SJÄLVSTÄNDIGA ARBETEN I MATEMATIK

MATEMATISKA INSTITUTIONEN, STOCKHOLMS UNIVERSITET

A Study on Linear Systems with Uncertainty

av

Julian Kiik

2021 - No K36

MATEMATISKA INSTITUTIONEN, STOCKHOLMS UNIVERSITET, 106 91 STOCKHOLM

A Study on Linear Systems with Uncertainty

Julian Kiik

Självständigt arbete i matematik 15 högskolepoäng, grundnivå

Handledare: Yishao Zhou

2021

Acknowledgements

I would like to thank my supervisor, professor Yishao Zhou for all her help with this thesis.

I also wish to thank my brother and parents for their support.

Abstract

This thesis studies how uncertainty affects solutions to linear equations. In particular, the Robust Least Square problem is studied in terms of the worst case residual. The robust analysis and solutions are investigated through a second order cone program and its duality. Numerical examples are also presented to illustrate the relation between the conditioning and the robustness of the problem.

1 Introduction

The background to studying how uncertainty affects solutions to robust max-norm minimization lies in the theory of numerical analysis.

Research into this topic has focused on fields such as engineering with problems of the form [8]

$$\begin{aligned}&\text{minimize } \|H(x)\|\\ &\text{where } H(x) = H_0 + \sum_{i=1}^{m} x_i H_i\end{aligned}$$

where H_i, $i = 1, \dots, n$ are given $p \times q$ matrices, but also branches into topics in Linear Algebra with robust eigenvalue minimization. Such examples will only be mentioned briefly in this thesis.

Throughout the thesis we will use different notations, namely

A^+: The $m \times n$ matrix A^+ is the Moore-Penrose inverse of the $m \times n$ matrix A satisfying the Penrose conditions, namely $AA^+A = A$, $A^+AA^+ = A^+$, A^+A and AA^+ are hermitian.

S^n: The set of all symmetric matrices, i.e. the set of all matrices equal to its transpose.

$\mathbb{C}$: The set of complex numbers.

$\mathbb{R}$: The set of real numbers.

$\kappa(A)$: The condition number of a matrix A. The condition number measures how the output value of a function is affected for a small change in the input argument.

$\mathcal{C}(A)$: Set of all vectors of the form Ax.

$\triangleq$: For example, $A \triangleq B$ means "A is defined to be B ".

Furthermore, the definitions used are

i. Subordinate matrix norm: A matrix norm that is said to be the natural norm induced by, or subordinate to, a previously defined vector norm.

ii. Least-squares solution: A least-squares solution solves the equation $Ax = b$ as close as possible, such that the sum of the squares of the difference $b - Ax$ is minimized.

iii. Feasible solution: A set of values for the decision variables satisfying all constraints in an optimization problem.

iv. Primal problem: The Conic Programming, or CP, problem.

v. Strictly feasible: All constraints are satisfied and nonlinear constraints are satisfied with strict inequalities.

vi. Condition Number: See $\kappa(A)$ above.

vii. Structured problems: Problems which can be solved by repeating examination and testing on the problem.

Note that this thesis is not meant to delve into the subject of numerical analysis or find novel methods of solving subject equations even though the main idea of our thesis is taken from numerical analysis. Instead, it is meant to study the subject of Least Squares from a theoretical aspect, hence why many theorems, propositions and remarks are referenced from sources that have studied this topic extensively.

We will discuss how uncertainty affects solutions to robust max-norm minimization with the help of a worked expression. The first section shall state theorems and proofs pertaining to norms for both vectors and matrices. The second section will treat the topic of error bounds and robustness for linear systems. The third section shall develop the previous and deal with error analysis for least squares problems. The fourth section will develop conic programming and the fifth section, based on all previous sections, will regard the main question of this thesis by computing an example. The sixth and final section shall mention robust polynomial interpolation, but will only regard it in the most basic terms since it is beyond the scope of this thesis.

2 Matrix and vector norms

When numerically solving a system of equations $Ax = b$ we tend to use direct methods. These include, for example, Gaussian Elimination on the Triangular Decomposition of a Matrix, Gauss-Jordan algorithm or the Cholesky

Decomposition.

However, when using these methods we may sometimes receive an approximation $\tilde{x}$ to the true solution x. To measure how accurate $\tilde{x}$ is, we need to measure the error

$$x - \tilde{x}$$

by measuring the size of a vector, or rather its norm. Hence, a norm $\|x\|$ is introduced on the vector space $\mathbb{C}^n$ over the complex numbers, or rather a function

$$\|\cdot\| : \mathbb{C}^n \to \mathbb{R}$$

which assigns to each vector $x \in \mathbb{C}^n$ a real value $\|x\|$ which is a measure of the size of x. This function must satisfy three properties:

i. Positivity: $\|x\| > 0$ for all $x \in \mathbb{C}^n$, $x \neq 0$.

ii. Homogeneity: $\|\alpha x\| = |\alpha|\|x\|$ for all $\alpha \in \mathbb{C}$, $x \in \mathbb{C}^n$.

iii. Triangle Inequality: $\|x + y\| \leq \|x\| + \|y\|$ for all $x, y \in \mathbb{C}^n$.

Theorem 1. *Each norm $\|\cdot\|$ on either $\mathbb{R}^n$ (or $\mathbb{C}^n$) is a uniformly continuous function with respect to the metric $\rho(x, y) = \max_i |x_i - y_i|$ on $\mathbb{R}^n$ (or $\mathbb{C}^n$)*

Our second theorem is thus

Theorem 2. *All norms on $\mathbb{R}^n$ (or $\mathbb{C}^n$) are equivalent in the sense that for each pair of norms $p_1(x)$ and $p_2(x)$ there are positive constants m and M satisfying*

$$mp_2(x) \leq p_1(x) \leq Mp_2(x) \textit{ for all } x$$

The proofs for these theorems are left out, however they can be found in Stoer and Bulirsch [2].

Now we prove that for the vector $x = (x_1, ..., x_n)^T$, and $p \geq 1$,

$$\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p}$$

defines a norm.

Obviously $\|x\|_p \geq 0$, and for $\|x\|_p$ to be zero all $|x_i| = 0$ and so $x = 0$. Hence the first axiom holds. The second axiom is also true, since

$$\|\alpha x\|_p^p = \sum_{i=1}^n |\alpha x_i|^p = \sum_{i=1}^n |\alpha|^p |x_i|^p = |\alpha|^p \sum_{i=1}^n |x_i|^p = |\alpha|^P \|x\|_p^p.$$

Finally, the triangle inequality follows from Minkowski's inequality. [1]

Note that $\|x\|_2$, called the Euclidean norm, is often considered as the length of the vector x.

Next we extend the vector norm to the matrix norm. This is important, since we need to measure the size of a matrix, for example in solving linear equations $Ax = b$. Let $M(m,n)$ be a vector space of all $m \times n$ matrices. We can use the vector p-norms defined above for matrices as well.

If $A \in M(m,n)$ we have the matrix norm

$$\|A\|_{(p)} = \left(\sum_{i,j} |a_{ij}|^p \right)^{1/p} .$$

The frequently used matrix norms are 2, which is also called the Frobenius norm:

$$\|A\|_F = \|A\|_{(2)} = \sqrt{\sum_{i,j} |a_{ij}|^2}.$$

The following properties show why this norm is interesting.

- $\|x\|_F = \|x\|_2$ for any column or row vector x.
- $\|AB\|_F \leq \|A\|_F \|B\|_F$ (the size of A and B can be different as far as AB is well-defined).
- $\|Ax\|_2 \leq \|A\|_F \|x\|_2$ for any column vector x.

It is trivial that the first property holds by definition. The last property follows from the first and the second. It remains to show that the second property holds. Let now $C = AB$, then

$$|c_{ij}|^2 = \left| \sum_k a_{ik} b_{kj} \right|^2 \leq \left(\sum_k |a_{ik}||b_{kj}| \right)^2 \leq \left(\sum_k |a_{ik}|^2 \right) \left(\sum_k |b_{ij}|^2 \right)$$

[1] $\left(\sum_{i=1}^k |x_i + y_i|^p \right)^{1/p} \leq \left(\sum_{i=1}^k |x_i|^p \right)^{1/p} + \left(\sum_{i=1}^k |y_i|^p \right)^{1/p}$. The proof of it is based on Hölder's inequality, a generalization of the Cauchy-Schwarz' inequality. See e.e. `https://www.comm.utoronto.ca/frank/notes/ineq.pdf` using convexity arguments.

where we applied the Cauchy-Schwarz inequality in the last step. Hence

$$\begin{aligned}\|AB\|_F^2 = \sum_{i,j} |c_{ij}|^2 \le \sum_{i,j} \left(\sum_k |a_{ik}|^2\right)\left(\sum_k |b_{ij}|^2\right)\\ =\sum_j \left(\sum_{i,k} |a_{ik}|^2\right)\left(\sum_k |b_{ij}|^2\right) = \sum_j \|A\|_F \left(\sum_k |b_{ij}|^2\right)\\ =\|A\|_F \left(\sum_{j,k} |b_{ij}|^2\right) = \|A\|_F\|B\|_F\end{aligned}$$

To summarize, like the vector norm the Frobenius norm, in general $\|A\|_{(p)}$, satisfies

1. Positivity : $\|A\|_F < 0$ for all $A \neq 0$;
2. Homogeneity: $\|\alpha A\|_F = |\alpha|\|A\|_F$;
3. Triangle inequality: $\|A + B\|_F \le \|A\|_F + \|B\|_F$.

Moreover it satisfies

Submultiplicity: $\|AB\|_F \le \|A\|_F\|B\|_F$.

However, we learned in the basic linear algebra course that any matrix is a linear transformation, i.e. if we have a linear transformation we would like the matrix norm to have the same value for any matrix representing the linear transformation and the Frobenius norm are definitely not such norms, which can be seen from the following two matrices

$$\begin{bmatrix}2 & 0\\ 0 & 1\end{bmatrix} \quad \begin{bmatrix}2 & 3\\ 0 & 1\end{bmatrix}.$$

They are similar matrices but the Frobenius norms are $\sqrt{5}$ and $\sqrt{14}$ respectively. Therefore we aim to define a matrix norm which can overcome this drawback.

We consider $A \in M(n, n)$. Remember that it is a matrix representation of the linear operator on the vector space V of dimension n with norm $\|x\|$. We define

$$\mathrm{lub}(A) := \max_{x\neq 0} \frac{\|Ax\|}{\|x\|}.$$

This is called the *subordinate matrix norm* in some literature (e.g. [2]) or the operator norm. A few arguments are now in order.

First, we have to prove that maximum exists, since the set for which $x \neq 0$ is open and we know that not every continuous function has a maximum on an open set. But if $x \neq 0$ then $\|x\| \neq 0$. Therefore, using the properties of the vector norm on V, we get

$$\frac{\|Ax\|}{\|x\|} = \left\| A\frac{x}{\|x\|} \right\| = \|Ay\|.$$

where $y = \frac{x}{\|x\|}$, whose norm is equal to 1. This shows that the maximum in question exists, since the set $\{x : \|x\| = 1\}$ is compact and $\|\cdot\|$ is continuous by Theorem 1. Consequently

$$\text{lub}(A) = \max_{x\neq 0} \frac{\|Ax\|}{\|x\|} = \max_{\|x\|=1} \|Ax\|.$$

Next we show that it is a norm. It is apparent that this definition satisfies the positivity and homegeneity of the norm. To prove the triangle inequality, we note first that by definition

$$\frac{\|Ax\|}{\|x\|} \leq \text{lub}(A)$$

that is,

$$\|Ax\| \leq \text{lub}(A)\|x\|.$$

This means that such a matrix norm is consistent with the vector norm. Pick any vector x with unit norm and any two matrices A and B using triangle inequality for the vector norm and we have

$$\|(A+B)x\| = \|Ax + Bx\| \leq \|Ax\| + \|Bx\| \leq \text{lub}(A) + \text{lub}(B).$$

Consequently, maximum over $\|x\| = 1$ gives

$$\text{lub}(A+B) \leq \text{lub}(A) + \text{lub}(B).$$

Finally, let $x \neq 0$, using the consistency property twice we have

$$\|ABx\| = \|A(Bx)\| \leq \text{lub}(A)\|Bx\| \leq \text{lub}(A)\text{lub}(B)\|x\|.$$

Thus

$$\frac{\|ABx\|}{\|x\|} \leq \text{lub}(A)\text{lub}(B)$$

Maximizing over $x \neq 0$ yields the submultiplicity

$$\text{lub}(AB) \leq \text{lub}(A)\text{lub}(B)$$

Now we consider some simple examples.

If $A = 0$, then $\mathrm{lub}(A) = 0$ for any choice of norm on V. If $A = I$ then

$$\mathrm{lub}(I) = \max_{x \neq 0} \frac{\|Ix\|}{\|x\|} = \max_{x \neq 0} \frac{\|x\|}{\|x\|} = 1$$

for any subordinate matrix norm. But $\|I\|_{(p)} = n^{1/p}$ when n is the dimension of V. If $A = a$ is a scalar then $\mathrm{lub}(A) = |a|$.

Although we defined the subordinate matrix norm for square matrices and operators, it can be reformulated to non-square matrices. We will show this with a case that is of interest to us in answering the question on how to calculate and use this norm for any concrete operator or matrix. Since we will study the least square problems, the natural norm is the Euclidean vector norm or 2-norm $\|x\|_2 = \sqrt{\langle x, x\rangle_2}$, where $\langle x, x\rangle_2$ is the standard inner product. Related to this norm we have

Proposition 3. *For $A \in M(m, n)$, related to the Euclidean vector norm the subordinate matrix norm*

$$\mathrm{lub}_2(A) = \sigma_1$$

the largest singular value of A. In particular we denote $\|A\|_2 = \mathrm{lub}_2(A)$.

Proof. We show this via Singular Value Decomposition, or SVD, of A, i.e. there are an $m \times m$ unitary matrix U and an $n \times n$ unitary matrix V such that $A = U\Sigma V^*$, where $V^* = \bar{V^t}$ and

$$\Sigma = \begin{bmatrix} \Sigma_1 & 0 \\ 0 & 0 \end{bmatrix}$$

which is an $m \times n$ matrix with $\Sigma_1 = \mathrm{diag}(\sigma_1, \ldots, \sigma_r)$, $\sigma_1 \geq \ldots \geq \sigma_r > 0$ being singular values of A, and $r \leq \min(m, n)$, the rank of A. Thus,

$$\begin{aligned} \|Ax\|_2^2 = \|U\Sigma V^* x\|_2^2 &= \langle U\Sigma V^* x, U\Sigma V^* x\rangle_2 \\ &= \langle \Sigma V^* x, U^* U\Sigma V^* x\rangle_2 \\ &= \langle \Sigma V^* x, \Sigma V^* x\rangle_2 \end{aligned}$$

By variable change, $y = V^* x \in \mathbb{C}^n$ we receive

$$\begin{aligned} \|Ax\|_2^2 = \langle \Sigma y, \Sigma y\rangle &= \langle y, \Sigma^* \Sigma y\rangle \\ &= \sum_{i=1}^{r} \sigma_i^2 |y_i|^2 \leq \sigma_1^2 \|y\|_2 \end{aligned}$$

and $\|x\|_2 = \|Vy\|_2 = \|y\|_2$. Hence,

$$\max_{x\neq 0}\frac{\|Ax\|_2}{\|x\|_2} = \max_{y\neq 0}\frac{\left(\sum_{i=1}^{r}\sigma_i^2|y_i|^2\right)^{\frac{1}{2}}}{\|y\|_2} \leq \sigma_1.$$

However, for $y = (1, 0, \ldots, 0)^t \Leftrightarrow x = v_1$ with $Av_1 = \sigma_1 u_1$, where u_1 is the first column of the matrix U and v_1, the first column for matrix V, the maximum is attained at $x = v_1$ and therefore

$$\max_{x\neq 0}\frac{\|Ax\|_2}{\|x\|_2} = \sigma_1.$$

□

Remark 4. *We can equivalently use* $\|A\|_2 = \max\limits_{\|x\|_2=1}\|Ax\|_2$.

Note that this is just one possible subordinate matrix norm. For example, we have

- For $\|x\|_\infty$, $\|A\|_\infty := \mathrm{lub}_\infty(A) = \max_i \sum_j |A_{ij}|$;
- For $\|x\|_1$, $\|A\|_1 := \mathrm{lub}_1(A) = \max_j \sum_i |A_{ij}|$.

Following [9] we define the Euclidean condition number $\kappa_2(A)$ for a rectangular $m \times n$ matrix A with linearly independent columns, i.e. A^{-1} exists if $m = n$,

$$\kappa_2(A) = \frac{\max\limits_{\|x\|_2=1}\|Ax\|_2}{\min\limits_{\|x\|_2=1}\|Ax\|_2}.$$

If $m = n$, then by SVD $\max\limits_{\|x\|_2=1}\|Ax\|_2 = \sigma_1$, $\min\limits_{\|x\|_2=1}\|Ax\|_2 = \sigma_n = \|A^{-1}\|_2$, the smallest singular value of A. We recover therefore the ordinary definition in 2-norm

$$\kappa_2(A) = \|A\|_2\|A^{-1}\|_2.$$

Generally we have

Proposition 5. $\kappa_2^2(A) = \kappa_2(A^TA) = \frac{\max\limits_i \lambda_i}{\min\limits_i \lambda_i}$ *where* λ_i *are eigenvalues of* A^TA, *where* A *is a rectangular* $m \times n$ *matrix.*

Proof. By definition of the Euclidean condition number we have

$$\begin{aligned}
\kappa_2^2(A) &= \left(\frac{\max\limits_{\|x\|_2=1}\|Ax\|_2}{\min\limits_{\|x\|_2=1}\|Ax\|_2}\right)^2 && \text{(Proposition 3)}\\
&= \left(\frac{\sigma_1}{\sigma_n}\right)^2 && \text{(definition of Singular values)}\\
&= \left(\frac{\lambda_1^{\frac{1}{2}}}{\lambda_n^{\frac{1}{2}}}\right)^2 = \frac{\lambda_1}{\lambda_n},
\end{aligned}$$

where λ_1 is the maximum eigenvalue of A^TA and λ_n, the minimum eigenvalue of A^TA. □

As stated before the Frobenius norm is not a subordinate matrix norm. But we can show

Proposition 6. *Let A be a $m \times n$ real matrix.*

1. $\|A\|_F = \sqrt{tr(A^TA)}$;
2. *Let A be a real rank-1 matrix. Then $\|A\|_F = \|A\|$.*

Proof. The first part follows from a straightforward calculation. We now show the second part.

First we prove that if A is a rank one matrix then there are vectors $u \in \mathbb{R}^n$ and $v \in \mathbb{R}^n$ such that $A = uv^T$.

If A is rank 1, then the columns of A, $a_1, \ldots, a_n$, have maximum one linearly independent vector. Without loss of generality, assume $a_1 \neq 0$. Then $a_i = \alpha_i a_i$, $\alpha_i \in \mathbb{R}$, $i = 2, \ldots, n$. Thus

$$A = (a_1, \alpha_2 a_1, \ldots, \alpha_n a_1) = a_1(1, \alpha_2, \ldots, \alpha_n).$$

Let $u = a_1 \in \mathbb{R}^m$, $v = (1, \alpha_2, \ldots, \alpha_n)^T \in \mathbb{R}^n$. Then $A = uv^T$. Furthermore, by definition (Remark 4)

$$\|A\| = \max_{\|x\|=1}\|Ax\| = \max_{\|x\|=1}\|uv^Tx\| = \max_{\|x\|=1}\|u\||v^Tx| = \|u\|\max_{\|x\|=1}|v^Tx| = \|u\|\cdot\|v\|.$$

Now we compute

$$\|A\|_F^2 = \mathrm{tr}(A^TA) = \mathrm{tr}\left((uv^T)^T(uv)^T\right) = \mathrm{tr}(vu^Tuv^T) = \mathrm{tr}(v^Tvu^Tu) = \|v\|^2\|u\|^2$$

proving that $\|A\|_F = \|A\|$. □

3 Error Bounds and Sensitivity of Square Linear Systems

We have shown the properties of the norm and in this section we shall investigate approximate solutions to a system $Ax = b$, where A is a square matrix. The following results will be useful in such a study. The matrix norm in this section can be any subordinate matrix norm, although our primary interest will be the subordinate matrix to the Euclidean vector norm, i.e.

Lemma 7. *If F is an $n \times n$ matrix with $\|F\| < 1$, then $(I + F)^{-1}$ exists and satisfies*

$$\|(I + F)^{-1}\| \leq \frac{1}{1 - \|F\|}.$$

Recall the condition number of an invertible matrix A is defined as $\kappa(A) := \|A\|\|A^{-1}\|$. Then we have

Theorem 8. *Let A be a nonsingular $n \times n$ matrix, $B = A(I + F)$, $\|F\| < 1$ and x and Δx be defined by $Ax = b$, $B(x + \Delta x) = b$. It follows that*

$$\frac{\|\Delta x\|}{\|x\|} \leq \frac{\|F\|}{1 - \|F\|},$$

as well as

$$\frac{\|\Delta x\|}{\|x\|} \leq \frac{\kappa(A)}{1 - \kappa(A)\frac{\|B-A\|}{\|A\|}} \frac{\|B - A\|}{\|A\|}$$

if $\kappa(A) \cdot \|B - A\|\|A^{-1}\| < 1$.

The proofs are omitted but can be found in Stoer and Bulirsch [2]. Our task here is to study the sensitivity of changes if the solutions of A or b are perturbed when solving a system of the form $Ax = b$. We divide this into two cases. First, b is replaced by $b + \Delta b$ and the second A is replaced by $A + \Delta A$.

Let us now assume $\|x\|$ is an arbitrary vector norm and $\|A\|$ a consistent submultiplicative matrix norm. If the solution $x + \Delta x$ corresponds to the right hand side $b + \Delta b$ then the relation $\Delta x = A^{-1}\Delta b$ follows from $A\Delta x = \Delta b$, and by the norm property it follows that

$$\|\Delta x\| \leq \|A^{-1}\|\|\Delta b\|.$$

Hence relative change $\|\Delta x\|/\|x\|$ is bounded as follows:

$$\frac{\|\Delta x\|}{\|x\|} \leq \|A\|\|A^{-1}\|\frac{\|\Delta b\|}{\|b\|} = \kappa(A)\frac{\|\Delta b\|}{\|b\|}$$

since $\|b\| = \|Ax\| \leq \|A\|\|x\|$. Here, the condition number of A, $\kappa(A)$, measures the sensitivity of the relative error in the solution to changes in the right hand side of b.

Next we want to know the sensitivity of the solution to changes in the matrix. If $\tilde{x}$ is an approximate solution to $Ax = b$ with the residual

$$r(\tilde{x}) = b - A\tilde{x}$$

then $\tilde{x}$ is the exact solution of

$$A\tilde{x} = b - r(\tilde{x}).$$

Since $A\Delta x = r(\tilde{x})$, or equivalently $\Delta x = A^{-1}r(\tilde{x})$, we have that the estimate

$$\|\Delta x\| \leq \|A^{-1}\|\|r(\tilde{x})\|$$

must hold for the error Δx defined by $\Delta x = x - \tilde{x}$.

Theorem 8 states that $\kappa(A)$ measures the sensitivity of the solution x to changes in the matrix A. By considering the relations

$$\begin{aligned} C &= (I+F)^{-1} = B^{-1}A \\ F &= A^{-1}B - I \end{aligned}$$

it follows from Lemma 5 that

$$\|B^{-1}A\| \leq \frac{1}{1 - \|I - A^{-1}B\|}.$$

Switching A and B, we find a new condition given rise from $A^{-1} = A^{-1}BB^{-1}$,

$$\|A^{-1}\| \leq \|A^{-1}B\|\|B^{-1}\| \leq \frac{\|B^{-1}\|}{1 - \|I - B^{-1}A\|}$$

and then using the residual estimate from above, $\|\Delta x\| \leq \|A^{-1}\|\|r(\tilde{x})\|$, we find the bound

$$\|\tilde{x} - x\| \leq \frac{\|B^{-1}\|}{1 - \|I - B^{-1}A\|}\|r(\tilde{x})\|, \ r(\tilde{x}) = b - A\tilde{x}$$

where B^{-1} is an approximate inverse to A with $\|I - B^{-1}A\| < 1$. The estimates that we have found give bounds on the error $\tilde{x} - x$, however evaluation of these bounds requires a basic knowledge of the inverse to A, namely A^{-1}.

We now discuss a different set of estimates that do not require any knowledge of A^{-1}, based on Prager and Oettli, 1964 [10].

The given data A_0, b_0 of an equation system $A_0 x = b_0$ tends to be inexact if it has been affected by measurement errors ΔA, Δb. Thus, it is reasonable to assume an approximate solution $\tilde{x}$ to the above equation system is correct if $\tilde{x}$ is the exact solution to a neighbouring system of equations $A\tilde{x} = b$ with

$$\begin{aligned} A \in \mathfrak{A} &:= \{A \mid |A - A_0| \leq \Delta A\} \\ b \in \mathfrak{B} &:= \{b \mid |b - b_0| \leq \Delta b\} \end{aligned}$$

Let α_{ij} be the component of A and β_i be the component of b. That is $A = (\alpha_{ij})$, $b = (\beta_1, \cdots, \beta_n)^T$. Then denote $|A| = (|\alpha_{ij}|)$, $|b| = (|\beta_1|, \cdots, |\beta_n|)^T$.

Theorem 9. *Let $\Delta A \geq 0$, $\Delta b \geq 0$, and let $\mathfrak{A}$, $\mathfrak{B}$ be defined as above. Then, for any approximate solution $\tilde{x}$ of the system $A_0 x = b_0$ there exists a matrix $A \in \mathfrak{A}$ and a vector $b \in \mathfrak{B}$ satisfying $A\tilde{x} = b$ if and only if*

$$|r(\tilde{x})| \leq \Delta A|\tilde{x}| + \Delta b$$

where $r(\tilde{x}) := b_0 - A_0\tilde{x}$ is the residual of $\tilde{x}$.

Proof. We divide the proof in two steps.

1) We first assume $A\tilde{x} = b$ holds for some $A \in \mathfrak{A}$, $b \in \mathfrak{B}$. Since

$$\begin{aligned} A &= A_0 + \delta A \text{ where } |\delta A| \leq \Delta A \\ b &= b_0 + \delta b \text{ where } |\delta b| \leq \Delta b \end{aligned}$$

we have

$$\begin{aligned}|r(\tilde{x})| &= |b_0 - A_0\tilde{x}| = |b - \delta b - (A - \delta A)\tilde{x}| \\ &= |-\delta b + (\delta A)\tilde{x}| \\ &\leq |\delta b| + |\delta A||\tilde{x}| \\ &\leq \Delta b + \Delta A|\tilde{x}|.\end{aligned}$$

2) If $|r(\tilde{x})| \leq \Delta b + \Delta A|\tilde{x}|$, and if r and s stand for the vectors

$$\begin{aligned}&r := r(\tilde{x}) = (\rho_1, \ldots, \rho_n)^T \\ &s := \Delta b + \Delta A|\tilde{x}| \geq 0,\ s = (s_1, \ldots, s_n)^T.\end{aligned}$$

Further, set

$$\begin{aligned}&\delta A = (\delta\alpha_{ij}),\ \delta b = \begin{bmatrix}\delta\beta_1 \\ \vdots \\ \delta\beta_n\end{bmatrix},\ \tilde{x} = \begin{bmatrix}\zeta_1 \\ \vdots \\ \zeta_n\end{bmatrix} \\ &\delta\alpha_{ij} := \rho_i \Delta\alpha_{ij} \cdot \operatorname{sign}(\zeta_j)/s_i \\ &\delta\beta_i := -\rho_i \Delta\beta_i/s_i,\ \text{where } \rho_i/s_i := 0 \text{ if } s_i = 0\end{aligned}$$

From $|r(\tilde{x})| \leq \Delta b + \Delta A|\tilde{x}|$ we have that $|\rho_i/s_i| \leq 1$ and thus

$$\begin{aligned}A = A_0 + \delta A \in \mathfrak{A} \\ b = b_0 + \delta b \in \mathfrak{B}\end{aligned}$$

and, for $i = 1, 2, \ldots, n$,

$$\begin{aligned}\rho_i = \beta_i - \sum_{j=1}^{n} \alpha_{ij}\zeta_j &= \left(\Delta\beta_i + \sum_{j=1}^{n} \Delta\alpha_{ij}|\zeta_j|\right)\frac{\rho_i}{s_i} \\ &= -\delta\beta_i + \sum_{j=1}^{n} \delta\alpha_{ij}\zeta_j\end{aligned}$$

which can be rewritten as

$$\sum_{j=1}^{n}(\alpha_{ij} + \delta\alpha_{ij})\zeta_j = \beta_i + \delta\beta_i.$$

In other words,

$$A\tilde{x} = b$$

which is what we wanted to show.

□

The requirements for the above theorem allows us to investigate a solution from the smallness of its residual.

For example, if all components of A_0 and b_0 have the same relative accuracy ϵ, that is

$$\Delta A = \epsilon|A_0|$$
$$\Delta b = \epsilon|b_0|$$

then the condition of the theorem is satisfied when

$$|A_0\tilde{x} - b_0| \leq \epsilon(|b_0| + |A_0||\tilde{x}|)$$

Using this inequality, the smallest ϵ is computed for which a given $\tilde{x}$ can still be accepted as a solution. This is in fact a robust analysis because we can consider the data (A, b) with uncertainty.

4 Error Analysis of Least Squares Problem

Least squares are used in a variety of applications, for example in regression analysis to approximate the solution of overdetermined systems by minimizing the sum of the squares of the residuals made in the results of every single equation. An overdetermined system is a system containing more equations than unknown variables.

Let us regard least squares for the matrix equation $Ax = b$. First we show that the solution exists

Theorem 10. *Let A be a $m \times n$ matrix with linearly independent columns and b, a column vector with (m) components. The least-squares solution of $Ax = b$ is the solution to the matrix equation*

$$A^TAx = A^Tb$$

Proof. Let $W = \mathcal{C}(A)$, that is W is the set of all vectors of the form Ax. Let $b = b_W + b_{W^\perp}$ be the orthogonal decomposition with respect to W. By definition, b_W lies in W and hence there exists a vector x in $\mathbb{R}^n$ such that $Ax = b_W$. Choosing any such x, we have $b - b_W = b - Ax$ lying in $W^\perp$. This belongs to $\mathcal{N}(A^T)$, the null space of A^T, and thus $0 = A^T(b - Ax) = A^Tb - A^TAx$, so $A^TAx = A^Tb$.

□

Moreover, this implies that $A^TAx = A^Tb$ is consistent, i.e the system has at least one solution.

Note that $A^TAx = A^Tb$, called the normal equation, is the necessary condition for x to be the minimum of $\|b - Ax\|_2$ which can be checked by a straightforward computation of the gradient to $\|b - Ax\|_2$.

Example 1

Find the least-squares solutions of $Ax = b$ where:

$$A = \begin{bmatrix} 2 & 0 \\ -1 & 1 \\ 0 & 2 \end{bmatrix} \qquad b = \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix}$$

Solution

We first calculate

$$A^TA = \begin{bmatrix} 2 & -1 & 0 \\ 0 & 1 & 2 \end{bmatrix} \begin{bmatrix} 2 & 0 \\ -1 & 1 \\ 0 & 2 \end{bmatrix} = \begin{bmatrix} 5 & -1 \\ -1 & 5 \end{bmatrix}$$

$$A^Tb = \begin{bmatrix} 2 & -1 & 0 \\ 0 & 1 & 2 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix} = \begin{bmatrix} 2 \\ -2 \end{bmatrix}.$$

We form the augmented matrix and row reduce to

$$\left[\begin{array}{cc|c} 5 & -1 & 2 \\ -1 & 5 & -2 \end{array}\right] \sim \left[\begin{array}{cc|c} 1 & 0 & \frac{1}{3} \\ 0 & 1 & -\frac{1}{3} \end{array}\right].$$

Hence, the only least square solution is $x = \frac{1}{3}\begin{bmatrix} 1 \\ 1 \end{bmatrix}$.

However, there are some issues in using the normal equation. We illustrate this by the next example.

Example 2

Let $y(s) = \frac{x_1}{s} + \frac{x_2}{s^2} + \frac{x_3}{s^3}$ with $x_1 = x_2 = x_3 = 1$. If we have a data set $(s_i, y(s_i))$, $i = 1, \ldots, 10$ with a machine that deals with the rounding error ϵ.

We wish to recover the solution $(1,1,1)^t$ by determining x_1, x_2 and x_3 given data $(s_i, y(s_i))$. We consider two cases:

a) $y_i = y_i(s)$, $i = 1, \ldots, 10$ exact.

b) y_i are not exactly $y_i(s)$.

a) We have the exact value $y_i = y(s_i)$, and since it is exact we know that there is an x such that the residual satisfies

$$r(x) = b - Ax = 0$$

where $b = (y_1, y_2, \ldots, y_{10})^T$ and

$$A = \begin{bmatrix} \frac{1}{s_1} & \frac{1}{s_1^2} & \frac{1}{s_1^3} \\ \frac{1}{s_2} & \frac{1}{s_2^2} & \frac{1}{s_2^3} \\ \cdots & \cdots & \cdots \\ \frac{1}{s_{10}} & \frac{1}{s_{10}^2} & \frac{1}{s_{10}^3} \end{bmatrix}$$

and $x = (x_1, x_2, x_3)$ should be determined through $Ax = b$, given $s_i = s_0 + i$, $i = 1, \ldots, 10$ where s_0 is a given number.

First, we solve it by directly solving the normal equation

$$A^T A x = A^T b$$

To this end we use the computer program Mathematica, and so we enter $s_0 = 10, 50, 100, 150$ and 200 to estimate the condition numbers of A and error norms $\Delta x = x - \hat{x}$, where $\hat{x} = (1,1,1)^t$ is the exact solution:

s_0	10	50	100	150	200
$\kappa_2(A)$	6.6×10^3	1.3×10^6	1.7×10^7	8.0×10^7	2.5×10^8
$\lVert\Delta x\rVert_2$	2.4×10^{-11}	8.8×10^{-8}	1.2×10^{-5}	2.5×10^{-5}	2.1×10^{-4}

Note that the condition number increases as s increases and Mathematica warns when $s = 50$. To avoid ill-conditioning we can use orthogonalization.

Let us decompose $A = QR$ into matrices Q, a 10×10 orthogonal matrix where $R = \begin{bmatrix} R_0 \\ 0 \end{bmatrix}$, with R_0 being a 3×3 invertible upper triangular matrix. Partition $Q = \begin{pmatrix} Q_1 & Q_2 \end{pmatrix}$ according to R and we see that

$$A^T A = (QR)^T QR = R^T (Q^T Q) R = R^T R = [R_0^T 0] \begin{bmatrix} R_0 \\ 0 \end{bmatrix} = R_0^T R_0$$

and

$$A^T b = \begin{bmatrix} R_0^T & 0 \end{bmatrix} \begin{bmatrix} Q_1^T \\ Q_2^T \end{bmatrix} b = R_0^T Q_1^T b.$$

Thus, the normal equation becomes

$$R_0^T R_0 x = R_0^T Q_1^T b \Leftrightarrow R_0 x = Q_1^T b.$$

Using the same linear solver as above in Mathematica, we receive the error norm $\|\Delta x_{\mathrm{orth}}\|_2$ in the following table together with the error obtained above.

s_0	10	50	100	150	200
$\|\Delta x\|_2$	2.4×10^{-11}	8.8×10^{-8}	1.2×10^{-5}	2.5×10^{-5}	2.1×10^{-4}
$\|\Delta x_{(orth)}\|_2$	4.0×10^{-14}	1.8×10^{-11}	3.3×10^{-10}	8.4×10^{-10}	8.5×10^{-9}

Hence, orthogonalization gives better results and Mathematica does not complain about ill conditioning.

b) Assume there is a disturbance in the data set y_i which is replaced by $y_i + \lambda v_i$, $\lambda \in \mathcal{R}$, where v satisfies $A^T v = 0$, $v = (v_1, \ldots, v_{10})$. Theoretically, the solution should remain unchanged

$$(A^T A)x = A^T (b + \lambda v) = A^T b$$

On the other hand, the residual $r(x) = b + \lambda v - Ax = \lambda v$. We vary $\lambda = 0,\ 10^{-6},\ 10^{-4},\ 10^{-2},\ 10^0,\ 10^2$ and we let $s_0 = 10$, $v =$

$(0.1331, -0.5184, 0.6591, 0.2744, 0, 0, 0, 0, 0, 0)^T$. With $s_0 = 10$ we have $\|A\|_2 \approx 0.22$ as shown above.

Like a), we determine x by directly solving the normal equation and orthogonalising, respectively. Using Mathematica we receive

λ	0	10^{-6}	10^{-4}	10^{-2}	10^{0}	10^{2}
$\|r(x)\|_2$	0	9×10^{-7}	9×10^{-5}	9×10^{-3}	9×10^{-1}	9×10^{1}
$\|\Delta x\|_2$	2.4×10^{-11}	2.4×10^{-11}	3.5×10^{-11}	2.4×10^{-11}	1.6×10^{-10}	4.4×10^{-10}
$\|\Delta x_{orth}\|_2$	4.0×10^{-14}	4.0×10^{-14}	4.0×10^{-14}	8.1×10^{-13}	4.7×10^{-11}	4.8×10^{-9}

From both a) and b) we see our errors are different. We can prove the general result [2]

Theorem 11. *The relative error using QR is*

$$\frac{\|\Delta x_{orth}\|_2}{\|x\|_2} \leq \kappa(R)\frac{\|\Delta A\|_2}{\|A\|_2} + \kappa(R)^2 \frac{\|r\|_2}{\|A\|_2\|x\|_2}\frac{\|\Delta A\|_2}{\|A\|_2} + \kappa(R)\frac{\|b\|_2}{\|A\|_2\|x\|_2}\frac{\|\Delta b\|_2}{\|b\|_2}$$

where A is replaced by $A + \Delta A$ and b, by $b + \Delta b$, with ΔA and Δb being small in relation to A and b respectively. If A^TA is replaced by $A^TA + G$ and A^Tb by $A^Tb + \tilde{\Delta b}$ in the normal equation, we have

$$\frac{\|\Delta x\|_2}{\|x\|_2} \leq \kappa(A)^2 \left(\frac{\|G\|_2}{\|A^TA\|_2} + \frac{\|\tilde{\Delta b}\|_2}{\|A^Tb\|_2} \right)$$

where G and $\tilde{\Delta b}$ are small relative to A^TA and A^Tb, respectively.

Proof. In this proof $\|\cdot\| = \|\cdot\|_2$. If A^TA, A^Tb in the normal equation are perturbed in the form $ATA + G$ and $A^tb + \tilde{\Delta b}$, then the solution is also perturbed to $x + \Delta x$. Thus we have the equation

$$(A^TA + G)(x + \Delta x) = A^Tb + \tilde{\Delta b}$$

To a first order approximation, this becomes

$$A^TA\Delta x + Gx = \tilde{\Delta b}$$

or equivalently, since A is assumed to have linearly independent columns,

$$\Delta x = (A^TA)^{-1}(\tilde{\Delta b} - Gx).$$

Then

$$\begin{aligned} \|\Delta x\| \leq& \|(A^TA)^{-1}\|(\|\tilde{\Delta b}\| + \|G\|\|x\|) \\ =& \|(A^TA)^{-1}\|\|A^TA\| \left(\frac{\|\tilde{\Delta b}\|}{\|A^TA\|} + \frac{\|G\|}{\|A^TA\|}\|x\| \right). \end{aligned}$$

Thus

$$\frac{\|\Delta x\|}{\|x\|} \leq \|(A^TA)^{-1}\|\|A^TA\| \left(\frac{\|\tilde{\Delta b}\|}{\|A^TA\|\|x\|} + \frac{\|G\|}{\|A^TA\|} \right).$$

From $A^TAx = A^Tb$,

$$\|A^Tb\| \leq \|A^TA\|\|x\|.$$

So

$$\frac{\|\Delta x\|}{\|x\|} \leq \|(A^TA)^{-1}\|\|A^TA\| \left(\frac{\|\tilde{\Delta b}\|}{\|A^Tb\|} + \frac{\|G\|}{\|A^TA\|} \right) = \kappa(A^TA) \left(\frac{\|\tilde{\Delta b}\|}{\|A^Tb\|} + \frac{\|G\|}{\|A^TA\|} \right).$$

By Proposition 5,

$$\frac{\|\Delta x\|}{\|x\|} \leq \kappa(A)^2 \left(\frac{\|\tilde{\Delta b}\|}{\|A^Tb\|} + \frac{\|G\|}{\|A^TA\|} \right).$$

We turn to orthogonalization. The matrix A and the vector b are replaced by $A + \Delta A$ and $b + \Delta b$, respectively. Then the resulting normal equation is

$$(A + \Delta A)^T(A + \Delta A)(x + \Delta x) = (A + \Delta A)^T(b + \Delta b).$$

To a first order approximation, this becomes

$$A^TAx + A^TA\Delta x + (\Delta A)^TAx + A^T\Delta Ax = A^Tb + A^T\Delta b + (\Delta A)^Tb.$$

Now that $A^TAx = A^Tb$ we get

$$\Delta x = (A^TA)^{-1} \left(A^T\Delta b + (\Delta A)^Tb - (\Delta A)^TAx - A^T\Delta Ax \right).$$

Substituting $r = b - Ax$ in above equation yields

$$\Delta x = (A^TA)^{-1} \left(A^T\Delta b + (\Delta A)^Tr - A^T\Delta Ax \right).$$

Then

$$\begin{aligned}\|\Delta x\| \leq& \|(A^TA)^{-1}A^T\|\|\Delta A\|\|x\| + \|(A^TA)^{-1}A^T\|\|\Delta b\| + \|(A^TA)^{-1}\|\|(\Delta A)^T\|\|r\| \\ =& \|(A^TA)^{-1}A^T\|\|A\|\frac{\|\Delta A\|}{\|A\|}\|x\| + \|(A^TA)^{-1}A^T\|\|A\|\frac{\|b\|}{\|A\|}\frac{\|\Delta b\|}{\|b\|} \\ &+ \|(A^TA)^{-1}\|\|A^T\|^2\frac{\|(\Delta A)^T\|}{\|A^T\|}\frac{\|r\|}{\|A^T\|}.\end{aligned}$$

Next using the QR factorization of A: $A = Q\begin{bmatrix} R \\ 0 \end{bmatrix}$ where Q is an $m \times m$ orthogonal matrix and R is an $n \times n$ invertible upper triangular matrix, it follows that

$$A^TA = R^TR, (A^TA)^{-1} = R^{-1}(R^T)^{-1}, (A^TA)^{-1}A^T = \begin{bmatrix} R^{-1} & 0 \end{bmatrix} Q^T.$$

Note that Q is orthogonal. The Euclidean norm is preserved. So

$$\begin{aligned}\|(A^TA)^{-1}A^T\|\|A\| &= \left\|\begin{bmatrix}R^{-1} & 0\end{bmatrix}Q^T\right\|\left\|Q\begin{bmatrix}R\\0\end{bmatrix}\right\| \\ &= \left\|\begin{bmatrix}R^{-1} & 0\end{bmatrix}\right\|\left\|\begin{bmatrix}R\\0\end{bmatrix}\right\| = \|R^{-1}\|\|R\| = \kappa(R),\end{aligned}$$

and

$$\begin{aligned}\|(A^TA)^{-1}\|\|A^T\|^2 =& \|(A^TA)^{-1}\|\|A\|^2 = \|(R^TR)^{-1}\|\|R\|^2 \\ \leq& \|R^{-1}\|^2\|R\|^2 = \kappa(R)^2.\end{aligned}$$

By substituting these two estimates in the previous inequality divided by $\|x\|$ we obtain

$$\frac{\|\Delta x\|}{\|x\|} \leq \kappa(R)\frac{\|\Delta A\|}{\|A\|} + \kappa(R)^2\frac{\|r\|}{\|A\|\|x\|}\frac{\|\Delta A\|}{\|A\|} + \kappa(A)\frac{\|b\|}{\|A\|\|x\|}\frac{\|\Delta b\|}{\|b\|}.$$

The proof is complete. □

Remark 12. *One way to measure ΔA and Δb relatively small to A and b can be*

$$\begin{aligned}\|\Delta A\|_2 \leq f(m)\cdot\epsilon\cdot\|A\|_2 \\ \|\Delta b\|_2 \leq f(m)\cdot\epsilon\cdot\|b\|_2\end{aligned}$$

with $f(m) = O(m)$, a slightly increasing function of m, and eps the machine precision.

Similarly, we can make an interpretation for G and $\tilde{\Delta b}$ where

$$\begin{aligned}\|G\|_2 \leq g(n)\cdot\epsilon\cdot\|A^TA\|_2 \\ \|\tilde{\Delta b}\|_2 \leq g(n)\cdot\epsilon\cdot\|A\|_2\|b\|_2\end{aligned}$$

where $g(n) \approx O(n)$.

Remark 13. *If both A and b are perturbed with ΔA and Δb, we would get a solution perturbed by $z = x + \Delta x$, that is $(A + \Delta A)Z = b + \Delta b$. This is equivalent to determining z where*

$$\min_z \|(b + \Delta b) - (A + \Delta A)z\|_2$$

Although orthogonalization can improve the accuracy of numerical computations, it needs other techniques to deal with input errors. Therefore, we consider next ΔA and Δb as uncertainty in the data. Thus we wish to find a minimum z in the worst case assuming ΔA and Δb have some constraints. We assume in this thesis that the augmented matrix $\begin{bmatrix}\Delta A & \Delta b\end{bmatrix}$, an $m \times (n+1)$ matrix, satisfies $\|\,[\Delta A \Delta b]\,\| \leq \rho_1$, where $\rho_1 \geq 0$ is given.

5 A brief review of theory on conic programming problems

To study robustness of the least square problem we will make use of second-order conic programming problems. To obtain some intuition we consider the following three optimization problems

a) A classic Linear Programming problem:

$$\begin{aligned} \text{minimize } & 2x_1 + x_2 + x_3 \\ \text{subject to } & x_1 + x_2 + x_3 = 1 \\ & (x_1; x_2; x_3) \geq 0 \end{aligned}$$

b) A second order Cone Linear Programming problem

$$\begin{aligned} \text{minimize } & 2x_1 + x_2 + x_3 \\ \text{subject to } & x_1 + x_2 + x_3 = 1 \\ & x_1 - \sqrt{x_2^2 + x_3^2} \geq 0 \end{aligned}$$

where the bottom constraint puts the variables in an ice-cream cone, or rather a second-order cone.

c) A semidefinite Cone Linear Programming problem

$$\begin{aligned} \text{minimize } & 2x_1 + x_2 + x_3 \\ \text{subject to } & x_1 + x_2 + x_3 = 1 \\ & \begin{bmatrix} x_1 & x_2 \\ x_2 & x_3 \end{bmatrix} \succeq 0 \end{aligned}$$

where the symbol $\succeq 0$ implies the left-side symmetric matrix must be positive semidefinite.

Even though the objective function and the first constraint are identical the last constraint distinguishes them as a different optimization problem. Thus, for example, the simplex method which works for LP does not work for the other two problems. Note, however, that interior-point methods developed for LP are naturally applied to solving the other two problems. To see this we look at the last constraint more closely. First, they can be viewed as the following three cones respectively.

$$\begin{aligned} &\text{the non-negative orthant:} \quad \mathbb{R}^3_+ := \{x \in \mathbb{R}^3 : x_1 \geq 0, x_2 \geq 0, x_3 \geq 0\}, \\ &\text{the second-order cone:} \quad L^3 := \{x \in \mathbb{R}^3 : x_1 \geq \sqrt{x_2^2 + x_3^2}\}, \\ &\text{the semi-definite cone:} \quad \mathcal{S}^2_+ := \{X \in \mathcal{S}^2 : X \succeq 0\}, \end{aligned}$$

where $\mathcal{S}^2$ is the set of all positive semi-definite (2×2) matrices. Note further that these cones are nested in the sense that we can view the non-negative orthant as the projection of a direct product of second-order cones on a subspace (by imposing $x_2 = x_3 = 0$) in the second-order cone L^3. Similarly, a projection of the semi-definite cone on a specific subspace gives the second-order cone, since

$$\left\| \begin{bmatrix} x_2 \\ x_3 \end{bmatrix} \right\| \leq x_1 \Longleftrightarrow X = \begin{bmatrix} x_1 & x_2 & x_3 \\ x_2 & x_1 & 0 \\ x_3 & 0 & x_1 \end{bmatrix} \succeq 0. \tag{1}$$

The proof of (1) is as follows. Decompose the matrix X into four blocks

$$X = \left[\begin{array}{c|cc} x_1 & x_2 & x_3 \\ \hline x_2 & x_1 & 0 \\ x_3 & 0 & x_1 \end{array} \right]$$

By a straightforward matrix manipulation we obtain that X is congruent to

$$\left[\begin{array}{c|c} x_1 - [x_2 \; x_3] x_1^{-1} \begin{bmatrix} x_2 \\ x_3 \end{bmatrix} & 0 \\ \hline 0 & x_1 I_2 \end{array} \right] = \left[\begin{array}{c|c} x_1 - x_1^{-1}(x_2^2 + x_3^2) & 0 \\ \hline 0 & x_1 I_2 \end{array} \right].$$

Therefore $X \succeq 0$ is equivalent to $x_1 \geq 0$ and $x_1 - x_1^{-1}(x_2^2 + x_3^2) \geq 0$ which is $\left\| \begin{bmatrix} x_2 \\ x_3 \end{bmatrix} \right\| \leq x_1$.

Note that this interpretation is valid also for larger problems. Moreover, notice that the matrix manipulation above can be carried out for a general positive semidefinite matrix if one of the main diagonal blocks is invertible. The diagonal blocks in the resulting matrix are commonly called Schur complement. This technique will be used later in reformulation of robust polynomial interpolation.

From the above argument we can cast the above three types of problem in a uniform formulation, called conic programming problem:

$$\begin{aligned} \min_{x \in \mathbb{R}^n} \quad & \langle c, x \rangle_2 \\ \text{s.t.} \quad & Ax = b, \\ & \quad x \in K, \end{aligned} \tag{2}$$

where $c \in \mathbb{R}^n$, $b \in \mathbb{R}^m$, $A \in \mathbb{R}^{m \times n}$, $\langle \cdot, \cdot \rangle_2$ the inner product and $K \subset \mathbb{R}^n$ is a convex cone. In other words K is a convex set with the property that for all $x \in K$, $\lambda x \in K$ for all $\lambda > 0$. In practice, the cone K is the direct product $K = K_1 \times K_2 \times \ldots \times K_l$ where $K_1, \ldots, K_l$ are cones, which is what we have later.

In this report we need the duality theory for the second-order programming problem. To this end we need a notion of dual cone which covers the above three cases.

5.1 Generalized inequalities

We start with the constraint inequality $Ax \geq b$ in LP to see how the inequality can be generalized. Given two vectors $a, b \in \mathbb{R}^m$, we say $a \geq b$ if the coordinates of a majorise the corresponding coordinates of b, i.e:

$$a \geq b \Leftrightarrow \forall i \in \{1, \ldots, m\} : a_i \geq b_i$$

Here, the latter relation uses the arithmetic $\geq$ - a relation between real numbers. This "coordinate-wise" partial ordering of vectors in $\mathbb{R}^m$ satisfies the following basic properties. For all vectors $a, b, c, d, \ldots \in \mathbb{R}^m$ we have:

1. Reflexivity: $a \geq a$;
2. Anti-symmetry: if both $a \geq b$ and $b \geq a$ then $a = b$;
3. Transitivity: if both $a \geq b$ and $b \geq c$ then $a \geq c$;
4. Compatibility with linear operations
   (a) Positive Homogeneity: if $a \geq b$ and λ is a nonnegative real, then $\lambda a \geq \lambda b$;
   (b) Additivity: if both $a \geq b$ and $c \geq d$ then $a + c \geq b + d$.

Consider vectors from a finite-dimensional Euclidean space **E** with an inner product $\langle \cdot, \cdot \rangle$ and assume that **E** has a partial ordering, in other words a vector inequality, denoted by $\succeq$. We denote this as a good ordering if it obeys the above axioms.

Moreover, a vector inequality $\succeq$ which we can consider to be good is fully identified by the set **K** of $\succeq$-nonnegative vectors

$$K = \{a \in E : a \succeq 0\}.$$

That is

$$a \succeq b \Leftrightarrow a - b \succeq 0 \ [\Leftrightarrow a - b \in \mathbf{K}].$$

However, this set **K** in the above observation cannot be any arbitrary set. It has to satisfy the following conditions

1. **K** is convex, nonempty and closed under addition

$$a, a' \in \mathbf{K} \Rightarrow a + a' \in \mathbf{K}$$

2. **K** is a conic set

$$a \in \mathbf{K}, \lambda \geq 0 \Rightarrow \lambda a \in \mathbf{K}$$

3. **K** is pointed

$$a \in \mathbf{K} \text{ and } -a \in \mathbf{K} \Rightarrow a = 0$$

Hence, **K** must be a pointed cone.

Every pointed cone **K** in **E** induces a partial ordering on **E** satisfying the above axioms. This ordering is denoted $\geq_{\mathbf{K}}$:

$$a \geq_{\mathbf{K}} b \Leftrightarrow a - b \geq_{\mathbf{K}} 0 \Leftrightarrow a - b \in \mathbf{K}$$

The cone responsible for the standard coordinate-wise ordering $\geq$ on $\mathbf{E} = \mathbb{R}^m$ is the cone comprised of vectors with nonnegative entries, the nonnegative orthant

$$\mathbb{R}^m_+ = \{x = (x_1, \ldots, x_m)^T \in \mathbb{R}^m : x_i \geq 0, i = 1, \ldots, m\}.$$

The pointed cone that is the nonnegative orthant satisfies two properties

1. The cone is closed: if a sequence of vectors a^i from the cone has a limit point, the latter also belongs to the cone.
2. The cone possesses a nonempty interior : there exists a vector such that a ball of positive radius centered at the vector is contained in the cone.

When mentioning vector inequalities $\geq_{\mathbf{K}}$ we assume that the underlying set $\mathbf{K}$ is a pointed and closed cone with a nonempty interior.

Allowing $\mathbf{K}$ to be a regular cone in $\mathbf{E}$, meaning the cone is convex, pointed, closed and with a nonempty interior, and given an objective $c \in \mathbb{R}^n$, a linear mapping $\mathbb{R}^n \to \mathbf{E} : x \to Ax$ and a right hand side $b \in \mathbf{E}$, we have the optimization problem

$$\min_x \{c^T x : Ax \geq_{\mathbf{K}} b\}.$$

We refer to the above problem as CP, the conic problem. The main difference between CP and LP is that the latter deals with the particular choice $\mathbf{E} = \mathbb{R}^m$, $\mathbf{K} = \mathbb{R}^m_+$.

5.2 Dual cones

The same reasoning that applies to the dual problem in LP can be applied to the dual problem in CP. We wish to know the "admissible" weight vectors λ, that is the vectors such that the scalar inequality $\langle \lambda, Ax \rangle \geq \langle \lambda, b \rangle$ is a consequence of the vector inequality $Ax \geq_{\mathbf{K}} b$.

Answering this question is the same as saying what the weight vectors λ are s.t

$$\forall a \geq_{\mathbf{K}} 0 : \langle \lambda, a \rangle \geq 0.$$

When λ has the above property the scalar inequality $\langle \lambda, a \rangle \geq \langle \lambda, b \rangle$ is a consequence of the vector inequality $a \geq_{\mathbf{K}} b$:

$$\begin{aligned} & a \geq_{\mathbf{K}} b \\ \Leftrightarrow\ & a - b \geq_{\mathbf{K}} 0 \\ \Rightarrow & \langle \lambda, a - b \rangle \geq 0 \\ \Leftrightarrow & \langle \lambda, a \rangle \geq \lambda^T b. \end{aligned}$$

If λ is an admissible weight vector for the partial ordering $\geq_{\mathbf{K}}$:

$$\forall (a, b : a \geq_{\mathbf{K}} b) : \ \langle \lambda, a \rangle \geq \langle \lambda, b \rangle$$

then λ satisfies the above property.

The weight vectors λ which are admissible for a partial ordering $\geq_{\mathbf{K}}$ are exactly the vectors satisfying the above property, or the vectors from the set

$$\mathbf{K}_* = \{\lambda \in \mathbf{E} : \langle \lambda, a \rangle \geq 0 \ \forall a \in \mathbf{K}\}.$$

The set $\mathbf{K}_*$ is comprised of vectors whose inner products with all vectors from $\mathbf{K}$ are nonnegative. This set $\mathbf{K}_*$ is called the cone dual to $\mathbf{K}$:

Theorem 14 (Properties of the dual cone)**.** *Let $\boldsymbol{E}$ be a finite dimensional Euclidean space with inner product $\langle \cdot, \cdot \rangle$ and let $K \subset \boldsymbol{E}$ be a nonempty set. Then*

i) The set

$$K_* = \{\lambda \in \boldsymbol{E}^m : \langle \lambda, a \rangle \geq 0 \ \forall a \in K\}$$

is a closed cone

ii) If int $K \neq \emptyset$, that is the interior of the cone $K \neq \emptyset$, then K_* is pointed

iii) If K is a closed convex pointed cone, then int $K_* \neq \emptyset$

iv) If K is a closed cone, then so is K_*, and the cone dual to K_* is K itself:

$$(K_*)_* = K$$

For example $\mathbb{R}^n_+ := \{x \in \mathbb{R}^n : x \geq 0\}$, $L^n := \{(x,t) \in \mathbb{R}^n_+ : t \geq \|x\|_2\}$ and $\mathcal{S}^n_+ := \{X \in \mathcal{S}^n : X \succeq 0\}$ are self dual cones.

A summarized comment on the proofs for the above is required. First, for $\mathbb{R}^n_+$ we see that

$$y^T x \geq 0 \ \forall \ x \succeq 0 \ \Leftrightarrow \ y \succeq 0$$

Hence, the cone is its own dual, or rather self dual.

If we assume $\|y\|_2 \geq s$, $s \geq 0$ we see by using Cauchy-Schwarz reversed that $(x, y) + st \geq -\|x\|_2 \|y\|_2 + st \geq 0$, for all $\|x\|_2 \leq t$ showing (y, s) is in the dual cone of L^n. We can also show that $(y, s) \in L^n$ by assuming conversely that (y, s) is in the dual cone. Hence, the so called Lorentz cone is self dual.

Moreover, for the positive semidefinite cone $\mathcal{S}_{+}^{n}$ we use the inner product $tr(X,Y) = \sum_{i,j=1}^{n} X_{ij}Y_{ij}$. Supposing that $Y \notin \mathcal{S}_{+}^{n}$ there exists a $q \in \mathbb{R}^n$ with

$$q^TYq = tr(qq^TY) < 0$$

Since the positive semidefinite matrix $X = qq^T$ satisfies $tr(XY) \leq 0$ we have that $Y \notin \mathcal{S}_{+}^{n}$. Assuming conversely that $X, Y \in \mathcal{S}_{+}^{n}$ we can use eigenvalue decomposition to show that $Y \in \mathcal{S}_{+}^{n}$. Hence, this cone is self dual[1].

5.3 The Dual of the Conic Programming Problem

Now we derive the dual of the CP problem following the standard Lagrangian theory [11].

Associate with the CP given in (2) a Lagrangian of the form $L : \mathbb{R}^n \times \mathbb{R}^m \times K_* \to \mathbb{R}$

$$L(x, y, s) = c^Tx + y^T(b - Ax) - s^Tx.$$

Minimizing $L(x, y, s)$ over x yields

$$\begin{aligned}\phi(y,s) &= \min_x c^Tx + y^T(b - Ax) - s^Tx = \min_x (c - A^Ty - s)^Tx + b^Ty \\ &= \begin{cases} b^Ty & \text{if } c - A^Ty - s = 0 \\ -\infty & \text{otherwise.}\end{cases}\end{aligned}$$

We maximize $\phi(y, s)$ over all $(y, s) \in \mathbb{R}^m \times K^*$ to acquire the conic dual, or CD, problem as

$$\begin{aligned} \max \quad & b^Ty \\ \text{s.t.} \quad & c - A^Ty = s \\ & s \in K_*. \end{aligned} \tag{3}$$

Similar to LP duality we have

Theorem 15 (Weak Conic Duality Theorem)**.** *The optimal value of CD is a lower bound on the optimal value of CP*

Proof. The proof is straightforward. Let $\bar{x}$ be a feasible solution to (2) and $(\bar{y}, \bar{s})$ be the feasible solution of (3)

$$b^T\bar{y} = (A\bar{x})^T\bar{y} = \bar{x}^T(A^T\bar{y}) = \bar{x}^T(c - \bar{s}) = c^T\bar{x} - \bar{s}^T\bar{x} \leq c^T\bar{x}.$$

The last inequality holds due to the fact that $\bar{s} \in K_*$, i.e. $\bar{s}^T\bar{x} \geq 0$. □

Note that the gap between the dual and primal problem is called the duality gap. In duality theory of LP, the duality gap is 0. Unfortunately, this is not the case for CP although geometrically there is a similarity between CP and LP. However, this can be recovered if we assume a constrained qualification condition. In the case at hand it is in fact the so-called Slater's condition. More precisely we have

Theorem 16. *[6] Consider a conic problem*

$$c^* = \underset{x}{min}\{c^T x : Ax \geq_{\boldsymbol{K}} b\}$$

along with its conic dual

$$b^* = max\{\langle b, \lambda\rangle : A^*\lambda = c, \lambda \geq_{\boldsymbol{K}^*} 0\}.$$

1. *The duality is symmetric: the dual problem is conic, and the problem dual to dual is the primal.*
2. *The value of the dual objective at every dual feasible solution λ is $\leq$ the value of the primal objective at every primal feasible solution x, so that the duality gap*

   $$c^T x - \langle b, \lambda\rangle$$

   is nonnegative at every primal-dual feasible pair (x, λ).
3. (a) *If the primal CP is a bounded below and strictly feasible (i.e. $Ax >_{\boldsymbol{K}} b$ for some x), then the dual CD is solvable and the optimal values in the problems are equal to each other, i.e. $c^* = b^*$.*

   (b) *If the dual CD is bounded above and strictly feasible (i.e. exists $\lambda >_{\boldsymbol{K}} 0$ such that $A^*\lambda = c$, then the primal CP is solvable and $c^* = b^*$.*
4. *Assume that at least one of the problems CP, CD is bounded and strictly feasible. Then a primal-dual feasible pair (x, λ) is a pair of optimal solutions to the respective problems*

   (a) *if and only if*

   $$\langle b, \lambda\rangle = c^T x$$

   and

   (b) *if and only if*

   $$\langle \lambda, Ax - b\rangle = 0$$

This proof is more involved but is not encompassed in this thesis.

5.4 Derivation of the Dual Problem of SOCP

We are interested in the second-order cone program of the form

$$\begin{aligned}&\min\ c^T x\\&\text{s.t. }\ \|A_i x + b_i\|_2 \leq c_i^T x + d_i,\ i = 1, \ldots, m\end{aligned}$$

where $c \in \mathbb{R}^n$, $A_i \in n_i \times n$, $b_i \in \mathbb{R}^{n_i}$, $c_i \in \mathbb{R}^n$, $d_i \in \mathbb{R}$, $i = 1, \ldots, m$ and $x \in \mathbb{R}^n$ is a variable. Here we derive two methods such that the dual can be expressed as

$$\begin{aligned}\max\ &\sum_{i=1}^{m}(b_i^T u_i + d_i v_i)\\ \text{s.t. }\ &\sum(A_i^T u_i + c_i v_i) + c = 0\\ &\|u_i\|_2 \leq v_i,\ i = 1, \ldots, m\end{aligned}$$

with variable $u_i \in \mathbb{R}^{n_i}$, $v_i \in \mathbb{R}$, $i = 1, \ldots, m$. Note that this is once more a second-order conic programming problem.

Method 1 (in terms of conic dual)

We begin from the conic form of the SOCP and use its conic dual, with the fact that the second order cone is self dual, which allows us to express the SOCP as a conic form

$$\begin{aligned}&\min_x\ c^T x && \text{(SOCP)}\\&\text{s.t. }\ -\langle c_i^T x + d_i, A_i x + b_i\rangle \preceq_{K_i} 0,\ i = 1, \ldots, m\end{aligned}$$

where K_i is the second order cone for each i. The Lagrangian is given by

$$\begin{aligned}L(x, u_i, v_i) &= c^T x - \sum_{i=1}^{m}(A_i x + b_i)^T u_i - \sum_{i=1}^{m}(c_i^T x + d_i)v_i\\ &= \left(c - \sum_{i=1}^{m}\left(A_i^T u_i + c_i v_i\right)^T\right)x - \sum_{i=1}^{m}(b_i^T u_i + d_i v_i)\end{aligned}$$

for $\langle u_i, v_i\rangle \succeq_{K_i^*} 0$, or equivalently $v_i \geq \|u_i\|_2$. Minimizing the Lagrangian over x, the dual objective function is

$$\theta(\nu, \mu) = \begin{cases} -\sum_{i=1}^{m}(b_i^T \nu_i + d_i \mu_i) & \text{if } \sum(A_i^T \nu_i + \mu_i c_i) = c\\ -\infty & \text{otherwise}\end{cases}$$

Thus, the dual of the (SOCP) is

$$\begin{aligned} &\max_{\nu_i,\mu_i} -\sum_{i=1}^{n}(b_i^T\nu_i + d_i\mu_i) \\ &\text{s.t. } \sum(A_i^T\nu_i + \mu_i c_i) = c \\ &\qquad \langle \nu_i, \mu_i \rangle \succeq_{K_i} 0,\, i = 1, \ldots, m \end{aligned}$$

Method 2 (reformulation)

Introduce new variables $y_i \in \mathbb{R}^{n_i}$ and $t_i \in \mathbb{R}$ as well as equalities $y_i = A_i x + b_i$ and $t_i = c_i^T x + d_i$. The (SOCP) can be reformulated as

$$\begin{aligned} &\min\, c^T x \\ &\quad \text{s.t } \|y_i\|_2 \leq t_i,\, i = 1, \ldots, m \\ &\qquad y_i = A_i x + b_i,\, t_i = c_i^T x + d_i,\, i = 1, \ldots, m \end{aligned}$$

The Lagrangian with dual variable $\lambda \geq 0$, since it is associated with the inequality constraints ν, μ is

$$\begin{aligned} &L(x, y, t, \lambda, \nu, \mu) \\ &= c^T x + \sum_{i=1}^{m} \lambda_i(\|y_i\|_2 - t_i) + \sum_{i=1}^{m} \nu_i^T(y_i - A_i x - b_i) + \sum_{i=1}^{m} \mu_i(t_i - c_i^T x - d_i) \\ &= (c - \sum_{i=1}^{m} A_i^T \nu_i - \sum_{i=1}^{m} \mu_i c_i)^T x + \sum_{i=1}^{m}(\lambda_i\|y_i\|_2 + \nu_i^T y_i) \\ &\quad + \sum_{i=1}^{m}(-\lambda_i + \mu_i)t_i - \sum_{i=1}^{m}(b_i^T \nu_i + d_i \mu_i). \end{aligned}$$

Since the variables x, t, y are separated in the Lagrangian, the minimization of the Lagrangian over x, t, y can be carried out by three suboptimal problems over x, t and y independently. So we solve each of the optimization problems.

i. The minimum over x is bounded below if and only if

$$\sum_{i=1}^{m}(A_i^T \nu_i + \mu_i c_i) = c.$$

ii. The minimum over t is bounded below if and only if

$$\lambda_i = \mu_i.$$

iii. To minimize over y, we assume $\|\nu_i\|_2 \leq \lambda_i$ and by the Cauchy-Schwarz inequality we receive

$$-\nu_i^T y_i \leq \|\nu_i\|_2 \|y_i\|_2 \leq \lambda_i \|y_i\|_2$$

which implies that

$$\lambda_i \|y_i\|_2 + \nu_i^T y_i \geq 0.$$

Thus, it follows that

$$\inf_{y_i} \lambda_i \|y_i\|_2 + \nu_i^T y_i = 0.$$

On the other hand, assume $\|\nu\|_2 > \lambda_i \geq 0$. Taking $y_i = -s\nu_i$ for some $s > 0$, we have

$$\lambda_i \|y_i\|_2 + \nu_i^T y_i = \lambda_i s \|\nu_i\|_2 - s\|v\|_2^2 = (\lambda_i - \|\nu_i\|_2) s \|v_i\|_2 < 0.$$

To minimize this we let s be very large so that

$$\inf_{y_i} \lambda_i \|y_i\|_2 + \nu_i^T y_i = -\infty.$$

Combining i-iii we have

$$\inf_{y_i} (\lambda_i \|y_i\|_2 + \nu_i^T y_i) = \begin{cases} 0 & \text{if } \|\nu_i\|_2 \leq \lambda_i \\ -\infty & \text{otherwise.} \end{cases}$$

Thus, the dual objective function is

$$\phi(\lambda, \nu, \mu) = \begin{cases} -\sum (b_i^T \nu_i + d_i \mu_i) & \text{if } \sum_{i=1}^m (A_i^T \nu_i + \mu_i c_i) = c, \\ -\infty & \text{otherwise} \end{cases}$$

which leads to the dual problem

$$\begin{aligned} \max \ & -\sum_{i=1}^{m}(b_i^t\nu_i + d_i\mu_i) \\ \text{s.t.} \ & \sum_{i=1}^{m}(A_i^T\nu_i + \lambda_i c_i) = c \\ & \|\nu_i\|_2 \leq \lambda_i, i = 1, \ldots, m \end{aligned}$$

We have now mentioned Conic Programming, its prerequisites as well as how it connects to Linear Programming, and Conic Duality with its theorem and proofs.

The most effective methods for SOCP are interior-point methods, much like any other LP problem. Moreover, having been extensively researched, SOCP is found in numerous applications ranging from finance to engineering to control [7].

6 Unstructured Robust Least Square Problems

Previously we studied the numerical rounding error in solving the least square problem. However, the input data in many problems can have errors. It is therefore natural to ask the following questions:

i. What is a least square solution if there are input data errors?

ii. How robust is the least square problems with uncertainties?

The task of this section is to answer these questions.

For an unknown matrix $[\Delta A, \Delta b]$ define the worst-case residual:

$$\phi(A, b, \rho) \triangleq \max_{\|[\Delta A, \Delta b]\|_F \leq \rho} \|(A + \Delta A)x - (b + \Delta b)\|_2.$$

The robust least square (RLS) problem is to determine the minimum of the worst case residual, i.e. we have to solve the min-max problem

$$\min_x \max_{\|[\Delta A, \Delta b]\|_F \|e\rho\|} \|(A + \Delta A)x - (b + \Delta b)\|_2.$$

When $\rho = 0$, we find a standard least square problem, and when $\rho > 0$, $\phi(A, b, \rho) = \rho\phi(A/\rho, b/\rho, 1)$ why $\rho = 1$ and $\phi(A, b)$ are taken for the remainder of this section. To simplify notation in this section the norms $\|\cdot\|$ are 2-norms unless explicitly stated.

Theorem 17. *When $\rho = 1$, the worst case residual is*

$$r(A, b, x) = \|Ax - b\| + \sqrt{\|x\|^2 + 1}.$$

There is a unique solution to the problem of minimizing $r(A, b, x)$ over $x \in \mathbb{R}^m$.

Proof. For fixed $x \in \mathbb{R}^m$,

$$r(A, b, x) \leq \|Ax - b\| + \sqrt{\|x\|^2 + 1}.$$

Choose $\Delta := [\Delta A, \Delta b]$ as

$$[\Delta A, \Delta b] = \frac{u}{\sqrt{\|x\|^2 + 1}} \begin{bmatrix} x^T & 1 \end{bmatrix}, \text{ where } u = \begin{cases} \frac{Ax-b}{\|Ax-b\|} \text{ if } Ax \neq b \\ \text{any unit-norm vector otherwise} \end{cases}$$

Since the rank of Δ is one, by Proposition 6 we have $\|\Delta\|_F = \|\Delta\| = 1$. Moreover, we have

$$\|(A + \Delta A)x - (b + \Delta b)\| = \|Ax - b\| + \sqrt{\|x\|^2 + 1}$$

implying Δ is a worst case perturbation and the equality always holds in $r(A, b, x)$. Finally, uniqueness of the minimiser x follows from the strict convexity of the worst case residual which is a sum of two strictly convex functions. □

Theorem 16 only says that the RLS problem has a unique solution. Note that the problem of minimizing $r(A, b, x)$ can be formulated as a second order cone. The next theorem provides an exact solution,

Theorem 18. *When $\rho = 1$, the unique solution x_{RLS} to the RLS problem is*

$$x_{RLS} = \begin{cases} (\mu I + A^T A)^{-1} A^T b & \text{if } \mu \overset{\Delta}{=} (\lambda - \tau)/\tau > 0 \\ A^+ b & \text{otherwise.} \end{cases}$$

where (λ, τ) is the unique optimal pair for the second order cone program problem

$$\text{minimize } \lambda \text{ subject to } \|Ax - b\| \leq \lambda - \tau, \quad \left\| \begin{bmatrix} x \\ 1 \end{bmatrix} \right\| \leq \tau. \tag{4}$$

Proof. Using the duality result shown in Section 5.3, the dual problem of (4) is

$$\text{maximize } b^T z - v \text{ subject to } A^T z + u = 0, \|z\| \leq 1, \quad \left\| \begin{bmatrix} u \\ v \end{bmatrix} \right\| \leq 1.$$

All constraints are satisfied and the nonlinear constraints are satisfied with strict inequalities for the primal problem, so both the primal as well as the dual are strictly feasible. By strong duality, this implies that optimal points exist for both. If $\lambda = \tau$ at the optimum, then $Ax = b$ as well as

$$\lambda = \tau = \sqrt{\|x\|^2 + 1}.$$

Here, the optimal x is the unique minimum norm solution to $Ax = b$, that is $x = A^+ b$.

Assume instead that $\lambda > \tau$. Since the same condition for the primal and dual holds here as above, the primal and dual optimal objectives are equal

$$\|Ax - b\| + \left\| \begin{bmatrix} x^T & 1 \end{bmatrix} \right\| = \lambda = b^T z - v = -(Ax - b)^T z - \begin{bmatrix} x^T & 1 \end{bmatrix} \begin{bmatrix} -A^T z \\ v \end{bmatrix}.$$

With $\|z\| \leq 1$, $\| \begin{bmatrix} u^T & v \end{bmatrix}^T \| \leq 1$, $u = -A^T z$, by inspection we have

$$z = -\frac{Ax - b}{\|Ax - b\|} \text{ and } \begin{bmatrix} u^T & v \end{bmatrix} = -\frac{\begin{bmatrix} x^T & 1 \end{bmatrix}}{\sqrt{\|x\|^2 + 1}}.$$

Thus

$$\begin{aligned} u &= -\frac{x}{\sqrt{\|x\|^2 + 1}} \\ v &= -\frac{1}{\sqrt{\|x\| + 1}}. \end{aligned}$$

Plugging expressions for z and u in the constraint $A^T z + u = 0$ yields

$$\frac{A^T (Ax - b)}{\|Ax - b\|} + \frac{x}{\sqrt{\|x\|^2 + 1}} = 0.$$

Rearranging terms in this equation we receive

$$(A^TA + \mu I)x = A^Tb$$

with

$$\mu = \frac{\|Ax - b\|}{\sqrt{\|x\|^2 + 1}} = \frac{\lambda - \tau}{\tau}.$$

Hence

$$x = (A^TA + \mu I)^{-1}A^Tb$$

□

Now we try to answer the question on robustness of LS. To this end we make use of the SVD of the matrix A, i.e. there are unitary matrices U and V such that

$$A = U \begin{bmatrix} \Sigma_1 & 0 \\ 0 & 0 \end{bmatrix} V^*$$

with $\Sigma_1 = \text{diag}(\sigma_1, ..., \sigma_r)$ and $\sigma_1 \geq \cdots \geq \sigma_r > 0$ being singular values of A. Note that Σ_1 is an $r \times r$ invertible matrix.

Let $U^*b = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}$. Assume $\lambda > \tau$ at the optimal solution of (4). Since the primal and dual optimal values are the same we have

$$\lambda = b^Tz - v = \frac{b^T(b - Ax)}{\|Ax - b\|} + \frac{1}{\sqrt{\|x\|^2 + 1}} = \frac{b^T(b - Ax)}{\tau} + \frac{1}{\tau},$$

as shown in the proof of Theorem 17. Now we compute the numerator of the first term using the SVD and the formula for the RLS solution x_{RLS} of

$$\begin{aligned}
&b^T(b - Ax) = b^T(b - A(A^TA + \mu I)^{-1}A^Tb) = b^T(I - A(A^TA + \mu I)^{-1}A^T)b \\
=&b^T \left(I - U \begin{bmatrix} \Sigma_1 & 0 \\ 0 & 0 \end{bmatrix} V^* \left(V \begin{bmatrix} \Sigma_1 & 0 \\ 0 & 0 \end{bmatrix}^T U^*U \begin{bmatrix} \Sigma_1 & 0 \\ 0 & 0 \end{bmatrix} V^* + \mu I \right)^{-1} V \begin{bmatrix} \Sigma_1 & 0 \\ 0 & 0 \end{bmatrix}^T U^* \right) b \\
=&b^T \left(I - U \begin{bmatrix} \Sigma_1 & 0 \\ 0 & 0 \end{bmatrix} \left(\begin{bmatrix} \Sigma_1 & 0 \\ 0 & 0 \end{bmatrix}^T \begin{bmatrix} \Sigma_1 & 0 \\ 0 & 0 \end{bmatrix} + \mu I \right)^{-1} \begin{bmatrix} \Sigma_1 & 0 \\ 0 & 0 \end{bmatrix}^T U^* \right) b \qquad . \\
=&b^TU^* \left(I - \begin{bmatrix} \Sigma_1(\mu I + \Sigma_1^2)^{-1}\Sigma_1 & 0 \\ 0 & 0 \end{bmatrix} \right) U^*b \\
=& \begin{bmatrix} b_1^* & b_2^* \end{bmatrix} \begin{bmatrix} I - \Sigma_1(\mu I + \Sigma_1^2)^{-1}\Sigma_1 & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} \\
=& \begin{bmatrix} b_1^* & b_2^* \end{bmatrix} \begin{bmatrix} (I + \mu^{-1}\Sigma_1^2)^{-1} & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} = b_2^*b_2 + b_1^*(I - \mu^{-1}\Sigma_1^2)^{-1}b_1
\end{aligned}$$

Hence

$$\frac{b^T(b-Ax)}{\tau} = \frac{b_2^* b_2}{\lambda-\tau} + b_1^*((\lambda-\tau)I + \tau\Sigma_1^2)^{-1}b_1.$$

Consequently

$$\lambda = \frac{1}{\tau} + \frac{b_2^* b_2}{\lambda-\tau} + b_1^*((\lambda-\tau)I + \tau\Sigma_1^2)^{-1}b_1.$$

This leads to

$$\lambda^2 = \frac{\lambda}{\tau} + \frac{\lambda b_2^* b_2}{\lambda-\tau} + \lambda b_1^*((\lambda-\tau)I + \tau\Sigma_1^2)^{-1}b_1$$

because $\lambda = 0$ is not feasible. Introduce now $\theta = \frac{\tau}{\lambda}$. Then

$$\lambda^2 = \frac{1}{\theta} + \frac{b_2^* b_2}{1-\theta} + b_1^*((1-\theta)I + \theta\Sigma_1^2)^{-1}b_1 =: f(\theta) \tag{5}$$

From the constraints of (4) we have $\tau \geq 1$ and $\lambda \leq \|b\| + 1$, which implies that $\theta \geq \theta_{\min} := \frac{1}{\|b\|+1}$. That is $\theta_{\min} \leq \theta \leq 1$.

These computations show that we have reduced the optimization problem to determine the worst-case residual to the following optimization problem of a one variable function

$$\inf_{\theta_{\min}\leq\theta\leq 1} f(\theta)$$

Taking a closer look at the function $f(\theta)$, the point $\theta = 1$ needs more attention. However, $\theta = 1$ is equivalent to $\lambda = \tau$ and we know that the optimum $\lambda = 1 + \|x\|^2 = 1 + \|A^+ b\|^2$ for $b \in \mathcal{C}(A)$ by the proof of Theorem 17. It is clear that $f(\theta) \to \infty$ if $b \notin \mathcal{C}(A)$. Furthermore, if we backward substitute SVD of A into $f(\theta)$ we obtain for $\theta_{\min} \leq \theta < 1$

$$f(\theta) = \frac{1}{\theta} + b^T((1-\theta)I + \theta AA^T)^{-1}b.$$

Observe that

$$f(\theta) = \frac{1}{\theta} + \frac{\|b_2\|^2}{1-\theta} + \sum_{i=1}^{r} \frac{b_{1,i}^2}{1+\theta(\sigma_i^2-1)}$$

is a twice differentiable function and its first and the second derivatives are

$$\frac{df}{d\theta} = -\frac{1}{\theta^2} + \frac{\|b_2\|^2}{(1-\theta)^2} + \sum_{i=1}^{r} \frac{b_{1,i}^2((1-\sigma_i^2)}{(1+\theta(\sigma_i^2-1))^2}$$

and

$$\frac{d^2f}{d\theta^2} = \frac{2}{\theta^3} + \frac{2}{(1-\theta)^3} + \sum_{i=1}^{r} \frac{2b_{1,i}^2((1-\sigma_i^2)^2}{(1+\theta(\sigma_i^2-1))^3}.$$

We know that $\theta < 1 \Leftrightarrow (1-\theta)^3 > 0$ and $1+\theta(\sigma_i^2-1) = (1-\theta)+\theta\sigma_i^2 \geq 0$ for all $i = 1, ..., r$ proving that $\frac{d^2 f}{d\theta^2} > 0$. This shows that $f(\theta)$ is a strictly convex function for $\theta_{\min} \leq \theta < 1$ and consequently the necessary and sufficient condition for the minimizer θ^* is that θ^* is a solution (if any) to the equation

$$-\tfrac{1}{\theta^2} + \tfrac{\|b_2\|^2}{(1-\theta)^2} + \textstyle\sum_{i=1}^{r} \tfrac{b_{1,i}^2((1-\sigma_i^2)}{(1+\theta(\sigma_i^2-1))^2} = 0 \;\Leftrightarrow\; \tfrac{1}{\theta^2} = \tfrac{\|b_2\|^2}{(1-\theta)^2} + \textstyle\sum_{i=1}^{r} \tfrac{b_{1,i}^2((1-\sigma_i^2)}{(1+\theta(\sigma_i^2-1))^2}.$$

Hence we have proved the following theorem.

Theorem 19. *Assume $\rho = 1$. Then the solution of the unstructured RLS can be derived from the solution to the convex optimization problem of one variable*

$$\inf_{\theta_{\min} \leq \theta \leq 1} f(\theta)$$

where

$$f(\theta) = \begin{cases} \frac{1}{\theta} + b^T((1-\theta)I + \theta AA^T)^{-1} b & \text{if } \theta_{\min} \leq \theta < 1 \\ \begin{cases} \infty & \text{if } b \notin \mathcal{C}(A) \\ 1 + \|A^+ b\|^2 & \text{if } b \in \mathcal{C}(A) \end{cases} & \text{if } \theta = 1 \end{cases}.$$

Moreover the worst-case residual is

$$\phi(A, b)^2 = \inf_{\theta_{\min} \leq \theta \leq 1} f(\theta).$$

Some remarks are in order. If an SVD is available then Theorem 19 gives an alternative solution to the RLS and it can be solved, for example, by Newton's algorithm efficiently. As commonly known, an SVD requires a cost of about $O(nm^2 + m^3)$ which is not much cheaper than the SOCP method. Moreover, the SVD does not extend to the structured RLS, for example the robust polynomial interpolation as discussed later. One of the advantages of this solution is that it provides an easier robustness analysis as we will now show.

Intuitively, when RLS and LS solutions coincide we can say that the LS solution is robust. This occurs if and only if $\theta = 1$ and $b \in \mathcal{C}(A)$, In this case f has to be differentiable at $\theta = 1$ and its minimum over $\theta_{\min} \leq \theta \leq 1$ is at $\theta = 1$ if and only if $\frac{df}{d\theta}(1) \leq 0$.

Now

$$\begin{aligned} \frac{df}{d\theta} &= -\frac{1}{\theta^2} - b^T((1-\theta)I + \theta AA^T)^{-1} \frac{d((1-\theta)I + \theta AA^T)}{d\theta} ((1-\theta)I + \theta AA^T)^{-1} b \\ &= -\frac{1}{\theta^2} - b^T((1-\theta)I + \theta AA^T)^{-1}(AA^T - I)((1-\theta)I + \theta AA^T)^{-1} b. \end{aligned}$$

Then

$$\frac{df}{d\theta}(1) = -1 - b^T(AA^T)^+(AA^T - I)(AA^T)^+b.$$

It requires that it be ≤ 0 for $b \in \mathcal{C}(A)$, that is,

$$b \in \mathcal{C}(A),\ b^T((AA^T)^2)^+b \leq 1 + b^T(AA^T)^+b. \tag{6}$$

If (6) holds then the RLS and LS coincide. Otherwise the optimal $\theta < 1$ and $x = x_{RLS}$ given in Theorem 17.

If not, $\theta < 1$ and x is given by x_{RLS} that we defined above. We can write this latter condition in the case when the norm-bound of the perturbation ρ is different from 1 as $\rho > \rho_{min}$, where

$$\rho_{min}(A,b) \triangleq \begin{cases} \frac{\sqrt{1+\|A^+b\|^2}}{\|(AA^T)^+b\|} & \text{if } b \in \mathcal{C}(A),\ A \neq 0,\ b \neq 0 \\ 0 & \text{otherwise} \end{cases}$$

Hence, ρ_{min} can be thought of as the perturbation level that the LS solution allows. Thus, the LS and RLS solutions coincide whenever the norm-bound on the perturbation matrix ρ satisfies $\rho \leq \rho_{min}(A,b)$, where $\rho_{min}(A,b)$ is defined as above. Hence, $\rho_{min}(A,b)$ can be thought of as the robustness measure of the LS solution.

We regard the above mentioned example in numerical fashion.

A numerical example

We shall now compute the perturbation level of Example 2. If $y_i = y(s_i)$, $s_i = s_0 + i$, $i = 1, \ldots, 10$. Since $Ax = b$, $b \in \mathcal{C}(A)$, the LS is naturally robust with

$$\rho_{\min}(A,b) = \frac{\sqrt{1+\|A^+b\|^2}}{\|(AA^T)^+b\|}$$

We now calculate $\rho_{\min}$ for $s_0 = 10,\ 50,\ 100,\ 150$, and 200 as well as the solution to x_{RLS} for the RLS. To get this solution we need to solve the SOCP problem

$$\begin{aligned} &\min \lambda \\ &\text{s.t } \|Ax - b\| \leq \lambda - \tau,\ \left\| \begin{bmatrix} x \\ 1 \end{bmatrix} \right\| \leq \tau \end{aligned}$$

With the help of Mathematica we obtain

s_0	10	50	100	150	200
λ	1.20847	1.05663	1.02984	1.02027	1.01535
τ	1.02416	1.00164	1.00045	1.00021	1.00045
$\mu = \frac{\lambda-\tau}{\tau}$	0.1800	0.054897	0.029381	0.0200602	0.014893
ρ_{min}	7.6×10^{-5}	9.8×10^{-8}	1.5×10^{-5}	4.8×10^{-6}	2.1×10^{-6}
$\mathcal{X}_{RLS}$	$\begin{bmatrix} 0.2206 \\ 0.0159 \\ 0.0012 \end{bmatrix}$	$\begin{bmatrix} 0.05730 \\ 0.00104 \\ 0.00002 \end{bmatrix}$	$\begin{bmatrix} 3.0 \times 10^{-2} \\ 2.9 \times 10^{-4} \\ 2.7 \times 10^{-6} \end{bmatrix}$	$\begin{bmatrix} 2.0 \times 10^{-2} \\ 1.3 \times 10^{-4} \\ 8.4 \times 10^{-7} \end{bmatrix}$	$\begin{bmatrix} 1.6 \times 10^{-2} \\ 7.6 \times 10^{-5} \\ 3.7 \times 10^{-7} \end{bmatrix}$

From above we see that ρ_{min}, considered as the perturbation level which the LS solution allows, is decreasing as the condition number of A increases, except for $s_0 = 50$ when λ is very close to τ.

Simplifying the above observation, we find a gradually less robust solution as the condition number increases.

7 Robust Polynomial Interpolation

In this section we will show that some least square problems are structured and hence it is desired to take the structure into account. We illustrate this by polynomial interpolation.

For some integer k and $n \geq 1$, we search for a polynomial of degree $n-1$ such that

$$p(t) = x_1 + x_2 t + \ldots + x_n t^{n-1}$$

that interpolates given data (a_i, b_i), $i = 1, \ldots, k$, or more precisely,

$$p(a_i) = b_i,\ i = 1, \ldots k.$$

Let us first assume that (a_i, b_i) are known. So the above interpolation conditions gives the linear equations

$$\begin{bmatrix} 1 & a_1 & \ldots & a_1^{n-1} \\ 1 & a_2 & \ldots & a_2^{n-1} \\ \vdots & \ldots & \ldots & \vdots \\ 1 & a_k & \ldots & a_k^{n-1} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ \vdots \\ \vdots \\ b_k \end{bmatrix}$$

where $x = (x_1, \ldots, x_n)^T$ is unknown.

If $n = k$, then this becomes a system of equations with a Vandermonde matrix as its coefficient matrix A. Assume further that the a_i's are distinct. Then det(A) $= \prod_{1 \leq i \leq j \leq n}(a_j - a_i) \neq 0$ if $a_i \neq a_j$ for $i \neq j$. So we obtain a unique solution x and hence a unique interpolating polynomial $p(t)$.

If instead $k > n$, then we solve this by the standard least square method. Important to note is that the condition numbers often tend to be very poor if some a_i's are close to one another.

Here, we consider the case where the interpolation data is known but not exact. For example, assume b_i are known exactly, but a_i are parameter dependent of the form

$$a_i(\delta) = a_i + \delta_i,\ i = 1, \ldots, k$$

where the δ_i's are bounded but unknown: $|\delta_i| \leq \rho$, $i = 1, \ldots k$ where $\rho \geq 0$ is given. When $k = n$ we apply Theorem 8 if the δ_i's are considered round-off errors and ρ as machine-epsilon. When $k \geq n$ in general, we apply robust least square above. Thus, we look for a robust interpolant $p(t)$ such that x minimises the worst case residual

$$\max_{\|\delta\|_\infty \leq \rho} \|A(\delta)x - b\|$$

where

$$A(\delta) = \begin{bmatrix} 1 & a_1(\delta) & a_1(\delta)^2 & \ldots a_1(\delta)^{n-1} \\ 1 & a_2(\delta) & a_2(\delta)^2 & \ldots a_2(\delta)^{n-1} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & a_k(\delta) & a_k(\delta)^2 & \ldots a_k(\delta)^{n-1} \end{bmatrix}.$$

We see that this matrix has a Vandermonde structure. We regard the structure of this particular least square problem. We can rewrite $A(\delta)x - b$ as

$$A(\delta)x - b = [A(\delta), b] \begin{bmatrix} x \\ -1 \end{bmatrix}$$

We show that $[A(\delta), b]$ can be rewritten into linear fractional form

$$[A(\delta), b] = [A(0), b] + L\Delta(I - D\Delta)^{-1}[R_A, 0]$$

where L, Δ, D, R_A shall be given precisely.

For a fixed row index i, we have

$$\begin{aligned}&(1,(a_i+\delta_i),(a_i+\delta_i)^2,\dots,(a_i+\delta_i)^{n-1},b_i)\\ =&(1,a_1,a_i^2,\dots,a_i^{n-1},b_i)+L_i\Delta_i(I_{n-1}-D_i\Delta_i)^{-1}(R_i,0)\end{aligned}$$

where

$$D_i=\begin{bmatrix}0 & 1 & a_i & \dots & a_i^{n-3}\\ & \ddots & \ddots & \ddots & \vdots\\ & & \ddots & \ddots & a_i\\ & & & \ddots & 1\\ & & & & 0\end{bmatrix}\in\mathbb{R}^{(n-1)\times(n-1)},\ i=1,\dots k,$$

$$R_i=\begin{bmatrix}0 & 1 & a_i & \dots & a_i^{n-2}\\ & \ddots & \ddots & \ddots & \vdots\\ & & \ddots & \ddots & a_i\\ & & & \ddots & 1\\ & & & & 0\end{bmatrix}\in\mathbb{R}^{(n-1)\times n},\ i=1,\dots k.$$

Stack rows, we get

$$[A(\delta),b]=[A(0),b]+L\Delta(I-D\Delta)^{-1}[R_A,0]$$

where $L=\operatorname{diag}(L_1,\dots,L_k)$, $\Delta=\operatorname{diag}(\Delta_1,\dots,\Delta_k)$, $D=\operatorname{diag}(D_1,\dots D_k)$, $R_A=\begin{bmatrix}R_1\\ \vdots\\ R_k\end{bmatrix}$ and $L_i=(1,a_i,\dots a_i^{n-2})$, $\Delta_i=\delta_i I_{n-1}$.

Note that since D is strictly upper triangular, $(I-D\Delta)$ is invertible. Hence, this is the linear fractional structured robust least square problem. The computation of the worst-case residual is NP-complete, however positive semidefinite programming can provide upper bounds, see [8].

References

[1] L. Vandenberghe and S. Boyd. *Convex Optimization* Cambridge University Press, 2004

[2] J. Stoer and R. Bulirsch. *Introuduction to Numerical Analysis.* Springer, New York, NY, 1993.

[3] Y. Ye. *Conic Linear Programming.* `https://web.stanford.edu/class/msande314/sdpmain.pdf`

[4] T. Anderson and N. D'Addio. *Duality in Conic Programming* `https://people.smp.uq.edu.au/YoniNazarathy/teaching_projects/studentWork/Duality.pdf`

[5] L. El Ghaoui and H. Lebret. *Robust Solutions to Least-Squares Problems with Unceratin Data.* SIAM Journal on Matrix Analysis and Applications vol 18 (1997), pp. 1035-1064

[6] A. Ben-Tal and A. Nemirovksi. *Lectures on Modern Convex Optimization* `https://www2.isye.gatech.edu/~nemirovs/LMCO_LN.pdf` H. Milton Stewart School of Industrial & Systems Engineering, Georgia Institute of Technology, 2019

[7] *Second-order Cone Programming (SOCP)* `https://www.nag.com/content/second-order-cone-programming-socp-0`

[8] El Ghaoui, Oustry and Leberet. *Robust solutions to uncertain semidefinite programs.* SIAM Journal of Optimization, vol 9 (1998), pp. 33-52

[9] Dahlquist and Björck. *Numerical Methods.* Dover Publications, New York, 1974

[10] W. Prager and W. Oettli. *Compatibility of approximate solution of linear equations with given error bounds for coefficients and right hand sides.* Numerische Matematik 6, pp. 405-409, 1964

[11] Bazaraa et. al. *Nonlinear Programming, Theory & Algorithms.* John Wiley and Sons Ltd., 1979