matrix. The segments belonging to the foreground (sparse part) forms the salient part of the image. This is thus an RPCA problem. The cost function is the usual RPCA cost function, mentioned in equations 1 and 2 of the paper.