Linear Least Squares Problem Conditioning

ana.lovejoy0209

June 2020

1 Introduction

This document provides proofs for conditioning numbers and their bounds for Linear Least Squares Problem as stated in Chapter 18 of Numerical Linear Algebra by Trefethen and Bau [6]. When providing these proofs I rely heavily on work by Stewart [5]. I adapt the results from [5] to match the ones in [6] for perturbations of A and provide proofs similar to those in [5] for perturbations of b. [6] provides concise proofs for perturbations of b, while here I provide slightly different take on proving the same facts.

The document is organized as follows. Section 1 summarizes the purpose and the intention behind the document. It also presents the structure of the document. Section 2 introduces linear least squares problem and the definition of condition number. It also summarizes the results regarding the condition numbers that are analyzed here. These results are then proven in Section 3. There are four different condition numbers that are analyzed in Section 3: two condition numbers arising from the perturbations of b and their effect on the outputs x and y and two more condition numbers describing the effect of perturbations of A to x and y. The proofs regarding the perturbations of A are preceded with some preliminary theorems and lemmas with accompanying proofs. Section 4 is the Appendix. The Appendix contains some minor proofs that are used in Section 3 that might be of interest to a novice student like me. The difference between the preliminary proofs in Section 3 and the ones in the Appendix is that the former are more substantial results that are necessary to fully understand the effect of perturbations of A on the outputs whereas the latter are more common results probably familiar to a mature student of Linear Algebra.

This document was motivated by my desire to fully understand the results regarding condition numbers presented in Chapter 18 of [6]. Often the proofs presented here will be more verbose than is perhaps necessary and some minor facts will be proven in detail. This is because I wanted to understand the derivations fully so I elaborate on a lot of things in a way I explained them to myself. This comes at a cost of lack of conciseness throughout the document. My ideal was to provide proofs that go all the way 'to the bottom' so to speak to the very basics of linear algebra used in them. I do not provide proofs for everything though as that is harder than I would like but I try to at least mention why something is true so that the interested reader can follow up using the right keywords or I quote a source that provides a detailed proof.

2 Condition Numbers

2.1 Linear Least Squares Problem

Here is the statement of linear least squares problem (see [6]):

Given $A \in \mathbb{C}^{m \times n}$ of full rank, $m \geq n$, $b \in \mathbb{C}^m$, find $x \in \mathbb{C}^n$ such that $\|b - Ax\|_2$ is minimized.

The solution x and the corresponding point $y = Ax$ that is closest to b in $range(A)$ are given by:

$$x = A^+ b \qquad y = Pb \tag{1}$$

where $A^+ \in \mathbb{C}^{n \times m}$ is the pseudoinverse of A and $P = AA^+ \in \mathbb{C}^{m \times m}$ is the orthogonal projector onto $range(A)$ (see [6], Lecture 11 and Appendix, Lemma 6).

We are going to observe four condition numbers. Our inputs are A, b and solutions are x and y. We shall observe how x changes with perturbations of b and then with perturbations of A separately. We shall do the same for y.

2.2 Definition of the Relative Condition Number

The definition is taken from [6].

Definition 2.1. Relative Condition Number. Let $f : X \to Y$ be a function from a normed vector space X of data to a normed vector space Y of solutions. Let δx denote a small perturbation of x and $\delta f = f(x + \delta x) - f(x)$.

We define *relative condition number* $\kappa = \kappa(x)$ as:

$$\kappa = \lim_{\delta \to 0} \sup_{\|\delta x\| \leq \delta} \left(\frac{\|\delta f\|}{\|f(x)\|} \Big/ \frac{\|\delta x\|}{\|x\|} \right) \tag{2}$$

Relative condition number captures some properties of how the function f behaves when its input $x \in X$ is perturbed. κ depends on the input values, i.e. it is not the same across different input values.

2.3 Results

In this document we observe condition numbers for linear least squares problem. This section will introduce the results we will prove later on.

First we need to introduce three parameters:

- $\kappa(A)$ is the condition number of A:

$$\kappa(A) = \|A\|_2 \|A^+\|_2 = \sigma_1 / \sigma_n \tag{3}$$

where σ_1 is the largest singular value of A and σ_n is the smallest nonzero singular value of A.

- Angle θ as a measure of the closeness of fit:

$$\theta = cos^{-1}(\|y\|_2/\|b\|_2) \tag{4}$$

- The last parameter is η:

$$\eta = \|A\|_2\|x\|_2/\|y\|_2 = \|A\|_2\|x\|_2/\|Ax\|_2 \tag{5}$$

It should be noted that any norm could be used to define a condition number. However, in this document we use the 2-norm, same as [6].

And now onto the results:

Theorem 1. *(Theorem 18.1 from [6]) Let $b \in \mathbb{C}^m$ and $A \in \mathbb{C}^{m \times n}$ of full rank be fixed. The least squares problem has the following 2-norm relative condition numbers describing the sensitivities of y and x to perturbations in b and A:*

	y	x
b	$\dfrac{1}{\cos\theta}$	$\dfrac{\kappa(A)}{\eta\cos\theta}$
A	$\dfrac{\kappa(A)}{\cos\theta}$	$\kappa(A) + \dfrac{\kappa(A)^2\tan\theta}{\eta}$

The results in the first row are exact, being attained for certain perturbations δb, and the results in the second row are upper bounds.

3 Proofs and Derivations

3.1 Sensitivity of y to Perturbations in b

We start with the simplest of the results: relative condition number for changes in y depending on perturbations of b.

$$y + \delta y = P(b + \delta b) \tag{6}$$

We know that $y = Pb$, orthogonal projection of b into $range(A)$. We can write:

$$\delta y = Pb - P\delta b - y = P\delta b \tag{7}$$

Let us bound the ratio of the relative change of y with respect to the relative change of b:

$$\begin{aligned}\frac{\|\delta y\|_2}{\|y\|_2}\Big/\frac{\|\delta b\|_2}{\|b\|_2} &= \frac{\|P\delta b\|_2}{\|y\|_2}\Big/\frac{\|\delta b\|_2}{\|b\|_2} \\ &\le \frac{\|b\|_2}{\|y\|_2}\frac{\|P\|_2\|\delta b\|_2}{\|\delta b\|_2}\end{aligned} \tag{8}$$

Above we use the fact that $\|P\delta b\|_2 \le \|P\|_2\|\delta b\|_2$ which arises from the fact that matrix norm $\|\cdot\|_2$ is a norm induced on vector norm and the properties of induced norms (see pages 18, 19 in [6]).

P is an orthogonal projector and it only has singular values 0 and 1 (see proof of Theorem 6.1. in [6]) which means that $\|P\|_2 = 1$, since 2-norm of a matrix corresponds to its highest singular value. Hence we have:

$$\frac{\|\delta y\|_2}{\|y\|_2}\Big/\frac{\|\delta b\|_2}{\|b\|_2} \le \frac{1}{\cos\theta} \tag{9}$$

Equation (9) holds no matter how small we make $\|\delta b\|_2$. Also, no matter how small the bound δ on $\|\delta b\|_2$ is we can always choose a δb which achieves the equality in (9). The equality is achieved for any δb that is in range of P because in that case we have $\|P\delta b\|_2 = \|\delta b\|_2$:

$$\frac{\|\delta y\|_2}{\|y\|_2}\Big/\frac{\|\delta b\|_2}{\|b\|_2} = \frac{\|P\delta b\|_2}{\|y\|_2}\Big/\frac{\|\delta b\|_2}{\|b\|_2} = \frac{\|\delta b\|_2}{\|y\|_2}\Big/\frac{\|\delta b\|_2}{\|b\|_2} = \frac{1}{\cos\theta} \tag{10}$$

We can obtain a sufficiently small δb in $range(P)$ by selecting any vector a in $range(A) = range(P)$ and multiplying it by factor $\delta/\|a\|_2$. Then $\delta b = a\cdot\delta/\|a\|_2$ and $\|\delta b\|_2 = \delta \le \delta$.

Bearing all this in mind we can write:

$$\kappa_{b\to y} = \frac{1}{\cos\theta} \tag{11}$$

3.2 Sensitivity of x to Perturbations in b

First we give the bounds on relative perturbation of x depending on the perturbation of b and then we proceed to give the condition number. We observe the following:

$$x + \delta x = A^{+}(b + \delta b) \tag{12}$$

From there we get an expression for δx:

$$\delta x = A^{+}b + A^{+}\delta b - x = A^{+}\delta b \tag{13}$$

We can write the following for $\|x\|_2$:

$$\|x\|_2 = \eta\|Ax\|_2/\|A\|_2 = \eta\|y\|_2/\|A\|_2 \tag{14}$$

From equations (13) and (14) we get:

$$\frac{\|\delta x\|_2}{\|x\|_2} = \frac{\|A^+\delta b\|_2}{\eta\|y\|_2/\|A\|_2} \leq \frac{\|A^+\|_2\|\delta b\|_2\|A\|_2}{\eta\|y\|_2} = \frac{\kappa(A)\|\delta b\|_2}{\eta\|y\|_2} \tag{15}$$

In order to get a bound that can help us obtain the relative condition number (2) we divide both sides of the inequality by $\|\delta b\|_2/\|b\|_2$. We get:

$$\frac{\|\delta x\|_2}{\|x\|_2} \bigg/ \frac{\|\delta b\|_2}{\|b\|_2} \leq \frac{\kappa(A)\|b\|_2}{\eta\|y\|_2} = \frac{\kappa(A)}{\eta\cos\theta} \tag{16}$$

The above equation is true for all $\|\delta b\|_2$, no matter how small. There is also always such a value of δb for which the equality is achieved: any δb for which $\|A^+\delta b\|_2 = \|A^+\|_2\|\delta b\|_2$. We know that there exists vector v for which $\|A^+v\|_2 = \|A^+\|_2\|v\|_2$. We can choose $\delta b = v \cdot C$ where C is some sufficiently small constant, similar to what we did in Section 3.1.

Since we can find a perturbation of b for which equality in (16) is achieved no matter how small the bound on the norm of the perturbation of b is, the condition number for changes in x invoked by perturbations in b is equal to:

$$\kappa_{b\to x} = \frac{\kappa(A)}{\eta\cos\theta} \tag{17}$$

3.3 Preliminaries for Analysis of Perturbations of A

3.3.1 Convergent Matrices

The definition of convergent matrix and the related theorems and proofs are taken from an online lecture on convergent matrices [2].

Definition 3.1. Convergent Matrix Matrix $A \in \mathbb{C}^{n\times n}$ is said to be convergent if

$$\lim_{k\to\infty} A^k = 0 \tag{18}$$

Theorem 2. *Matrix $A \in \mathbb{C}^{n\times n}$ is convergent iff $|\lambda_i| < 1$, $\forall\lambda_i$ eigenvalues of A.*

Proof. Each matrix $A \in \mathbb{C}^{n\times n}$ can be written in Jordan canonical form (for proof see [1]):

$$A = S^{-1}JS \tag{19}$$

where J is a diagonal block matrix where each block looks like this:

$$J_B = \begin{pmatrix} \lambda_i & 1 & & 0 \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ 0 & & & \lambda_i \end{pmatrix} \tag{20}$$

with λ_i being an eigenvalue of A.

We will use the Jordan canonical norm to prove the theorem.

Let us first prove that if A is convergent that $|\lambda_i|< 1$ for all its eigenvalues λ_i.

We can write:

$$A^k = S^{-1}J^kS \tag{21}$$

From there we see that for A to be convergent J must be convergent. J consists of Jordan blocks and J^k consists of powers of Jordan blocks J_B^k which means that each J is convergent iff each Jordan block is convergent. We will observe a single Jordan block.

$$J_B^k = (\lambda_i \cdot I + N_l)^k \tag{22}$$

where

$$N_l = \begin{pmatrix} 0 & 1 & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ \vdots & & & 1 \\ 0 & \cdots & \cdots & 0 \end{pmatrix}_{l\times l} \tag{23}$$

N_l is a nilpotent matrix meaning $N_l^p = 0$ for some $p \leq l$ with l being number of rows and columns of the Jordan block associated with N_l.

Let us expand the expression above using the binomial theorem:

$$J_B^k = \sum_{j=0}^{k} \binom{k}{j} \lambda_i^j N_l^{k-j} = \sum_{j=k-p+1}^{k} \binom{k}{j} \lambda_i^j N_l^{k-j} \tag{24}$$

All the powers of N_l^{k-j} where $k-j \neq 0$ have 0s on the diagonal. Hence the only values on the diagonal of J_B^k are λ_i^k. We know that all the entries in J_B^k must go to 0 as $k \to \infty$, so $\lambda_i^k \to \infty$ too. This only happens if $|\lambda_i|< 1$. Thus we have proven the first direction of the theorem: if matrix A is convergent, then all for all its eigenvalues λ_i, $|\lambda_i|< 1$.

The second direction is proving that $|\lambda_i|< 1 \Rightarrow J_B$ (and thus J and A) is convergent.

We can write out the expression for J_B^k in slightly differently from (24). The following is true as well:

$$J_B^k = \sum_{j=0}^{k} \binom{k}{k-j} \lambda_i^{k-j} N_l^j = \sum_{j=0}^{p-1} \binom{k}{k-j} \lambda_i^{k-j} N_l^j \tag{25}$$

For $\lim_{k\to\infty} J_B^k = 0$ to be true it must be the case that $\binom{k}{k-j}\lambda_i^{k-j} \to 0$ for $j = 0, 1, ..., p-1$ when $k \to \infty$ because those are the sum parts where $N_l^j \neq 0$.

We can write:

$$|\begin{pmatrix} k \\ k-j \end{pmatrix} \lambda_i^{k-j}|=|\frac{k!\,\lambda^k}{(k-j)!\,j!\,\lambda^j}|\leq|\frac{k^j\lambda^k}{j!\,\lambda^j}| \tag{26}$$

The only part of the right side of the inequality above that depends on k is $|k^j\lambda^k|=k^j|\lambda|^k$. If we prove that that term goes to 0 when $k \to \infty$ we are done. We can apply the logarithm on the term:

$$\lim_{k\to\infty} \log k^j|\lambda_i|^k = \lim_{k\to\infty} j\log k + k\log|\lambda_i| = \lim_{k\to\infty} k(j\frac{\log k}{k}+\log|\lambda_i|) = -\infty \tag{27}$$

Above we used $\lim_{k\to\infty}\frac{\log k}{k}=0$ and the fact that $|\lambda_i|<1 \Rightarrow \log|\lambda_i|<0$. From (27) we infer $\lim_{k\to\infty} k^j|\lambda_i|^k=0$ for $j=0,...,p-1$. We have:

$$\forall j=0,...,p-1, \lim_{k\to\infty} k^j|\lambda_i|^k=0 \Rightarrow \forall j=0,...,p-1, \lim_{k\to\infty}|\begin{pmatrix} k \\ k-j \end{pmatrix}\lambda_i^{k-j}|=0$$
$$\Rightarrow \lim_{k\to\infty} J_B^k = 0 \tag{28}$$

This applied on all the blocks of matrix J proves it is convergent and by extension so is matrix A. This finishes the proof in the second direction.

□

Corollary 2.1. *If $\|A\|<1$ for any norm $\|\cdot\|$ induced on a vector norm then A is a convergent matrix.*

Proof. Let λ_1 be an eigenvalue of A with largest absolute value. By the definition of the eigenvalues we know that there exists a vector v such that:

$$Av=\lambda_1 v \Rightarrow \|Av\|=|\lambda_1|\|v\|\leq\|A\|\|v\|\Rightarrow\|\lambda_1\|\leq\|A\|<1 \tag{29}$$

Since λ_1 is the biggest eigenvalue in absolute value we have $|\lambda_i|<1$ for all eigenvalues λ_i of A. By Theorem 2 A is a convergent matrix. □

3.3.2 Sufficient Condition for Nonsingular Perturbations

These proofs were taken, with small adaptations from videos [2] and [3].

Lemma 1. *Let $E \in \mathbb{C}^{n\times n}$ be a matrix such that $\|E\|<1$, where $\|\cdot\|$ denotes any subordinate/induced matrix norm. Then $I+E$ is invertible and:*

1. $(I+E)^{-1}=I-E+E^2-E^3+\cdots+(-1)^kE^k+\cdots$
2. $\|(I+E)^{-1}\|\leq 1/(1-\|E\|)$
3. $\|I-(I+E)^{-1}\|\leq\frac{\|E\|}{1-\|E\|}$

Proof. $\|E\| < 1 \Rightarrow \lim_{k\to\infty} E^k = 0$ by Corollary 2.1. This in turn implies by Theorem 2 that $|\lambda_i| < 1, \forall i = 1, 2, ..., n$, where λ_i is an eigenvalue of E.

Eigenvalues of $I + E$ are $1 + \lambda_i, \forall i = 1, 2, \cdots, n$ (see Appendix, Lemma 3). We have:

$$|1 + \lambda_i| \geq 1 - |\lambda_i| > 0, \forall i = 1, 2, ..., n \tag{30}$$

Since all the eigenvalues of $I + E$ are bigger than 0, $I + E$ is non-singular.

Note that:

$$(I + E)(I - E + E^2 - \cdots + (-1)^k E^k) = I + (-1)^k E^{k+1} \tag{31}$$

Since $I + E$ is non-singular we can write:

$$I - E + \cdots + (-1)^k E^k = (I + E)^{-1} + (I + E)^{-1}(-1)^k E^{k+1} \tag{32}$$

Since E is convergent for $k \to \infty$ we have:

$$\sum_{i=0}^{\infty} (-1)^i E^i = (I + E)^{-1} \tag{33}$$

which proves *1.*

Let us observe the norm of the inverse:

$$\|(I + E)^{-1}\| \leq \sum_{i=0}^{\infty} |(-1)^i| \|E\|^i = \sum_{i=0}^{\infty} \|E\|^i = \frac{1}{1 - \|E\|} \tag{34}$$

which proves *2.* The final step was calculated using the formula for the sum of a geometric series.

Next we prove *3.*

$$(I + E)(I + E)^{-1} = I \tag{35}$$

$$I - (I + E)^{-1} = E(I + E)^{-1} \tag{36}$$

$$\|I - (I + E)^{-1}\| \leq \|E\| \|(I + E)^{-1}\| \leq \frac{\|E\|}{1 - \|E\|} \tag{37}$$

□

Theorem 3. *Let A be non singular and suppose that*

$$\|A^{-1}\| \|E\| < 1. \tag{38}$$

Then $A + E$ is nonsingular and:

$$\|(A + E)^{-1}\| \leq \frac{\|A^{-1}\|}{\gamma} \tag{39}$$

and

$$\frac{\|(A + E)^{-1} - A^{-1}\|}{\|A^{-1}\|} \leq \frac{\kappa(A)}{\gamma} \frac{\|E\|}{\|A\|} \tag{40}$$

where $\kappa(A) = \|A\|\|A^{-1}\|$ and $\gamma = 1 - \|A^{-1}\|\|E\| > 0$.

Proof. Since A is nonsingular we can write $A + E = A(I + A^{-1}E)$. By Theorem condition $\|A^{-1}E\| \leq \|A^{-1}\|\|E\| < 1$. This implies, by Lemma 1 that $(I + A^{-1}E)$ is nonsingular. Since $A + E$ is a product of two nonsingular matrices, it is also nonsigular. Let us look at the norm of $(A + E)^{-1}$:

$$\|(A + E)^{-1}\| = \|(I + A^{-1}E)^{-1}A^{-1}\| \leq \|A^{-1}\|\|(I + A^{-1}E)^{-1}\| \tag{41}$$

By Lemma 1, part *3.* we get:

$$\|(A + E)^{-1}\| \leq \frac{\|A^{-1}\|}{1 - \|A^{-1}E\|} \leq \frac{\|A^{-1}\|}{1 - \|A^{-1}\|\|E\|} = \frac{\|A^{-1}\|}{\gamma} \tag{42}$$

We have proven the first part of the Theorem. Now onto the second part:

$$(A + E)^{-1} - A^{-1} = (I + A^{-1}E)^{-1}A^{-1} - A^{-1} = ((I + A^{-1}E)^{-1} - I)A^{-1} \tag{43}$$

$$\|(A + E)^{-1} - A^{-1}\| \leq \|A^{-1}\|\|I - (I + A^{-1}E)^{-1}\| \leq \|A^{-1}\|\frac{\|A^{-1}E\|}{1 - \|A^{-1}E\|} \tag{44}$$

$$\frac{\|(A + E)^{-1} - A^{-1}\|}{\|A^{-1}\|} \leq \frac{\|A^{-1}E\|}{1 - \|A^{-1}E\|} \leq \frac{\|A^{-1}\|\|E\|}{1 - \|A^{-1}E\|} \leq \frac{\|A^{-1}\|\|E\|}{1 - \|A^{-1}\|\|E\|} \tag{45}$$

$$\frac{\|(A + E)^{-1} - A^{-1}\|}{\|A^{-1}\|} \leq \frac{\kappa(A)}{\gamma}\frac{\|E\|}{\|A\|} \tag{46}$$

□

3.3.3 Reduced Form

We are going to be using 2-norm when calculating the condition numbers. 2-norm is unitarily invariant, meaning it does not change when a vector or matrix is multiplied by a unitary matrix. To make our calculations simpler we will use a so-called reduced form of A in our proofs. In this section we introduce the necessary terminology and various relationships between A in reduced form and its perturbed version B.

We know that every matrix has a singular value decomposition:

$$A = U\Sigma V^* \tag{47}$$

where U and V are unitary matrices and Σ is a diagonal matrix with singular values $\sigma_i \geq 0$ on the diagonal. We know:

$$\|A\|_2 = \|\Sigma\|_2 = \sigma_1 \tag{48}$$

where $\sigma_1 = \max_i \sigma_i$.

The reduced form of *A* is:

$$U^*AV = \Sigma = \begin{pmatrix} A_{11} & 0 \\ 0 & 0 \end{pmatrix} \tag{49}$$

where A_{11} is a diagonal matrix with nonzero elements on the diagonal. It is the part of Σ with nonzero singular values on the diagonal.

In [6] assumption is made that A has full rank. We follow that assumption in this document. When A has **full rank reduced form of *A*** becomes:

$$U^*AV = \begin{pmatrix} A_1 \\ 0 \end{pmatrix} \tag{50}$$

Here A_1 is a diagonal matrix with all non-zero diagonal entries.

We are observing perturbations of A and the effect they have on the solutions of the least squares problem. We will have to deal with matrix $B = A+E$ where E represents the perturbations. **In the reduced form *E*** becomes:

$$U^*EV = \begin{pmatrix} E_1 \\ E_2 \end{pmatrix}. \tag{51}$$

And *B* becomes:

$$U^*BV = \begin{pmatrix} B_1 \\ B_2 \end{pmatrix} = \begin{pmatrix} A_1 + E_1 \\ E_2 \end{pmatrix}.$$

In the rest of the document we will assume A and B are already in the reduced form. Since we are working with a unitarily invariant norm, our results can be generalized to the original, non-reduced matrices.

Next we list several facts that result from the use of the reduced form of A that will be used later in the document. We write $b = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}$ where b_1 contains the first n entries of b and b_2 contains the remaining $m - n$ entries.

1. The **pseudoinverse of the reduced form of *A*** is:

$$\boldsymbol{A^+ = \begin{pmatrix} A_1^{-1} & 0 \end{pmatrix}}. \tag{52}$$

   This can easily be confirmed by verifying the four Moore-Penrose conditions for the pseudoinverse.

2. $\boldsymbol{y = \begin{pmatrix} b_1 \\ 0 \end{pmatrix}}$:

$$y = Pb = P\begin{pmatrix} b_1 \\ b_2 \end{pmatrix} = AA^+\begin{pmatrix} b_1 \\ b_2 \end{pmatrix} = \begin{pmatrix} I_{n\times n} & 0 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} b_1 \\ b_2 \end{pmatrix} = \begin{pmatrix} b_1 \\ 0 \end{pmatrix} \tag{53}$$

3. $\|\boldsymbol{A^+}\|_{\mathbf{2}}\boldsymbol{=}\|\boldsymbol{A^{-1}}\|_{\mathbf{2}}$. Let us prove this. We write $x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$:

$$A^+x = \begin{pmatrix} A_1^{-1} & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = A_1^{-1}x_1 \tag{54}$$

$$\frac{\|A^+x\|_2}{\|x\|_2} = \frac{\|A_1^{-1}x_1\|_2}{\sqrt{\|x_1\|_2^2 + \|x_2\|_2^2}} \leq \frac{\|A_1^{-1}x_1\|_2}{\|x_1\|_2} \leq \|A_1^{-1}\|_2 \tag{55}$$

Since (55) holds for all x we have $\|A^+\|_2 \leq \|A_1^{-1}\|_2$.

We also have:

$$\|A_1^{-1}\|_2 = \max_{\|x\|_2=1}\|A_1^{-1}x\|_2 = \max_{\|x\|_2=1}\|A^+ \begin{pmatrix} x \\ 0 \end{pmatrix} \|_2 \leq \max_{\|y\|_2=1}\|A^+y\|_2 = \|A^+\|_2 \tag{56}$$

Which leads us to:

$$\begin{aligned} &\|A^{-1}\|_2 \leq \|A^+\|_2 \leq \|A^{-1}\|_2 \Rightarrow \\ &\qquad\qquad \Rightarrow \|A^+\|_2 = \|A^{-1}\|_2 \end{aligned} \tag{57}$$

4.

$$\boldsymbol{b_1 = A_1 x} \qquad \boldsymbol{x = A_1^{-1} b_1} \tag{58}$$

Proof:

$$\begin{aligned} &Ax = \begin{pmatrix} A_1 \\ 0 \end{pmatrix} x = \begin{pmatrix} A_1 x \\ 0 \end{pmatrix} = y = \begin{pmatrix} b_1 \\ 0 \end{pmatrix} \\ &\Rightarrow A_1 x = b_1 \Rightarrow x = A_1^{-1} b_1 \end{aligned} \tag{59}$$

3.3.4 Bounds on Error Components

Here we give some bounds regarding the error term E and its components. We can write $E_1 = \begin{pmatrix} I_n & 0 \end{pmatrix} E$ and $E_2 = \begin{pmatrix} 0 & I_{m-n} \end{pmatrix} E$. From there we get:

$$\|E_1\|_2 \leq \| \begin{pmatrix} I_n & 0 \end{pmatrix} \|_2 \|E\|_2 = \|E\|_2 \tag{60}$$

and similarly:

$$\|E_2\|_2 \leq \|E\|_2. \tag{61}$$

We use the fact that $\| \begin{pmatrix} I_n & 0 \end{pmatrix} \|_2 = \| \begin{pmatrix} 0 & I_{m-n} \end{pmatrix} \|_2 = 1$. Here is a quick proof:

We know that $\begin{pmatrix} I_n & 0 \end{pmatrix} \begin{pmatrix} b_{1(n\times1)} \\ 0 \end{pmatrix} = b_1$ and that $\| \begin{pmatrix} b_1 \\ 0 \end{pmatrix} \|_2 = \|b_1\|_2$. Therefore we have $\| \begin{pmatrix} I_n & 0 \end{pmatrix} \|_2 \geq 1$. We also know that $\begin{pmatrix} I_n & 0 \end{pmatrix} \begin{pmatrix} b_{1(n\times1)} \\ b_{2(m-n\times1)} \end{pmatrix} = b_1$ which means $\forall b$, $\| \begin{pmatrix} I_n & 0 \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} \|_2 = \|b_1\|_2 \leq \|b\|_2$, so we have $\| \begin{pmatrix} I_n & 0 \end{pmatrix} \|_2 \leq 1$. Tying it together we get $\| \begin{pmatrix} I_n & 0 \end{pmatrix} \|_2 = 1$. Similarly we can prove $\| \begin{pmatrix} 0 & I_{m-n} \end{pmatrix} \|_2 = 1$.

3.3.5 Pseudoinverse of $\begin{pmatrix} I \\ F \end{pmatrix}$

$\begin{pmatrix} I \\ F \end{pmatrix}_{m \times n}$ has full rank because it contains an identity matrix $I_{n \times n}$. From [6] we know that the pseudoinverse of a matrix $A \in \mathbb{C}^{m \times n}$, $m \geq n$, of full rank is $A^+ = (A^*A)^{-1}A^*$. This can also be proven by verifying Moore-Penrose conditions for the pseudoinverse. By applying this to $\begin{pmatrix} I \\ F \end{pmatrix}$ we get:

$$\begin{pmatrix} I \\ F \end{pmatrix}^+ = (I + F^*F)^{-1} \begin{pmatrix} I & F^* \end{pmatrix} \tag{62}$$

3.3.6 B^+ Decomposition

Theorem 4. *Let $B = A + E$ where $A \in \mathbb{C}^{m \times n}$ is a full rank matrix and $m > n$. Let E be such that $\|A^+\| \|E_1\| < 1$. Then*

$$B^+ = B_1^{-1} \begin{pmatrix} I \\ F_2 \end{pmatrix}^+$$

where $F_2 = E_2 B_1^{-1}$.

Proof.

$$B = A + E = \begin{pmatrix} A_1 \\ 0 \end{pmatrix} + \begin{pmatrix} E_1 \\ E_2 \end{pmatrix} = \begin{pmatrix} B_1 \\ E_2 \end{pmatrix}$$

where $B_1 = A_1 + E_1$.

By Theorem 3 and condition $\|A^+\|_2 \|E_1\|_2 = \|A_1^{-1}\|_2 \|E_1\|_2 < 1$ (see expression (57)) B_1 is nonsingular so we can write:

$$B = \begin{pmatrix} I \\ E_2 B_1^{-1} \end{pmatrix} B_1 = \begin{pmatrix} I \\ F_2 \end{pmatrix} B_1$$

We can prove that $B^+ = B_1^{-1} \begin{pmatrix} I \\ F_2 \end{pmatrix}^+$ is pseudoinverse of B by proving it satisfies the four Moore-Penrose conditions for pseudoinverse. In the calculations below we use expression (62) for $\begin{pmatrix} I \\ F_2 \end{pmatrix}^+$ in part 4.

1. $B^+BB^+ = B_1^{-1} \begin{pmatrix} I \\ F_2 \end{pmatrix}^+ \begin{pmatrix} I \\ F_2 \end{pmatrix} B_1 B_1^{-1} \begin{pmatrix} I \\ F_2 \end{pmatrix}^+ = B_1^{-1} \begin{pmatrix} I \\ F_2 \end{pmatrix}^+ = B^+$

2. $BB^+B = \begin{pmatrix} I \\ F_2 \end{pmatrix} B_1 B_1^{-1} \begin{pmatrix} I \\ F_2 \end{pmatrix}^+ \begin{pmatrix} I \\ F_2 \end{pmatrix} B_1 = \begin{pmatrix} I \\ F_2 \end{pmatrix} B_1 = B$

3. $(BB^+)^* = (\begin{pmatrix} I \\ F_2 \end{pmatrix} B_1 B_1^{-1} \begin{pmatrix} I \\ F_2 \end{pmatrix}^+)^H = \begin{pmatrix} I \\ F_2 \end{pmatrix} \begin{pmatrix} I \\ F_2 \end{pmatrix}^+ =$
   $\begin{pmatrix} I \\ F_2 \end{pmatrix} B_1 B_1^{-1} \begin{pmatrix} I \\ F_2 \end{pmatrix}^+ = BB^+$

4. $(B^+B)^* = (B_1^{-1}\begin{pmatrix} I \\ F_2 \end{pmatrix}^+ \begin{pmatrix} I \\ F_2 \end{pmatrix} B_1)^* =$
$= (B_1^{-1}(I + F_2^* F_2)^{-1} \begin{pmatrix} I & F_2^* \end{pmatrix} \begin{pmatrix} I \\ F_2 \end{pmatrix} B_1)^* =$
$= (B_1^{-1}(I + F_2^* F_2)^{-1}(I + F_2^* F_2)B_1)^* =$
$= I^* = I = B^+B$

□

3.4 Sensitivity of y to Perturbations in A

We can write:

$$y + \delta y = P_B b \rightarrow \delta y = (P_B - P_A)b \tag{63}$$

where P_A is an orthogonal projector to the column space of A and P_B is an orthogonal projector to the column space of B.

$$\|\delta y\|_2 = \|(P_B - P_A)b\|_2 \leq \|P_B - P_A\|_2 \|b\|_2 = \|P_B - P_A\|_2 \frac{\|y\|_2}{\cos\theta} \tag{64}$$

Let us find a bound for $\|P_B - P_A\|_2$ now. The derivation of the bound was taken from [5].

We are going to limit our analysis to perturbations for which $\|A^{-1}E\|_2 < 1$. We will be able to use these results to bound the condition number because the condition number is defined in the limit, when $\|E\|_2 \rightarrow 0$ and thus $\|A^{-1}E\|_2$ goes to 0. We also know that $\|A^{-1}E\|_2 < 1 \Rightarrow \|A^{-1}E_1\|_2 < 1$ from Section 3.3.4.

We have proven that when $\|A^{-1}E_1\|_2 < 1$ that $B_1 = A_1 + E_1$ is nonsingular (Theorem 3). Thus we can write:

$$B = \begin{pmatrix} B_1 \\ E_2 \end{pmatrix} = \begin{pmatrix} I \\ E_2 B_1^{-1} \end{pmatrix} B_1 = \begin{pmatrix} I \\ F_2 \end{pmatrix} B_1 \tag{65}$$

with $F_2 = E_2 B_1^{-1}$.

We use the fact that an orthogonal projector onto a column space of a matrix B is $P_B = BB^+$ where B^+ is the pseudoinverse of B (see [5], Introduction and Lemma 6 in the Appendix).

We have:

$$\begin{aligned} P_B = BB^+ &= \begin{pmatrix} I \\ F_2 \end{pmatrix} B_1 B_1^{-1} \begin{pmatrix} I \\ F_2 \end{pmatrix}^+ \\ &= \begin{pmatrix} I \\ F_2 \end{pmatrix} (I + F_2^* F_2)^{-1} \begin{pmatrix} I & F_2^* \end{pmatrix} \end{aligned} \tag{66}$$

We are using the reduced form of A so we have:

$$P_A = AA^+ = \begin{pmatrix} A_1 \\ 0 \end{pmatrix} \begin{pmatrix} A_1^{-1} & 0 \end{pmatrix} = \begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix} \tag{67}$$

We have:

$$P_B - P_A = \begin{pmatrix} (I+F_2^*F_2)^{-1} - I & (I+F_2^*F_2)^{-1}F_2^* \\ F_2(I+F_2^*F_2)^{-1} & F_2(I+F_2^*F_2)^{-1}F_2^* \end{pmatrix} \tag{68}$$

We know that singular values $\sigma_i(P_B - P_A)$ squared are equal to eigenvalues of $(P_B - P_A)^*(P_B - P_A) = (P_B - P_A)^2$.

In order to get the expression for $(P_B-P_A)^2$ in (71) we used the substitutions (69) and (70) as needed:

$$(I+F_2^*F_2)^{-1} - I = (I - (I+F_2^*F_2))(I+F_2^*F_2)^{-1} = -F_2^*F_2(I+F_2^*F_2)^{-1} \tag{69}$$

$$(I+F_2^*F_2)^{-1} - I = (I+F_2^*F_2)^{-1}(I - I - F_2^*F_2) = -(I+F_2^*F_2)^{-1}F_2^*F_2 \tag{70}$$

$$(P_B - P_A)^2 = \begin{pmatrix} F_2^*F_2(I+F_2^*F_2)^{-1} & 0 \\ 0 & F_2(I+F_2^*F_2)^{-1}F_2^* \end{pmatrix} \tag{71}$$

Since $(P_B - P_A)^2$ is Hermitian, its singular values are equal to the absolute values of its eigenvalues (see Real Spectral Theorem, 7.29 in [1], and Theorem 5.5 in [6]). In this case, all the eigenvalues are non-negative because nonzero eigenvalues of $(P_B - P_A)^2$ are squares of singular values of $P_B - P_A$ meaning singular values and eigenvalues of $(P_B - P_A)^2$ are the same.

Let us now look at the singular values of $(P_B - P_A)^2$. By Lemma 2 we know that singular values of a block diagonal matrix are equal to the singular values of the blocks. We can thus observe the singular values of the two blocks of $(P_B - P_A)^2$: $F_2^*F_2(I+F_2^*F_2)^{-1}$ and $F_2(I+F_2^*F_2)^{-1}F_2^*$. Below we give an expression for the largest singular value of $F_2^*F_2(I+F_2^*F_2)^{-1}$:

$$\begin{aligned} \sigma_1(F_2^*F_2(I+F_2^*F_2)^{-1}) &= \|F_2^*F_2(I+F_2^*F_2)^{-1}\|_2 \\ &\leq \|F_2^*\|_2\|F_2\|_2\|(I+F_2^*F_2)^{-1}\|_2 \\ &= \frac{\sigma_1(F_2)^2}{1+\sigma_{m-n}(F_2)^2} \end{aligned} \tag{72}$$

with $\sigma_1(F_2)$ being the largest singular value of F_2 and $\sigma_{m-n}(F_2)$ being the smallest singular value of F_2.

We obtained (72) by using $\|(I+F_2^*F_2)^{-1}\|_2 = \frac{1}{1+\sigma_{m-n}(F_2)^2}$ which we derive below.

Since $(I+F_2^*F_2)^* = I+F_2^*F_2$, $I+F_2^*F_2$ is Hermitian and so is its inverse. This means that for both $I+F_2^*F_2$ and $(I+F_2^*F_2)^{-1}$ their singular values are equal to the absolute values of their eigenvalues (see [6], Theorem 5.5). Since $F_2^*F_2$ is a positive semi-definite matrix, all of its eigenvalues are non-negative and thus all the eigenvalues of $I+F_2^*F_2$ (see Lemma 3) are non-negative too. Since eigenvalues of $I+F_2^*F_2$ are nonnegative and equal in absolute value to its singular values we have the equality of singular values and eigenvalues. By Lemma 4 we can conclude that all the eigenvalues of $(I+F_2^*F_2)^{-1}$ are also non-negative and thus equal to the singular values of $(I + F_2^* F_2)^{-1}$. By the equality of singular values to eigenvalues and Lemma 4 we get:

$$\sigma_1((I + F_2^* F_2)^{-1}) = \frac{1}{\sigma_{m-n}(I + F_2^* F_2)} \tag{73}$$

where σ_{m-n} is the smallest singular value of $I + F_2^* F_2$.

We know that eigenvalues of $F_2^* F_2$ equal square of singular values of F_2. By that fact and Lemma 3 we get that eigenvalues of $I + F_2^* F_2$ are $1 + \sigma_i(F_2)^2$. By equality of singular values and eigenvalues of $I + F_2^* F_2$ we know that the smallest singular value of $I + F_2^* F_2$ is $1 + \sigma_{m-n}(F_2)^2$. By applying this to the expression (73) we get:

$$\sigma_1((I + F_2^* F_2)^{-1}) = \|(I + F_2^* F_2)^{-1}\|_2 = \frac{1}{1 + \sigma_{m-n}(F_2)^2} \tag{74}$$

which is what we use in equation (72).

Similarly we have:

$$\begin{aligned} \sigma_1(F_2(I + F_2^* F_2)^{-1} F_2^*) &= \|F_2(I + F_2^* F_2)^{-1} F_2^*\|_2 \\ &\leq \|F_2\|_2 \|(I + F_2^* F_2)^{-1}\|_2 \|F_2^*\|_2 \\ &= \frac{\sigma_1(F_2)^2}{1 + \sigma_{m-n}(F_2)^2} \end{aligned} \tag{75}$$

We know that:

$$\begin{aligned} \sigma_1((P_B - P_A)^2) &= \max(\sigma_1(F_2^* F_2(I + F_2^* F_2)^{-1}), \sigma_1(F_2(I + F_2^* F_2)^{-1} F_2^*)) \\ &\leq \frac{\sigma_1(F_2)^2}{1 + \sigma_{m-n}(F_2)^2} \leq \sigma_1(F_2)^2 \end{aligned} \tag{76}$$

From there we get:

$$\|P_B - P_A\|_2 = \sigma_1(P_B - P_A) = \sqrt{\sigma_1((P_B - P_A)^2)} \leq \sqrt{\sigma_1(F_2)^2} = \sigma_1(F_2) \tag{77}$$

We have:

$$\sigma_1(F_2) = \|F_2\|_2 = \|E_2 B_1^{-1}\|_2 \leq \|B_1^{-1}\|_2 \|E_2\|_2 \tag{78}$$

By combining expressions (77) and (78) and using Theorem 3 we get:

$$\|P_B - P_A\|_2 \leq \frac{\|A_1^{-1}\|_2}{\gamma} \|E_2\|_2 = \frac{\kappa(A)}{\gamma} \frac{\|E_2\|_2}{\|A\|_2} \tag{79}$$

with $\gamma = 1 - \|A^+\|_2 \|E_1\|_2 = 1 - \|A_1^{-1}\|_2 \|E_1\|_2$ and $\kappa(A) = \|A^+\|_2 \|A\|_2 = \|A_1^{-1}\|_2 \|A\|_2$.

Now by putting (79) and (64) together and by the fact that $\|E_2\|_2 \leq \|E\|_2$ we get:

$$\frac{\|\delta y\|_2}{\|y\|_2} \Big/ \frac{\|E\|_2}{\|A\|_2} \leq \frac{\kappa(A)}{\gamma} \frac{1}{\cos\theta} \tag{80}$$

We can use this to bound the condition number. $\gamma \to 1$ when $\|E\| \to 0$ so we have:

$$\kappa_{A\to y} = \lim_{\delta\to 0} \sup_{\|E\|_2 \leq \delta} \frac{\|\delta y\|_2}{\|y\|_2} \Big/ \frac{\|E\|_2}{\|A\|_2} \leq \lim_{\delta\to 0} \sup_{\|E\|_2 \leq \delta} \frac{\kappa(A)}{\gamma} \frac{1}{\cos\theta} = \frac{\kappa(A)}{\cos\theta} \tag{81}$$

We have thus obtained the upper bound for $\kappa_{A\to y}$: $\frac{\kappa(A)}{\cos\theta}$.

3.5 Sensitivity of x to Perturbations in A

Let us have $x = A^+ b$ and $x + \delta x = B^+ b$ where $B = A + E$ and E is chosen so $\|A^+\|\|E\| < 1$. First we will prove that we can bound the relative perturbation of x like this:

$$\frac{\|\delta x\|_2}{\|x\|_2} \leq \frac{\kappa(A)}{\gamma} \frac{\|E_1\|_2}{\|A\|_2} + \frac{\kappa(A)^2}{\gamma^2} \frac{\|E_2\|_2}{\|A\|_2} \left(\frac{1}{\eta} \frac{\|b_2\|_2}{\|b_1\|_2} + \frac{1}{\gamma} \frac{\|E_2\|_2}{\|A\|_2} \right) \tag{82}$$

where $\kappa(A) = \|A^+\|_2 \|A\|_2$, $\gamma = 1 - \|A^+\|_2 \|E_1\|_2$ and $\eta = \|A\|_2 \|x\|_2 / \|Ax\|_2$. The proof and the claim are taken and slightly adapted from [5].

Since our end goal is to obtain a conditioning number we are focusing on the behaviour of the output for small perturbations of the input (condition $\|A^+\|_2 \|E\|_2 < 1$). Because of this condition we have $\|A^+\|_2 \|E_1\|_2 \leq 1$ (section 3.3.4) so we can use the derivation of B^+ from Theorem 4.

Let us rewrite δx as:

$$\delta x = B^+ b - x = B^+ b - A^+ b = B_1^{-1} \begin{pmatrix} I \\ F_2 \end{pmatrix}^+ b - A_1^{-1} b_1 \tag{83}$$

$$\begin{aligned} \delta x &= B_1^{-1} \begin{pmatrix} I \\ F_2 \end{pmatrix}^+ b - B_1^{-1} b_1 + B_1^{-1} b_1 - A_1^{-1} b_1 \\ &= B_1^{-1} \begin{pmatrix} I \\ F_2 \end{pmatrix}^+ b - B_1^{-1} \begin{pmatrix} I & 0 \end{pmatrix} b + (B_1^{-1} - A_1^{-1}) b_1 \\ &= B_1^{-1} (\begin{pmatrix} I \\ F_2 \end{pmatrix}^+ - \begin{pmatrix} I & 0 \end{pmatrix}) b + (B_1^{-1} - A_1^{-1}) b_1 \end{aligned} \tag{84}$$

Let us observe the two terms separately. By Theorem 3 we have the following expression for the inverse of B_1:

$$B_1^{-1} = (I - A_1^{-1} E) A_1^{-1} \tag{85}$$

We can write:

$$\begin{aligned}\|(B_1^{-1}-A_1^{-1})b_1\|_2 &= \|((I-A_1^{-1}E)^{-1}A_1^{-1}-A_1^{-1})A_1x\|_2\\ &= \|((I-A_1^{-1}E)^{-1}-I)x\|_2\\ &\le \|((I-A_1^{-1}E)^{-1}-I)\|_2\|x\|_2\end{aligned} \tag{86}$$

By Lemma 1 this translates to:

$$\|(B_1^{-1}-A_1^{-1})b_1\|_2 \le \frac{\|A_1^{-1}E_1\|_2}{1-\|A_1^{-1}E_1\|_2}\|x\|_2 \le \frac{\|A_1^{-1}\|_2\|E_1\|_2}{1-\|A_1^{-1}\|_2\|E_1\|_2}\|x\|_2 \tag{87}$$

We have $\|A^+\|_2 = \|A^{-1}\|_2$ (see section 3.3.3 for proof) so we can write:

$$\|(B_1^{-1}-A_1^{-1})b_1\|_2 \le \frac{\kappa(A)}{\gamma}\frac{\|E_1\|_2}{\|A\|_2}\|x\|_2 \tag{88}$$

The first term is more complicated. We will split it up further in two parts:

$$\begin{aligned}B_1^{-1}(\begin{pmatrix} I \\ F_2\end{pmatrix}^+ - \begin{pmatrix} I & 0\end{pmatrix})b &= B_1^{-1}((I+F_2^*F_2)^{-1}\begin{pmatrix} I & F_2^*\end{pmatrix} - \begin{pmatrix} I & 0\end{pmatrix})\begin{pmatrix} b_1 \\ b_2\end{pmatrix}\\ &= B_1^{-1}((I+F_2^*F_2)^{-1}(b_1+F_2^*b_2)-b_1)\\ &= B_1^{-1}((I+F_2^*F_2)^{-1}-I)b_1 + B_1^{-1}(I+F_2^*F_2)^{-1}F_2^*b_2\end{aligned} \tag{89}$$

Let us observe the first part of (89):

$$\|B_1^{-1}((I+F_2^*F_2)^{-1}-I)b_1\|_2 = \|B_1^{-1}(I+F_2^*F_2)^{-1}F_2^*F_2b_1\|_2 \tag{90}$$

From expression (74) we know that $\|(1+F_2^*F_2)^{-1}\|_2 = \frac{1}{1+\sigma_{m-n}(F_2)^2} \le 1$ and from (58) we have $x = A_1^{-1}b_1$ so we can write:

$$\begin{aligned}\|B_1^{-1}((I+F_2^*F_2)^{-1}-I)b_1\|_2 &\le \|B_1^{-1}\|_2\|F_2\|_2\|F_2b_1\|_2\\ &\le \|B_1^{-1}\|_2^2\|E_2\|_2\|E_2B_1^{-1}b_1\|_2\\ &\le \|B_1^{-1}\|_2^2\|E_2\|_2^2\|(I+A_1^{-1}E_1)^{-1}A_1^{-1}b_1\|_2\\ &\le \|B_1^{-1}\|_2^2\|E_2\|_2^2\|(I+A_1^{-1}E_1)^{-1}\|_2\|x\|_2\end{aligned} \tag{91}$$

By Lemma 1 and Theorem 3 we can further write:

$$\begin{aligned}\|B_1^{-1}((I+F_2^*F_2)^{-1}-I)b_1\|_2 &\le \|A_1^{-1}\|_2^2\|E_2\|_2^2\|x\|_2/\gamma^3\\ &= \frac{\kappa(A)^2}{\gamma^3}\frac{\|E_2\|_2^2}{\|A\|_2^2}\|x\|_2\end{aligned} \tag{92}$$

Onto the second term of (89) where we use some of the same substitutions:

$$\begin{aligned}\|B_1^{-1}(I+F_2^*F_2)^{-1}F_2^*b_2\|_2 &\leq \frac{\|A_1^{-1}\|_2^2}{\gamma^2}\|E_2\|_2\|b_2\|_2 \\ &= \frac{\kappa(A)^2}{\gamma^2}\frac{\|E_2\|_2}{\|A\|_2^2}\|b_2\|_2\frac{1}{\eta}\frac{\|A\|_2\|x\|_2}{\|b_1\|_2} \\ &= \frac{1}{\eta}\frac{\kappa(A)^2}{\gamma^2}\frac{\|b_2\|_2}{\|b_1\|_2}\|x\|_2\frac{\|E_2\|_2}{\|A\|_2}\end{aligned} \tag{93}$$

Putting together expressions in (88), (92) and (93) we get the expression (82).

Now we shall adapt the expression (82) to get an upper bound for condition number for changes in x given perturbations of A.

By using the fact that $\|E_1\|_2\leq \|E\|_2$ and $\|E_2\|_2\leq \|E\|_2$ as well as making the substitution $\tan\theta = \|b_2\|_2/\|b_1\|_2$ in expression (82) we get the upper bound:

$$\frac{\|\delta x\|_2}{\|x\|_2}\Big/\frac{\|E\|_2}{\|A\|_2} \leq \frac{\kappa(A)}{\gamma} + \frac{\kappa(A)^2}{\gamma^2}(\frac{1}{\gamma}\frac{\|E\|_2}{\|A\|_2} + \frac{1}{\eta}\tan\theta) \tag{94}$$

We get the bound for the condition number $\kappa_{A\to x}$ by observing the bound (94) for small $\|E\|_2$:

$$\begin{aligned}\kappa_{A\to x} &= \lim_{\delta\to 0}\sup_{\|E\|_2\leq\delta}\frac{\|\delta x\|_2}{\|x\|_2}\Big/\frac{\|E\|_2}{\|A\|_2} \\ &\leq \lim_{\delta\to 0}\sup_{\|E\|_2\leq\delta}(\frac{\kappa(A)}{\gamma} + \frac{\kappa(A)^2}{\gamma^2}(\frac{1}{\gamma}\frac{\|E\|_2}{\|A\|_2} + \frac{1}{\eta}\tan\theta)) \\ &= \kappa(A) + \frac{\kappa(A)^2}{\eta}\tan\theta\end{aligned} \tag{95}$$

We get the final sum in (95) because $\gamma \to 0$ when $\|E\|_2\to 0$. We also lose the first part of the second term because it contains the factor $\|E\|_2\to 0$.

4 Appendix

Lemma 2. *Singular values of a block diagonal matrix are equal to the singular values of individual blocks.*

Proof. Let $A = \begin{pmatrix} B_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & B_k \end{pmatrix}$ with B_i being matrix blocks and $B_i = U_i\Sigma_i V_i^*$ being a singular value decomposition of each block. We can write:

$$
\begin{aligned}
A &= \begin{pmatrix} U_1\Sigma_1V_1^* & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & U_k\Sigma_kV_k^* \end{pmatrix} \\
&= \begin{pmatrix} U_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & U_k \end{pmatrix} \begin{pmatrix} \Sigma_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \Sigma_k \end{pmatrix} \begin{pmatrix} V_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & V_k \end{pmatrix}^*
\end{aligned} \tag{96}
$$

By setting $U = \begin{pmatrix} U_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & U_k \end{pmatrix}$, $\Sigma = \begin{pmatrix} \Sigma_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \Sigma_k \end{pmatrix}$ and $V = \begin{pmatrix} V_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & V_k \end{pmatrix}$ we get a singular value decomposition of $A = U\Sigma V^*$. We can do this because U and V are unitary matrices. We would also need to order singular values in descending order and reorder the columns of U and V accordingly. From this decomposition we see that the singular values of A are equal to the singular values of individual blocks B_i.

□

Lemma 3. *Let A be a matrix of dimensions $n \times n$. For any $\mu \in \mathbb{C}$, $\lambda_i + \mu$ is an eigenvalue of $A + \mu I$ if and only if λ_i is an eigenvalue of A.*

Proof. Let us look at the eigenvalues of A. We have $Av_i = \lambda_i v_i$. From there we get $(A + \mu I)v_i = Av_i + \mu v = (\lambda_i + \mu)v$ for every eigenvalue λ_i-eigenvector v_i pair of A. So we have proven that if λ_i is an eigenvalue of A then $\lambda_i + \mu$ is an eigenvalue of $A + \mu I$.

Now onto the other direction. Let us observe the eigenvalues of $A + \mu I$. We have $(A + \mu I)v_i = \delta_i v_i$ for every eigenvalue δ_i-eigenvector v_i pair of $A + \mu I$. We can write $Av_i = \delta_i v_i - \mu v_i = (\delta_i - \mu)v_i$. We see that for every eigenvalue δ_i of $A + \mu I$ we have an eigenvalue of A that is equal to $\delta_i - \mu$. If we name that eigenvalue $\lambda_i = \delta_i - \mu$ we have $\delta_i = \lambda_i + \mu$ and we have thus proven the second direction of the iff statement.

□

Lemma 4. *Let A be a nonsingular matrix. λ_i is an eigenvalue of A if and only if $1/\lambda_i$ is an eigenvalue of A^{-1}.*

Proof. For all eigenvalues of A we have $Av_i = \lambda_i v_i$. If we multiply both sides of the equation by A^{-1} we get $v_i = \lambda_i A^{-1} v_i$. We can rewrite that as $A^{-1}v_i = (1/\lambda_i)v_i$. From here we see that if λ_i is an eigenvalue of A then $1/\lambda_i$ is an eigenvalue of A^{-1}. Since A is an inverse of A^{-1} this implies that for every eigenvalue δ_i of A^{-1} there exist an eigenvalue $1/\delta_i$ of A. If we set $\lambda_i = 1/\delta_i$ we get $\delta_i = 1/\lambda_i$. Thus we have proven that for every eigenvalue $1/\lambda_i$ of A^{-1} there exists an eigenvalue λ_i of A. We have now proven both directions of the iff statement. □

Lemma 5. *Let P_1 and P_2 be orthogonal projectors onto some subspace $S \subseteq \mathbb{C}^{n\times n}$. Then, $P_1 = P_2$.*

Proof. The idea for the proof was taken from [4]. We have:

$$\forall x \quad P_1 x = x_S, P_2 x = x_S \tag{97}$$

where x_S is an orthogonal projection of vector x onto subspace S. We can write:

$$(P_1 - P_2)x = x_S - x_S = 0 \qquad \forall x \in \mathbb{C}^n \tag{98}$$

We want to prove that (98) implies that $P_1 - P_2 = 0$. Let us assume $P_1 - P_2 \neq 0$. This means that there is a nonzero column in $P_1 - P_2$. We now pick any $x \in \mathbb{C}^n$. We have $(P_1 - P_2)x = 0$. Let us now observe y that is equal to x for all entries except for an entry i for which i-th column of $P_1 - P_2$ is nonzero. Let us set that entry to $y_i = x_i + 1$. Let us denote i-th column of $P_1 - P_2$ as $(P_1 - P_2)_i$. We get:

$$(P_1 - P_2)y = (P_1 - P_2)x + (P_1 - P_2)_i = (P_1 - P_2)_i \neq 0 \tag{99}$$

Since $(P_1 - P_2)y$ must be 0, we got contradiction. Hence all the columns of $(P_1 - P_2)$ must be 0 and $P_1 - P_2$ must be a zero matrix. Therefore $P_1 = P_2$ □

Lemma 6. *Let A be a matrix and A^+ its pseudoinverse. Then $P_A = AA^+$ is a unique orthogonal projector onto the range of A.*

Proof. First we will prove that $P_A = AA^+$ is a projector:

$$P_A P_A = AA^+AA^+ = AA^+ = P_A \tag{100}$$

We used one of the Moore-Penrose conditions for pseudoinverses above ($AA^+A = A$).

Necessary and sufficient conditon for a projector P_A to be an orthogonal projector is $P_A^* = P_A$ (see [6], Lecture 6):

$$P_A^* = (AA^+)^* = AA^+ = P_A \tag{101}$$

Here we used another Moore-Penrose conditon for pseudoinverses: $(AA^+)^* = AA^+$.

We know that the every vector v in the $range(P_A)$ is also in the range of A:

$$v = P_A x = AA^+x = Ay, \; y = A^+x \tag{102}$$

We can also show that every vector in the range of A is also in the range of P_A. For that we use the previously mentioned Moore-Penrose condition for pseudoinverses: $AA^+A = A$.

$$P_A A = AA^+A = A \Rightarrow \forall a_i, \; P_A a_i = a_i \tag{103}$$

where a_i are columns of A. Since each vector in the range of A can be expressed using columns of P_A and vice versa we have $range(A) = range(P_A)$ so P_A is a projector onto the column space of A.

So far we have proven that $P_A = AA^+$ is an orthogonal projector onto $range(A)$. By Lemma 5 we know that it is also unique.

□

References

[1] Sheldon Jay Axler. *Linear Algebra Done Right.* Undergraduate Texts in Mathematics. Springer, New York, 1997.

[2] D.N. Pandey. Convergent matrices - i, 2018. Available at https://youtu.be/ucAIzQm9˙kg.

[3] D.N. Pandey. Convergent matrices - ii, 2018. Available at https://youtu.be/WFEJPkyzZ88.

[4] rschwieb. Uniqueness of orthogonal projector, 2016. Available at https://math.stackexchange.com/questions/1937327/uniqueness-of-orthogonal-projector.

[5] G.W. Stewart. *On the Perturbation of Pseudo-Inverses, Projections and Linear Least Squares Problems.* Defense Technical Information Center, 1975.

[6] Lloyd N. Trefethen and David Bau. *Numerical Linear Algebra.* SIAM, 1997.