Assignment 3 - Advanced Image Processing

Kartik Gokhale, Pulkit Agarwal

March 2022

Contents

Note - To obtain the results reported in this document, run the MATLAB files specified alongside the section headers above.

1 Problem 1

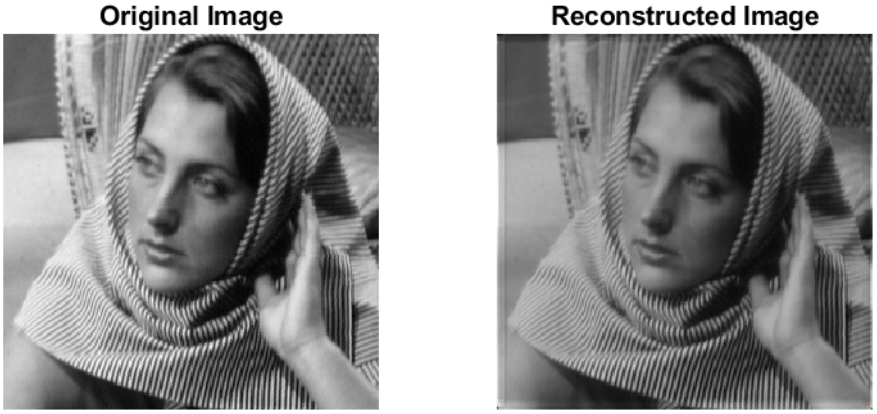
1.1 Part A - Q1a.m

The obtained Root Mean Squared Error for the reconstructed image is $\boxed{0.011744}$.

Note that, since the measurement matrix is simply the 64×64 identity matrix, so we have that A is the DCT matrix. Then all eigenvalues of $A^T A$ are 1, and so is the largest eigenvalue. But if we take $\alpha = 1$, then θ_{k+1} will no longer depend on θ_k. So I have taken $\alpha = 1.1$. Also, there is no change in the error after 10 iterations, so the ISTA algorithm has only been run for 10 iterations.

1.2 Part B - Q1b.m

The obtained Root Mean Squared Error for the reconstructed image is $\boxed{0.40953}$.



We have again used a slightly higher value than the largest eigenvalue of $A^T A$ for α. The ISTA algorithm was run for a total of 50 iterations, after which the error had almost become constant.

2 Problem 2

2.1 Part A

Take $\alpha > 0$. Consider a subset $S \subseteq \{1, 2 \dots d\}$ where S is not the empty set. Also, let

$$C_\alpha S = \{x \in \mathcal{R}^d : ||x_{S^c}||_1 \leq \alpha ||x_S||_1\}$$

where x_S denotes the subvector indexed by elements from S, and similarly for x_{S^c}. Then an $n \times d$ matrix A satisfies the restricted eigenvalue condition (α, κ) with respect to S if

$$\frac{1}{n}||Ax||^2 \geq \kappa ||x||^2 \ \forall x \in C_\alpha S \text{ and for } \kappa > 0$$

If we take the vector $X = \hat{\beta} - \beta^*$,we cannot conclude that $||X||_2^2$ is small. Hence, we cannot obtain a bound on the error. However, If we have that $\lambda_{min}(X^T X)/n$ is bounded away from zero, say in fact that we have enough curvature and our function is adequately convex, then it allows us to obtain an upper bound on the error. This is what the restricted eigenvalue condition says.

2.2 Part B

Consider equation (11.20).

$$G(\hat{\nu}) = \frac{1}{2N}||\boldsymbol{y} - \boldsymbol{X}(\beta^* + \hat{\nu})||_2^2 + \lambda_N ||(\beta^* + \hat{\nu})||_1$$

We know that β^* is a feasible solution and $\hat{\beta}$ is an optimal solution,

$$\begin{aligned} G(\hat{\nu}) &= \frac{1}{2N}||\boldsymbol{y} - \boldsymbol{X}(\beta^* + \hat{\nu})||_2^2 + \lambda_N ||(\beta^* + \hat{\nu})||_1 \\ &= \frac{1}{2N}||\boldsymbol{y} - \boldsymbol{X}\hat{\beta}||_2^2 + \lambda_N ||\hat{\beta}||_1 \\ &\leq \frac{1}{2N}||\boldsymbol{y} - \boldsymbol{X}\beta^*||_2^2 + \lambda_N ||\beta^*||_1 = G(0) \end{aligned}$$

Thus, we can write $G(\hat{\nu}) \leq G(0)$

2.3 Part C

Using the inequality obtained in previous part

$$\frac{||\boldsymbol{y} - \boldsymbol{X}(\beta^* + \hat{\nu})||_2^2}{2N} + \lambda_N ||(\beta^* + \hat{\nu})||_1 \leq \frac{||\boldsymbol{y} - \boldsymbol{X}\beta^*||_2^2}{2N} + \lambda_N ||\beta^*||_1$$

Now since $\boldsymbol{y} = \boldsymbol{X}\beta^* + \boldsymbol{w}$, we substitute $\boldsymbol{w}$ in the above inequality to get

$$\frac{||\boldsymbol{w} - \boldsymbol{X}\hat{\nu}||_2^2}{2N} - \frac{||\boldsymbol{w}||_2^2}{2N} \leq \lambda_N (||\beta^*||_1 - ||(\beta^* + \hat{\nu})||_1)$$

$$\frac{(\boldsymbol{w} - \boldsymbol{X}\hat{\nu})^T(\boldsymbol{w} - \boldsymbol{X}\hat{\nu})}{2N} - \frac{\boldsymbol{w}^T\boldsymbol{w}}{2N} \leq \lambda_N (||\beta^*||_1 - ||(\beta^* + \hat{\nu})||_1)$$

After simplification, this gives

$$-\frac{\boldsymbol{w}^T\hat{\nu}\boldsymbol{X}}{N} + \frac{\hat{\nu}^T\boldsymbol{X}^T\boldsymbol{X}\hat{\nu}}{2N} \leq \lambda_N (||\beta^*||_1 - ||(\beta^* + \hat{\nu})||_1)$$

The above inequality becomes our required expression after rearranging

$$\frac{||\boldsymbol{X}\hat{\nu}||_2^2}{2N} \leq \frac{\boldsymbol{w}^T\boldsymbol{X}\hat{\nu}}{N} + \lambda_N (||\beta^*||_1 - ||(\beta^* + \hat{\nu})||_1)$$

2.4 Part D

Since the expression $\frac{\boldsymbol{w}^T\boldsymbol{X}\hat{\nu}}{N}$ can also be written as $\frac{\hat{\nu}^T\boldsymbol{X}^T\boldsymbol{w}}{N}$ because this quantity is a scalar and transpose of scalar is same as the scalar, we get

$$\hat{\nu}^T\boldsymbol{X}^T\boldsymbol{w} =< \hat{\nu}, \boldsymbol{X}^T\boldsymbol{w} >$$

We will now obtain a sub-result:
Consider a vector y. Thus, the dot product $\langle \nu, y \rangle$ becomes

$$< \hat{\nu}, \boldsymbol{y} >= \sum_{i=1}^{p} \hat{\nu}_i y_i \leq \sum_{i=1}^{p} |\hat{\nu}_i y_i| \leq \max_i\{|y_i|\} \sum_{i=1}^{p} |\hat{\nu}_i|$$

Writing the above equation in norm form we get ($||a||_\infty = \max_n |a_n|$)

$$< \hat{\nu}, \boldsymbol{y} >= \sum_{i=1}^{p} \hat{\nu}_i y_i \leq ||\boldsymbol{y}||_\infty ||\hat{\nu}||_1$$

Hence, after dividing by N both sides we get

$$\frac{\hat{\nu}^T y}{N} \leq \frac{||y||_\infty}{N} ||\hat{\nu}||_1$$

And now, say $y = \boldsymbol{X}^T\boldsymbol{w}$

$$\frac{\hat{\nu}^T\boldsymbol{X}^T\boldsymbol{w}}{N} \leq \frac{||\boldsymbol{X}^T\boldsymbol{w}||_\infty}{N} ||\hat{\nu}||_1$$

Let $x_S \in R^{|S|}$ denote the subvector indexed by elements of set S, with x_{S^c} defined in an analogous manner. For the other term we have $\beta^*_{S_c} = 0$, as β^* is assumed to be $|S| - sparse$ vector and

$$||\beta^* + \hat{\nu}||_1 = ||\beta^*_S + \hat{\nu}_S||_1 + ||\hat{\nu}_{S^c}||_1 \geq ||\beta^*_S||_1 - ||\hat{\nu}_S||_1 + ||\hat{\nu}_{S^c}||_1$$

Using above two inequalities, we can obtain eq. 11.22 as

$$\frac{||\boldsymbol{X}\hat{\nu}||_2^2}{2N} \leq \frac{||\boldsymbol{X}^T\boldsymbol{w}||_\infty}{N} ||\hat{\nu}||_1 + \lambda_N(||\hat{\nu}_S||_1 - ||\hat{\nu}_{S^c}||_1)$$

2.5 Part E

Firstly, it is assumed that $\frac{||\boldsymbol{X}^T\boldsymbol{w}||_\infty}{N} \leq \frac{\lambda_N}{2}$. Using this

$$\frac{||\boldsymbol{X}\hat{\nu}||_2^2}{2N} \leq \frac{\lambda_N}{2}(||\hat{\nu}_S||_1 - ||\hat{\nu}_{S^c}||_1) + \lambda_N(||\hat{\nu}_S||_1 + ||\hat{\nu}_{S^c}||_1)$$

Using Cauchy-Schwartz inequality

$$\frac{\lambda_N}{2}(||\hat{\nu}_S||_1 - ||\hat{\nu}_{S^c}||_1) + \lambda_N(||\hat{\nu}_S||_1 + ||\hat{\nu}_{S^c}||_1) = \frac{3}{2}\lambda_N||\hat{\nu}_S||_1 - \frac{1}{2}\lambda_N||\hat{\nu}_{S^c}||_1 \leq \frac{3}{2}\lambda_N||\hat{\nu}_S||_1 \leq \frac{3}{2}\sqrt{k}\lambda_N||\hat{\nu}||_2$$

2.6 Part F

Using the restricted Eigenvalue condition

$$v^T\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}v \geq \gamma||v||_2^2 \ \forall \text{ non-zero } v \ \in \ \boldsymbol{C}$$

which gives:

$$\frac{||\boldsymbol{X}\hat{v}||_2^2}{2N} = \frac{\hat{v}^T\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}\hat{v}}{2N} \geq \frac{\gamma||\hat{v}||_2^2}{2}$$

From the bound derived in previous part, we can say that

$$\frac{\gamma||\hat{v}||_2^2}{2} \leq \frac{||\boldsymbol{X}\hat{\nu}||_2^2}{2N} \leq \frac{3}{2}\sqrt{k}\lambda_N||\hat{\nu}||_2$$

Thus, we have

$$||\hat{\beta} - \beta^*||_2 = ||\hat{v}||_2 \leq \frac{3}{\gamma}\sqrt{\frac{k}{N}}\sqrt{N}\lambda_N$$

2.7 Part G

It is an assumption that we make in order to prove the cone constraint that $||\hat{\nu}_{S^c}||_1 \leq \alpha||\hat{\nu_S}||_1$. Specifically, we used the inequality to prove the bound in part e. Then the inequality in part e can be used to get

$$\frac{||\boldsymbol{X}\hat{\nu}||_2^2}{2N} \leq \frac{3}{2}\lambda_N||\hat{\nu}_S||_1 - \frac{1}{2}\lambda_N||\hat{\nu}_{S^c}||_1$$

Since the LHS is always greater than or equal to 0, this simplifies to

$$0 \leq \frac{3}{2}\lambda_N||\hat{\nu}_S||_1 - \frac{1}{2}\lambda_N||\hat{\nu}_{S^c}||_1 \implies ||\hat{\nu}_{S^c}||_1 \leq 3||\hat{\nu}_S||_1 \text{ (Cone constraint)}$$

2.8 Part H

From the restricted eigenvalue condition, we can say that

$$\frac{1}{n}||Ax||^2 \geq \kappa||x||^2 \ \forall x \in C_\alpha S \text{ and for } \kappa > 0$$

If we take the vector $X = \hat{\beta} - \beta^*$,we cannot conclude that $||X||_2^2$ is small. Thus, we saw that in order to minimize the loss function it must be strictly convex whereas in practice it is convex in some direction while non-concave in others. Thus, we relaxed the condition and allowed strict convexity only on a subset $\boldsymbol{C} \subset \boldsymbol{R}^p$. The lasso regression constraint that $||\beta||_1 \leq R$ with or without regularization leads the subset to be cone, i.e,

$$|\hat{\nu_S}||_1 \leq \alpha||\hat{\nu}_{S^c}||_1$$

2.9 Part I

In both the theorems we are required to calculate the minimum eigenvalue of our dictionary matrix but in this theorem our search for the minimum eigenvalue is deterministic, i.e., in a set with cone constraint and moreover here we do not need the explicit calculation of the RIC. One fascinating thing to observe here is that this theorem also emphasizes on the interaction of the dictionary with noise vector, greater the correlation greater is the error whereas theorem 3 doesn't take any such thing into account. Even with knowledge of the support set, since the model has k free parameters, no method can achieve squared '2-error that decays more quickly than k/N . Thus, apart from the logarithmic factor, the lasso rate matches the best possible that one could achieve. Apart from this, the rate this theorem achieves, including the logarithmic factor, is known to be minimax optimal, meaning that it cannot be substantially improved upon by any estimator.

The biggest drawback of this theorem is that it doesn't consider compressible signals. The tightness conditions do not allow near 0 elements to be considered as 0. Thu,s representation of this signal in the standard basis does not work. as the sparsity constraint for the sensed signal is very tight and does not allow room for signals which are less sparse but some elements are very small. Theorem 3 handles such situation better.

2.10 Part J

For 'Dantzig selector' the maximum correlation $||A^T e||_\infty$ is assumed to be less than some constant λ. Thus, the common thread between 'Dantzig selector' and the LASSO is accounting for the interaction of dictionary with the noise vector in the error bound. A similar kind of constraint was assumed for LASSO also, as done while proving the sub part e.

3 Problem 3

In all the images below, the left one corresponds to the reconstructed slice 50, and the right one is the reconstructed slice 51.

3.1 Part A - Q3a.m

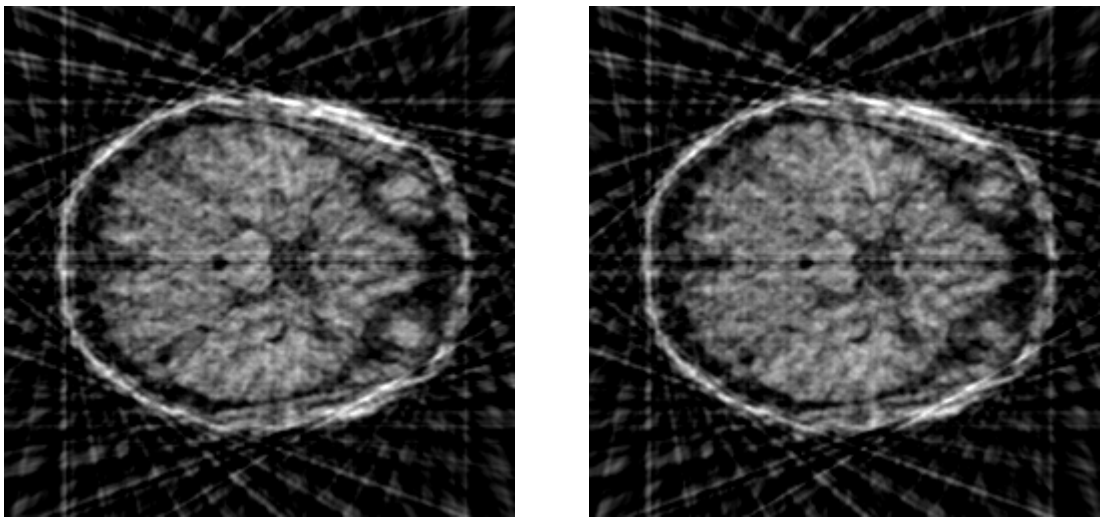


Figure 1: Ram-Lak Filtered Back-Propagation - Slice 50 and 51 (using 18 uniformly separated angles)

3.2 Part B - Q3b.m

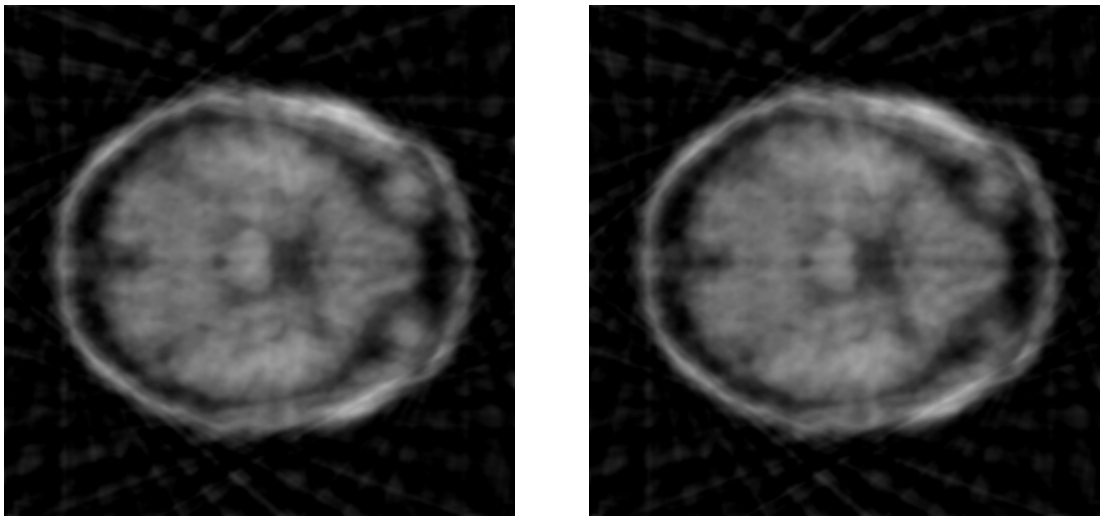


Figure 2: Independent CS based reconstruction - Slice 50 and 51 (using 18 uniformly separated angles)

3.3 Part C - Q3c.m

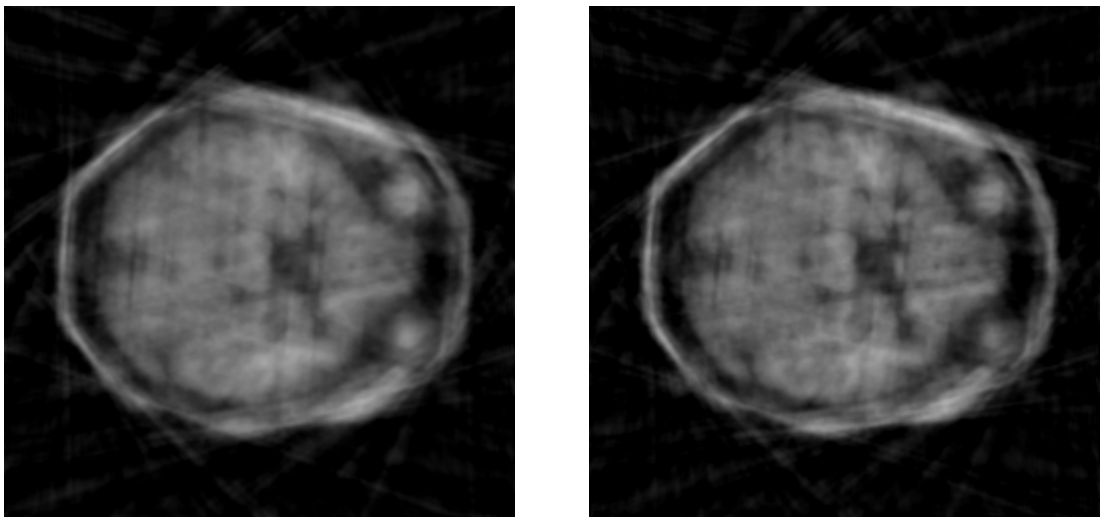


Figure 3: Coupled CS based reconstruction - Slice 50 and 51 (using 18 randomly chosen angles)

3.4 Part D

In case of 3 consecutive slices, we can perform coupled CS based reconstruction with the help of the following objective function:

$$E(\beta 1, \beta 2, \beta 3) = \left|\left| \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} - \begin{bmatrix} R_1U & 0 & 0 \\ R_2U & R_2U & 0 \\ R_3U & 0 & R_3U \end{bmatrix} \begin{bmatrix} \beta_1 \\ \Delta\beta_1 \\ \Delta\beta_2 \end{bmatrix} \right|\right|_2^2 + \lambda \left|\left| \begin{bmatrix} \beta_1 \\ \Delta\beta_1 \\ \Delta\beta_2 \end{bmatrix} \right|\right|_1$$

Here,

- x_1, x_2, x_3 are the 3 consecutive slices (each slice is a 2D image).
- y_1, y_2, y_3 are the tomographic projections of the 3 slices x_1, x_2, x_3 under consideration, represented as 1D-vectors.
- β_1 is the vectorized 2D-DCT of the first slice x_1.
- $\beta_1 + \Delta\beta_1$ is the vectorized 2D-DCT of the second slice x_2.
- $\beta_1 + \Delta\beta_2$ is the vectorized 2D-DCT of the third slice x_3.
- R_1 is the radon transform matrix for the random angles used to compute tomographic projections of x_1.
- R_2 is the radon transform matrix for the random angles used to compute tomographic projections of x_2.
- R_3 is the radon transform matrix for the random angles used to compute tomographic projections of x_3.
- U is the 2D-DCT basis matrix

4 Problem 4

4.1 Details of the Paper

Title: **High-resolution seismic tomography of Long Beach, CA using machine learning**
Venue: Scientific Reports volume 9, Article number: 14987 (2019)
Date: 18 October 2019
The paper can be accessed here.

4.2 Introduction

This paper is essentially about a machine learning-based tomography method to obtain high-resolution subsurface geophysical structure from seismic noise. The method used, called locally sparse travel time tomography (LST), uses unsupervised machine learning to exploit the dense sampling obtained by ambient noise processing on large arrays. Dense sampling permits the LST method to learn a dictionary of local, or small-scale, geophysical features directly from the data. The features are the small scale patterns of Earth structure most relevant to the given tomographic imaging scenario. The results show promise for LST in obtaining detailed geophysical structure in travel time tomography studies.

4.3 Problem

The proposed locally-sparse travel time tomography (LST) approach obtains high resolution by assuming that small patches of discrete phase speed maps are repetitions of few elemental patterns from a dictionary of patterns. Relative to conventional tomography methods, the sparsity of the dictionary representation permits smooth and discontinuous, high-resolution features where warranted by the data. In LST, surface wave propagation is approximated as straight ray paths through an $N = W_1 \times W_2$ pixel phase speed map, and the travel time perturbations $t \in R^M$ from a known reference for M rays are modeled as

$$t = As_g + \epsilon$$

where $A \in R^M \times N$ is the tomography matrix, $s_g \in R^N$ is the perturbation global slowness (inverse of speed), and $\epsilon \in R^M$ is Gaussian noise.
Now, the paper also talks about a local model, where Dictionary learning is added to the local problem by optimizing D (the dictionary) in small patches across the signal. The optimisation problem is easy to express in terms of the radon transform and the local models as it is combined to a Bayesian maximum a posteriori (MAP) objective

$$\{s_g, s_s, \hat{X}_i\} = \min_{s_g, s_s, \hat{X}_i} \left(\frac{1}{\sigma^2}||t - As||_2^2 + \frac{1}{\sigma_g^2}||s_s - s_g||_2^2 + local \right)$$

where σ is variance of noise and the 'local' stands for the objective related to the local model. The tomographic image used for geophysical interpretation from the LST algorithm is $\hat{s}_s$. They also implement conventional tomography using a Bayesian approach, which regularizes the inversion with a global smoothing (non-diagonal) covariance. Considering the measurements, the MAP estimate of the slowness is

$$s_g = (A^T A = \eta \Sigma_L^{-1})^{-1} A^T t$$

4.4 Optimization Technique

The above objective functions are optimised in a variety of ways. The Dictionary learning procedure is here solved using iterative thresholding (ISTA) and signed k-means (ITKM) algorithms.
They find the MAP estimates $\{s_g, s_s, \hat{X}_i\}$ using a block-coordinate minimization algorithm by decoupling the local and global models via substitution, which gives

$$s_g = \min \left(\frac{1}{\sigma^2}||t - As||_2^2 + \frac{1}{\sigma_g^2}||s_s - s_g||_2^2 \right)$$

This expression, along with the others are solved iteratively until convergence

5 Problem 5

By definition, the Radon transform of a 2-dimensional signal is computed as

$$R_\theta(f) = g_1(\rho, \theta) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y)\delta(x\cos\theta + y\sin\theta - \rho)dxdy$$

Thus, for the scaled image, $f(ax, ay)$, we have

$$g_2(\rho, \theta) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(ax, ay)\delta(x\cos\theta + y\sin\theta - \rho)dxdy$$

Putting $p = ax, q = ay$, we obtain

$$\begin{aligned} g_2(\rho, \theta) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \frac{1}{a^2} f(p, q)\delta\left(\frac{p}{a}\cos\theta + \frac{q}{a}\sin\theta - \rho\right) dpdq \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \frac{|a|}{a^2} f(p, q)\delta(p\cos\theta + q\sin\theta - a\rho)dpdq \\ &= \boxed{\frac{1}{|a|} g_1(a\rho, \theta)} \text{ , where } g_1(\rho, \theta) = R_\theta(f) \end{aligned}$$

Here we use the fact that for any non-zero scalar α, $\delta(\alpha x) = \frac{1}{|\alpha|}\delta(x)$.

6 Problem 6

Consider the impulse $f(x, y) = \delta(x, y)$. By definition, the Radon transform of a 2-dimensional signal is computed as

$$R_\theta(f) = g(\rho, \theta) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y)\delta(x\cos\theta + y\sin\theta - \rho)dxdy$$

Substituting $f(x, y) = \delta(x, y)$, we obtain.

$$\begin{aligned} R(f) = g(\rho, \theta) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y)\delta(x\cos\theta + y\sin\theta - \rho)dxdy \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \delta(x, y)\delta(x\cos\theta + y\sin\theta - \rho) \end{aligned}$$

Using the sifting property of the Dirac-Delta

$$R(f) = g(\rho, \theta) = \boxed{\delta(\rho)} = \begin{cases} 0, \rho \neq 0 \\ \infty, \rho = 0 \end{cases}$$

Now, for the shifted impulse, Substituting $f(x, y) = \delta(x - x_0, y - y_0)$, we obtain.

$$\begin{aligned} R(f) = g(\rho, \theta) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y)\delta(x\cos\theta + y\sin\theta - \rho)dxdy \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \delta(x - x_0, y - y_0)\delta(x\cos\theta + y\sin\theta - \rho) \end{aligned}$$

Using the sifting property of the Dirac-Delta

$$R(f) = g(\rho, \theta) = \boxed{\delta(x_0\cos\theta + y_0\sin\theta - \rho)} = \begin{cases} 0, \rho \neq x_0\cos\theta + y_0\sin\theta \\ \infty, \rho = x_0\cos\theta + y_0\sin\theta \end{cases}$$