CS754 Assignment 2 Report

Arpon Basu
Shashwat Garg

Spring 2022

Contents

Introduction

Welcome to our report on CS754 Assignment 2. 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. Feel free to keep our code script alongside to know the exact implementation of our tasks.

Hope you enjoy reading the report. Here we go!

1 Problem 1

1.1 Given $\delta_{2s} = 1$

In this case, our RIP theorem inequality becomes-

$$0 \leq \|\Phi h\|_2^2 \leq 2\, \|h\|_2^2$$

where h is a 2s-sparse vector.

We know that δ_{2s} is the smallest number such that the above holds for all s-sparse vectors for the matrix Φ.

This means that, unless the $\delta_{2s} = 1$ is present to take care of the upper bound on $\|\Phi h\|_2^2$, there will exist an s-sparse vector such that $\|\Phi h\|_2^2 = 0$, which implies that $\Phi h = \mathbf{0}$. Denoting the columns of Φ as $C_1, C_2 \cdots C_n$ and the entries of h as $a_1, a_2, \cdots a_n$, we can get that-

$$\sum_{i=1}^{n} C_i \cdot a_i = 0$$

If we focus only on the non-zero entries of h, we can get a linear combination of at-most 2s columns of Φ which sums up to 0. Thus, in this case, we can always say that 2s columns of Φ are linearly dependent.

Hence, when $\delta_{2s} = 1$, 2s columns of Φ may be linearly dependent.

1.2 Triangle Inequality

We are trying to solve the following-

$$\min_{\tilde{x} \in \mathbb{R}^n} \|\tilde{x}\|_1 \;\; where \;\; \|y - \Phi\tilde{x}\|_2 \leq \epsilon$$

The solution we receive is called x^* and the actual, true, unknown solution is called x. Now, note that our reconstruction demands that $\|y - \Phi x_{estimated}\|_2 \leq \epsilon$

This implies that x^* will also obey the same property.

Secondly, in the premise of the Theorem, we assume that the noise magnitude is less than ϵ. This is one of the factors we take into account while choosing the ϵ value. Hence, by choice of ϵ in noise, we get that $\|y - \Phi x\|_2 \leq \epsilon$.

Finally, with these two tools, we can use the Triangle Inequality as follows-

$$\|\Phi(x^* - x)\|_2 = \|(\Phi x^* - y) + (y - \Phi x)\|_2 \leq \|\Phi x^* - y\|_2 + \|y - \Phi x\|_2$$

Using the two arguments above, we can say that-

$$\|\Phi x^* - y\|_2 + \|y - \Phi x\|_2 \leq \epsilon + \epsilon = 2\epsilon$$

Thus, finally, we get that

$$\|\Phi(x^* - x)\|_2 \leq \|\Phi x^* - y\|_2 + \|y - \Phi x\|_2 \leq 2\epsilon$$

Hence, proved.

1.3 Relation between Norms

We know that l_∞ norm gives out the element with the maximum magnitude in the vector. Let us assume entries of h_{T_j} to be $a_1, a_2, \cdots a_n$. WLOG, let us assume that, since h_{T_j} is s-sparse, $a_1, a_2, \cdots a_s$ are non-zero and rest are 0. The re-ordering won't affect the solution since we would be taking the norm anyways. With this knowledge, we can see that-

$$\forall a_i : i \leq s, (a_i)^2 \leq \left\|h_{T_j}\right\|_\infty^2$$

Adding all these s inequalities up, since all a_i for $i > s$ are zero, we get-

$$\left\|h_{T_j}\right\|_2^2 \leq s \left\|h_{T_j}\right\|_\infty^2$$

From this, we directly get that-

$$\left\|h_{T_j}\right\|_2 \leq s^{1/2} \left\|h_{T_j}\right\|_\infty$$

Secondly, by construction, we know that every element of h_{T_j} is smaller than the every element of $h_{T_{j-1}}$. Thus, we can say the following-

$$\left\|h_{T_j}\right\|_\infty \leq |b_i|$$

where $b_1, b_2, \cdots b_s$ are the s non-zero elements of $h_{T_{j-1}}$. Adding these up for the s inequalities, and using the fact that all b_i for $i > s$ are zero, we get-

$$s \left\|h_{T_j}\right\|_\infty \leq \left\|h_{T_{j-1}}\right\|_1$$

This is equivalent to-

$$s^{1/2} \left\|h_{T_j}\right\|_\infty \leq s^{-1/2} \left\|h_{T_{j-1}}\right\|_1$$

Hence, finally, we have obtained the required relations-

$$\left\|h_{T_j}\right\|_2 \leq s^{1/2} \left\|h_{T_j}\right\|_\infty \leq s^{-1/2} \left\|h_{T_{j-1}}\right\|_1$$

1.4 Summing up the norms

By just summing up the above inequality, we can say that-

$$\left\|h_{T_j}\right\|_2 \leq s^{-1/2}\left\|h_{T_{j-1}}\right\|_1$$
$$\implies \sum_{j\geq 2}\left\|h_{T_j}\right\|_2 \leq s^{-1/2}\left(\|h_{T_1}\|_1 + \|h_{T_2}\|_1 + \dots\right)$$

Also, we know that-

$$\|h_{T_0}\|_1 + \|h_{T_1}\|_1 + \|h_{T_2}\|_1 + \dots = \|h\|_1$$

Thus, we can say that

$$\left\|h_{T_0^c}\right\|_1 = \|h\|_1 - \|h_{T_0}\|_1 = \|h_{T_1}\|_1 + \|h_{T_2}\|_1 + \dots$$

Hence, finally we have-

$$\sum_{j\geq 2}\left\|h_{T_j}\right\|_2 \leq s^{-1/2}\left(\|h_{T_1}\|_1 + \|h_{T_2}\|_1 + \dots\right) \leq s^{-1/2}\left\|h_{T_0^c}\right\|_1$$

1.5 Extended Triangle Inequality

By definition of h_{T_i}, we can say that, $h_{(T_0\cup T_1)^c} = \sum_{j\geq 2} h_{T_j}$ Thus, obviously

$$\left\|h_{(T_0\cup T_1)^c}\right\|_2 = \left\|\sum_{j\geq 2} h_{T_j}\right\|_2$$

Also, from previous parts, we know that

$$\sum_{j\geq 2}\left\|h_{T_j}\right\|_2 \leq s^{-1/2}\left\|h_{T_0^c}\right\|_1$$

Now, by Triangle inequality, we know that-

$$\|h_{T_2} + h_{T_3} + \cdots\|_2 \leq \|h_{T_2}\|_2 + \|h_{T_2}\|_2 + \cdots$$

This is equivalent to-

$$\left\|\sum_{j\geq 2} h_{T_j}\right\|_2 \leq \sum_{j\geq 2}\left\|h_{T_j}\right\|_2$$

Combining everything, we get-

$$\left\|h_{(T_0\cup T_1)^c}\right\|_2 = \left\|\sum_{j\geq 2} h_{T_j}\right\|_2 \leq \sum_{j\geq 2}\left\|h_{T_j}\right\|_2 \leq s^{-1/2}\left\|h_{T_0^c}\right\|_1$$

1.6 Reverse Triangle Inequality, for L1 norm

Note that x^* is the minimizer of the l1 norm. Thus, if we have chosen a correct error bound, we will get that-

$$\|x\|_1 \geq \|x^*\|_1 \implies \|x\|_1 \geq \|x+h\|_1$$

Now from reverse triangle inequality, I can say that, $\|a+b\|_1 \geq \|a\|_1 - \|b_1\|$. Also I can split the L1-norm of (x+h) into terms in T_0 and those in T_{0^c}.

Thus finally, can say the following, by repeatedly applying triangle Inequality. Note that this is not the L2 norm inequality, but it works similarly, due to behaviour of the || function. We are considering (x+h) to be (x-(-h)), while applying the inequality.

$$\sum_{i \in T_0} |x_i + h_i| \geq \|x_{T_0}\|_1 - \|h_{T_0}\|_1$$

$$\sum_{i \in T_0^c} |x_i + h_i| \geq \left\|h_{T_0^c}\right\|_1 - \left\|x_{T_0^c}\right\|_1 \text{ [1]}$$

Adding these two, along with our previous results, we get-

$$\|x\|_1 \geq \|x+h\|_1 = \sum_{i \in T_0} |x_i + h_i| + \sum_{i \in T_0^c} |x_i + h_i| \geq \|x_{T_0}\|_1 - \|h_{T_0}\|_1 + \left\|h_{T_0^c}\right\|_1 - \left\|x_{T_0^c}\right\|_1$$

1.7 Rearranging

Now, we have the following-

$$\|x\|_1 \geq \|x_{T_0}\|_1 - \|h_{T_0}\|_1 + \left\|h_{T_0^c}\right\|_1 - \left\|x_{T_0^c}\right\|_1$$

$$\implies \left\|h_{T_0^c}\right\|_1 \leq \|h_{T_0}\|_1 + \|x\|_1 - \|x_{T_0}\|_1 + \left\|x_{T_0^c}\right\|_1$$

Now, note that $x_{T_0^c} = x - x_{T_0}$. Thus, using triangle inequality, we can say that-

$$\|x\|_1 - \|x_{T_0}\|_1 \leq \|x - x_{T_0}\|_1 = \left\|x_{T_0^c}\right\|_1$$

Combining the above two, we get-

$$\left\|h_{T_0^c}\right\|_1 \leq \|h_{T_0}\|_1 + \|x\|_1 - \|x_{T_0}\|_1 + \left\|x_{T_0^c}\right\|_1 \leq \|h_{T_0}\|_1 + \left\|x_{T_0^c}\right\|_1 + \left\|x_{T_0^c}\right\|_1 = \|h_{T_0}\|_1 + 2\left\|x_{T_0^c}\right\|_1$$

Hence, proved.

[1] (PS- note how the inequality was applied in the opposite sense)

2 Problem 2

(a) First of all, note that $\boldsymbol{\Phi x} = \boldsymbol{\Phi_S \widetilde{x}}$, since all other entries of $\boldsymbol{x}$ are zero, and hence all the corresponding columns of $\boldsymbol{\Phi}$ get nullified. Now, we introduce the notion of **pseudo-inverse** as mentioned in the question: For any matrix A (which can be rectangular too), there exists a unique **Moore-Penrose pseudo-inverse** $A^\dagger$ which mimics many of the properties that the actual inverse satisfies. In particular, if it so happens that the columns of A are linearly independent, then $A^\dagger A = I$. Also note that since we are assuming $\boldsymbol{\Phi}$ satisfies RIC of order S, it automatically implies that any S columns of $\boldsymbol{\Phi}$ are linearly independent. Thus hereon we can assume WLOG that $\boldsymbol{\Phi_S^\dagger \Phi_S = I}$.
Armed with this knowledge, it's very easy to determine the oracular solution: Note that while deriving any **deterministic** theoretical solution, we must ignore the noise $\boldsymbol{\eta}$ as we have no precise way of quantifying it apart from an upper bound on it's norm. Thus, $\boldsymbol{y_{\text{theoretical}}} = \boldsymbol{\Phi x} = \boldsymbol{\Phi_s \widetilde{x}}$, and thus

$$\boldsymbol{\widetilde{x}} = \boldsymbol{\Phi_S^\dagger y_{\text{theoretical}}}$$

Note that we don't actually know the value of $\boldsymbol{y_{\text{theoretical}}}$. This is a shortcoming that is fixed below.
(b) Clearly

$$\boldsymbol{y} = \boldsymbol{\Phi x} + \boldsymbol{\eta} = \boldsymbol{\Phi_S \widetilde{x}} + \boldsymbol{\eta}$$
$$\Rightarrow \boldsymbol{\Phi_S^\dagger y} = \boldsymbol{\Phi_S^\dagger \Phi_S \widetilde{x}} + \boldsymbol{\Phi_S^\dagger \eta} = \boldsymbol{\widetilde{x}} + \boldsymbol{\Phi_S^\dagger \eta}$$

Treating $\boldsymbol{\Phi_S^\dagger y}$ as the "actual" $\boldsymbol{x}$, we get that

$$\boldsymbol{x} = \boldsymbol{\widetilde{x}} + \boldsymbol{\Phi_S^\dagger \eta}$$
$$\Rightarrow \|\boldsymbol{x} - \boldsymbol{\widetilde{x}}\|_2 = \|\boldsymbol{\Phi_S^\dagger \eta}\|_2$$

Now, from the theory of **Singular Value Decomposition** we recall that the largest singular value of any matrix A could be framed as the maximum of the following optimization problem

$$\boldsymbol{\sigma_{\max}} = \|\boldsymbol{A}\|_2 = \mathbf{max}_{\|\boldsymbol{x}\|_2=1}\|\boldsymbol{Ax}\|_2$$

Then it's a direct consequence of the above maximization that $\|\boldsymbol{\Phi_S^\dagger \eta}\|_2 \leq \|\boldsymbol{\Phi_S^\dagger}\|_2 \|\boldsymbol{\eta}\|_2$, as desired.
(c) From the theory of **Singular Value Decomposition** we know that the singular values of any matrix A are the **square roots of the eigenvalues of** AA^* (note that the eigenvalues of AA^* are non-negative by the Spectral theorem since AA^* is Hermitian). Thus, to calculate the singular values of $\boldsymbol{\Phi_S^\dagger}$, we first evaluate the matrix $\boldsymbol{\Phi_S^\dagger \Phi_S^{\dagger^*}}$

$$\boldsymbol{\Phi_S^\dagger \Phi_S^{\dagger^*}} = ((\boldsymbol{\Phi_S^* \Phi_S})^{-1}\boldsymbol{\Phi_S})(\boldsymbol{\Phi_S^*}(\boldsymbol{\Phi_S^* \Phi_S})^{-1}) = (\boldsymbol{\Phi_S^* \Phi_S})^{-1}$$

Thus, $\boldsymbol{\Phi_S^\dagger \Phi_S^{\dagger^*}} = (\boldsymbol{\Phi_S^* \Phi_S})^{-1}$. Now, we utilise a linear algebra property which says that if λ is an eigenvalue of an invertible matrix A, then $\frac{1}{\lambda}$ will be an eigenvalue of the matrix A^{-1}.
Since $\boldsymbol{\Phi_S^* \Phi_S}$ is obviously invertible, the eigenvalues of $(\boldsymbol{\Phi_S^* \Phi_S})^{-1}$ are the reciprocals of the eigenvalues of $\boldsymbol{\Phi_S^* \Phi_S}$.

Now, since $\boldsymbol{\Phi}$ is RIC with order $|S| = 2k$, from the slides (`CS_Theory, Pg.101/109`) we have that $\boldsymbol{\delta_{2k}} = \max(\boldsymbol{\lambda_{\max} - 1, 1 - \lambda_{\min}})$ where $\boldsymbol{\lambda_{\max}}$ is the maximum eigenvalue of $\boldsymbol{\Phi_\Gamma \Phi_\Gamma^*}$ and $\boldsymbol{\lambda_{\min}}$ is the minimum eigenvalue of $\boldsymbol{\Phi_\Gamma \Phi_\Gamma^*}$ over all possible sets $\boldsymbol{\Gamma \subseteq [n]}$, $\boldsymbol{|\Gamma| \leq |S|}$ (note that $\boldsymbol{\lambda_{\max}}$ and $\boldsymbol{\lambda_{\min}}$ both don't necessarily come from some same set $\boldsymbol{\Gamma}$). Then, considering the fact that $\boldsymbol{\Phi_\Gamma \Phi_\Gamma^*}$ and $\boldsymbol{\Phi_\Gamma^* \Phi_\Gamma}$ have the same non-zero eigenvalues (this is again a linear algebra property: For any matrix A, AA^* and A^*A have the same non zero eigenvalues), we see that every eigenvalue of $\boldsymbol{\Phi_S^* \Phi_S}$ lies in $[\boldsymbol{1 - \delta_{2k}, 1 + \delta_{2k}}]$ (because note that S is just an instance of $\boldsymbol{\Gamma}$ and $\boldsymbol{\delta_{2k}} = \max(\boldsymbol{\lambda_{\max} - 1, 1 - \lambda_{\min}})$ implies that all eigenvalues, are first bounded by $\boldsymbol{\lambda_{\max}}$ and $\boldsymbol{\lambda_{\min}}$ which are in turn bounded by $\boldsymbol{1 - \delta_{2k}}$ and $\boldsymbol{1 + \delta_{2k}}$. Refer to the solution of Q3 for an elaboration of the same), and consequently **EVERY** singular value of $\boldsymbol{\Phi_S^\dagger}$, which as we showed above were the square roots of the reciprocals of the eigenvalues of $\boldsymbol{\Phi_S^* \Phi_S}$, must lie within $[\boldsymbol{\frac{1}{\sqrt{1+\delta_{2k}}}, \frac{1}{\sqrt{1-\delta_{2k}}}}]$, as desired.
(d) Note that Thm. 3 said that $\boldsymbol{\|x^* - x\|_2 \leq \frac{C_0}{\sqrt{S}} \|x - \widetilde{x}\|_2 + C_1 \epsilon}$. But we already know that $\boldsymbol{\|x - \widetilde{x}\|_2 \leq \frac{\epsilon}{\sqrt{1-\delta_{2k}}} \leq \frac{\epsilon}{\sqrt{2-\sqrt{2}}}}$, since $\boldsymbol{\delta_{2k} \leq \sqrt{2} - 1}$ for Thm. 3 to hold. Thus, $\boldsymbol{\|x^* - x\|_2 \leq (\frac{C_0}{\sqrt{(2-\sqrt{2})S}} + C_1)\epsilon}$, and since C_0 and C_1 are known bounded constants we obtain that (the errors in) both the solutions are some constant factors of $\boldsymbol{\epsilon}$, and consequently some constant factors of each other.

3 Problem 3

We establish a certain lemma first before proving the asked for result.

3.1 The Submatrix Lemma

Claim: Let $\boldsymbol{A} \in \mathbb{K}^{\boldsymbol{m \times n}}$ be a matrix and let $\boldsymbol{A_s} \in \mathbb{K}^{\boldsymbol{m \times s}}$ be a submatrix of $\boldsymbol{A}$. Then all the eigenvalues of $\boldsymbol{A_s^* A_s}$ are also eigenvalues of $\boldsymbol{A^* A}$.
Proof: Notice that the sampling of $\boldsymbol{A_s}$ from $\boldsymbol{A}$ can be done via a **selection matrix** $\boldsymbol{B} \in \mathbb{K}^{\boldsymbol{n \times s}}$ such that $\boldsymbol{A_s = AB}$. Notice that the columns of $\boldsymbol{B}$ are just "one-hot" columns corresponding to the columns of $\boldsymbol{A}$ which were sampled to make $\boldsymbol{A_s}$. One of the consequences of this is that $\boldsymbol{BB^* = I_n} \in \mathbb{K}^{\boldsymbol{n \times n}}$. We shall use this below.
Also, WLOG assume that $\boldsymbol{A_s}$ are the first s columns of $\boldsymbol{A}$ (if not, then we can always swap columns without affecting "global" matrix properties such as eigenvalues).
Then, note that

$$\begin{aligned}
\boldsymbol{A_s^* A_s x = \lambda x}&, \text{ where } x \neq 0 \text{ is an eigenvector of } A_s^* A_s \\
\Rightarrow \boldsymbol{(AB)^*(AB)x = \lambda x}& \\
\Rightarrow \boldsymbol{B^* A^* A B x = \lambda x}& \\
\Rightarrow \boldsymbol{B B^* A^* A B x = \lambda B x}& \\
\Rightarrow \boldsymbol{A^* A (Bx) = \lambda (Bx)}&
\end{aligned}$$

Thus, if $\boldsymbol{Bx \neq 0}$, then $\boldsymbol{\lambda}$ is an eigenvalue of $\boldsymbol{A^* A}$ with eigenvector $\boldsymbol{Bx}$. But note that (since we assumed that $\boldsymbol{A_s}$ comprised of the first s columns of $\boldsymbol{A}$) $\boldsymbol{B}$ is basically $\boldsymbol{I_s}$ padded below with $(n - s)$ zero row vectors, and consequently $\boldsymbol{Bx} = [\boldsymbol{x}^T\ \boldsymbol{0}_{n-s}^T]^T$, where $\boldsymbol{0}_{n-s} \in \mathbb{K}^{n-s}$ is the zero vector. Since $\boldsymbol{x \neq 0}$, $\boldsymbol{Bx}$, whose first s entries are identical to

$\boldsymbol{x}$'s, can't be zero either, and consequently $\boldsymbol{Bx}$ is a valid eigenvector for $\boldsymbol{A^*A}$ and hence $\boldsymbol{\lambda}$ is an eigenvalue of $\boldsymbol{A^*A}$.
But now we're almost done: As in Q2, we can say that $\boldsymbol{\delta_s} = \max(\boldsymbol{\lambda_{\max} - 1, 1 - \lambda_{\min}})$ implying that every non-zero eigenvalue of $\boldsymbol{\Phi_\Gamma^*\Phi_\Gamma}$ ($|\boldsymbol{\Gamma}| = \boldsymbol{s}$) [2] for any s column sampling of $\boldsymbol{\Phi}$ must lie within $[\boldsymbol{1 - \delta_s, 1 + \delta_s}]$. In particular, for some s-sampling, either the largest eigenvalue of $\boldsymbol{\Phi_\Gamma^*\Phi_\Gamma}$ is $(\boldsymbol{1 + \delta_s})$ or the smallest eigenvalue is $(\boldsymbol{1 - \delta_s})$. In either case, take that $\boldsymbol{\Phi_\Gamma}$ and append any $(t - s)$ new columns of $\boldsymbol{\Phi}$ to it. For the resulting $m \times t$ matrix, by the submatrix lemma we have that it either has $(\boldsymbol{1 + \delta_s})$ or $(\boldsymbol{1 - \delta_s})$ as it's eigenvalue, and once again by the RIC property (that all eigenvalues are bounded by $(1 \pm \delta)$), we either have that $(\boldsymbol{1 - \delta_s}) \geq (\boldsymbol{1 - \delta_t})$ or $(\boldsymbol{1 + \delta_s}) \leq (\boldsymbol{1 + \delta_t})$, both of which yield $\boldsymbol{\delta_s} \leq \boldsymbol{\delta_t}$, as desired.

4 Problem 4

5 Problem 5

Let

$$\boldsymbol{x^*} := \min_{\boldsymbol{x} \in \mathbb{K}^n}(\|\boldsymbol{y} - \boldsymbol{\Phi x}\|_2^2 + \boldsymbol{\lambda}\|\boldsymbol{x}\|_1)$$
$$\boldsymbol{\epsilon'} := \|\boldsymbol{y} - \boldsymbol{\Phi x^*}\|_2$$
$$\boldsymbol{A_t} := \{\boldsymbol{x_0} : \boldsymbol{x_0} \in \mathbb{K}^n;\ \|\boldsymbol{y} - \boldsymbol{\Phi x_0}\|_2 \leq \boldsymbol{t}\}$$

Thus the sample space of the problem P1 for any given $\boldsymbol{\epsilon}$ is $\boldsymbol{A_\epsilon}$. Now, choose $\boldsymbol{\epsilon} = \boldsymbol{\epsilon'}$. Thus, $\|\boldsymbol{y} - \boldsymbol{\Phi x}\|_2 \leq \boldsymbol{\epsilon'}\ \forall\ \boldsymbol{x} \in \boldsymbol{A_{\epsilon'}}$. But since the value of $\|\boldsymbol{y} - \boldsymbol{\Phi x}\|_2$ over $\boldsymbol{A_{\epsilon'}}$ is lesser than or equal to $\boldsymbol{\epsilon'}$, we have that their L1 norms should be greater than or equal to $\|\boldsymbol{x^*}\|_1$ because otherwise if there existed $\boldsymbol{x'} \in \boldsymbol{A_{\epsilon'}}$ such that $\|\boldsymbol{y} - \boldsymbol{\Phi x'}\|_2 \leq \boldsymbol{\epsilon'}$ and $\|\boldsymbol{x'}\|_1 < \|\boldsymbol{x^*}\|_1$, then $\boldsymbol{J(x')} < \boldsymbol{J(x^*)}$, contradicting the minimality of $\boldsymbol{x^*}$.
Thus $\|\boldsymbol{x'}\|_1 \geq \|\boldsymbol{x^*}\|_1\ \forall\ \boldsymbol{x'} \in \boldsymbol{A_{\epsilon'}}$ (note that $\boldsymbol{x^*} \in \boldsymbol{A_{\epsilon'}}$ too). Thus, the set of vectors with minimum L1-norms in $\boldsymbol{A_{\epsilon'}}$ includes $\boldsymbol{x^*}$, and consequently $\boldsymbol{x^*}$ is a minimizer for problem P1 too (although note that without additional information we can't say if $\boldsymbol{x^*}$ is the unique minimizer of P1 or not).

6 Problem 6

We use the linearity of expectation to find out the number of tests. Thus, the expected number of tests required for the entire population is equal to the (Number of groups) multiplied by the expected number of tests required for a single group.
Now, let $p = \frac{k}{n}$ be the probability that a randomly chosen person is infected. Then, the expected number of tests that need to be carried out can be calculated as follows:

- Suppose none of our g pool members are infected. The probability of this happening is $(1 - p)^g$, and thus the contribution of this scenario to the expected value is $(1 - p)^g \cdot 1$, ie:- after carrying out the initial pool test itself we can declare all of the pool members to be negative.

[2] once again, CS Theory actually talks of the eigenvalues of $\boldsymbol{\Phi_\Gamma\Phi_\Gamma^*}$, but since we know that the non-zero eigenvalues of $\boldsymbol{\Phi_\Gamma^*\Phi_\Gamma}$ and $\boldsymbol{\Phi_\Gamma\Phi_\Gamma^*}$ are the same, we swap the $*$ according to our wishes

- Suppose at least one of our pool members is infected. The probability of this happening is $1-(1-p)^g$, and in this case we need to carry out $(g+1)$ tests according to the Dorfman algorithm (one initial pool test, and then g individual tests). Thus, the contribution of this scenario to the expected value is $(1-(1-p)^g)\cdot(g+1)$.

Thus the total expected value of the number of tests that need to be carried out for a single pool is $(1-p)^g+(1-(1-p)^g)\cdot(g+1) = 1+g(1-(1-p)^g)$, and consequently the **number of tests required for the entire population is** $\frac{n}{g}\cdot(1+g(1-(1-p)^g)) =$ $\boldsymbol{\frac{n}{g}+n(1-(1-p)^g)}$, where $p=\frac{k}{n}$.
SHASHWAT CLARIFY: WHAT IS MEANT BY WORST CASE? WORST CASE WRT WHAT? p **OR** g**?**
For simplifying our further calculations, we shall assume that $p=\frac{k}{n} \ll 1$ is a very small number and hence the binomial approximation is applicable on $(1-p)^g$, which yields $(1-p)^g \approx 1-gp$, and consequently the expected value of the number of tests that need to be carried out becomes $\approx (\frac{n}{g}+ngp) = n\cdot(\frac{1}{g}+gp)$. Minimizing this expression w.r.t g to obtain the optimal group size yields $\frac{1}{(g^*)^2}=p \Rightarrow g^*=\frac{1}{\sqrt{p}}=\frac{1}{\sqrt{k/n}}=\sqrt{\frac{n}{k}}$, ie:- the **optimal group size is** $\boldsymbol{\sqrt{\frac{n}{k}}}$**, where** k **is the number of people infected and** n **is the population size**.