# CS754 Assignment 5 Report

Arpon Basu
Shashwat Garg

Spring 2022

## Contents

# Introduction

Welcome to our report on CS754 Assignment 5. We have tried to make this report comprehensive and self-contained. We hope reading this would give you a proper flowing description of our work, methods used and the results obtained.

Also note that we installed the `Image Processing Toolbox` in MATLAB for this assignment. Thus the grader is urged to install it if she wishes to run the code on her on her machine. Also note that some of our code may take a while to run because of the intensive nature of the computations involved.

Hope you enjoy reading the report. Here we go!

# 1 Problem 1

## 1.1 (a)

## 1.2 (b)

The likelihood was Laplacian.
The terms involving the likelihood were

$$\sum_{i,k} \rho(f_{i,k} \cdot I_1) \text{ and } \rho(f_{i,k} \cdot (I - I_1))$$

The log-histogram of derivative filters was the prior. The terms involving the prior were

$$\sum_{i,k} \rho(f_{i,k} \cdot I_1) \text{ and } \rho(f_{i,k} \cdot (I - I_1))$$

## 1.3 (c)

Note that since we have to remove reflections from natural images, we have to somehow distinguish between natural images and artificial reflections. That is done by exploiting the property of natural images that the log-histograms of the derivatives of the image (obtained through some derivative filter) are sparse, and thus they decay rapidly away from zero (of the x-axis), thus giving them a **convex nature**.
Thus, **Gaussian likelihoods aren't chosen because they are concave in nature**, ie:- they decay more slowly than one would expect from a natural image. On the other hand, for a Laplacian likelihood, the log histogram is straight decreasing line, which basically delineates the boundary between sparse and non-sparse likelihoods. In general, one may say that likelihoods of the form $e^{-x^\alpha}$ are better than likelihoods of the form $e^{-x^\beta}$ if $\alpha < \beta$, and thus since $1 < 2$, a Laplacian is a better likelihood than the Gaussian.

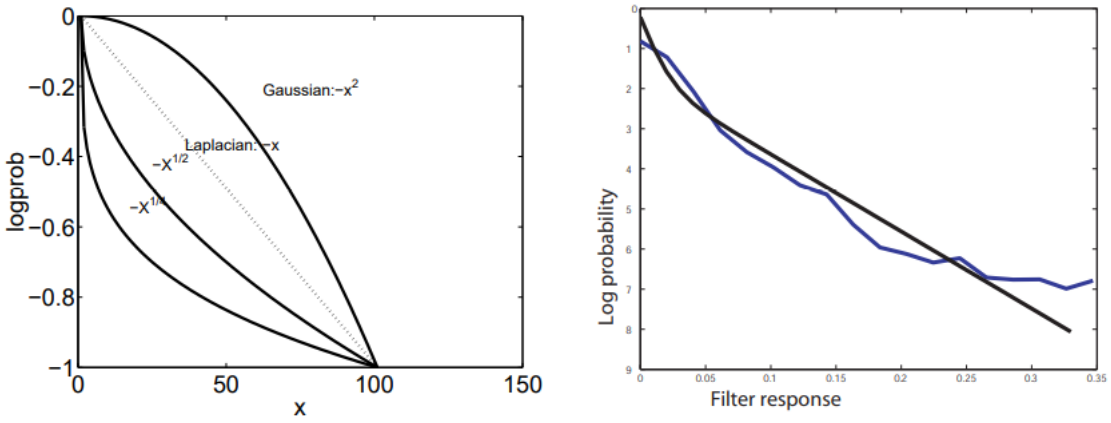


Figure 1: Graph highlighting better sparsity of Laplacian vis-a-vis Gaussian

## 2 Problem 2

From the machinery of Baye's theorem we know that the **Maximum A Posteriori Estimate** $\boldsymbol{x}$ is given by

$$\boldsymbol{x} = \arg\max_x p(x|y)$$

But we also know that

$$p(x|y) \propto p(x)p(y|x)$$

Now, since $x \sim \mathcal{N}(0, \Sigma_x)$, we have that (we ignore differentials and just quote the value of the PDF of the multivariate Gaussian)

$$p(x) = \frac{1}{\sqrt{(2\pi)^n |\Sigma_x|}} \exp(-\frac{1}{2} x^T \Sigma_x^{-1} x)$$

Also, note that that $p(y|x)$ denotes how likely $y$ is to be obtained if $x$ is the input vector from which it was created. But note that this is equivalent to asking how likely is $\eta = y - \Phi x$ to have come from the given noise distribution $\mathcal{N}(0, \sigma^2 I_{m\times m})$. Thus

$$p(y|x) = p(\eta) = \frac{1}{\sqrt{(2\pi)^m \sigma^{2m}}} \exp(-\frac{1}{2\sigma^2} \|\eta\|_2^2) = \frac{1}{\sqrt{(2\pi)^m \sigma^m}} \exp(-\frac{1}{2\sigma^2} \|y - \Phi x\|_2^2)$$

where we plugin the scalar (a multiple of the identity matrix) covariance matrix to derive the neat form seen above. Thus, when we plugin the expressions (and remove all the constants), we get

$$p(x|y) \propto \exp(-\frac{1}{2}(\frac{1}{\sigma^2} \|y - \Phi x\|_2^2 + x^T \Sigma_x^{-1} x))$$

Thus maximizing $p(x|y)$ is equivalent to minimizing $(\frac{1}{\sigma^2} \|y - \Phi x\|_2^2 + x^T \Sigma_x^{-1} x)$ given $y$ and $\Phi$. Differentiating the expression w.r.t the vector $\boldsymbol{x}$, we get

$$\frac{\partial(\frac{1}{\sigma^2} \|y - \Phi x\|_2^2 + x^T \Sigma_x^{-1} x)}{\partial x} = 0$$

$$\implies \frac{1}{\sigma^2}(2\Phi^T \Phi x - 2\Phi^T y) + 2\Sigma_x^{-1} x = 0$$

$$\implies (\Phi^T \Phi + \sigma^2 \Sigma_x^{-1}) x = \Phi^T y$$

$$\implies \boldsymbol{x}_{\text{MAP}} = (\Phi^T \Phi + \sigma^2 \Sigma_x^{-1})^{-1} \Phi^T y$$

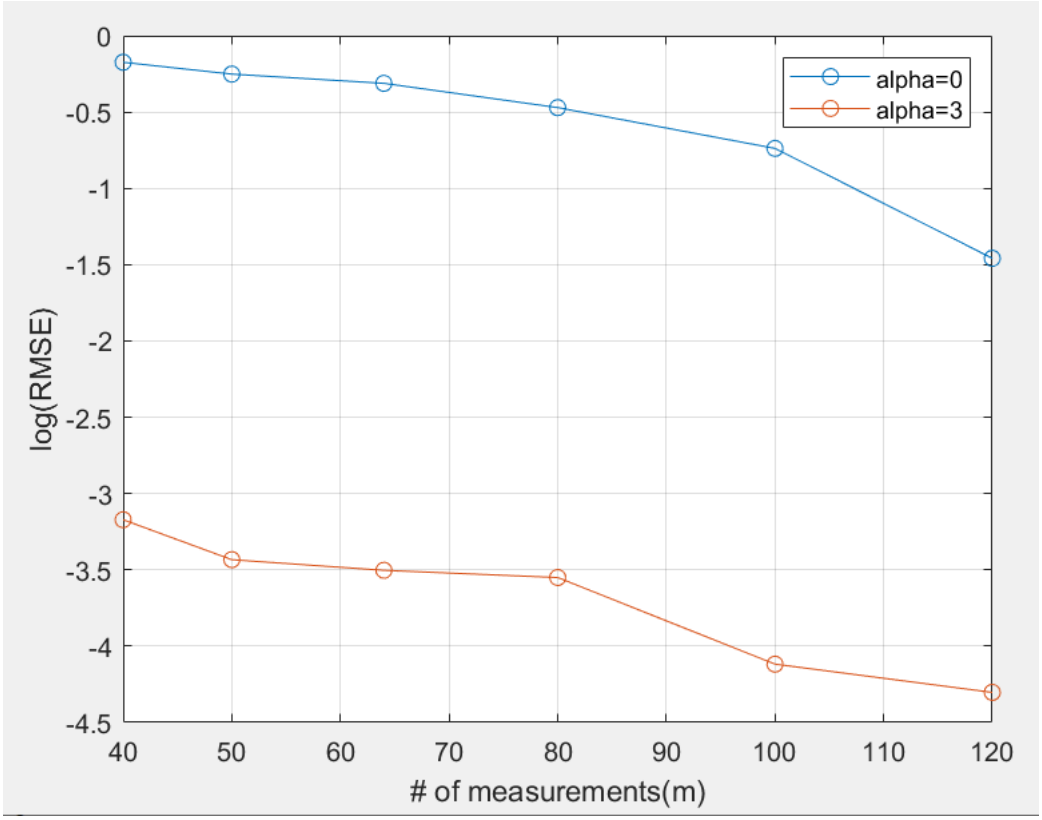


Figure 2: Errors obtained for various $m$'s for $\alpha = 0$ and $\alpha = 3$

Note that the errors here have been plotted on log-scale to better highlight the differences between them. As one can see, there is indeed a noticeable difference in the reconstruction accuracy for $\alpha = 0$ and $\alpha = 3$. We would like to explain this phenomena as a combination of two observations:

- RMSE decreases as $m$ increases for a given $\alpha$: This is quite intuitive since more samples of $\boldsymbol{x}$ will obviously increase the accuracy of our Bayesian estimate.
- RMSE decreases as $\alpha$ increases for a given $m$: Note that for higher $\alpha$, the eigenvalues of our covariance matrix $\Sigma_x$ fall of very rapidly, and thus, a random variable $\boldsymbol{x}$ sampled from a probability distribution with $\Sigma_x$ as the covariance matrix will vary very little in the directions of the principal eigenvectors of those small eigenvalues, thus essentially "reducing" the "effective" dimension of the vector $\boldsymbol{x}$ (one may note that this is the very insight of PCA, which recognizes that one may effectively set the very small eigenvalues to 0). Thus, for larger $\alpha$, we effectively have to interpolate "smaller" $\boldsymbol{x}$, and hence the better reconstruction accuracy.

# 3 Problem 3

In this problem we'll deal with the problem below.
Let $X^*$ be the solution of the optimization problem given below

$$X^* := \arg\min_X \|X\|_* \quad \text{subject to } \mathcal{A}(X) = b$$

and let $X_0$ be *any* matrix of rank $r$ such that $\mathcal{A}(X_0) = b$. We will be investigating under what conditions will $X^*$ necessarily equal $X_0$.

## 3.1 3(a)

$X^*$ is the matrix of minimum $\|\cdot\|_*$ norm subject to the constraint $\mathcal{A}(X) = b$. Now, $X_0$ belongs to our search space since by the hypothesis of the theorem $\mathcal{A}(X_0) = b$. And since we optimize the $\|\cdot\|_*$ norm over our search space to get $X^*$, we have that

$$\|X^*\|_* \leq \|X_0\|_*$$

## 3.2 3(b)

By the very definition of $R_c$, as also written two lines above where this question was posed, we have that $X_0'R_c$ and $X_0R_c'$ are both zero (matrices). Thus, applying Lemma 2.3 verbatim (which states that if two matrices $A$ and $B$ of equal dimensions satisfy $A'B = 0$ and $AB' = 0$, then $\|A+B\|_* = \|A\|_* + \|B\|_*$), we get that $\|X_0 + R_c\|_* = \|X_0\|_* + \|R_c\|_*$, since $X_0$ and $R_c$ are clearly of equal dimensions ($R = X^* - X_0$, $R = R_0 + R_c$, thus implying that $X_0$ and $R_c$ are addable, and thus of equal dimensions) and satisfy $X_0'R_c = 0$ and $X_0R_c' = 0$.

## 3.3 3(c)

WLOG assume that the singular values of $R_c$ are arranged in descending order, ie:- $\sigma_i \leq \sigma_j$ for any $i \geq j$. Then note that if $k \in I_{i+1}$, then $\sigma_k \leq \sigma_j \ \forall j \in I_i$ since $k$ is a larger index. Thus

$$3r \cdot \sigma_k \leq \sum_{j \in I_i} \sigma_j \ \forall \ k \in I_{i+1}$$

by adding up all the inequalities individually, and thus, finally

$$\sigma_k \leq \frac{1}{3r} \sum_{j \in I_i} \sigma_j \ \forall \ k \in I_{i+1}$$

## 3.4 3(d)

For this problem, we recall a linear algebra property which says that for any matrix $M$, we have that

$$\|M\|_F^2 = \sum_{i=1}^{n} \sigma_i^2$$

where $\sigma_1$, $\sigma_2$, ..., $\sigma_n$ are the singular values of $M$. One can easily see this by observing that $\|M\|_F^2 = \text{tr}(MM')$. But the trace of a matrix is equal to the sum of it's eigenvalues, and the eigenvalues of $MM'$ are the squares of the singular values of $M$.
Also note that the $\|\cdot\|_*$ norm is defined as

$$\|X\|_* = \sum_{\sigma \text{ is a singular value of X}} \sigma$$

Thus

$$\sigma_k \leq \frac{1}{3r} \sum_{j \in I_i} \sigma_j = \frac{1}{3r} \|R_i\|_* \; \forall \; k \in I_{i+1}$$

$$\implies \sigma_k^2 \leq \frac{1}{9r^2} \|R_i\|_*^2 \; \forall \; k \in I_{i+1}$$

$$\implies \sum_{k \in I_{i+1}} \sigma_k^2 \leq 3r \frac{1}{9r^2} \|R_i\|_*^2$$

But then we also know that

$$\|R_{i+1}\|_F^2 = \sum_{k \in I_{i+1}} \sigma_j^2$$

Thus

$$\|R_{i+1}\|_F^2 \leq \frac{1}{3r} \|R_i\|_*^2$$

as desired.

## 3.5 3(e)

We know from the previous part that

$$\|R_{i+1}\|_F^2 \leq \frac{1}{3r} \|R_i\|_*^2$$

$$\implies \|R_{i+1}\|_F \leq \frac{1}{\sqrt{3r}} \|R_i\|_*$$

$$\implies \sum_{i=1}^{t-1} \|R_{i+1}\|_F \leq \sum_{i=1}^{t-1} \frac{1}{\sqrt{3r}} \|R_i\|_* \leq \sum_{i=1}^{t} \frac{1}{\sqrt{3r}} \|R_i\|_*$$

$$\implies \sum_{i \geq 2} \|R_i\|_F \leq \sum_{i \geq 1} \frac{1}{\sqrt{3r}} \|R_i\|_*$$

as desired, where there are $t$ matrices $\{R_i\}_{1 \leq i \leq t}$.

## 3.6 3(f)

Note that the justification we gave for 3(b) was basically to prove that $\|R_0\|_* \geq \|R_c\|_*$. We proceed on that as follows

$$\|X_0\|_* \geq \|X^*\|_* = \|X_0 + R\|_*$$

where $R := X_0 - X^*$. Then

$$\|X_0\|_* \geq \|X_0 + R\|_* = \|X_0 + R_c + R_0\|_* \geq \|X_0 + R_c\|_* - \|R_0\|_*$$

where the last inequality follows by the triangle inequality. But in part (b) we just proved that $\|X_0 + R_c\|_* = \|X_0\|_* + \|R_c\|_*$, and thus

$$\|X_0\|_* \geq \|X_0\|_* + \|R_c\|_* - \|R_0\|_*$$

$$\implies \|R_0\|_* \geq \|R_c\|_*$$

as desired.

## 3.7 3(g)

We use a relation from linear algebra which establishes a bound between Frobenius norm and the nuclear norm

$$\boldsymbol{\|M\|_* \leq \sqrt{\text{rank } M}\|M\|_F}$$

But from it's construction itself (the construction of $R_0$ entails (through Lemma 3.4) that $\text{rank}(R_0) \leq 2{\cdot}\text{rank}(X_0) = 2r$, since $\text{rank}(X_0) = r$) we know that $\text{rank}(R_0) \leq 2r$, thus

$$\sum_{i\geq 2}\|R_i\|_F \leq \sum_{i\geq 1}\frac{1}{\sqrt{3r}}\|R_i\|_* = \frac{1}{\sqrt{3r}}\|R_c\|_* \leq \frac{1}{\sqrt{3r}}\|R_0\|_* \leq \frac{\sqrt{\text{rank } R_0}}{\sqrt{3r}}\|R_0\|_F \leq \frac{\sqrt{2r}}{\sqrt{3r}}\|R_0\|_F$$

as desired.

## 3.8 3(h)

As we have seen above, $\text{rank}(R_0) \leq 2r$, and also note that $R_1$ was constructed such that it's rank was atmost $3r$ ("$R_c = \sum_{i=1}^{t} R_i$, $\text{rank}(R_i) \leq 3r$").
Now, we also know that

$$\text{rank}(A + B) \leq \text{rank}(A) + \text{rank}(B)$$

Thus

$$\text{rank}(R_0 + R_1) \leq \text{rank}(R_0) + \text{rank}(R_1) \leq 2r + 3r = 5r$$

as desired.

## 3.9 3(i)

Note that $\mathcal{A}$ was defined was a linear mapping (from matrices to vectors), and thus

$$\|\mathcal{A}(R)\| = \|\mathcal{A}(R_0 + R_c)\| = \|\mathcal{A}(R_0 + R_1 + R_2 + \ldots)\| = \|\mathcal{A}(R_0 + R_1) + \mathcal{A}(R_2 + \ldots)\|$$

$$\geq \|\mathcal{A}(R_0 + R_1)\| - \sum_{j\geq 2}\|\mathcal{A}(R_j)\|$$

where the last inequality follows by the triangle inequality (note that $\|\cdot\|$ is a metric, and hence the triangle inequality applies on it).

### 3.10 3(j)

By the definition of Restricted Isometry constants, we have

$$(1-\delta(\mathcal{A}))\|X\|_F \leq \|\mathcal{A}(X)\| \leq (1+\delta(\mathcal{A}))\|X\|_F$$

where $\delta$ is a function of the rank of the mapping $\mathcal{A}$.
Thus

$$\|\mathcal{A}(R_0+R_1)\| \geq (1-\delta_{5r})\|R_0+R_1\|_F$$
$$\sum_{j\geq 2}\|\mathcal{A}(R_j)\| \leq \sum_{j\geq 2}(1+\delta_{3r})\|R_j\|_F$$

and thus

$$\|\mathcal{A}(R)\| \geq \|\mathcal{A}(R_0+R_1)\| - \sum_{j\geq 2}\|\mathcal{A}(R_j)\| \geq (1-\delta_{5r})\|R_0+R_1\|_F + \sum_{j\geq 2}(1+\delta_{3r})\|R_j\|_F$$

as desired.

### 3.11 3(k)

Note that

$$\mathcal{A}(R) = \mathcal{A}(X^* - X_0) = \mathcal{A}(X^*) - \mathcal{A}(X_0) = b - b = 0$$

Note that since $\mathcal{A}$ is a linear mapping, that's why $\mathcal{A}(X^* - X_0) = \mathcal{A}(X^*) - \mathcal{A}(X_0)$.
Thus we get our desired result.

### 3.12 3(l)

The condition is

$$(1-\delta_{5r}) - \frac{9}{11}(1+\delta_{3r}) \geq 0$$
$$\Leftrightarrow 11(1-\delta_{5r}) - 9(1+\delta_{3r}) \geq 0$$
$$\Leftrightarrow 11 - 9 - 11\delta_{5r} - 9\delta_{3r} \geq 0$$
$$\Leftrightarrow 11\delta_{5r} + 9\delta_{3r} \leq 2$$

as desired.

## 4 Problem 4

### 4.1 (1)

Note that any algorithm for matrix completion will basically specify a sampling of entries to "see" to predict the other entries. Now, if the singular vectors of the matrix aren't incoherent with canonical basis vectors, then one can intuitively say that they're quite "concentrated", and consequently there is a high chance of them lying in the null space of our sampling operator, and if that happens, then it becomes impossible to reconstruct the matrix without seeing all of it's entries, which defeats the purpose of matrix prediction. The "spread" of singular vectors w.r.t the canonical basis vectors can be quantified as the coherence.

### 4.2 (2)

If the problem is changed to that in Eq. 1.13, ie:-

$$\begin{gathered}\text{minimize rank}(X)\\ \text{subject to: } f_i^* X g_j = f_i^* M g_j \ \forall (i,j) \in \Omega\\ \text{where } \Omega \text{ is to be determined by our algorithm}\end{gathered}$$

then according to the paper itself we can propose the following modification:
Let $\{f_i\}$ and $\{g_i\}$ be the bases of our modified rank minimization problem. Then note the similar with the previous vanilla problem, where both $\{f_i\}$ and $\{g_i\}$ were $\{e_i\}$. **Thus, for this modified problem, it's uniquely solvable if our matrix $M$ is incoherent w.r.t $\{f_i\}$ and $M'$ is incoherent w.r.t $\{g_i\}$ respectively**.
One may also formally see this in the following way: Let

$$M = \sum_{k \geq 1} \sigma_k u_k v_k^*$$

be the SVD expansion of $M$. Then

$$f_i^* M g_j = \sum_{k \geq 1} \sigma_k f_i^* u_k v_k^* g_j$$

Clearly, incoherence of $f_i$ with $\{u_k\}_{k \geq 1}$ and $g_j$ with $\{v_k\}_{k \geq 1}$ is necessary for good reconstruction.

### 4.3 (3)

The example given in the paper is the follows:
Let $M \in \mathbb{R}^{n \times n}$ be the matrix which has a 1 on it's top right corner, and is zero everywhere else. Clearly, the rank of this matrix is 1. However, for this matrix, most samplings of it's entries would just be a set of zeros, and it's impossible to reasonably guess via any algorithm that the matrix is non-zero. This is due to the fact, as mentioned above, that the singular vectors of $M$ aren't sufficiently "spread" out (uncorrelated with the columns of $I$) for our convex optimization program to predict the matrix successfully.

## 5 Problem 5

### 5.1 (1)

The title of the paper we chose was **Low-Rank Sparse Learning for Robust Visual Tracking**, and it was presented in the venue **European Conference on Computer Vision (ECCV), 2012**. Click here for the paper,

### 5.2 (2)

The problem being solved in the paper is that of **particle tracking**, ie:- we have to track the trajectory of an object across a series of snapshots, say as in a surveillance footage, or robotics, or human computer interaction. The primary challenges in this problem is how we deal with **occlusion (ie:- our target getting blocked by some other object), background clutter (many objects in the background, thus making it difficult to keep track of our object in their midst), varying viewpoints (the object might look different from different viewpoints), and illumination and scale changes (these are also factors which changes how the object is perceived)**. The solution to this problem is motivated by similar advances made in the field of Computer Vision in things such as robust face recognition, subspace clustering, background subtraction.

## 5.3 (3)

The optimization problem which was formulated to solve this problem is presented below

$$\min_{Z,E} \lambda_1\|Z\|_* + \lambda_2\|Z\|_{1,1} + \lambda_3\|E\|_{1,1}$$
$$\text{such that: } X = D_tZ + E$$

where $D_t$ is our dictionary, $Z$ represents the set of our particles (each column of $Z$ is a particle), and $X$ is the actual observation of particle pixels in space. $E$ represents a sparse error (denoting an occluding object, for example), which may assume large values on it's (small) support.
The assumptions (pertaining to low rank recovery) taken are that every column vector $x_i$ of $X$ is expressible as a sparse combination of column vectors of $Z$, ie:- $z_i$. Now, since the objective of the paper is to minimize the rank of the representations of all particles together, the **low rank matrix recovery problem that kicks into action is the minimization of the nuclear norm of $Z$, ie:- $\|Z\|_*$, since the nuclear norm of a matrix represents the convex envelope of the matrix rank, we achieve the minimization of matrix rank indirectly through the minimization of $\|Z\|_*$, and hence the $\lambda_1\|Z\|_*$ term**.
This paper also aims to achieve sparse representations (along with low rank, as mentioned above) of the matrix $Z$ (and hence the $\lambda_2\|Z\|_{1,1}$ term) and $E$ (and hence the $\lambda_3\|E\|_{1,1}$ term). Note that $E$ is sparse because the disturbances to the object are highly localized, and thus a sparse representation serves us well. Note that we forced $Z$ to be both low rank and sparse, thus employing machinery from both low rank recovery and compressed sensing.
Thus, in the following ways, the low rank matrix recovery problem helps us solve the problem outlined above.