Assignment 3: CS 754, Advanced Image Processing

Due: 23rd March 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: A3-IdNumberOfFirstStudent-IdNumberOfSecondStudent.zip. (If you are doing the assignment alone, the name of the zip file is A3-IdNumber.zip). Upload the file on moodle BEFORE 11:55 pm on the due date. The cutoff (beyond which no submission will be accepted) is mentioned on moodle. 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. Download the book 'Statistical Learning with Sparsity: The Lasso and Generalizations' from `https://web.stanford.edu/~hastie/StatLearnSparsity_files/SLS_corrected_1.4.16.pdf`, which is the website of one of the authors. (The book can be officially downloaded from this online source). Your task is to trace through the steps of the proof of Theorem 11.1(b). This theorem essentially derives error bounds on the minimum of the following objective function: $J(\boldsymbol{\beta}) = \frac{1}{2N}\|\boldsymbol{y} - \boldsymbol{X\beta}\|^2 + \lambda_N\|\boldsymbol{\beta}\|_1$ where λ_N is a regularization parameter, $\boldsymbol{\beta} \in \mathbb{R}^p$ is the unknown sparse signal, $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{w}$ is a measurement vector with N values, $\boldsymbol{w}$ is a zero-mean i.i.d. Gaussian noise vector whose each element has standard deviation σ and $\boldsymbol{X} \in \mathbb{R}^{N\times p}$ is a sensing matrix whose every column is unit normalized. This particular estimator (i.e. minimizer of $J(\boldsymbol{x})$ for $\boldsymbol{x}$) is called the LASSO in the statistics literature. The theorem derives a statistical bound on λ also. Your task is split up in the following manner:

   (a) Define the restricted eigenvalue condition (the answer's there in the book and you are allowed to read it, but you also need to understand it).
   Answer: Refer equation 11.10 of the book.

   (b) Starting from equation 11.20 on page 309 - explain why $G(\hat{v}) \leq G(0)$.
   Answer: The true signal is β^* and the estimate $\hat{\beta}$ is the minimum of $G(\hat{\nu})$ where $\hat{\nu} = \hat{\beta} - \beta^*$. Since $\hat{\beta}$ is the minimum, we clearly must have $G(\hat{\nu}) = \frac{1}{2N}\|\boldsymbol{y} - \boldsymbol{X\hat{\beta}}\|^2 + \lambda_N\|\boldsymbol{\hat{\beta}}\|_1 \leq \frac{1}{2N}\|\boldsymbol{y} - \boldsymbol{X\beta^*}\|^2 + \lambda_N\|\boldsymbol{\beta^*}\|_1 = G(0)$.

   (c) Do the algebra to obtain equation 11.21.
   Answer: We have $y = X\beta^* + w$. Hence from $G(\hat{v}) \leq G(0)$, we have $\frac{1}{2N}\|w - X\hat{\nu}\|^2 + \lambda_N\|\beta^* + \hat{\nu}\|_1 \leq \frac{1}{2N}\|w\|^2 + \lambda_N\|\beta^*\|_1$. Opening out the brackets, $\frac{1}{2N}\|X\hat{\nu}\|^2 \leq \frac{1}{N}w^tX\hat{\nu} + \lambda_N(\|\beta^*\|_1 - \|\beta^* + \hat{\nu}\|_1)$ which is eqn. 11.21.

   (d) Do the algebra in more detail to obtain equation 11.22 (state the exact method of application of Holder's inequality - check the wiki article on it, if you want to find out what this inequality states).
   Answer: We use the reverse triangle inequality to prove that $\|\beta^* + \hat{\nu}\|_1 \geq \|\beta^*_S\|_1 - \|\hat{\nu}_S\|_1 + \|\hat{\nu}_{S_c}\|_1$.

Plugging this into Eqn. 11.21, we get $\frac{1}{2N}\|X\hat{\nu}\|^2 \leq \frac{w^t X\hat{\nu}}{N} + \lambda_N(\|\hat{\nu}_S\|_1 - \|\hat{\nu}_{S_c}\|_1) \leq \frac{\|X^T w\|_\infty \|\hat{\nu}\|_1}{N} + \lambda_N(\|\hat{\nu}_S\|_1 - \|\hat{\nu}_{S_c}\|_1)$. The last step follows Holder's inequality which states that for vectors a, b, we have $a^t b \leq \|a\|_p \|b\|_q$ where $p, q \in [1, \infty]$ and $1/p + 1/q = 1$. here we have $p = \infty, q = 1$.

(e) Derive equation 11.23.
Answer: This is straightforward, starting from Eqn. 11.22.

(f) Assuming Lemma 11.1 is true and now that you have derived equation 11.23, complete the proof for the final error bound for equation 11.14b.

Answer: Eqn. 11.23 yields $\|X\hat{\nu}\|^2/(2N) \leq \frac{3\sqrt{k}\lambda_N\|\hat{\nu}\|_2}{2}$. But assuming that Lemma 11.1 is true, the restricted eigenvalue condition states that $\|X\hat{\nu}\|^2/(N) \geq \gamma\|\hat{\nu}\|^2$. Combining this with the earlier equation yields the correct bounds.

(g) In which part of the proof does the bound $\lambda_N \geq 2\frac{\|\boldsymbol{X}^T\boldsymbol{w}\|_\infty}{N}$ show up? Explain.
Answer: In two place. The first is when going from Eqn. 11.22 to Eqn. 11.23. This essentially tells us how to choose the regularization parameter in terms of the noise vector w. The second is in proving Lemma 11.1, which proves that the solution to Lasso obeys the cone constraint, needed to apply the RE condition.

(h) Why is the cone constraint required? You may read the rest of the chapter to find the answer.
Answer: Strong convexity is required for uniqueness of the solution of the given problem. Strong convexity is not possible in the given setting because the matrix $X^T X$ is low rank since $N < p$. Hence, we consider a restricted version of strong convexity, which is restricted to vectors that lie in some constraint set $\mathcal{C}$. It turns out that solutions to the Lagrangian Lasso problem obey a cone constraint of the form $\|\hat{\nu}_{S_c}\|_1 \leq 3\|\hat{\nu}_S\|_1$ where $\hat{\nu} \triangleq \hat{\beta} - \beta^*$. In fact, the proof we just did, has an important step that proves this cone constraint. It is called a cone constraint because the condition gives you the equation of a cone. The set S is the support of β^* which is assumed to be sparse. Thus, we are restricting our requirement of strong convexity not to arbitrary vectors, but to vectors $\hat{\nu}$ that lie inside this cone.

(i) Read example 11.1 which tells you how to put a tail bound on λ_N assuming that the noise vector $\boldsymbol{w}$ is zero-mean Gaussian with standard deviation σ. Given this, state the advantages of this theorem over Theorem 3 that we did in class. You may read parts of the rest of the chapter to answer this question. What are the advantages of Theorem 3 over this particular theorem?
Answer: The advantage of this theorem over that of Theorem 3 is that the error here in the case of noise from $\mathcal{N}(0, \sigma^2)$ is upper bounded by $\mathcal{O}(\sigma\sqrt{k \log p})$, whereas for Theorem 3 it was upper bounded by $\mathcal{O}(\sigma\sqrt{N})$. Therefore this theorem effectively predicts a tighter upper bound for sparser signals, something which was missing in theorem 3. Also, the upper bound on the error does not scale with the number of measurements in this new theorem. The advantage of Theorem 3 over this one is that Theorem 3 handles the case of signals that are not exactly sparse. However this new theorem does have an extension to handle the compressible case as seen in Eqn. 11.24 of the book. Moreover, this theorem gives bounds that are minimax optimal, i.e. there is no other algorithm which can yield a substantially better error bound.

(j) Now read Theorem 1.10 till corollary 1.2 and comments on it concerning an estimator called the 'Dantzig selector', in the tutorial 'Introduction to Compressed Sensing' by Davenport, Duarte, Eldar and Kuttyniok. You can find it here: `http://www.ecs.umass.edu/~mduarte/images/IntroCS.pdf`. What is the common thread between the bounds on the 'Dantzig selector' and the LASSO?
Answer: The common thread is that bounds for both estimators scale as $\mathcal{O}(\sigma\sqrt{k \log p})$. Morover, both are proved to be minimax estimators, i.e. you cannot invent another estimator whose bounds are

substantially better than these bounds (apart from constant factors). The part on minimax estimators is not required for the grade for this sub-question.

(k) Read the abstract and introduction (section 1) of the paper 'Square-root lasso: pivotal recovery of sparse signals via conic programming' by Belloni et al, published in the journal Biometrika. See `https://www.jstor.org/stable/23076172`. This paper proposes an estimator called the square-root LASSO. What is the advantage of the square-root LASSO over the LASSO?
Answer: The optimal regularization parameter in the square-root LASSO does not depend on σ (noise standard deviation) unlike in the LASSO. This is because $\|\nabla L\|_\infty$ in case of the square-root LASSO is independent of σ unlike the case in LASSO, where L stands for the data fidelity term (i.e. $\|\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\|_2^2$ in case of LASSO, and $\|\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\|_2$ in case of square-root LASSO). In many applications, σ may not be known in advance, and hence the square-root LASSO is a more practical estimator. Of course, the optimal regularization parameter can be chosen by cross-validation. But that can be expensive. In square-root LASSO, the regularization parameter λ is chosen to be $cN^{-0.5}\Phi^{-1}(1-\alpha/2p)$ where $\Phi^{-1}(x)$ refers to the inverse CDF of a zero-mean Gaussian distribution with standard deviation 1 at value x. Here α is a value between 0 and 1, for which the proven bounds hold with probability $1-\alpha$. [2 x 8 + 6 + 4 + 4 = 30 points]

2. In this task, you will you use the well-known package L1_LS from `https://stanford.edu/~boyd/l1_ls/`. This package is often used for compressed sensing solution, but here you will use it for the purpose of tomographic reconstruction. The homework folder contains images of two slices taken from an MR volume of the brain. Create measurements by parallel beam tomographic projections at any 18 randomly angles chosen from a uniform distribution on $[0, \pi)$. Use the MATLAB function 'radon' for this purpose. Now perform tomographic reconstruction using the following method: (a) filtered back-projection using the Ram-Lak filter, as implemented in the 'iradon' function in MATLAB, (b) independent CS-based reconstruction for each slice by solving an optimization problem of the form $J(\boldsymbol{x}) = \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|^2 + \lambda\|\boldsymbol{x}\|_1$, (c) a coupled CS-based reconstruction that takes into account the similarity of the two slices using the model given in the lectures notes on tomography. For parts (b) and (c), use the aforementioned package from Stanford. For part (c), make sure you use a different random set of 18 angles for each of the two slices. The tricky part is careful creation of the forward model matrix $\boldsymbol{A}$ or a function handle representing that matrix, as well as the corresponding adjoint operator $\boldsymbol{A}^T$. Use the 2D-DCT basis for the image representation. Modify the objective function from the lecture notes for the case of three similar slices. Carefully define all terms in the equation and re-implement it. [3+7+8+7 = 25 points]
Answer: The operator A acting on θ is basically RU where U is the 2D-DCT basis, i.e. it computes the Radon of the inverse 2D-DCT of θ. The adjoint operator A^T acting on the projection vector y basically computes the 2D-DCT of the inverse Radon (with a Ram-Lak filter as stated here, though this filter is not really required) of y. The modified function for three slices is as follows:
$J(\boldsymbol{\beta_1}, \boldsymbol{\Delta\beta_{12}}, \boldsymbol{\Delta\beta_{13}}) = \|\boldsymbol{y_1} - \boldsymbol{R_1 U \beta_1}\|^2 + \|\boldsymbol{y_2} - \boldsymbol{R_2 U}(\boldsymbol{\beta_1} + \boldsymbol{\Delta\beta_{12}})\|^2 + \|\boldsymbol{y_3} - \boldsymbol{R_3 U}(\boldsymbol{\beta_1} + \boldsymbol{\Delta\beta_{13}})\|^2 +$
$\lambda_1\|\boldsymbol{\beta_1}\|_1 + \lambda_2\|\boldsymbol{\Delta\beta_{12}}\|_1 + \lambda_3\|\boldsymbol{\Delta\beta_{13}}\|_1$.
Another modified form, which is also correct, is: $J(\boldsymbol{\beta_1}, \boldsymbol{\Delta\beta_{12}}, \boldsymbol{\Delta\beta_{23}}) = \|\boldsymbol{y_1} - \boldsymbol{R_1 U \beta_1}\|^2 + \|\boldsymbol{y_2} - \boldsymbol{R_2 U}(\boldsymbol{\beta_1} + \boldsymbol{\Delta\beta_{12}})\|^2 + \|\boldsymbol{y_3} - \boldsymbol{R_3 U}(\boldsymbol{\beta_1} + \boldsymbol{\Delta\beta_{12}} + \boldsymbol{\Delta\beta_{23}})\|^2 +$
$\lambda_1\|\boldsymbol{\beta_1}\|_1 + \lambda_2\|\boldsymbol{\Delta\beta_{12}}\|_1 + \lambda_3\|\boldsymbol{\Delta\beta_{23}}\|_1$.
In general, better results are expected with the coupling effects as compared to without. This requires careful selection of regularization parameters. If the student has presented results with more than 18 angles, that is fine too.

3. Prove the following properties of the Radon transform:

(a) Shifting: $R(g(x - x_0, y - y_0))(\rho, \theta) = R(g(x, y))(\rho - x_0\cos\theta - y_0\sin\theta, \theta)$. Here ρ is the offset and θ is the angle of projection.
Answer: $R(g(x - x_0, y - y_0))(\rho, \theta) = \int\int g(x - x_0, y - y_0)\delta(\rho - x\cos\theta - y\sin\theta)dxdy$. A change of variables $x' = x - x_0, y' = y - y_0$ yields $\int\int g(x', y')\delta(\rho - (x' + x_0)\cos\theta - (y' + y_0)\sin\theta)dx'dy' = \int\int g(x', y')\delta(\rho - x_0\cos\theta - y_0\sin\theta - x'\cos\theta - y'\sin\theta)dx'dy' = R(g(x, y))(\rho - x_0\cos\theta - y_0\sin\theta, \theta)$.

(b) Rotation: Let $g'(r, \psi) = g(r, \psi - \psi_0)$. Then prove that $R(g')(\rho, \theta) = R(g)(\rho, \theta - \psi_0)$. ~~$R(g')(\rho, \theta) = R(g)(\rho, \psi_0 - \theta)$~~
Answer: We know that $R(g)(\rho, \theta) = \int\int g(x, y)\delta(\rho - x\cos\theta - y\sin\theta)dxdy = \int\int g(r\cos\psi, r\sin\psi)\delta(\rho -$

$r\cos\theta\cos\psi - r\sin\theta\sin\psi)rdrd\psi = \int\int g(r\cos\psi, r\sin\psi)\delta(\rho - r\cos(\theta-\psi))rdrd\psi$ by changing from (x,y) to (r,ψ), i.e. to polar coordinates. Suppose I rotate image g by ψ_0 to get image g'. Then I have $Rg'(\rho,\theta) = \int\int g(r\cos(\psi-\psi_0), r\sin(\psi-\psi_0))\delta(\rho - r\cos(\theta-\psi))rdrd\psi$. Now, we do a change of variables to $\psi' = \psi - \psi_0$, which yields $R_\theta g'(\rho) = \int\int g(r\cos\psi', r\sin\psi')\delta(\rho - r\cos(\theta-\psi_0-\psi'))rdrd\psi' = R(g)(\rho, \theta-\psi_0)$.
Apart from algebra, there is also an intuitive answer here. The Radon transform of an object rotated by angle α at angle θ is equal to the Radon transform of the un-rotated object at angle $\theta - \alpha$.

(c) Convolution: Given image $f(x,y)$ and kernel $k(x,y)$, show that $R_\theta(f*k) = R_\theta(f)*R_\theta(k)$, where $*$ is the convolution operation. In other words, the Radon transform of the convolution of two signals is equal to the convolution of the Radon transform of the individual signals (in the same angles). [8+5+7=20 points]
Answer: The convolution is given as $h(x,y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x_1,y_1)k(x-x_1,y-y_1)dx_1dy_1$. Now we have $R_\theta(h)(\rho) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x_1,y_1)\Big(k(x-x_1,y-y_1)\delta(\rho - x\cos\theta - y\sin\theta)dxdy\Big)dx_1dy_1$.
The double integral over dx, dy is the Radon transform of k shifted by x_1, y_1, and hence we can use the Radon shift theorem and replace it by $R_\theta(k)(\rho - x_1\cos\theta - y_1\sin\theta)$. This yields:
$R_\theta(h)(\rho) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x_1,y_1)R_\theta(k)(\rho - x_1\cos\theta - y_1\sin\theta)dx_1dy_1$.
We now introduce another Dirac delta over a dummy variable $\bar{\rho}$ as follows:
$R_\theta(h)(\rho) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x_1,y_1)R_\theta(k)(\rho-\bar{\rho}))\delta(\bar{\rho} - x_1\cos\theta - y_1\sin\theta)dx_1dy_1d\bar{\rho}$.
The inner integral over dx_1, dy_1 is basically $R_\theta(f)(\bar{\rho})$. Substituting that, we now get: $R_\theta(h)(\rho) = \int_{-\infty}^{\infty} R_\theta(f)(\bar{\rho})R_\theta(k)(\rho-\bar{\rho})d\bar{\rho}$ which is the 1D convolution of the Radon transforms of f and k, over the bin indices in $\bar{\rho}$.

marking scheme: Correct substitution of the Radon transform of the modified function should fetch 2 points in each of the three cases. The rest of the points are to be given for correct steps to be followed. Give partial credit for sensible attempts even if there is a mistake.

4. For a sensing matrix with columns having unit magnitude, with s-order restricted isometry constant δ_s and mutual coherence μ, prove that $\delta_s \leq (s-1)\mu$. (Hint: Use Gershgorin's disc theorem.) [10 points]
Answer: Look at the definition of the RIC in our lecture slides, where we have $\delta_s = \max(1-\lambda_{min}, \lambda_{max}-1)$ where $\lambda_{max} = \max_{\boldsymbol{\theta}\in\mathbb{R}^{|S|}, S\subset[n], |S|\leq s}\frac{\|\boldsymbol{A_S\theta}\|^2}{\|\boldsymbol{\theta}\|^2}$ where $[n] = \{1,2,...,n\}$ and $\boldsymbol{A_S}$ is a sub-matrix of sensing matrix $\boldsymbol{A}$ with columns whose are indices are strictly from the set S. λ_{min} is defined analogously. Here, λ_{max} and λ_{min} are by definition the largest and smallest eigenvalues of the matrix $\boldsymbol{A_S}^T\boldsymbol{A_S}$ (over different subsets denoted by S). Note that the matrix $\boldsymbol{A_S}^T\boldsymbol{A_S}$ has ones on its diagonal (as the columns of $\boldsymbol{A}$ have unit norm) and the off-diagonal elements have the form $\boldsymbol{a_i}\cdot\boldsymbol{a_j}$. Clearly, each off-diagonal element is upper-bounded by μ. By Gershgorin's disk theorem, each eigenvalue of $\boldsymbol{A_S}^T\boldsymbol{A_S}$ must lie in the range $[1-\mu(s-1), 1+\mu(s-1)]$ since the disks are centered at 1 and there are only $s-1$ off-diagonal elements in any row (or column). This shows that $\delta_s \leq \mu(s-1)$.
Marking scheme: Methods that do not use Gershgorin's theorem are allowed, and should be given full points if they are correct (but this result is generally more difficult to prove without the theorem). 3 points for correct definition of RIC, 4 points for correct usage of the theorem and 3 points for concluding the proof correctly.

5. Here is our Google search question again. You know of the applications of tomography in medicine (CT scanning) and virology/structural biology. Your job is to search for a journal paper from any other field which requires the use of tomographic reconstruction (examples: seismology, agriculture, gemology). State the title, venue and year of publication of the paper. State the mathematical problem defined in the paper. Take care to explain the meaning of all key terms clearly. State the method of optimization that the paper uses to solve the problem. [15 points]
Answer: A very nice (and unexpected) application of tomography is in non-parametric probability density estimation, as seen in the paper 'Multi-dimensional density estimation by tomography' published in the

Journal of the Royal Statistical Society (B) in 1993. Let $X = (X_1, X_2)$ be a two-dimensional random variable. Instead of estimating the density of X directly, the paper estimates the density of various quantities of 1D quantities of the form $\phi_1 X_1 + \phi_2 X_2$ where $\phi_1^2 + \phi_2^2 = 1$ and ϕ_1, ϕ_2 are chosen randomly from a unit circle. It turns out (see Theorem 1) that the density of these 1D quantities is equal to the Radon transform of the original 2D density taken in the direction (ϕ_1, ϕ_2). Thus given such 1D densities, the 2D density is computed by filtered backprojection with a Ram-Lak filter (equation 18).

Marking scheme: There are many applications of tomography in agriculture, food science, seismology, minerology. One example in agriculture is `https://plantmethods.biomedcentral.com/articles/10.1186/s13007-019-0468-y`. Many of these papers do not define the mathematical problem. In such cases, merely stating that so and so algorithm was used for reconstruction is enough.