UNIVERSITY of CALIFORNIA
Santa Barbara

A New Look at Robust Estimation and Identification

A Dissertation submitted in partial satisfaction of the requirements for the degree

Doctor of Philosophy
in
Electrical and Computer Engineering

by

Keith Evan Schubert

Committee in charge:

Professor Shivkumar Chandrasekaran, Chair
Professor Ian Rhodes
Professor John Shynk
Professor Roy Smith

September 2003

The dissertation of Keith Evan Schubert is approved.

Professor Ian Rhodes

Professor John Shynk

Professor Roy Smith

Professor Shivkumar Chandrasekaran, Committee Chair

July 2002

A New Look at Robust Estimation and Identification

Acknowledgements

I want to thank my committee for all their advice that improved the quality of this dissertation tremendously. I cannot thank them enough for all the time and help they have given me. In particular I would like to thank Shiv for all of the long talks we have had on research and teaching. I would also like to thank Roy for his careful editing of this text.

Thanks to Terry Kupfer, who read and corrected my spelling and grammatical errors.

My wife Kym, and my children Carl and Victoria, have my deepest thanks for encouraging me and giving me the time to finish this.

All this would be for naught if it were not for the guiding hand of the Almighty, without whom I can do nothing.

Curriculum Vitæ

Keith Evan Schubert

Education

B.S. General Engineering, University of Redlands, 1991

M.S. Electrical Engineering, UCLA, 1992

Ph.D. Electrical and Computer Engineering, UCSB, 2003 (expected)

Professional

1992–1996	Associate Engineer, Northrop-Grumman
1996–2000	Teaching Assistant & Graduate Research Assistant, UCSB
2000–2002	Visiting Lecturer in Mathematics and Computer Science, University of Redlands
2002–	Assistant Professor in Computer Science, CSUSB

Publications

S. Chandrasekaran and M. Gu and A. H. Sayed and K. E. Schubert, "The Degenerate Bounded Errors-In-Variables Model", SIMAX, 23(1):138-166, 2001.

S. Chandrasekaran and K. E. Schubert, "Models for Robust Estimation and Identification", in S. Van Huffel and P. Lemmerling, editors, *Total Least Squares and Errors-In-Variables Modeling*, pages 199-208, Kluwer Academic Publishers, Dordrecht, 2001.

Abstract

A New Look at Robust Estimation and Identification

by

Keith Evan Schubert

Estimation and identification are important areas of almost every problem in science and engineering. A typical way of stating an estimation or identification problem is that there is a system, described by a matrix, A, with inputs, x, and outputs, b. The inputs and outputs could either be matrices or vectors. The equation which describes this is thus $Ax = b$. The outputs of the system are considered measurable, and from them and the matrix A, it is desired to find the unknown inputs, x. In real systems the equality rarely holds because b is never measured perfectly, modeling and identification do not produce an exact A, and the basic equation $Ax = b$ is a linear approximation. The fundamental problem considered is thus $Ax \approx b$, where both A and b are assumed to have errors associated with them.

This Dissertation proposes five regression methods to handle the fundamental problem of $Ax \approx b$. In particular, let the "true" system, A_{true}, be related to the nominal model, A, by an error matrix E_A. Similarly let the true outputs, b_{true} be related to the measured outputs, b, by E_b. Since the true system is not a mathematical model, the resulting equation is still approximate, $(A + E_A)x \approx (b + E_b)$, but is the best approximation possible. The goal is to find the best x, in the resulting minimization problem, $\min_x \|(A+E_A)x-(b+E_b)\|$. Each of the

five methods in this dissertation consider this problem and makes assumptions on the size and structure of the errors, E_A and E_b. All problems are solved using secular equation techniques, so finding the solution corresponds to finding the zero of a possibly multi-dimensional secular equation.

The first three methods are extensions of min max regression, which minimizes the cost over x and maximizes it over E_A and E_b. The fourth method is the degenerate case (multiple solutions) of min min regression, which minimizes over x, and the errors E_A and E_b. The fifth is actually a family of regression problems with rational cost functions based off the backward error criterion of Numerical Analysis.

Contents

List of Tables

List of Figures

Pilate therefore said unto him, Art thou a king then? Jesus answered, Thou sayest that I am a king. To this end was I born, and for this cause came I into the world, that I should bear witness unto the truth. Every one that is of the truth heareth my voice. Pilate saith unto him, What is truth? And when he had said this, he went out again unto the Jews, and saith unto them, I find in him no fault at all.

John 18:37-38 (KJV)

Chapter 1

Introduction

The subject of this dissertation is the estimation of unknowns that are related to some measurements by a linear model that is subject to uncertainty. This introduction will examine the mathematical preliminaries, the uncertainty to be considered, and the solution techniques proposed. The mathematical preliminaries needed to discuss the problems examined in this dissertation are presented in Section 1.1. A brief look at why uncertainty exists in all models is examined in Section 1.2. A simple example, which demonstrates the need to consider uncertainty, is provided in Section 1.3. Finally, the five methods proposed are introduced in Section 1.4.

1.1 Mathematical Preliminaries

The space of real numbers is denoted, $\mathbb{R}$, and the corresponding n dimensional space of real numbers is denoted by the n-tuple of real numbers, $\mathbb{R}^n$. The space of real valued matrices with m rows and n columns is denoted, $\mathbb{R}^{m\times n}$. Similarly, the space of complex numbers is denoted, $\mathbb{C}$, and the corresponding n dimensional space of complex numbers is denoted by the n-tuple of complex numbers, $\mathbb{C}^n$. The space of complex valued matrices with m rows and n columns is denoted, $\mathbb{C}^{m \times n}$. Since the extension from real to complex numbers is straightforward, this dissertation will describe the problems in terms of real numbers.

Consider the set of linear equations, $Ax = b$, where $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$ are given. The goal is to calculate the value of $x \in \mathbb{R}^n$. If the equation is exact and A is not singular, the solution can be readily found by a variety of techniques, such as taking the QR factorization of A.

$$
\begin{aligned}
Ax &= b \\
QRx &= b \\
Rx &= Q^T b
\end{aligned}
$$

The last equation can be solved for x by back-substitution, since R is upper triangular. Given errors in modeling, estimation, and numeric representation the equality rarely holds. The least squares technique directly uses techniques like the QR factorization, by considering all the errors to be present in b. A more realistic appraisal of the system, considers errors in both A and b. Numerous methods exist for describing the errors in A and b, such as

1. bounding the norms of the errors in A and b,
2. constraining the errors in A and b to some structure,
3. partitioning A, b, and their corresponding errors, then placing bounds on the norms of each partition of the errors.

Combinations of the methods to describe the errors are also considered by some techniques. The description of the errors is one of the two fundamental ways a technique is specified for the linear model $Ax \approx b$. The other fundamental way of describing a method is to specify the cost function used to select the best value for x.

Least squares considers the cost function, $\min_x \|Ax - b\|$. Total least squares minimizes the errors in A and b (say E_A and E_b) subject to the resulting system being consistent, thus

$$\min \left\| \begin{matrix} E_A & E_b \end{matrix} \right\|_F$$

s.t.

$$(A + E_A)x = b + E_b.$$

Other techniques consider directly minimizing the norm of $(A + E_A)x - (b + E_b)$ subject to some description of the errors as described above. For specialized situations other cost functions and error descriptions are considered, such as rational cost functions. Chapter 2 presents an overview of the current techniques in estimation, and how they relate to the techniques proposed.

1.2 Uncertain Models

Mathematical models of real world systems are used to quantify system parameters and simulate system behavior. This dissertation examines the problem of uncertainty in a model that is to be used to perform estimation or identification. Uncertainty is unavoidable, and to some extent, everyone who does estimation or identification must consider the effects of it. Lennart Ljung discusses this basic principle in [93], when he speaks of the fiction of a true system. Ljung says,

> The real-life actual system is an object of a different kind than our mathematical models. In a sense, there is an impenetrable but transparent screen between our world of mathematical descriptions and the

> real world. We can look through this window and compare certain aspects of the physical system with its mathematical description, but we can never establish any exact connection between them. The question of nature's susceptibility to mathematical description has some deep philosophical aspects, and in practical terms we have to take a more pragmatic view of models. Our acceptance of models should thus be guided by "usefulness" rather than "truth."

Ljung is noting that there is not an exact model of a real-life actual system but useful models exist. The challenge lies in defining what is meant by usefulness, and determining how to obtain a solution from the model, which of necessity contains uncertainty. Usefulness is dependent on the problem and goals. A necessary result of this challenge is that different models are needed, and no one technique will be best in every situation. This dissertation proposes five new techniques that are useful in solving ill-conditioned systems with special conditions on the errors.

1.3 A Simple Example

It is reasonable to ask how much can uncertainty affect a real system. Consider, for example the simple system described by

$$Ax = b,$$

with

$$\begin{aligned} A &= \begin{bmatrix} 0.11765 & 0.12909 \\ -0.24957 & -0.26919 \end{bmatrix} \\ b &= \begin{bmatrix} -0.074888 \\ 0.154728 \end{bmatrix}. \end{aligned}$$

For this exact system the solution is given by

$$x = \begin{bmatrix} 0.34 \\ -0.89 \end{bmatrix}.$$

This is a nice system with reasonable condition number, but if A matrix is rounded to two decimal places,

$$A = \begin{bmatrix} 0.12 & 0.13 \\ -0.25 & -0.27 \end{bmatrix},$$

the new solution is

$$x = \begin{bmatrix} 1.0505 \\ -1.5457 \end{bmatrix}.$$

The best that can be said about this is that the signs of the solution are correct. This illustrates that even innocent looking systems can exhibit bad behavior in normal situations. What can be done? Consider the normal equations which describe the least squares solution,

$$A^T Ax = A^T b.$$

The difficulty exists in the $A^T A$ term. The condition number of this term can be much worse than that of A, which causes problems when it is inverted to find the solution. A standard way of dealing with this is to use a regularized solution, in which a diagonal matrix is added to the $A^T A$ term to make the inversion easier and more accurate. For example, the general form of the regularized solution,

$$x(\psi) = \left(A^T A + \psi I\right)^{-1} A^T b, \tag{1.1}$$

with $\psi = 10^{-7}$ yields a solution of

$$x\left(10^{-7}\right) = \begin{bmatrix} 0.21515 \\ -0.77273 \end{bmatrix}.$$

This is closer to the true solution, but can the selection of the regularization parameter be automated? Examining the one parameter family given in Equation 1.1 to find the one closest to the true system would be ideal, but that requires knowing the answer a priori. Note also that frequently the exact solution is not as important as the residual, which needs to be taken into account.

1.4 Proposed Formulations

Five formulations are proposed in this dissertation, four are extensions of earlier works, and one is a new direction. One current technique that is extended is referred to as robust least squares in [62], bounded data uncertainties in [24], and for neutrality shall be referred to as the min max technique (since both a minimization and maximization is performed on the cost function). The other current technique that is extended is referred to as a bounded errors-in-variables model in [25], and will be referred to as the min min technique (as two minimizations are performed on the cost function) here to be more specific and consistent with the min max problem above. The five formulations are:

1. Multiple (block) column partitioned min max (multi-col min max)
2. Multiple (block) row partitioned min max (multi-row min max)
3. General (block) partitioned min max (general min max)
4. Degenerate min min
5. Min max backward error

The first problem is the multiple (block) column partitioning case for the min max, which is handled in Chapter 3. In this problem A and E_A are considered to be partitioned into block columns (A_j and $E_{A,j}$), and the norm of each partition of the error, E_A, is assigned a bound. The problem can be thought of as an order updating problem if partitioned into two blocks. For multiple block columns, this method is useful in tracking multiple targets on radar arrays, or working with inverse problems such as those in seismology. The problem, for p block partitions, is given as

$$\min_x \max_{\substack{\|E_j\|_2 \leq \eta_j \\ \|E_b\|_2 \leq \eta_b}} \left\| \begin{bmatrix} A_1 + E_1 & \cdots & A_p + E_p \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_p \end{bmatrix} - (b + E_b) \right\| \tag{1.2}$$

where $j = 1, 2, \ldots, p$. First, key simplifications from the original min max problem are shown to be no longer true, thus demonstrating the increased difficulty of the problem. Second, it will be shown that this problem can be expressed as an identical problem in which the maximization has already been done. A quadratically convergent algorithm is presented for finding the solution using the new formulation. The new formulation is then used to show the form of the solution, which allows the problem to be reduced to finding the zeros of a p dimensional secular equation. Several lemmas to characterize important properties of the problem are developed, and the existence and uniqueness of the solution is proven. A numerical example is given to show the performance.

The second problem is the multiple (block) row partitioning case for the min max, which is handled in Chapter 4. This is similar to the multiple column case above, with the obvious distinction that A and E_A have been partitioned into block rows (A_i and $E_{A,i}$) rather than block columns. The bounds are thus placed on the norms of the block row partitions of E_A. The problem, for q block partitions, is given as

$$\min_x \max_{\substack{\|E_i\|_2 \leq \eta_i \\ \|E_{b,i}\|_2 \leq \eta_{b,i}}} \left\| \begin{bmatrix} A_1 + E_1 \\ \vdots \\ A_q + E_q \end{bmatrix} [x] - \begin{bmatrix} b_1 + E_{b,1} \\ \vdots \\ b_q + E_{b,q} \end{bmatrix} \right\| \tag{1.3}$$

where $i = 1, 2, \ldots, q$. The maximization is performed to obtain an equivalent formulation and this is used to show the form of solution when the solution is at a differentiable point. The secular equation is designed for finding the regression parameter. The non-differentiable points are discussed and techniques for telling when the solution is at a non-differentiable point are covered. An algorithm showing how to implement the solution is presented and conclusions are drawn.

The third problem is the general (block) partitioning case for the min max, which is handled in Chapter 5. This problem combines the row and column partitioning as special sub-cases. In this problem both the columns and rows of A and E_A are partitioned into blocks ($A_{i,j}$ and $E_{i,j}$), and the norm of each $E_{i,j}$ are assigned distinct bounds. The problem is

$$\min_x \max_{\substack{\|E_{i,j}\|_2 \leq \eta_{i,j} \\ \|E_{b,i}\|_2 \leq \eta_{b,i}}} \left\| \begin{bmatrix} A_{1,1} + E_{1,1} & \ldots & A_{1,p} + E_{1,p} \\ \vdots & \ddots & \vdots \\ A_{q,1} + E_{q,1} & \ldots & A_{q,p} + E_{q,p} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_p \end{bmatrix} - \begin{bmatrix} b_1 + E_{b,1} \\ \vdots \\ b_q + E_{b,q} \end{bmatrix} \right\| \tag{1.4}$$

where $i = 1, 2, \ldots, q$ and $j = 1, 2, \ldots, p$. It allows perturbation bounds on individual blocks of the min max problem, which is particularly useful if the matrix, A comes from different sources, or has modeling errors that are non-uniform. Note that this case will not exactly handle structure but can approximate it. If the structure is the main goal, the structured LMI technique discussed in Section 2.10 is the best choice, as it can handle structure exactly. The maximization has been performed, the form of solution for differentiable points has been found, and the secular equation to calculate the regression parameter is provided. The differentiable points can be handled similarly to the row case, and the basic algorithm presented there holds for the general block case. A fixed point method for finding the solution is presented as an alternative, which performs well in practice. Results of many numerical runs on random matrices are presented to demonstrate the behavior of the method.

The fourth problem, covered in Chapter 6, is the degenerate version of the min min problem presented in [25]. The degenerate version is the more difficult and more general extension of that paper's problem. This problem has been completely solved and has been published by SIMAX as [26]. The problem is

$$\min_{x \in \mathcal{R}^n} \min_{\|E\| \leq \eta} \|(A + E_A)x - b\| \tag{1.5}$$

with

$$\eta\|x\| \geq \|Ax - b\|.$$

The assumption in this problem is that the errors can be used to improve the estimate, similar to the total least squares (TLS) problem (the norm of the x that minimizes the cost function is larger in general than the corresponding norm of the least squares solution). This problem can also actually yield a robust solution, which is impossible for other algorithms like TLS to do. It is even possible for the degenerate min min problem to coincide with the min max solution. This property is very encouraging given the usual trade-off between accuracy and robustness. This problem in essence picks the accurate solution, but should the conditions warrant it will pick a robust one. The solution method is shown to be of compatible complexity to the least squares problem (n^3 algorithm), and the data requirements are only one scalar (perturbation bound for A) larger.

The final formulation, covered in Chapter 7, is the most promising. The min max backward error problem, or backward error for short, seeks to correct for numerical conditioning. The backward error technique can yield increased accuracy or increased robustness as the degenerate min min problem can, but it does so on a more sophisticated criterion. The problem is thus stated

$$\min_{x} \max_{\|E\| \leq \eta} \frac{\|(A + E_A)x - b\|}{\|A\|\|x\| + \|b\|}.$$

In experimental runs, the backward error problem has consistently yielded the best results. The data requirements are the same as the degenerate min min problem, but the added superiority is bought at the cost of a non-convex cost function. The backward error technique is broken down into four related problems, three of which are solved completely, the final problem is solved but not completely characterized.

Chapter 2

Current Formulations

The regression problem has a long history, and many methods have been proposed to handle it. Linear regression models and techniques, in particular, have been studied for hundreds of years. Over the course of the last fifty years, methods have been introduced to address a variety of problems encountered, even, in some cases including uncertainty. A proper understanding of the formulations proposed in this dissertation requires a good familiarity of the material that has gone before. This chapter will examine some of the major techniques and formulations currently in use. Section 2.1 covers an overview of least squares, the most well known and used regression technique. Section 2.2 provides an overview of total least squares. Section 2.3 contains an overview of weighted least squares. Section 2.4 is an overview of constrained least squares, which encompasses a wide variety of problem formulations. Section 2.5 covers an overview of Ridge Regression, another method, which can fit under both constrained least squares and Tikhonov, but is worth discussing separately. Section 2.6 contains an overview of Tikhonov, another general technique with a variety of sub-techniques. Section 2.7 provides an overview of the min max problem, which is the basis of the multi-column min max criterion of Chapter 3, the multi-row min max criterion of Chapter 4, and the general block min max criterion of Chapter 5. Section 2.8, compares the min max problem to a suggested form of the Tikhonov regulator that has a closed form solution. Section 2.9 covers an overview of the non-degenerate min min problem, which is the basis of the degenerate min min problem discussed in Chapter 6. Section 2.10 contains an overview of the Linear Matrix Inequality (LMI) techniques of robust estimation, which is one of the most flexible techniques available.

2.1 Least Squares

Least squares assumes that the matrix, A is known exactly ($E_A = 0$), and thus all errors occur in the observations, b, only. It is sometimes also referred to as Errors-in-Observations, due to the assumption of the errors occurring only in b. Least squares has been studied heavily since Gauss introduced it to calculate orbits of objects in the solar system [56]. Numerous other works have covered solution methods [8, 89] and solved cases such as special structure [30, 32], sparse matrices [36, 57], and numerical issues [10, 28, 118]. The least squares formulation solves the problem of $Ax \approx b$ by minimizing the error of the estimates in the Euclidean sense. The problem can be stated as

$$\min_x \|Ax - b\|^2. \tag{2.1}$$

This is the least squares criterion, and it works well in most situations. The least squares problem has a simple solution, namely $x_{LS} = A^{\dagger}b$, where $A^{\dagger}$ is the pseudo-inverse of A. The solution is the same for a deterministic assumption, as for an assumption that additive zero mean gaussian noise is present in the measurements, b. The solution can be thought of as the projection of b into the

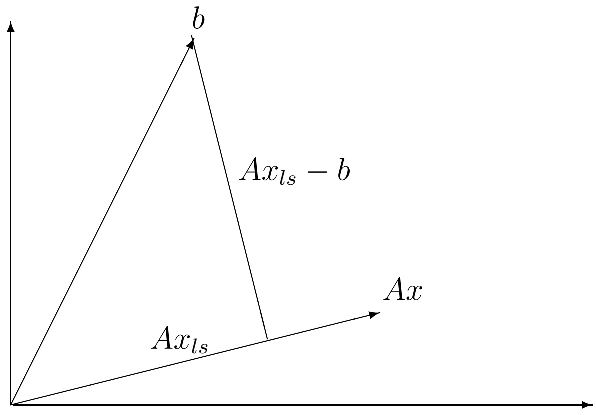

Figure 2.1. Geometric Interpretation of Least Squares Solution

range of A, as seen in Figure 2.1. The cost criterion is appealing to physical intuition, since it requires the solution to account for all of the measurements the system could have produced. The solution only requires the basic data (system matrix and measurements), and the complexity of the solution is the standard of comparison. It is easy see why the least squares criterion is popular, but it is not without its problems. The choice of independent variables, which will be shown below, and scaling problems, which are outlined in [99], are few of the well known problems with least squares.

Consider the problem of calibrating a gas thermometer. Gas thermometers are based on Charles' law, which states that the volume of a fixed mass of gas at a fixed pressure is proportional to its temperature. A simple gas thermometer can be made by trapping some gas with a mercury plug in a capillary tube that is open on only one end [98]. The volume is thus proportional to the height of the plug. The equation of the thermometer is thus $hc_1 = T$, where h is the height of the plug, c_1 is the unknown value, and T is the absolute temperature. The gas thermometer is placed in a stirred liquid bath with a known thermometer. The bath is heated and height and temperature measurements are taken at various times. The least squares solution yields $\hat{c}_1 = h^\dagger T$, but this minimizes only the error in the measured temperature, T, from the predicted temperature, $hh^\dagger T$. Alternately the relation $h = c_2 T$ could have been used, with $c_2 = \frac{1}{c_1}$. The least squares solution, $\hat{c}_2 = T^\dagger h$, thus minimizes the error between the measured height, h, and the predicted height $TT^\dagger h$. A problem arises with the least squares method in that generally $\hat{c}_1 \neq \frac{1}{\hat{c}_2}$. This can be seen easily in Figure 2.2. The slope of the line designated temperature errors, is $\hat{c}_1$, while the slope of the line designated height errors is $\frac{1}{\hat{c}_2}$. The line designated theoretical is the "true" system from which the estimates were generated. It is easy to see that the slopes are not the same, and thus $\hat{c}_1 \neq \frac{1}{\hat{c}_2}$. The least squares solution does not even perfectly handle the case where the system matrix is "known", which gives cause to be concerned as to how it will perform when there are perturbations to the system matrix.

2.2 Total Least Squares

One method to deal with the problem mentioned in the last section is to change the criterion from measuring either the error in the height or the error in the temperature, to measuring the sum of the squares of both. This gives rise to the total least squares criterion. The total least squares criterion has been examined in detail elsewhere, see for instance [67, 82]. The total least squares criterion is known to improve the accuracy of a particular problem at the cost of robustness. This can be seen by looking at the least squares and total least squares problems geometrically. The least squares problem assumes that the error occurs only in b and thus it projects b into $\mathcal{R}(A)$, see Figure 2.1. The total

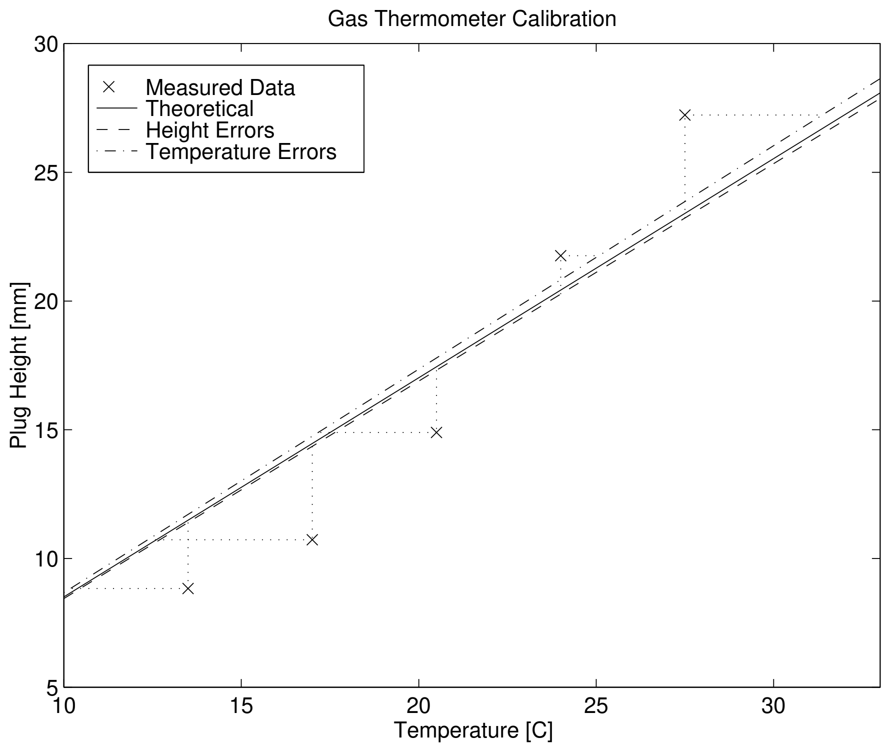

Figure 2.2. Gas Thermometer Example

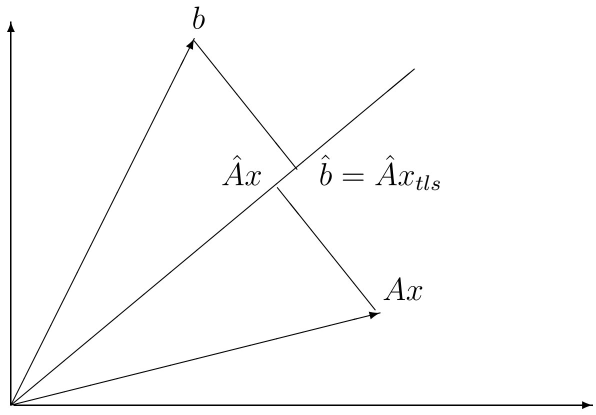

Figure 2.3. Geometric Interpretation of Total Least Squares Solution

least squares problem assumes the error occurs in both A and b and thus it finds the closest $\hat{A}$ and $\hat{b}$ to the original problem such that $\hat{b} \in \mathcal{R}(\hat{A})$, as is stated in [82]. See Figure 2.3 for an geometrical understanding of total least squares. One disadvantage with this is that there is no way to limit how much $\hat{A}$ is changed, and thus the solution could be for a very different problem than what was originally posed. A key disadvantage though, is that the total least squares problem picks the most optimistic $\hat{A}$ and $\hat{b}$ possible from a potentially large set of matrices. Such optimism is not realistic and results in the lack of robustness noted.

A key assumption of the total least squares problem is thus that both A and b have errors, so the true system is $\hat{A}$ and $\hat{b}$ with $\hat{b} \in \mathcal{R}(\hat{A})$. The problem is then to find $\hat{A}$ and $\hat{b}$ and solve $\hat{A}x = \hat{b}$. Another key assumption is that the perturbations in A and b are independent, which is not true for structured matrices so common in estimation and identification problems. The cost function has good physical correspondence, in that the errors are taken as the distance both A and b must be perturbed to get a consistent model, rather than what least squares does, which is to consider it to be the distance b must be perturbed to get a consistent model.

It is inherently more satisfactory to account for errors on both sides, rather than to try and lump all the errors into one side of the equation. The solution has a closed form, so it is trivially tractable. The data requirements are the same as the least squares problem, which is ideal. Finally the solution complexity is on the same order as the least squares problem (n^3). All in all, the total least squares problem is seen to be very useful, but it still has drawbacks.

2.3 Weighted Least Squares

Weighted least squares is a general technique which seeks to account for the relative importance or accuracy of a row in the equation $Ax - b$ by multiplying by a weighting matrix, which is usually diagonal. Both row and column weighting can be done [89, 138], though only the more common case of row weighting will be considered. Weighted least squares provides a simple way of controlling the influence of a row, although it unfortunately suffers from ill conditioning. Some major uses of weighted least squares are iterative improvement of a least squares solution [4, 5, 7, 13, 70], electrical networks, and finite elements [140]. The weighted least squares problem can be stated as

$$\min_x \left\| W^{\frac{-1}{2}} (Ax - b) \right\|^2 ,$$

with solution

$$\begin{aligned} A^T W^{-1} A x &= A^T W^{-1} b \\ x &= \left(A^T W^{-1} A \right)^{-1} A^T W^{-1} b. \end{aligned}$$

The main source of ill conditioning is the matrix W, which is by its very nature designed to have a large spread in its Eigenvalues. Vavasis [148] claims that this ill condition causes usual techniques for least squares type problems to yield highly inaccurate answers. Given the ill conditioning, it is reasonable to ask if x ever can become infinite due to W. Both Stewart [139] and Todd [144] were able to establish independently that for all positive-definite real diagonal matrices, W, the following supremums are finite:

1. $\sup\left\{\|(A^TW^{-1}A)^{-1}A^TW^{-1}\|\right\}$
2. $\sup\left\{\|A(A^TW^{-1}A)^{-1}A^TW^{-1}\|\right\}$

Since the supremums are finite, x must be finite. Another question to answer is if a stable method for solution exists. No method has been shown to be stable using the usual backward error analysis technique, but [81] shows one exists if stability is defined as

$$\|x_{True} - x_{Est}\| \leq \epsilon \cdot f(A) \cdot \|b\|,$$

where ϵ is machine precision and $f(A)$ is some function of A that does not depend on W. It is important to note again that a special definition of stability is needed for this problem, due to the conditioning problem in W. The basic solution of [81] can then be expressed as

1 QR factor (with pivoting) $A^TW^{\frac{-1}{2}}$:
$$A^TW^{\frac{-1}{2}} = Q_1R_1P$$
2 Reduced QR factor (without pivoting) R_1^T:
$$R_1^T = Q_{2,1}R_{2,1}$$
3 Solve by back substitution for z:
$$R_{2,1}z = Q_{2,1}^TPW^{\frac{-1}{2}}b$$
4 Multiply:
$$x = Q_1z$$

The first QR factorization is to provide stabilization. The second QR factorization is to solve the least squares problem. The net effect is to factor $A^TW^{\frac{-1}{2}}$ into $Q_1R_{2,1}^TQ_{2,1}^TP$, which is essentially a complete orthogonal decomposition [68].

Alternately, the solution can be stated in terms of the equilibrium system

$$\begin{bmatrix} W & A \\ A^T & 0 \end{bmatrix} \begin{bmatrix} y \\ x \end{bmatrix} = \begin{bmatrix} b \\ 0 \end{bmatrix},$$

and then found by any method desired, such as the QR factorization. While this has not been shown to be stable, the alternate formulation shows that the weighted least squares problem is essentially a special case of the generalized least squares problem [68, 111, 112, 113, 86, 3].

Finally, the weights can be difficult to select if they do not arise naturally. For instance, in simple DC electrical networks the weights can be considered as resistances, and the A matrix defines the adjacencies, b gives the voltage sources, and x is the desired node voltages. Such cases naturally give rise to the weights. What happens when this does not happen? The engineer is left trying to apply heuristics to select a weighting matrix. Given the drawbacks, weighted least squares will not be considered further.

2.4 Constrained Least Squares

Often constraints naturally arise in problems. A fitting function could have prescribed values, a physical system could have limits on its operation, or a solution in a particular set could be desired. Probably the most basic constrained problem is the least squares Quadratic Inequality problem described in [68]. The problem can be stated as

$$\min_x \|Ax - b\| \qquad \text{subject to} \qquad \|Bx - d\| \leq \alpha.$$

Often $d = 0$ and B is nonsingular, though this is not required. The problem can be solved by the method of Lagrange multipliers, which has the nice bonus of having connections to Tikhonov's method, which is dealt with in Section 2.6. The problem becomes

$$\min_{x,\lambda} \|Ax - b\|^2 + \lambda(\|Bx - d\|^2 - \alpha^2).$$

Taking derivative and setting equal to zero the solution is

$$\begin{aligned} x &= (A^T A + \lambda B^T B)^{-1}(A^T b + \lambda B^T d) & (2.2) \\ g(\lambda) &= \|B(A^T A + \lambda B^T B)^{-1}(A^T b + \lambda B^T d) - d\|^2 - \alpha^2 = 0. & (2.3) \end{aligned}$$

Equation 2.2 gives the one parameter family of solutions for the problem. When the value of the Lagrange multiplier, λ, is known the unique solution is specified. Equation 2.3 is called the secular equation in [68] and this designation will be used throughout the dissertation. The purpose of Equation 2.3 is to find the value of the Lagrange multiplier, λ. The multiplier is found by any root finding method desired, though typically Newton's method or bisection is used. The general procedure of finding a solution used in this dissertation follows this basic strategy.

A particular case of the least squares quadratic inequality, minimization over a sphere, is of particular importance and has been studied extensively, see [6, 40, 41, 42, 43, 52, 55, 68, 108, 130, 138]. The minimization of a least squares problem over a sphere has strong connections to robustness, and is strongly connected to the Ridge Regression and Cross-validation problems, which will be discussed in Section 2.5. The basic problem is

$$\min_{\|x\| \leq \alpha} \|Ax - b\|.$$

Following the procedure outlined above, the solution is found to be

$$x = \begin{cases} A^\dagger b & \text{if} \quad \|A^\dagger b\| \leq \alpha \\ (A^T A + \lambda I)^{-1} A^T b & \text{else} \end{cases}.$$

This special case covers the solution being confined to a particular set, and thus forces the solution to stay bounded. Since the solution is always bounded to a reasonable size it prevents one problem associated with lack of robustness, namely solutions being unstable and growing without bound. A major problem with this is how to know a priori the size of the true $\|x\|$. An error on the guess of the size of x can cause a reduction in the signal strength (as $\|x\|$ is forced to be smaller than the guess). Another problem is that λ can in general be quite large, but experience shows that a small value of λ is more desirable as large values tend to remove fine details (usually carried in the singular vectors associated with smaller singular values) first. The solution obtained from large values of λ tend to bear little resemblance to the true solution in all but the major details. This is a key area for this dissertation, how to get a good value for the regression parameter, λ so it neither becomes unstable nor loses data.

2.5 Ridge Regression

The Ridge Regression problem is an important special case of constrained least squares. Ridge Regression can also be considered a special case of Tikhonov regularization, which is covered in Section 2.6. Golub and Van Loan [68] describe the RR problem as

$$\min_x \|Ax - b\|^2 + \lambda\|x\|^2 \tag{2.4}$$

with the criterion for picking $\lambda > 0$ such as $\|x(\lambda)\| \leq \alpha$, i.e. minimization over a sphere as discussed in Section 2.4.

Other techniques for selecting the ridge parameter exist, such as the generalized cross-validation function [65, 44]. The cross-validation function seeks to reduce the dependence of the solution on any one experiment, and thus increases the robustness of the problem, as seen in [68]. The cost function for the cross-validation problem is given by

$$C(\lambda) = \frac{1}{m}\sum_{k=1}^{m} w_k \left[\frac{\bar{b}_k - \sum_{j=1}^{r} u_{kj}\bar{b}_j \left(\frac{\sigma_j^2}{\sigma_j^2+\lambda}\right)}{1 - \sum_{j=1}^{r} u_{kj}^2 \left(\frac{\sigma_j^2}{\sigma_j^2+\lambda}\right)} \right]^2$$

with

- w_k a weight on the importance of the k^{th} row (or experiment),
- the SVD of A given by $U\Sigma V^T$,
- u_{jk} is the j, k^{th} element of U,
- σ_j is the j^{th} diagonal element of the diagonal matrix Σ,
- and $\bar{b} = U^T b$.

Details on the minimization of this cost function are discussed in [65]. The case of the Ridge Regression problem with the cross-validation function used to select λ is often called the cross-validation problem. No matter how the value of λ is selected, the expression for $x(\lambda)$ is given by $x(\lambda) = (A^TA + \lambda I)^{-1}A^Tb$. It can be easily seen that each component of the RR solution is smaller than the corresponding component of the LS problem, and thus the robustness is gained at the cost of signal strength (or information content).

2.6 Tikhonov

The Tikhonov problem can be expressed as

$$\min_x \|Ax - b\|^2 + \lambda\|Lx\|^2. \tag{2.5}$$

Note L can be indefinite. Two parameters can be chosen by the designer to select the desired solution. The first parameter is L, which is used to specify conditions on x. For instance, a solution with a small norm could be desired, which would correspond to picking L to be the identity matrix. Alternately, a solution with a small derivative could be desired, which corresponds to picking L to be the discrete approximation of the derivative operator. Similar to weighted least squares, weights could be placed on particular portions of x to limit their sizes.

The second parameter is λ. Rather than chose λ directly, as is done for L, a requirement for λ in terms of the rest of the problem is usually chosen. For instance, the problem of minimizing a solution on a sphere used $\|x(\lambda)\| \leq \alpha$. What requirement should be used becomes the central discussion of Tikhonov regularization.

- In [80], it was shown that a non-zero λ produces smaller error on average.
- The discrepancy principle [103] assumes the true system has been corrupted by noise and uses the standard deviation of the noise, to find λ.
- The L-curve [75] assumes the system is corrupted by noise but does not require as much information on the noise properties as the discrepancy principle.
- Bounded variations for piecewise continuous functions with at most countably many discontinuities, are handled in [105].
- Generalized cross-validation [65], mentioned earlier tries to minimize the dependence on any one trial.
- Residual and singular value plots have also been suggested [123] to pick λ.

- Minimizing the lengths of confidence intervals [119] has been done.
- Even parameter choices for iterative solution methods exist [84].
- Most interesting though are the methods that attempt to minimize the distance to the true solution, such as [45, 58, 74, 121, 107].

In particular, consider the most recent method as covered in [107]. Let the SVD of A be $U\Sigma V^T$ and define $\beta = U^T b$. The Tikhonov solution to $Ax \approx b$, with $L = I$ is

$$\begin{aligned} x_{tik} &= V(\Sigma^T\Sigma + \lambda I)^{-1}\Sigma^T\beta \\ &= \sum_{i=1}^{n} \frac{\sigma_i \beta_i}{\sigma_i^2 + \lambda} v_i. \end{aligned}$$

The true system with noise ϵ can be expressed as

$$\begin{aligned} x_{true} &= V\Sigma^{\dagger}(\beta - \epsilon) \\ &= \sum_{i=1}^{n} \frac{\beta_i - \epsilon_i}{\sigma_i} v_i. \end{aligned}$$

Minimizing the distance between these two values gives the condition for λ. To compute the function exactly requires the knowledge of ϵ, which is not known. An approximation can be made if the system satisfies the discrete Picard condition (the data values $\beta_i - \epsilon_i$ goes to zero faster than the singular values) and β_i is a true value plus noise. With these assumptions and the standard deviation, s, of the noise, the root of the function

$$\sum_{i=1}^{n} \frac{\beta_i^2 \lambda}{(\sigma_i^2 + \lambda)^3} - \sum_{i=1}^{k-1} \frac{s^2}{(\sigma_i^2 + \lambda)^2} - \sum_{i=k}^{n} \frac{\beta_i^2}{(\sigma_i^2 + \lambda)^2}$$

gives the value of λ. While the value obtained is an approximation, O'Leary [107] shows that the resulting x value, $x_{tik}(\lambda)$ is close to the true (using the not approximated value of λ_{true}) value $x_{tik}(\lambda_{true})$, and that in particular

$$\frac{\|x_{tik}(\lambda_{true}) - x_{tik}(\lambda)\|}{\|x_{tik}(\lambda_{true})\|} \leq \frac{|\lambda_{true} - \lambda|}{\sigma_n^2 + \lambda}.$$

Additionally, as the standard deviation of the noise, s, goes to zero, $x_{tik}(\lambda)$ goes to x_{true}. A nice result. An alternate method of choosing λ is presented at the same time using

$$x_{alt} = \sum_{i=1}^{n} \frac{\beta_i}{\sigma_i + \lambda} v_i.$$

The choice was suggested for Hermitian positive definite matrices [53], convolution problems with reordering [39], and some additional cases [76, 31]. The alternate method proceeds similarly with a small alteration in the function, whose root must be found. The fact that two methods are suggested, indicates that no one best method exists. Both perform well however, and demonstrate robustness, which was a major goal.

Tikhonov regularization, works by damping out the terms that correspond to the smaller singular values, see for example [64]. This can be thought of geometrically as finding a worse model within some bounded region from the original model and solving the LS problem on this new model. Tikhonov regularization generates robust solutions, but the wealth of techniques to select λ shows that there is no obvious best technique. A drawback to Tikhonov regularization is thus also one of its strengths, namely the wide variety of techniques to pick λ. The criterion for picking λ is really dependent on the solution desired, for instance, the choice of the solution lying in a ball is usually done to fulfill a heuristic requirement for boundedness. In the end this method usually ends up being more based on the skill and experience of the engineer who sets up the problem. This is exactly how the problem is treated in [64, 107]. The damping of terms corresponding to smaller singular values, essentially means that data and thus accuracy will be lost. Most of the accuracy loss is due to the "waterbed effect", in that accuracy and robustness are competing goals, so advances in one area causes losses in another. Such competing goals thus are not so much a problem but rather a design decision based on the problem requirements. The assumptions are another matter. By adding an implicit heuristic element, the problem de facto includes the "gut feel" of the designer. While this may seem appealing, it is not rigorous, and does not allow for confidence in the final result. A good guess will give a good result, a bad one a bad result, but there is no way of assessing the guesses. The desire for a more philosophically pleasing and mathematically rigorous method for posing robust problems led to the development of the min max problem.

2.7 Min Max

The min max problem was proposed and solved separately in [24] by secular equation techniques and in [62] by Linear Matrix Inequality techniques. This section will concentrate on the secular equation formulation. The LMI techniques are discussed in Section 2.10

Simply stated the min max problem seeks to find the worst model in a bounded region, and then solve the problem based on this worst case scenario. Mathematically it is written as

$$\min_{x} \max_{\substack{\|E_A\| \leq \eta \\ \|E_b\| \leq \eta_b}} \|(A + E_A)x - (b + E_b)\|. \tag{2.6}$$

This problem can be shown to be equivalent to solving a problem with similar form to the Tikhonov problem, see [24]. Equation 2.6 can be interpreted geometrically by Figure 2.4. The maximization forms the hyperspheres around A and b. The cone around A is formed by varying the size of x. The solution, x, and the residual, R, are found by connecting the furthest points on the hyperspheres.

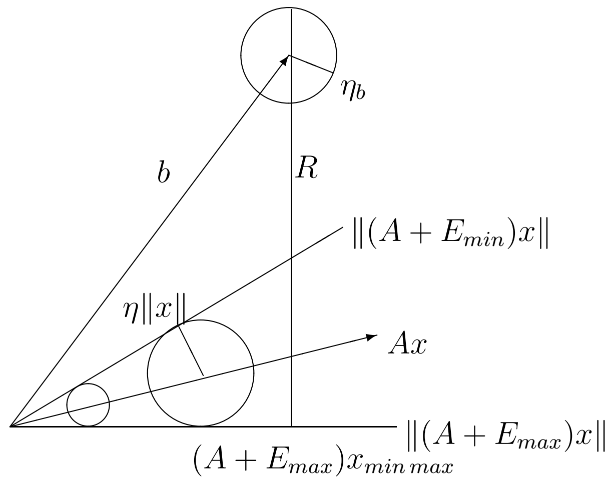

Figure 2.4. Geometric Interpretation of Min Max Solution

The maximization restricts the problem to the lower line of the cone. The minimization selects the point on the lower cone such that the line segment from the furthest point on the hypersphere around b to the lower cone is perpendicular to the lower cone. The norm used in [24] is the 2-norm, though [155] extends it to other norms. The min max problem becomes

$$\min_x(\|Ax - b\| + \eta\|x\| + \eta_b), \tag{2.7}$$

which differs from the typical Tikhonov problem in that the norms are not squared. As opposed to the Tikhonov problem, the term η now has a physical intuition also, that being the amount of uncertainty in the matrix. Computing the min max solution takes longer than computing the solution to a Tikhonov problem if a simple choice of regression parameter is chosen for the Tikhonov problem, so it is logical to ask why one would want to spend the extra operations to do so. The simple answer is that the two problems can give arbitrary differences, which we will examine in Section 2.8.

In the form of Equation 2.7 it is easy to see that the problem is continuous but non-smooth, since it is non-differentiable whenever $x = 0$ or when $Ax = b$. The solution to Equation 2.7 and thus Equation 2.6, is summarized in Table 2.1. For Table 2.1, let the SVD of A be given by

$$A = \begin{bmatrix} U_1 U_2 \end{bmatrix} \begin{bmatrix} \Sigma \\ 0 \end{bmatrix} V^T.$$

Partition the vector $U^T b$ into

$$\begin{bmatrix} U_1 U_2 \end{bmatrix}^T b = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix},$$

and introduce the secular equation

$$g(\psi) = b_1^T (\Sigma^2 - \eta^2 I)(\Sigma^2 + \psi I)^{-2} b_1 - \frac{\eta^2}{\psi^2} \|b_2\|^2,$$

which has a unique positive root, denoted $\bar{\psi}$ under the conditions noted in Table 2.1. Finally define

$$\tau_1 = \frac{\|\Sigma^{-1} b_1\|}{\|\Sigma^{-2} b_1\|} \qquad \text{and} \qquad \tau_2 = \frac{\|A^T b\|}{\|b\|}.$$

The solution is thus given below. Notice that the least squares solution, $A^\dagger b$, is the min max solution under special conditions. In one case a scaled family of the least squares solution solves the problem. In general though the solution is given by finding the unique root of the secular equation, $g(\psi)$ in the positive quadrant. When η is large the solution is zero.

2.8 Comparison of Min Max and Tikhonov

At this point, it is reasonable to ask if there is a similar, but simpler way to solve the problem, which exhibits the desired behavior of min max that can be solved instead of the min max methodology of [24, 155, 62]. One candidate

	$b \in \mathcal{R}(A)$	$b \notin \mathcal{R}(A)$
$\eta \geq \tau_2$	0	0
$\tau_1 < \eta < \tau_2$	$x = (A^T A + \bar{\psi} I)^{-1} A^T b$	$x = (A^T A + \bar{\psi} I)^{-1} A^T b$
$\eta \leq \tau_1$	$x = A^\dagger b$	$x = (A^T A + \bar{\psi} I)^{-1} A^T b$
$\eta = \tau_1 = \tau_2$	$x = \beta A^\dagger b$ with $0 \leq \beta \leq 1$	$x = (A^T A + \bar{\psi} I)^{-1} A^T b$

Table 2.1. Min Max Solution

solution that has been suggested is Tikhonov regulation. It has a large body of literature, such as [64, 107], and a closed form solution. Start by noting that a reasonable choice for the parameter λ in the Tikhonov problem is to chose it to be equal to the square of the uncertainty, since all the other terms are squared and this will account for the size of the uncertainty. In this case the model has a closed form solution which is given by

$$\hat{x} = \left(A^T A + \eta^2 I\right)^{-1} A^T b. \tag{2.8}$$

Note that this is clearly a regularized estimator, with the regularization parameter given by the bound in the error. Note also that for the min max problem that if $Ax \neq b$ and $x \neq 0$ then the min max problem also has a solution with a similar form given by

$$\begin{aligned} \hat{x} &= \left(A^T A + \alpha I\right)^{-1} A^T b && (2.9) \\ \alpha &= \eta \frac{\|Ax - b\|}{\|x\|}. && (2.10) \end{aligned}$$

The min max problem is also a regularized solution, with the regularization parameter given by α. Since α is dependent on unknown values it must be calculated, which is usually done by a secular equation. The logical question is, "Why not use the Tikhonov cost function, which has the closed form solution?" To answer this it must be seen if the Tikhonov problem's regularization parameter can be arbitrarily larger or smaller. If the Tikhonov problem's parameter can be arbitrarily larger, then the solution can be over regularized and thus valuable information can be lost. If the Tikhonov parameter can be arbitrarily smaller, then the solution can be under regularized and thus the solution might not be robust. Thus to compare the two, examine the ratio of the min max problem's regularization parameter, α, to the Tikhonov problem's regularization parameter, η^2. Doing so, obtain

$$\frac{\alpha}{\eta^2} = \frac{\|Ax_{mm} - b\|}{\eta \, \|x_{mm}\|}. \tag{2.11}$$

2.8.1 Over-Regularization

First, see if the Tikhonov problem can be over regularized, which is the more dangerous problem. This corresponds to the ratio being arbitrarily small. Note that $\|Ax_{mm} - b\| \leq \|b\|$ at the solution, by noting the cost at the solution must be less than the cost at the point $x = 0$. Thus,

$$\frac{\alpha}{\eta^2} \leq \frac{\|b\|}{\eta \, \|x_{mm}\|}. \tag{2.12}$$

It is clearly possible to pick A and b such that $\eta \|x_{mm}\| \gg \|b\|$. For example consider the following simple system,

$$A = \begin{bmatrix} 0.2 \\ 0 \end{bmatrix} \qquad b = \begin{bmatrix} 5 \\ 1 \end{bmatrix} \qquad \eta = 0.1. \tag{2.13}$$

For this simple system the min max problem has a solution of $x_{mm} = 22.11$ while the modified problem has a solution of $x_T = 20$. We note that for this problem the Tikhonov regularization parameter is twice as large as the min max problem. Clearly the over-regularization has also yielded a loss of information that is not warranted by the problem. We note that while this simple example does not show an arbitrarily large ratio difference, since it is used only as a numerical motivation. To see the arbitrary difference consider the following for $\delta \ll 1$,

$$A = \begin{bmatrix} \delta \\ 0 \end{bmatrix} \qquad b = \begin{bmatrix} \frac{1}{\delta} \\ \delta \end{bmatrix} \qquad \eta = \frac{\delta}{2}. \tag{2.14}$$

For this system note that the least squares (LS) solution is given by $x_{LS} = \frac{1}{\delta^2}$, and the min max system is $x_{mm} = \frac{1}{\delta^2} - \frac{1}{\delta\sqrt{3}}$. Note that since $\delta \ll 1$, the min max estimate is extremely close to the LS solution. The Tikhonov problem solution is given by $x_T = \frac{4}{5\delta^2}$, which is easily seen to be arbitrarily far from the desired solution, since for $\delta \ll 1$ the two candidate solutions differ by almost 20% of an arbitrarily large number. Moreover, the ratio of regularization parameters is approximately given by the arbitrarily small number,

$$\frac{\alpha}{\eta^2} \approx \frac{4}{\sqrt{3}}\delta^2. \tag{2.15}$$

2.8.2 Under-Regularization

The second area to be considered is if the Tikhonov problem can be under-regularized. This corresponds to the ratio of α over η^2 being arbitrarily large. Note that $\|Ax_{mm} - b\| \geq \|P_{A^\perp} b\|$, thus

$$\frac{\alpha}{\eta^2} \geq \frac{\|P_{A^\perp} b\|}{\eta\,\|x_{mm}\|}. \tag{2.16}$$

It is clearly possible to pick A and b such that $\|x_{mm}\| \ll \|P_{A^\perp} b\|$. For example consider the following simple system,

$$A = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \qquad b = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \qquad \eta = 1. \tag{2.17}$$

Note that since the perturbation is as large as the norm of the A matrix, $x_{mm} = 0$, which corresponds to $\alpha \to \infty$. This is intuitively pleasing, as it confirms the belief that no valid information exists for a system with uncertainty as large as the system. Note also that $x_{LS} = 1$. Now the Tikhonov problem has the solution $x_T = \frac{1}{2}$. Not only is this clearly too optimistic an answer, it is also unrealistic. The ratio is infinite and thus arbitrarily large, as was desired to be shown. Thus while the Tikhonov problem has nice properties for calculation, its estimator can be arbitrarily different than the min max problem. Additionally, the Tikhonov problem does not correspond to physical intuition as can be seen in the last example above. The min max problem can thus not be altered to an apparently similar problem and solved for that system.

2.9 Non-Degenerate Min Min

The non-degenerate min min problem was presented in [25] for the case of the 2-norm and extended to other norms in [155]. The essential idea is to assume, similar to total least squares, that the actual system $A + E_A$ and $b + E_b$ is such that $b + E_b$ is as close to being in the subspace defined by $A + E_A$ as possible. The residual is then minimized over all choices x. The problem is thus expressed as

$$\min_x \min_{\substack{\|E\| \le \eta \\ \|E_b\| \le \eta_b}} \|(A + E_A)x - (b + E_b)\|. \tag{2.18}$$

The geometric view is very similar to the min max problem and is provided in Figure 2.5. The cone around A and the ball around b are the same as before (i.e.: all possible values for the problem). The min min problem is thus to find

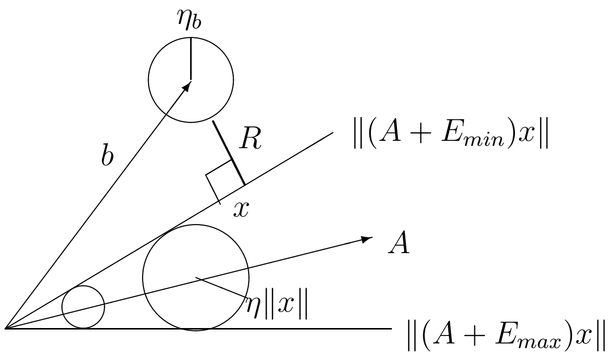

Figure 2.5. Geometric Interpretation of Min Min Solution

the smallest distance from the ball to the cone, which is shown in the figure. Note that it is possible for the ball and cone to have points in common, this is the degenerate case, and is covered in Chapter 6. This section will cover the non-degenerate condition. Two main issues are:

1. find a computable condition for checking degeneracy,

2. find a secular equation and region to find the solution.

First, find a computable condition for degeneracy. For the problem to be non-degenerate the residual must be greater than the possible perturbation. In equation form this is

$$\eta\|x\| < \|Ax - b\|. \tag{2.19}$$

This equation depends on the solution, x, so an equation with only A, b, and η is desired. By squaring the degeneracy condition, Equation 2.19, the condition becomes

$$x^T(A^TA - \eta^2 I)x - 2x^TA^Tb + b^Tb > 0. \tag{2.20}$$

For this to hold for all x, the minimum value of the function must be greater than zero. The function must have a finite minimum, and that minimum must be positive. To have a finite minimum, the following must hold

$$A^T A - \eta^2 I > 0,$$

or defining the minimum singular value of A to be σ_{min},

$$\sigma_{min} > \eta. \tag{2.21}$$

Noting that in practice $\eta > 0$ ($\eta = 0$ is least squares), this requires that A is full rank, which is assumed from now on. Provided Equation 2.21 holds, the minimum value of Equation 2.20 is

$$b^T \left[I - A(A^T A - \eta^2 I)^{-1} A^T \right] b > 0. \tag{2.22}$$

The problem is non-degenerate if A is full rank, and both Equation 2.21 and Equation 2.22 hold.

The computable condition is needed to see if the non-degenerate case applies (if it doesn't the non-degenerate case of Chapter 6 holds) and is useful in the proof. The proof of the solution is too lengthy to present here, readers are referred to [25] for a full treatment. Assume the problem is non-degenerate and let the SVD of A be

$$A = \begin{bmatrix} U_1 & U_2 \end{bmatrix} \begin{bmatrix} \Sigma \\ 0 \end{bmatrix} V^T,$$

with smallest singular value σ_n and corresponding left singular vector u_n. Define

$$\begin{bmatrix} b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} U_1 & U_2 \end{bmatrix}^T b.$$

Then for the secular equation,

$$g(\psi) = b_1^T(\Sigma^2 - \eta^2 I)(\Sigma^2 - \psi I)^{-2} b_1 - \frac{\eta^2}{\psi^2}\|b_2\|^2,$$

find the unique root of $g(\psi)$ in the interval (η^2, σ_n^2) and if it exists call it $\hat{\psi}$, otherwise let $\hat{\psi} = \sigma_n^2$. If $\hat{\psi} < \sigma_n^2$ then the solution is

$$x = (A^T A - \hat{\psi} I)^{-1} A^T b,$$

else there are two solutions

$$x = V \begin{bmatrix} (\bar{\Sigma}^2 - \sigma_n^2 I)^{-1}\bar{\Sigma}\bar{b}_1 \\ \pm\frac{\sigma_n}{\sqrt{\sigma_n^2 - \eta^2}}\sqrt{-\bar{g}(\sigma_n^2)} \end{bmatrix}$$

with

$$\begin{aligned} \bar{g}(\psi) &= g(\psi) - (u_n^T b)^2 \frac{\sigma_n^2 - \eta^2}{(\sigma_n^2 - \psi)^2} \\ \Sigma &= \begin{bmatrix} \bar{\Sigma} & 0 \\ 0 & \sigma_n \end{bmatrix} \\ b_1 &= \begin{bmatrix} \bar{b}_1 \\ b_{1,n} \end{bmatrix} = \begin{bmatrix} \bar{b}_1 \\ 0 \end{bmatrix}. \end{aligned}$$

The work in this dissertation completes the analysis of this problem by solving the degenerate case, which turns out to be the more general situation.

2.10 LMI Techniques

Of all the techniques presented the Linear Matrix Inequality (LMI) techniques are the most flexible. Most of the techniques both currently used and those in this dissertation, can be solved using LMI techniques. The Backward Error method is an example of a problem that does not fit into the LMI framework, due to its rational cost function. The principle concern of this section is to consider the LMI techniques that are similar to what is covered in this dissertation. In [17, 18, 60, 62, 90, 94, 95], the LMI methodology for solving the min max problem with and without structure were covered. This section will cover two principle areas of [62], that being the structured and unstructured case.

The unstructured perturbations are identical to the min max case. The problem is defined as

$$\min_{x} \max_{\|EE_b\|_F \leq 1} \|(A+E_A)x-(b+E_b)\|.$$

Note that the bound is 1 since the problem can always be normalized to this by dividing A and b by any other bound thus yielding a problem of the form above. The problem can be reformulated as a Second-Order Cone Programming (SOCP) problem of the form

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

Define ψ to be $\frac{(\lambda-\tau)}{\tau}$. The solution can then be shown to be

$$x = \begin{cases} (A^T A + \psi I)^{-1} A^T b & \text{if } \psi > 0 \\ A^{\dagger} b & \text{else} \end{cases}$$

with λ and τ are the unique optimal points for the system. The parameter ψ is the same as was found in the min max problem by secular equation techniques.

At this point it is reasonable to ask why further work should be done. The basic reason is speed. Each iteration of a SOCP is basically $O((m+n)n^2)$ and

[62] asserts that the number of iterations is almost constant and independent of the problem size, resulting in a reasonably sized constant multiplying the n^3. In contrast, solving a secular equation can be done in iterations that are n^2 and then the overall solution takes n^3 but has a smaller constant since it mostly comes from the calculation of the SVD (the "light" version of the SVD can be used further saving time). In [62], it is noted that both have the same order of complexity, which is true, but order is not the only determiner, the constant that is ignored when reporting order can greatly influence practical speed. The speed advantage of secular techniques is noted in [62], thus secular equation techniques have a slight advantage over SOCP. In [94] it is asserted that the secular equation technique is simpler and that LMI techniques only have advantage over secular techniques on robust regression problems when additional constraints need to be applied. Noting that SOCP problems can be solved faster than SDP (semi-definite programming) problems, the use of secular equation techniques becomes more apparent. On a different line, developing algorithms to solve problems in different ways is a valuable goal in and of itself, that yields benefits in many areas from theoretical understanding to programming implementations. The nice features of the LMI techniques thus do not rule out the other solution techniques, rather they complement each other and serve to give a more complete understanding.

The second area to cover is the structured perturbations problem. The structured perturbations can approximate the multi-column, multi-row, and general min max problems developed in this dissertation and some cases not expressible in one of the proposed techniques. Note that the structured case can approximate but not directly solve the cases covered in this dissertation. The structured perturbations are of the form

$$
\begin{aligned}
E(\delta_1, \ldots, \delta_p) &= \sum_{k=1}^{p} \delta_k E_k \\
E_b(\delta_1, \ldots, \delta_p) &= \sum_{k=1}^{p} \delta_k E_{b,k}
\end{aligned}
$$

where p is the number of basic perturbations, and E_k is a fixed basic perturbation on A and $E_{b,k}$ is a fixed basic perturbation on b. The problem with trying to directly solve the multi-column case (for instance) is that the basic perturbations would need to be,

$$
\begin{aligned}
E_1 &= \frac{Ax-b}{\|Ax-b\|} \frac{\begin{bmatrix} x_1^T & 0 & \ldots & 0 \end{bmatrix}}{\|x_1\|} \\
&\vdots \\
E_p &= \frac{Ax-b}{\|Ax-b\|} \frac{\begin{bmatrix} 0 & \ldots & 0 & x_p^T \end{bmatrix}}{\|x_p\|}.
\end{aligned}
$$

This requires E_k to be dependent on x and thus the LMI to solve the problem that is presented, is no longer linear. The multiple column case is used in [62] as a motivation for the linear-fractional case, which is shown to be NP-hard in the general case and for which an upper bound for the worst case residual was obtained. As special cases, the multiple column, multiple row, and general block perturbation cases are not NP-hard, and allow for the results in this dissertation. The three cases could be approximated with the structured problem (as opposed to the linear-fractional) by examining a series of problems with the values of E_k based off the previous problem's solution, starting with say $x = A^{\dagger}b$, though it is not obvious if this would converge. Optionally the problem could be approximated by some other column (row, or general also) structure, though it will not necessarily generate the same one found in this dissertation. Undoubtedly an LMI formulation can be found to solve the same problem, the key point is that the structured formulation as outlined in [62] will not. The formulation in [62] does permit important structures, such as Toeplitz, that cannot be handled in the formulations of this dissertation. Each formulation is useful and has a place, the needs of each individual problem under consideration determine which to use. Returning to the structured problem, first define the following:

$$\begin{aligned} M &= \begin{bmatrix} E_1 x - E_{b,1} & \dots & E_p x - E_{b,p} \end{bmatrix} \\ F &= M^T M \\ g &= M^T (Ax - b) \\ h &= \|Ax - b\|. \end{aligned}$$

The problem can be written as

$$\min_x \max_{\|\delta\| \leq 1} \begin{bmatrix} 1 \\ \delta \end{bmatrix}^T \begin{bmatrix} h & g^T \\ g & F \end{bmatrix} \begin{bmatrix} 1 \\ \delta \end{bmatrix}.$$

The problem can then be solved by the SDP,

$$\begin{aligned} &\min \lambda \\ &s.t. \quad \begin{bmatrix} \lambda - \tau & 0 & (Ax - b)^T \\ 0 & \tau I & M^T \\ Ax - b & M & I \end{bmatrix} \geq 0. \end{aligned}$$

The SDP formulation, while not as efficient as a SOCP formulation, is still polynomial time, and can be solved by interior point algorithms. The variety of structures that can be handled or approximated by the structured technique in [62] is tremendous and covers most of the cases of interest. Problems still exist which cannot be directly handled, or which could be solved more efficiently by specialized solvers, such as are covered in this dissertation.

2.11 Simple Comparative Example

The previous discussion includes three general groups of problem formulations that can be used in estimation. The following provides a feel for how these problems operate on a simple example. Consider a simple one dimensional “skyline” image that has been blurred. A “skyline” image is a one dimensional image that looks like a city skyline when graphed, and thus is the most basic image processing example. “Skyline” images involve sharp corners, and it is of key importance to accurately locate these corner transitions. Blurring occurs often in images, for example atmospheric conditions, dust or imperfections in the optics can cause a blurred image. Blurring is usually modelled as a gaussian blur, which is a smoothing filter. The gaussian blur causes greater distortion on the corners, which is exactly where we do not want it. The component of a gaussian blur with standard deviation, σ, in position, (i,j), is given by

$$G_{i,j} = e^{-\left(\frac{i-j}{\sigma}\right)^2}.$$

Going on the presumption that the exact blur that was applied is not known (σ unknown) the exact system cannot be recovered. While the original system cannot be perfectly extracted, some improvement on the blurred image is desirable. The blur is “known” to be small compared to the information so some improvement should be possible. The least squares solution fails completely, yielding a result that is about three orders of magnitude off in the scale and is oscillating badly, see Figure 2.6. Notice that the total least squares solution is better than the least squares solution (only off by an order of magnitude and the oscillations are slower), but still not acceptable. The Tikhonov solution works well due to its increased robustness. All of the methods of this dissertation can be seen to yield very good solutions to the problem.

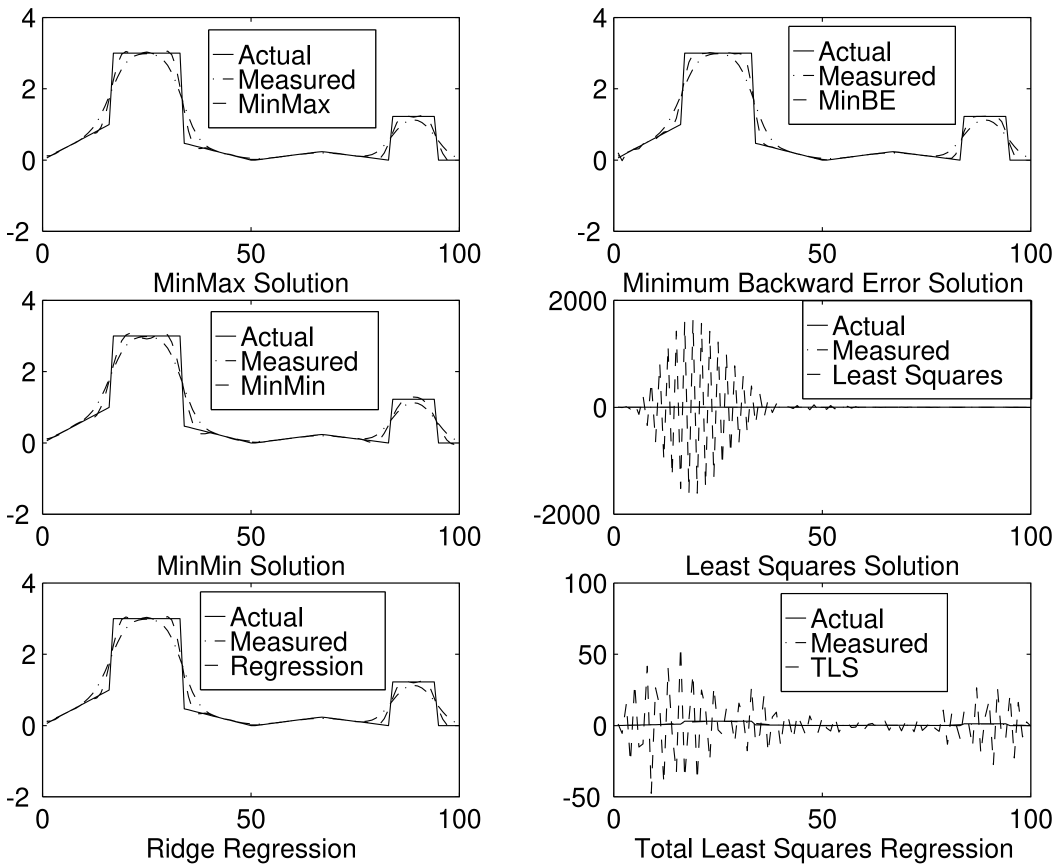

Figure 2.6. Skyline Problem

Chapter 3

Multi-Column Min Max Criterion

In Chapter 2 it was seen that the unstructured and unpartitioned min max problem, and the structured and unpartitioned min max problem have both been solved. In the next three chapters the unstructured and partitioned min max will be examined. The easiest case is the column partitioning problem, which will be examined in this chapter. This can arise if a new (block) column is added to A corresponding to an increase in the order of the filter. The new block column will not necessarily have the same uncertainty as the original block, thus partitioning is needed so different errors may be assigned to each block. Alternately, the column partitioning case could be used to model a series of geophones in a seismology problem that have different uncertainties due to geometry or surface geology conditions. Column partitioning also describes signal separation with different uncertainties associated with each signal. The column partitioning case also could be dealing with various polynomials in a polynomial fitting problem. In short, many problems satisfy the basic conditions of the multi-column problem.

A simplified case of this, where one of the columns was unperturbed is considered in [24]. In this section the general problem of partitioned and perturbed columns will be solved.

The chapter is divided as follows. In Section 3.1, the fundamental difference between the unstructured non-partitioned case and the unstructured column partitioned case is examined. The difference between the structured non-partitioned case (LMI) and the unstructured partitioned column case was already covered in Section 2.10. Section 3.2, shows how to perform the maximization and obtain an alternate formulation. Section 3.3 covers the quadratically convergent method of Overton, that solves the sum of Euclidean norms problem. The potential benefit of reducing the problem order from n, the number of columns, to p, the number of column partitions, motivates continuing. Therefore, the form of the solution at a differentiable point is shown in Section 3.4. The secular equation is developed in Section 3.5. Finally a numerical example is provided to demonstrate the results.

3.1 Column Dependence

Given the similarity of the problem structure to the non-partitioned case, some have concluded that the solution conditions should be the same. In particular, the non-partitioned problem has two simple conditions on x that do not carry into the partitioned case,

1. the solution, x, is non-zero if and only if $\|A^T b\| > \eta \|b\|$,

2. the solution, x, has a smaller norm than the least squares solution.

3.1.1 When x Is Zero

First consider the simple relation that the solution x is non-zero if and only if $\|A^T b\| > \eta\|b\|$. This is not true for the partitioned case, which can be seen by considering the following

$$A_1 = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} \quad A_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \quad b = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}.$$

It is readily apparent that $A_2^T b = 0$ and thus from the original problem, $x_2 = 0$ for all η_2. Now consider $\eta_1 = \eta_2 = \frac{1}{4}$, and consider the cost for $x_2 = 0$ and $x_1 \neq 0$.

$$\begin{aligned} J(x_1, x_2 = 0) &= \|A_1 x_1 - b\| + \frac{|x_1|}{4} \\ &= \sqrt{(1 - x_1)^2 + x_1^2 + 1} + \frac{|x_1|}{4} \\ &= \sqrt{2(x_1^2 - x_1 + 1)} + \frac{|x_1|}{4} \end{aligned}$$

The minimum can be found by taking the derivative of $J(x_1, x_2 = 0)$ and setting it equal to zero.

$$\begin{aligned} 0 &= \frac{\partial J(x_1, x_2 = 0)}{\partial x_1} \\ &= \frac{2x_1 - 1}{\sqrt{2(x_1^2 - x_1 + 1)}} + \frac{\operatorname{sgn}(x_1)}{4} \\ &= 8x_1 - 4 + \operatorname{sgn}(x_1)\sqrt{2(x_1^2 - x_1 + 1)} \end{aligned}$$

To simplify take the term with the square root to the other side and square both sides. Note that this will introduce a fictitious root into the equation, which will need to be removed from the solution.

$$\begin{aligned} -\operatorname{sgn}(x_1)\sqrt{2(x_1^2 - x_1 + 1)} &= 8x_1 - 4 \\ 2x_1^2 + 2x_1 - 2 &= 64x_1^2 - 64x_1 + 16 \\ 0 &= 62x_1^2 - 62x_1 + 14 \end{aligned}$$

The desired root is $x_1 = \frac{1}{2} - \sqrt{\frac{3}{124}}$. Thus the minimum cost for $x_2 = 0$ is

$$\begin{aligned}
J\left(\frac{1}{2} - \sqrt{\frac{3}{124}}, 0\right) &= \sqrt{2\left(\left(\frac{1}{2} - \sqrt{\frac{3}{124}}\right)^2 - \frac{1}{2} + \sqrt{\frac{3}{124}} + 1\right)} + \frac{1}{8} - \frac{1}{4}\sqrt{\frac{3}{124}} \\
&= (3.875)\sqrt{\frac{3}{31}} + \frac{1}{8} \\
&\approx 1.330.
\end{aligned}$$

Now consider the case when $x_2 \neq 0$. Note that when $x_2 \neq 0$, it must be that $x_1 \neq 0$ because if not, it is easily verified that $J(x_1 = 0, x_2 \neq 0) > J(x_1 = 0, x_2 = 0)$. To start, the expression for the cost is given by

$$\begin{aligned}
J(x_1, x_2 \neq 0) &= \|A_1 x_1 + A_2 x_2 - b\| + \frac{|x_1| + |x_2|}{4} \\
&= \sqrt{(1 - x_1)^2 + (x_1 + x_2)^2 + 1} + \frac{|x_1| + |x_2|}{4}.
\end{aligned}$$

A few things are readily apparent from looking at the square root term in the cost function. The first squared term in the square root is $(1 - x_1)$, which shows that the optimal x_1 is greater than zero and by considering the last term in the cost it can be concluded the optimal x_1 must lie between zero and one. The second squared term in the square root is $(x_1 + x_2)$, which shows that the optimal x_2 must be less than zero, and by considering the last term in the cost the conclusion is that the optimal x_2 must lie between zero and $-x_1$. These relations can be used to simplify the derivatives about to be taken. Since both x_1 and x_2 are not zero, take the derivative of the cost with respect to each variable in turn and set the result equal to zero,

$$\begin{aligned}
0 &= \frac{\partial J(x_1, x_2 \neq 0)}{\partial x_1} \\
&= \frac{2x_1 - 1 + x_2}{\sqrt{(1 - x_1)^2 + (x_1 + x_2)^2 + 1}} + \frac{\operatorname{sgn}(x_1)}{4} \\
&= 8x_1 - 4 + 4x_2 + \operatorname{sgn}(x_1)\sqrt{(1 - x_1)^2 + (x_1 + x_2)^2 + 1} \\
0 &= 8x_1 - 4 + 4x_2 + \sqrt{(1 - x_1)^2 + (x_1 + x_2)^2 + 1}, \qquad (3.1)
\end{aligned}$$

and

$$
\begin{aligned}
0 &= \frac{\partial J(x_1, x_2 \neq 0)}{\partial x_2} \\
&= \frac{x_1 + x_2}{\sqrt{(1-x_1)^2 + (x_1+x_2)^2 + 1}} + \frac{\operatorname{sgn}(x_2)}{4} \\
&= 4x_1 + 4x_2 + \operatorname{sgn}(x_2)\sqrt{(1-x_1)^2 + (x_1+x_2)^2 + 1} \\
0 &= 4x_1 + 4x_2 - \sqrt{(1-x_1)^2 + (x_1+x_2)^2 + 1}. \qquad (3.2)
\end{aligned}
$$

Now use Equation 3.1 and Equation 3.2 to obtain

$$
\begin{aligned}
0 &= 12x_1 - 4 + 8x_2 \\
&= 3x_1 - 1 + 2x_2 \\
x_1 &= \frac{1-2x_2}{3}.
\end{aligned}
$$

Substituting the equation for x_1 into Equation 3.2, obtain

$$
\begin{aligned}
0 &= 4\left(\frac{1-2x_2}{3}\right) + 4x_2 - \sqrt{\left(1 - \left(\frac{1-2x_2}{3}\right)\right)^2 + \left(\left(\frac{1-2x_2}{3}\right) + x_2\right)^2 + 1} \\
&= \frac{4}{3}(x_2+1) - \sqrt{\left(\frac{2}{3}(x_2+1)\right)^2 + \left(\frac{1}{3}(x_2+1)\right)^2 + 1} \\
&= \frac{4}{3}(x_2+1) - \sqrt{\frac{5}{9}(x_2+1)^2 + 1}.
\end{aligned}
$$

Now take the square root to the other side and square both sides. Note that this will create an extra root, but one is positive and the other negative. Since it is already known that the negative root is the correct one, this does not add any difficulty. Thus,

$$
\begin{aligned}
\frac{4}{3}(x_2+1) &= \sqrt{\frac{5}{9}(x_2+1)^2 + 1} \\
\frac{16}{9}(x_2+1)^2 &= \frac{5}{9}(x_2+1)^2 + 1 \\
\frac{11}{9}(x_2+1)^2 &= 1 \\
x_2 &= \frac{3}{\sqrt{11}} - 1.
\end{aligned}
$$

The solution is negative, which fits our requirement. From this obtain the value for x_1,

$$\begin{aligned} x_1 &= \frac{1 - 2\left(\frac{3}{\sqrt{11}} - 1\right)}{3} \\ &= 1 - \frac{2}{\sqrt{11}}. \end{aligned}$$

Observe that x_1 is positive and that $|x_1| > |x_2|$. Now substitute this into the cost function to obtain,

$$\begin{aligned} J\left(1 - \frac{2}{\sqrt{11}}, \frac{3}{\sqrt{11}} - 1\right) &= \sqrt{\left(\frac{2}{\sqrt{11}}\right)^2 + \left(\frac{1}{\sqrt{11}}\right)^2 + 1} + \frac{2 - \frac{5}{\sqrt{11}}}{4} \\ &= \sqrt{\frac{5}{11} + 1} + \frac{2 - \frac{5}{\sqrt{11}}}{4} \\ &= \frac{4}{\sqrt{11}} + \frac{2 - \frac{5}{\sqrt{11}}}{4} \\ &= \frac{11}{4\sqrt{11}} + \frac{1}{2} \\ &= \frac{\sqrt{11}}{4} + \frac{1}{2} \\ &\approx 1.329. \end{aligned}$$

Thus, the cost for $x_2 \neq 0$ is less than the cost for $x_2 = 0$, so while in the original problem it would have been predicted that $x_2 = 0$ this is not the case.

3.1.2 Size of $\|x\|$

The second relation to consider is that the size of the multi-column partitioned min max solution, x_Ψ, should be smaller than the least squares solution, x_{LS}, since both have the same numerator and the denominator of x_Ψ is larger. This is not always the case. To demonstrate this, consider a simple problem.

Let A and b be the matrices defined below with each column of A a separate partition.

$$A = \begin{bmatrix} 1 & 0 & 0.1 \\ 1 & -1 & 1 \\ 0 & 0 & 0.1 \\ 0 & 0 & 0 \end{bmatrix} \quad b = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 10 \end{bmatrix}$$

The least squares solution is given by

$$x_{LS} = \begin{bmatrix} 1 & 1 & 0 \end{bmatrix}^T.$$

Now consider the case when $\eta_1 = 2$,$\eta_2 = 0$, and $\eta_3 = 0$. The solution, x_Ψ is given by

$$x_\Psi = \begin{bmatrix} 0 & 5 & 5 \end{bmatrix}^T.$$

It is trivial to see that $\|x_{LS}\| < \|x_\Psi\|$, and thus the idea is disproved. The question remains then as to what can be said about the size of x_Ψ and thus where it lies. The following lemma is not tight in its bound but it does provide a good starting point for the analysis.

Lemma 3.1 *For a matrix A, a vector b, and scalars η_i, the solution to the multi-column partitioned min max problem, x_Ψ is contained within a ball, centered on the origin with radius $\frac{\sigma_1^2 \|x_{LS}\|}{\sigma_n^2}$, where σ_1 is the largest singular value of A, σ_n is the smallest singular value of A, and $x_{LS} = A^\dagger b$. This can be expressed simply as*

$$\|x_\Psi\| \leq \frac{\sigma_1^2}{\sigma_n^2} \|x_{LS}\|.$$

Proof:

$$
\begin{aligned}
x_\Psi &= (A^TA+\Psi)^{-1}A^Tb \\
\|x_\Psi\| &= \|(A^TA+\Psi)^{-1}A^Tb\| \\
\|x_\Psi\| &= \|(A^TA+\Psi)^{-1}A^TA(A^TA)^{-1}A^Tb\| \\
\|x_\Psi\| &= \|(A^TA+\Psi)^{-1}\|\|A^TA\|\|(A^TA)^{-1}A^Tb\| \\
\|x_\Psi\| &= \frac{1}{\sigma_{min}(A^TA+\Psi)}\sigma_1^2\|x_{LS}\| \\
\|x_\Psi\| &\leq \frac{1}{\sigma_n^2}\sigma_1^2\|x_{LS}\| \\
\|x_\Psi\| &= \frac{\sigma_1^2}{\sigma_n^2}\|x_{LS}\|
\end{aligned}
$$

$\diamondsuit$ SDG $\diamondsuit$

Other such bounds exist and can be used to tighten the starting condition. A key point of developing this lemma is that bounds exist on the size of the estimate, and can be calculated a priori. Such bounds could be used to start methods like the ellipsoidal algorithm. A tighter bound, that has connections to the original problem is:

Lemma 3.2 *For a matrix A, a vector b, and scalars η_i, the solution to the multi-column partitioned min max problem, x_Ψ is contained within a ball, centered on the origin with radius $\frac{\sigma_1\|P_Ab\|}{\sigma_n^2}$, where σ_1 is the largest singular value of A, σ_n is the smallest singular value of A, and P_A is the projection onto the range of A. This can be expressed simply as*

$$\|x_\Psi\| \leq \frac{\sigma_1}{\sigma_n^2}\|P_Ab\|.$$

Proof:

$$
\begin{aligned}
x_\Psi &= (A^TA+\Psi)^{-1}A^Tb \\
\|x_\Psi\| &= \|(A^TA+\Psi)^{-1}A^Tb\| \\
\|x_\Psi\| &= \|(A^TA+\Psi)^{-1}A^TP_Ab\| \\
\|x_\Psi\| &= \|(A^TA+\Psi)^{-1}\|\|A^T\|\|P_Ab\| \\
\|x_\Psi\| &= \frac{1}{\sigma_{min}(A^TA+\Psi)}\sigma_1\|P_Ab\| \\
\|x_\Psi\| &\leq \frac{1}{\sigma_n^2}\sigma_1\|P_Ab\| \\
\|x_\Psi\| &= \frac{\sigma_1}{\sigma_n^2}\|P_Ab\|
\end{aligned}
$$

◇ SDG ◇

3.2 An Equivalent Problem

It has been shown how this problem is different and what should not be done, but the algebra used to find the solution to the simple problem in Section 3.1 is hardly practical for general problems. This section will begin the solution of the column partitioning case, by finding an alternate cost function that is more amenable to solution.

Given a system matrix $A \in \mathbb{R}^{m\times n}$ with $m \geq n$ and $b \in \mathbb{R}^m$ be a given vector. An estimate of a vector of parameters, $x \in \mathbb{R}^n$, is desired, given that the system is linearly related by

$$b = Ax + \nu.$$

The vector $\nu \in \mathbb{R}^m$ denotes measurement noise in the system. The matrix, A, can be partitioned into $\begin{bmatrix} A_1 & \cdots & A_p \end{bmatrix}$ where $A_i \in \mathbb{R}^{m\times n_i}$ and $n = \sum_{i=1}^{p}(n_i)$, with a corresponding partitioning of x into $\begin{bmatrix} x_1^T & \cdots & x_p^T \end{bmatrix}^T$ with $x_i \in \mathbb{R}^{n_i}$. In addition, the true A matrix is not exactly known and is represented by $\begin{bmatrix} A_1 & \cdots & A_p \end{bmatrix} + \begin{bmatrix} E_{A_1} & \cdots & E_{A_p} \end{bmatrix}$ and only an upper bound on E_{A_i} is known,

$$\begin{aligned} \|E_{A_1}\|_2 &\leq \eta_1 \\ &\vdots \\ \|E_{A_p}\|_2 &\leq \eta_p. \end{aligned}$$

Similarly the true b vector is also not exactly known so that it is in reality $b + E_b$ and only an upper bound on E_b is known,

$$\|E_b\|_2 \leq \eta_b.$$

The problem can thus be stated as a min max problem,

$$\min_x \max_{\substack{\|E_{A_1}\|_2 \leq \eta_1 \\ \vdots \\ \|E_{A_p}\|_p \leq \eta_2 \\ \|E_b\|_2 \leq \eta_b}} \left\| \begin{bmatrix} A_1^T + E_{A_1}^T \\ \vdots \\ A_p^T + E_{A_p}^T \end{bmatrix}^T \begin{bmatrix} x_1 \\ \vdots \\ x_p \end{bmatrix} - (b + E_b) \right\|.$$

The cost function involves both a minimization and a maximization. The maximization will be handled first. Start by obtaining an upper bound on the maximization,

$$\begin{aligned}
&\left\| \begin{bmatrix} A_1^T + E_{A_1}^T \\ \vdots \\ A_p^T + E_{A_p}^T \end{bmatrix}^T \begin{bmatrix} x_1 \\ \vdots \\ x_p \end{bmatrix} - (b + E_b) \right\| \\
&\quad = \left\| \left(\begin{bmatrix} A_1^T \\ \vdots \\ A_p^T \end{bmatrix}^T \begin{bmatrix} x_1 \\ \vdots \\ x_p \end{bmatrix} - b \right) + \left(\sum_{i=1}^{p} E_{A_i} x_i - E_b \right) \right\| \\
&\quad \leq \|Ax - b\| + \sum_{i=1}^{p} \|E_{A_i} x_i\| + \|E_b\| \\
&\quad \leq \|Ax - b\| + \sum_{i=1}^{p} \eta_i \|x_i\| + \eta_b.
\end{aligned}$$

If there exists perturbations which make the cost function achieve the maximum, then those are the worst case perturbations and the maximization will be done. Consider the following perturbations,

$$\begin{aligned}
E_{A_i}^o &= \frac{\eta_i (Ax - b) x_i^T}{\|Ax - b\| \, \|x_i\|} \\
E_b^o &= \frac{-\eta_b (Ax - b)}{\|Ax - b\|}.
\end{aligned}$$

Note that,

$$\begin{aligned}
\left\| E_{A_i}^o \right\| &= \frac{\eta_i \left\| (Ax - b) x_i^T \right\|}{\|Ax - b\| \, \|x_i\|} \\
&\leq \eta_i \\
\|E_b^o\| &= \frac{\eta_b \left\| (Ax - b) \right\|}{\|Ax - b\|} \\
&\leq \eta_b.
\end{aligned}$$

Substituting this into the norm being maximized,

$$
\begin{aligned}
&\left\| \begin{bmatrix} A_1^T + E_{A_1}^T \\ \vdots \\ A_p^T + E_{A_p}^T \end{bmatrix}^T \begin{bmatrix} x_1 \\ \vdots \\ x_p \end{bmatrix} - (b + E_b) \right\| \\
&\qquad = \left\| (Ax - b) + \left(\sum_{i=1}^{p} \frac{\eta_i (Ax - b) \left\| x_i \right\|^2}{\left\| Ax - b \right\| \left\| x_i \right\|} + \frac{\eta_b (Ax - b)}{\left\| Ax - b \right\|} \right) \right\| \\
&\qquad = \left\| (Ax - b) \left(1 + \sum_{i=1}^{p} \frac{\eta_i \left\| x_i \right\|}{\left\| Ax - b \right\|} + \frac{\eta_b}{\left\| Ax - b \right\|} \right) \right\| \\
&\qquad = \left\| Ax - b \right\| \left(1 + \sum_{i=1}^{p} \frac{\eta_i \left\| x_i \right\|}{\left\| Ax - b \right\|} + \frac{\eta_b}{\left\| Ax - b \right\|} \right) \\
&\qquad = \left\| Ax - b \right\| + \sum_{i=1}^{p} \eta_i \left\| x_i \right\| + \eta_b.
\end{aligned}
$$

Thus, there are perturbations, which achieve the upper bound. Thus the perturbations above are the worst case perturbations and with these worst case perturbations, the maximum is achieved. Thus, the cost function can be simplified to

$$
\min_x \left(\left\| Ax - b \right\| + \sum_{i=1}^{p} \eta_i \left\| x_i \right\| + \eta_b \right).
$$

The cost function is clearly convex, as it is the sum of convex functions.

3.3 Quadratically Convergent Method

The cost function is not only convex, but it is also a sum of Euclidean norms. A large body of literature exists for solving the sum of Euclidean norms problem. The problem dates back to Fermat, who posed a special case of it. Various methods have been proposed which range from a sequence of linear least squares problems [158, 87, 37, 38, 150], successive over-relaxation [129], hyperbolic approximation procedure [46], subgradients [96, 29]. All of these have, at best, linear convergence, however there is a quadratically convergent method proposed by Michael Overton in [110]. Note that Overton's method is similar to [20].

Overton's method uses an active set and considers the projected objective function which is locally continuously differentiable. The basic idea of Overton's method is to use a relevant set of basis vectors to calculate the solution. An active set is then created in which the basis vectors that have solution components that are not near zero are kept. The basis vectors not in the active set are not differentiable and are thus temporarily inactive. The cost function is projected into the active set and a new solution is calculated. The new solution could cause members of the active set to become inactive or inactive members to become active. The process is continued until the solution converges, which Overton proved will happen quadratically. Interestingly, Overton notes that poor condition numbers actually increase the rate of convergence of his method. The actual numerical techniques of the algorithm are not relevant to this dissertation, as the reason for considering Overton's method is to survey the field and show that the problem is solvable in quadratic time.

A quadratically convergent method with good properties exists, so why look further? One major reason is that method operates on the size of the original problem (m), while a secular equation solution will operate on a smaller problem (p, with $p \ll m$ usually).

3.4 Form of Solution for Multiple Columns

The solution could be at either a differentiable point or a non-differentiable point. The non-differentiable points are located at $\|x_i\| \neq 0 \quad \forall i \in \{1, \cdots, p\}$ and $\|Ax - b\| \neq 0$. This section will consider the case when the solution is at a differentiable point. A necessary condition for a minimum at a differentiable point is obtained by taking the gradient and setting it equal to zero. Before taking the gradient, consider an $n \times n$ identity matrix that has been partitioned into columns like A. Define the i^{th} column of the partitioned identity matrix as $\mathcal{I}_i$, and note that $\mathcal{I}_i \in \mathbb{R}^{n \times n_i}$.

$$\begin{aligned} 0 &= \frac{A^T (A\hat{x} - b)}{\|A\hat{x} - b\|} + \sum_{i=1}^{p} \frac{\eta_i}{\|\hat{x}_i\|} \mathcal{I}_i \hat{x}_i \\ &= A^T (A\hat{x} - b) + \sum_{i=1}^{p} \frac{\eta_i \|A\hat{x} - b\|}{\|\hat{x}_i\|} \mathcal{I}_i \hat{x}_i. \end{aligned} \tag{3.3}$$

Now, define the following constants,

$$\begin{aligned} \psi_1 &= \frac{\eta_1 \|A\hat{x} - b\|}{\|\hat{x}_1\|} > 0 \\ &\vdots \\ \psi_p &= \frac{\eta_p \|A\hat{x} - b\|}{\|\hat{x}_p\|} > 0. \end{aligned}$$

Using these definitions in Equation 3.3 obtain,

$$\begin{aligned} 0 &= A^T (A\hat{x} - b) + \sum_{i=1}^{p} \psi_i \mathcal{I}_i \hat{x}_i \\ &= A^T A\hat{x} - A^T b + \begin{bmatrix} \psi_1 I & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \psi_p I \end{bmatrix} \hat{x}. \end{aligned}$$

For simplicity make the following notation,

$$\Psi = \begin{bmatrix} \psi_1 I & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \psi_p I \end{bmatrix}.$$

Thus the equations become, after taking $A^T b$ to the other side,

$$A^T b = \left(A^T A + \Psi\right) \hat{x}.$$

Note that A^TA is positive semi-definite and Ψ is positive definite, so that the matrix multiplying $\hat{x}$ is invertible. Since, $\left(A^TA+\Psi\right)$ is invertible, $\hat{x}$ can be solved for,

$$\begin{aligned}
\hat{x} &= \left(A^TA+\Psi\right)^{-1}A^Tb \\
\hat{x}_1 &= \mathcal{I}_1^T\left(A^TA+\Psi\right)^{-1}A^Tb \\
&\vdots \\
\hat{x}_p &= \mathcal{I}_p^T\left(A^TA+\Psi\right)^{-1}A^Tb.
\end{aligned}$$

The following calculations are independent of notation, so it will be solved for $\hat{x}_1$ and note that the solution for $\hat{x}_i$ will follow directly. At this point, consider the inverse needed to calculate $\hat{x}$. In particular, use a block inverse form from [88].

$$\begin{bmatrix} E & B \\ C & D \end{bmatrix}^{-1} = \begin{bmatrix} F & -FBD^{-1} \\ -D^{-1}CF & D^{-1}+D^{-1}CFBD^{-1} \end{bmatrix}$$

Where $F = (E - BD^{-1}C)^{-1}$. The blocks are defined as follows,

$$\begin{aligned}
H &= [A_2 \cdots A_p] \\
\Psi_1 &= \begin{bmatrix} \psi_2 I & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \psi_p I \end{bmatrix} \\
E &= A_1^T A_1 + \psi_1 I \\
B &= A_1^T \begin{bmatrix} A_2 & \cdots & A_p \end{bmatrix} \\
&= A_1^T H \\
C &= B^T \\
&= H^T A_1 \\
D &= \begin{bmatrix} A_2^T A_2 + \psi_2 I & A_2^T A_3 & \cdots & A_2^T A_p \\ A_3^T A_2 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & A_{p-1}^T A_p \\ A_p^T A_2 & \cdots & A_p^T A_{p-1} & A_p^T A_p + \psi_p I \end{bmatrix} \\
&= H^T H + \Psi_1.
\end{aligned}$$

Now using this form on the desired inverse and obtain,

$$\begin{aligned}
\left(A^T A + \Psi\right)^{-1} &= \begin{bmatrix} F & -F A_1^T H D^{-1} \\ -D^{-1} H^T A_1 F & D^{-1} + D^{-1} H^T A_1 F A_1^T H D^{-1} \end{bmatrix} \\
F &= \left(A_1^T A_1 + \psi_1 I - A_1^T H D^{-1} H^T A_1\right)^{-1} \\
&= \left(\psi_1 I - A_1^T \left(I - H D^{-1} H^T \right) A_1\right)^{-1} \\
\hat{x}_1 &= E_1^T \left(A^T A + \Psi\right)^{-1} A^T b \\
&= F A_1^T b - F A_1^T H D^{-1} H^T b \\
&= F A_1^T \left(I - H D^{-1} H^T\right) b.
\end{aligned}$$

Now note that $\left(I - HD^{-1}H^T\right)$ is in the form of the Sherman-Morrison-Woodbury formula, which is also known as the matrix inversion lemma [88].

$$\left(I - H\left(H^T H + \Psi_1\right)^{-1} H^T\right) = \left(I + H\Psi_1^{-1}H^T\right)^{-1}$$

Additionally, F is also in the form of the Sherman-Morrison-Woodbury formula by using this result,

$$\begin{aligned}
\left(E - BD^{-1}C\right)^{-1} &= \left(A_1^T A_1 + \psi_1 - A_1^T H D^{-1} H^T A_1 I\right)^{-1} \\
&= \left(A_1^T \left(I + HD^{-1}H^T\right)^{-1} A_1 + \psi_1 I\right)^{-1} \\
&= \left(A_1^T \left(I + H\Psi_1^{-1}H^T\right)^{-1} A_1 + \psi_1 I\right)^{-1} \\
&= \frac{1}{\psi_1}\left(I - \frac{1}{\psi_1}A_1^T\left(I + \frac{1}{\psi_1}A_1A_1^T + H\Psi_1^{-1}H^T\right)^{-1} A_1\right) \\
&= \frac{1}{\psi_1}\left(I - \frac{1}{\psi_1}A_1^T\left(I + \sum_{i=1}^{p}\frac{1}{\psi_i}A_iA_i^T\right)^{-1} A_1\right).
\end{aligned}$$

Now consider $\hat{x}_1$ in light of this,

$$\begin{aligned}
\hat{x}_1 &= \frac{1}{\psi_1}\left(I - \frac{1}{\psi_1}A_1^T\left(I + \sum_{i=1}^{p}\frac{1}{\psi_i}A_iA_i^T\right)^{-1} A_1\right) A_1^T \left(I + \sum_{i=2}^{p}\frac{1}{\psi_i}A_iA_i^T\right)^{-1} b \\
&= \frac{1}{\psi_1}A_1^T\left(I - \frac{1}{\psi_1}\left(I + \frac{1}{\psi_1}A_1A_1^T + \cdots + \frac{1}{\psi_p}A_pA_p^T\right)^{-1} A_1A_1^T\right) \\
&\qquad \left(I + \sum_{i=2}^{p}\frac{1}{\psi_i}A_iA_i^T\right)^{-1} b \\
&= \frac{1}{\psi_1}A_1^T\left(I + \sum_{i=1}^{p}\frac{1}{\psi_i}A_iA_i^T\right)^{-1}\left(I + \sum_{i=1}^{p}\frac{1}{\psi_i}A_iA_i^T - \frac{1}{\psi_1}A_1A_1^T\right) \\
&\qquad \left(I + \sum_{i=2}^{p}\frac{1}{\psi_i}A_iA_i^T\right)^{-1} b \\
&= \frac{1}{\psi_1}A_1^T\left(I + \sum_{i=1}^{p}\frac{1}{\psi_i}A_iA_i^T\right)^{-1} b.
\end{aligned}$$

Thus, in general the form of $\hat{x}_i$ is

$$\begin{aligned}
\hat{x}_i &= \frac{1}{\psi_i} A_i^T \left(I + \sum_{i=1}^{p} \frac{1}{\psi_i} A_i A_i^T \right)^{-1} b \\
&= \frac{1}{\psi_i} A_i^T \left(I + A\Psi^{-1} A^T \right)^{-1} b.
\end{aligned}$$

Given these simplifications it is easier to express $A\hat{x} - b$,

$$\begin{aligned}
A\hat{x} - b &= \sum_{i=1}^{p} A_i \hat{x}_i - b \\
&= \left(\sum_{i=1}^{p} \frac{1}{\psi_i} A_i A_i^T \right) \left(I + \sum_{i=1}^{p} \frac{1}{\psi_i} A_i A_i^T \right)^{-1} b - b \\
&= - \left(I + \sum_{i=1}^{p} \frac{1}{\psi_i} A_i A_i^T \right)^{-1} b \\
&= - \left(I + A\Psi^{-1} A^T \right)^{-1} b.
\end{aligned}$$

3.5 General Column Form Secular Equation

The secular equations for this problem are now developed. First, square the definition of ψ_i.

$$\psi_i^2 \left\| \hat{x}_i \right\|^2 = \eta_i^2 \left\| A\hat{x} - b \right\|^2$$

Then using the expressions derived for $\hat{x}_i$ and $A\hat{x} - b$, define the secular equations, G_i ($\forall i \in 1, \ldots, p$), to be

$$\begin{aligned}
G_i \left(\psi_1, \cdots, \psi_p \right) &= \psi_i^2 \left\| \hat{x}_i \right\|^2 - \eta_i^2 \left\| A\hat{x} - b \right\|^2 \\
&= b^T \left(I + A\Psi^{-1} A^T \right)^{-1} \left(A_i A_i^T - \eta_i^2 I \right) \left(I + A\Psi^{-1} A^T \right)^{-1} b.
\end{aligned}$$

For simplicity, make the following definitions, and note that the definition of F is positive definite for all positive values of ψ_i.

$$\begin{aligned}
F &= \left(I + A\Psi^{-1} A^T \right)^{-1} \\
N_i &= \left(A_i A_i^T - \eta_i^2 I \right)
\end{aligned}$$

The secular equations become

$$G_i(\psi) = b^T F N_i F b. \tag{3.4}$$

Note that the secular equations ($G_i(\psi)$,$i = 1, 2, \cdots, p$) have no singularities in the first quadrant and since the equations are rational expressions of ψ_i the functions are C^1 in the first quadrant. All that remains is to show the existence and uniqueness of the solution.

3.5.1 Uniqueness

First, the uniqueness of the solution will be shown. To do this, it will be shown that the cost function is strictly convex and thus any solution to the original problem is unique. Since the secular equations only have a root when the original problem has a solution, this will show that any solution to the secular equation is unique. To show the original problem is strictly convex in the region of interest for the problem, consider the Hessian of the cost, H,

$$\begin{aligned} H &= \frac{1}{\|Ax-b\|}\left(A^TA - \frac{A^T(Ax-b)(Ax-b)^TA}{\|Ax-b\|^2} + \Psi - \mathrm{diag}\left(\psi_i \frac{x_i x_i^T}{\|x_i\|}\right)\right) \\ &= \frac{1}{\|Ax-b\|}\left(A^T P_{Ax-b}^{\perp} A + \Psi \mathrm{diag}(P_{x_i}^{\perp})\right), \end{aligned}$$

where P is a projection matrix and its subscript specifies the space it projects onto. In order for the Hessian to be positive semi-definite there must be a column of A, say A_{i_k}, that is in the i^{th} partition and a corresponding element of x called x_{i_k} for which both

1. $P_{Ax-b}^{\perp} A_{i_k} = 0$,

2. $\psi_i e_{i_k}^T P_{x_i}^{\perp} e_{i_k} = 0$,

where e_{i_k} is a vector that is zero everywhere except the component in the i_k^{th} position, which is 1. In order for item 1 to hold, A_{i_k} must be in the direction of the residual, which means that $b \in \mathcal{R}(A)$. By assuming the standard condition that $b \notin \mathcal{R}(A)$, the first term is positive and thus the Hessian is positive definite. Note that if $b \in \mathcal{R}(A)$, then the secular equation has its root at $\psi_i = 0$ for all $i = 1, 2, \ldots$. This makes the solution to the multi-column problem the same as the least squares problem, and since $b \in \mathcal{R}(A)$ by assumption, this means the residual is zero. A zero residual makes the space perpendicular to the residual to be the identity matrix, so item 1 remains positive. The only remaining possibility is for $A_{i_k} = 0$, which means that the corresponding component of x, x_{i_k}, is zero. Since $x_{i_k} = 0$ this makes $e_{i_k}^T P_{x_i}^{\perp} e_{i_k} = 1$ and thus $\psi_i e_{i_k}^T P_{x_i}^{\perp} e_{i_k} = \psi_i$. For ψ_i to be zero, the solution must be the least squares solution, thus $b \in \mathcal{R}(A)$. The only way for the Hessian to be positive semi-definite is for both $b \in \mathcal{R}(A)$ and $A_{i_k} = 0$. Excluding this situation yields the desired uniqueness condition.

3.5.2 Existence

Two things must be shown for our general form of the secular equation. First, it must be shown that as $\psi_i \to 0$ that $G_i < 0$. Second, it must be shown that as $\psi_i \to \infty$ that $G_i > 0$. These two conditions will guarantee a solution in the first quadrant.

Lemma 3.3 (Negative Side Lemma)

$$\lim_{\psi_i \to 0} G_i = \lim_{\psi_i \to 0} b^T F N_i F b < 0$$

Proof:

For notational simplicity define the following

$$\begin{aligned}
A_{/i} &= \begin{bmatrix} A_1 & \cdots & A_{i-1} & A_{i+1} & \cdots & A_p \end{bmatrix} \\
\Psi_{/i} &= diag(\psi_1 I, \cdots, \psi_{i-1} I, \psi_{i+1} I, \cdots, \psi_p I) \\
F_{/i} &= \left(I + A_{/i} \Psi_{/i}^{-1} A_{/i}^T \right)^{-1} \\
&= M_{/i} M_{/i}^T \\
\tilde{A}_i &= M_{/i}^T A_i \\
\tilde{b} &= M_{/i}^T b.
\end{aligned}$$

Use the matrix inversion lemma to get an expression for F as $\psi_i \to 0$ in terms of these definitions,

$$\begin{aligned}
\lim_{\psi_i \to 0} F &= \lim_{\psi_i \to 0} \left(F_{/i}^{-1} + \frac{1}{\psi_i} A_i A_i^T \right)^{-1} \\
&= \lim_{\psi_i \to 0} F_{/i} - F_{/i} A_i \left(\psi_i I + A_i^T F_{/i} A_i \right)^{-1} A_i^T F_{/i} \\
&= M_{/i} \left(I - \tilde{A}_i \left(\psi_i I + \tilde{A}_i^T \tilde{A}_i \right)^{-1} \tilde{A}_i^T \right) M_{/i}^T \\
&= M_{/i} \left(I - \tilde{A}_i \tilde{A}_i^+ \right) M_{/i}^T.
\end{aligned}$$

The last expression contains the projection onto the space perpendicular to the range of $\tilde{A}_i$. Note that this means the $\lim_{\psi_i \to 0} F$ will project anything in the range of A_i to zero. Use this in the expression for $\lim_{\psi_i \to 0} G_i$,

$$\begin{aligned}
\lim_{\psi_i \to 0} G_i &= b^T F_{/i} (A_i A_i^T - \eta_i^2 I) F_{/i} b \\
&= -\eta_i^2 b^T F_{/i}^2 b \\
&\leq 0.
\end{aligned}$$

$\diamondsuit$ SDG $\diamondsuit$

The secular equations are always negative (or zero) when the corresponding value of ψ is zero. Now it remains to be shown that as ψ_i goes to infinity, that

$G_i \geq 0$ or that the solution can be found at one of the extrema. Before proving the first part, an intermediary result about F is needed.

Lemma 3.4 (Derivative of F is Positive Lemma)

$$\frac{\partial}{\partial \psi_i} F \geq 0$$

Proof:

$$\begin{aligned}
\frac{\partial}{\partial \psi_i} F &= \frac{\partial}{\partial \psi_i} \left(I + \sum_{i=1}^{p} \frac{1}{\psi_i} A_i A_i^T \right)^{-1} \\
&= \frac{1}{\psi_i^2} F A_i A_i^T F \\
&\geq 0
\end{aligned}$$

♢ SDG ♢

This will allow simplification of the argument that follows.

Lemma 3.5 (Positive Side Lemma)

$$\begin{aligned}
\lim_{\psi_i \to \infty} G_i &= \lim_{\psi_i \to \infty} b^T F N_i F b \\
&> 0
\end{aligned}$$

if

$$\eta_i \leq \frac{\left\| A_i^T b \right\|}{\|b\|}$$

Proof:

First examine what happens to F as $\psi_i \to \infty$,

$$\begin{aligned}
\lim_{\psi_i \to \infty} F &= \lim_{\psi_i \to \infty} \left(I + \sum_{i=1}^{p} \frac{1}{\psi_i} A_i A_i^T \right)^{-1} \\
&= \left(I + \sum_{i \neq i} \frac{1}{\psi_i} A_i A_i^T \right)^{-1} \\
&= F_{/i}.
\end{aligned}$$

Note that the singular values of F vary between zero and one. This can be easily seen as it is the inverse of the identity matrix plus some positive semi-definite matrices. Also note that since the size of F increases with increasing ψ_j, only consider the point where $\psi_j \to \infty$. This point is selected because it maximizes the negative term. When all $\psi \to \infty$ it forces $F = I$, so

$$\begin{aligned}
\lim_{\psi \to \infty} G_i &= b^T N_i b \\
&= b^T (A_i A_i^T - \eta_i^2 I) b.
\end{aligned}$$

For this to be non-negative it must be that

$$b^T A_i A_i^T b \geq \eta_i^2 b^T b.$$

Rearranging terms yields

$$\begin{aligned}
\eta_i^2 &\leq \frac{b^T A_i A_i^T b}{b^T b} \\
&= \frac{\left\| A_i^T b \right\|^2}{\|b\|^2}.
\end{aligned}$$

Taking the square root gives the solution

$$\eta_i \leq \frac{\left\| A_i^T b \right\|}{\|b\|}.$$

$\diamondsuit$ SDG $\diamondsuit$

The only thing left is to observe that the problem is strictly convex and does not have any place it is undefined, thus there is always a solution. Recall that earlier in the chapter it was shown that the requirement from Positive Side Lemma does not need to be fulfilled to have a non-zero answer. In that case the solution was at the extremum defined by the least squares condition for columns two and three. In other words, in that case the solution did not satisfy the Positive Side Lemma and thus the solution was at an extremum. The solution is thus characterized. Any multi-dimensional root finder can be used to calculate the actual location.

3.6 A Numerical Example

The following problem is based on an example of Dr. Ali Sayed in an unpublished paper entitled "Estimation in the Presence of Multiple Sources of Uncertainties with Applications". Assume that there are two different signals that need to be estimated from a series of three simultaneous observations. The relation between the signals and the observations are known approximately and are the A matrix. Additionally, assume the first signal is stronger and that the errors associated with the first signal are smaller.

First consider the case of singular A. This is shown in Figure 3.1. Least squares can only estimate the stronger signal, but does a reasonable job at it. The multi-column solution does quite well for the first signal, and gets basic features and is a reasonable scale for the second. Note that as is typical for a pessimistic problem, the multi-column min max tends to underestimate the size of the signal, but this underestimation is better than the alternatives. Total least squares is shown in Figure 3.2 because it is not even close, notice the order of magnitude is off by around 14.

Now consider the case of a near singular A. This is shown in Figure 3.3. Least squares and total least squares are almost identical for this problem, and off by a factor of two to seven. The multi-column solution is very good for first signal and reasonable for the second. Note the multi-column min max does not change significantly between the two cases. This is a result of the robustness of the solution. A solution for the min max problem works for nearby problems, so it tends not to change for small alterations in the problem, even when the change tends to cause a major change in other methods.

3.7 Summary

The multiple column min max problem has been posed and solved. Several techniques for solution are presented but the best technique is to use the secular equation because it is usually a much smaller problem. Overton's quadratically convergent method for the sum of Euclidean norms could be used, and can be faster if $m \approx p$ and the problem is ill-conditioned. Overton's method can converge faster when the problem is ill-conditioned, but note that for $m \approx p$ the problem must be at least nearly square with the partitions being individual columns of A. The conditions for the secular equation to work better are much more likely and thus are the advised solution technique.

The multiple column min max problem should be used instead of the regular min max problem if there is a significant difference in the bounds on some block columns. If the bounds are similar there is not a significant difference, but there is a processing cost difference. The usual case when the min max formulation has significant advantages over the least squares and total least squares is when

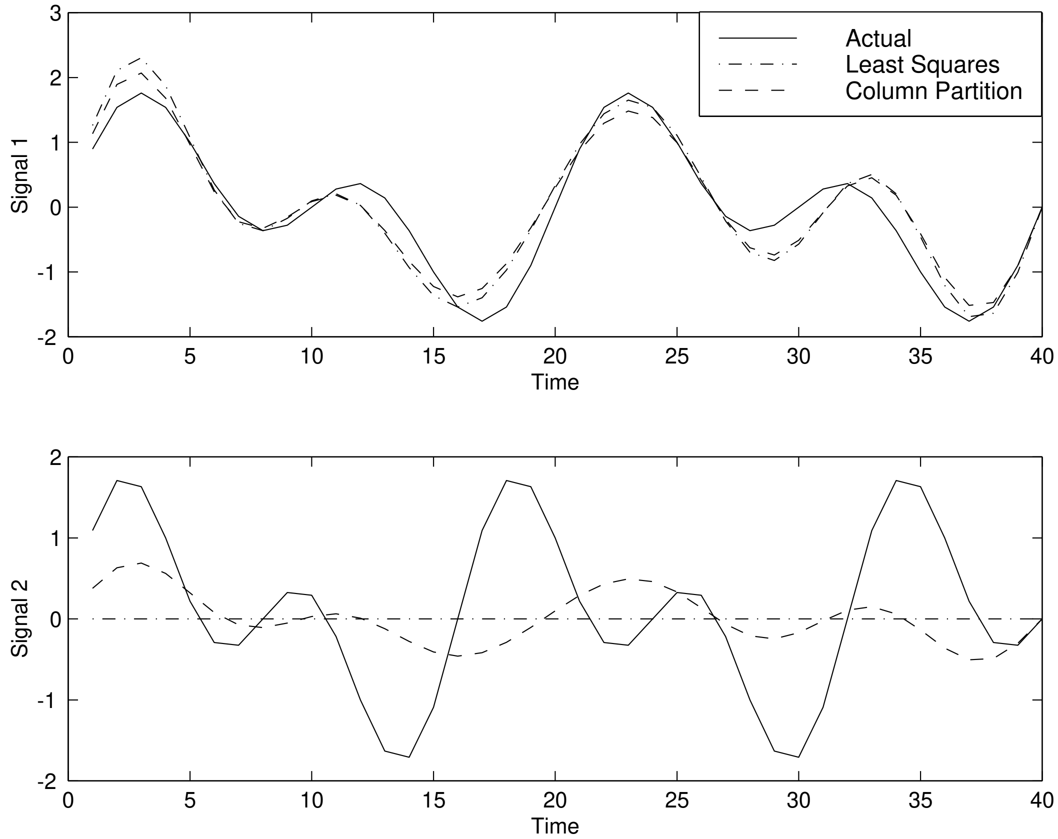

Figure 3.1. Singular Matrix Signal Separation Problem

the model has conditioning problems. Without conditioning problems the standard techniques give answers, which are reasonably close to the min max, and sometimes give better answers if the error bounds are over-estimated. When the conditioning problems exist, however the min max solution can maintain a reasonably good solution into areas where the other techniques are not capable. In cases where matrix structure is the key goal, and robustness is desired, the LMI techniques in Section 2.10 should be used. The min max solution has a viable use, and reasonable calculation time making it a very useful tool to have.

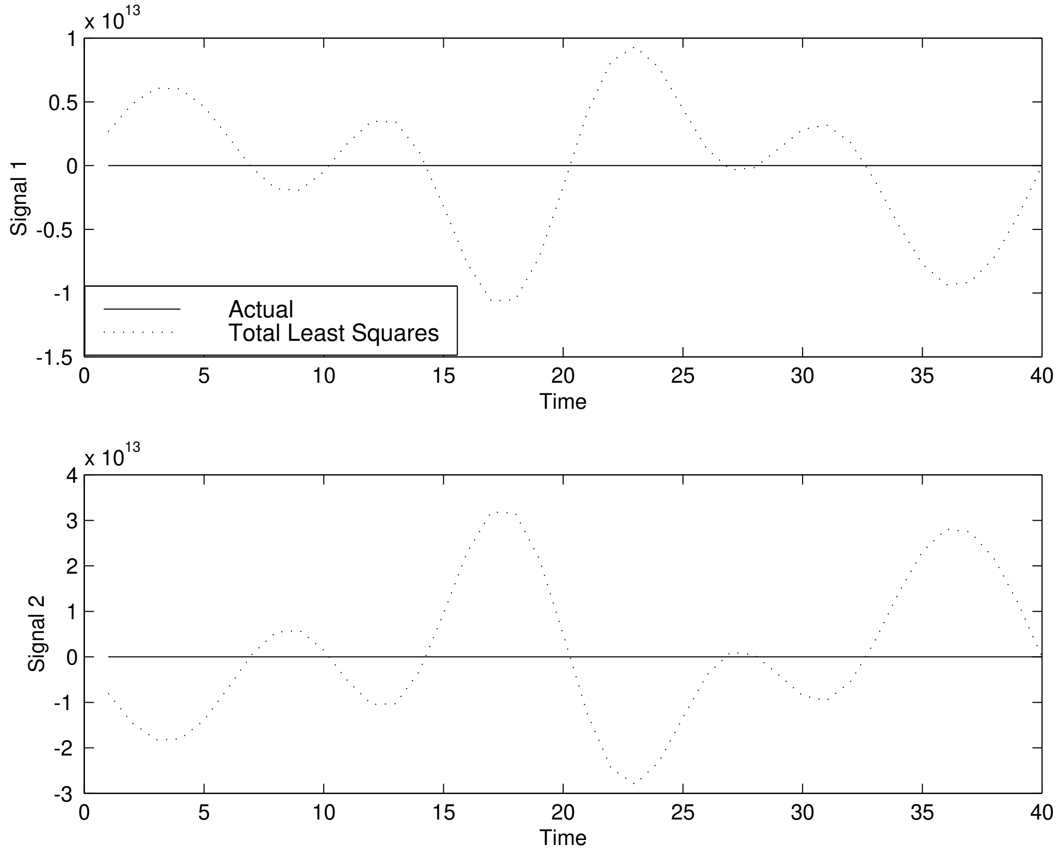

Figure 3.2. TLS Solution to Singular Matrix Signal Separation Problem

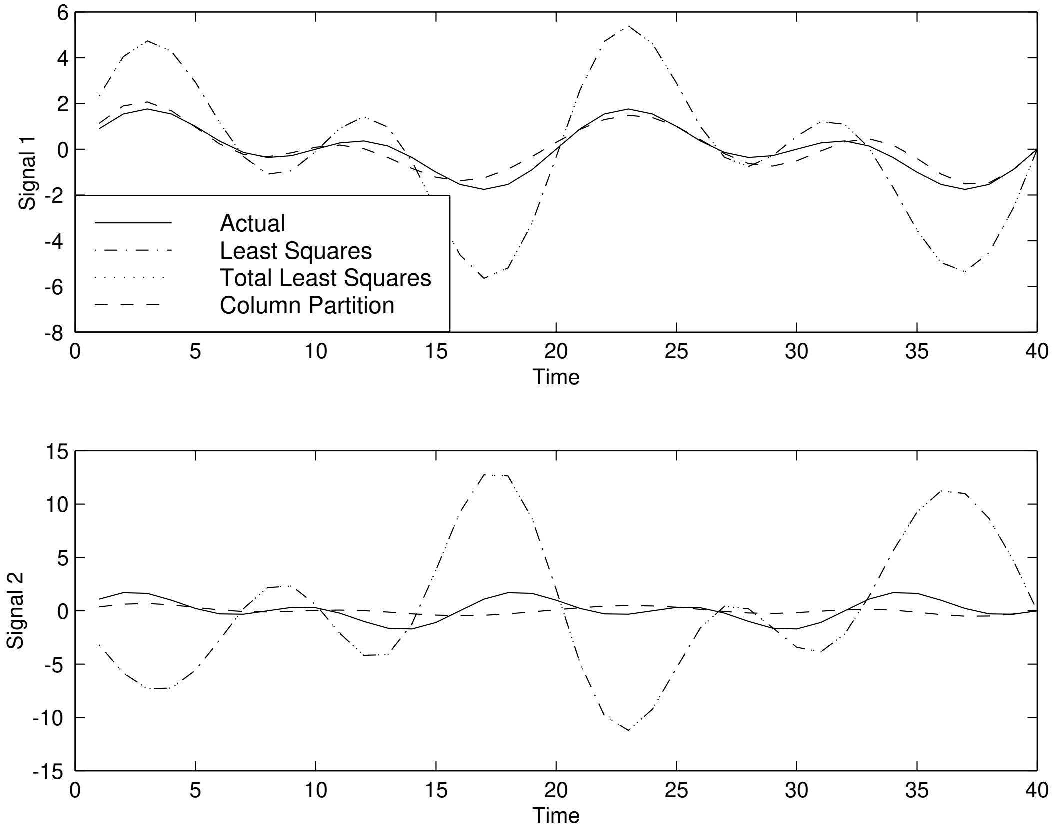

Figure 3.3. Near Singular Matrix Signal Separation Problem, TLS in on LS line

Chapter 4

Multi-Row Min Max Criterion

The multiple (block) column case has been solved. It is reasonable to ask if the multiple (block) row case can be solved. That will be the goal of this chapter. The basic problem can be stated as

$$C_{row} = \min_{x} \max_{\substack{\|E_i\|_2 \leq \eta_i \\ \|E_{b,i}\|_2 \leq \eta_{b,i}}} \frac{1}{2} \left\| \begin{bmatrix} A_1 + E_1 \\ \vdots \\ A_q + E_q \end{bmatrix} [x] - \begin{bmatrix} b_1 + E_{b,1} \\ \vdots \\ b_q + E_{b,q} \end{bmatrix} \right\|^2 \tag{4.1}$$

where $i = 1, 2, \ldots, q$ and the norm has been squared, and a factor of $\frac{1}{2}$ has been inserted to make later formulas neater.

The flow of the chapter is as follows. In Section 4.1, the maximization of the perturbations is solved. In Section 4.2 the form of solution is shown. In Section 4.3 the secular is found and basic properties are shown. In Section 4.4 the non-differentiable points are discussed and a sufficient condition on the η_i is developed for $x = 0$. Finally in Section 4.5 the algorithm to solve the problem is outlined.

4.1 An Equivalent Formulation

The multiple row case is similar to the multiple column case in many ways. One way is that it can be reduced to a simpler problem, where the maximization has already been done. Begin by considering the norm in Equation 4.1

$$\begin{aligned} C_{row} &= \left\| \begin{bmatrix} A_1 + E_1 \\ \vdots \\ A_q + E_q \end{bmatrix} [x] - \begin{bmatrix} b_1 + E_{b,1} \\ \vdots \\ b_q + E_{b,q} \end{bmatrix} \right\|^2 \\ &= \sum_{i=1}^{q} \|(A_i + E_i)x - (b_i + E_{b,i})\|^2 \\ &= \sum_{i=1}^{q} \|(A_i x - b_i) + (E_i x - E_{b,i})\|^2 \\ &\leq \sum_{i=1}^{q} \left(\|A_i x - b_i\| + \|E_i\|\|x\| + \|E_{b,i}\|\right)^2 . \end{aligned}$$

And by using the bounds on the perturbations

$$C_{row} \leq \sum_{i=1}^{q} \left(\|A_i x - b_i\| + \eta_i \|x\| + \eta_{b,i}\right)^2 . \tag{4.2}$$

Now consider the perturbations

$$\begin{aligned} E_i &= \eta_i \frac{(A_i x - b_i)x^T}{\|A_i x - b_i\|\|x\|} \\ E_{b,i} &= -\eta_{b,i} \frac{A_i x - b_i}{\|A_i x - b_i\|}. \end{aligned}$$

Note that

$$
\begin{aligned}
\|E_i\| &= \|\eta_i \frac{(A_i x - b_i)x^T}{\|A_i x - b_i\|\|x\|}\| \\
&= \eta_i \frac{\|(A_i x - b_i)x^T\|}{\|A_i x - b_i\|\|x\|} \\
&\leq \eta_i \frac{\|A_i x - b_i\|\|x^T\|}{\|A_i x - b_i\|\|x\|} \\
&\leq \eta_i \\
\|E_{b,i}\| &= \|\eta_{b,i} \frac{A_i x - b_i}{\|A_i x - b_i\|}\| \\
&= \eta_{b,i} \frac{\|A_i x - b_i\|}{\|A_i x - b_i\|} \\
&= \eta_{b,i}.
\end{aligned}
$$

Thus these perturbations meet the bounds. Substituting these bounds in the norm to be maximized yields

$$
\begin{aligned}
C_{row} &= \sum_{i=1}^{q} \|(A_i + E_i)x - (b_i + E_{b,i})\|^2 \\
&= \sum_{i=1}^{q} \|(A_i x - b_i) + (E_i x - E_{b,i})\|^2 \\
&= \sum_{i=1}^{q} \left\|(A_i x - b_i) + \left(\eta_i \frac{(A_i x - b_i)x^T}{\|A_i x - b_i\|\|x\|} x + \eta_{b,i} \frac{A_i x - b_i}{\|A_i x - b_i\|}\right)\right\|^2 \\
&= \sum_{i=1}^{q} \left\|(A_i x - b_i)\left(1 + \frac{\eta_i \|x\|}{\|A_i x - b_i\|} + \frac{\eta_{b,i}}{\|A_i x - b_i\|}\right)\right\|^2 \\
&= \sum_{i=1}^{q} \left(\|A_i x - b_i\|\left(1 + \frac{\eta_i \|x\|}{\|A_i x - b_i\|} + \frac{\eta_{b,i}}{\|A_i x - b_i\|}\right)\right)^2 \\
&= \sum_{i=1}^{q} \left(\|A_i x - b_i\| + \eta_i \|x\| + \eta_{b,i}\right)^2.
\end{aligned}
$$

Thus there exists a perturbation which reaches the upper bound in Equation 4.2 and so Equation 4.1 is equivalent to

$$
C_{row} = \min_x \frac{1}{2} \sum_{i=1}^{q} \left(\|A_i x - b_i\| + \eta_i \|x\| + \eta_{b,i}\right)^2. \tag{4.3}
$$

This is the cost function that will be used. First note that the norms are convex, thus so is their sum. The square of a convex function is convex, and so the cost function is convex. While the cost function is convex, it is no longer the sum of Euclidean norms (due to the square). The quadratically convergent method of Overton will not work on this form, and no rewriting has been found to get the equation into the correct form. Bundle methods should work as should any general convex solver that allows for discontinuities, but a better solution is desired.

4.2 Form of Solution

The non-differentiable points for the multiple row case are $\|A_i x - b_i\| = 0$ for all $i = 1, 2, \ldots, q$ and $x = 0$. The differentiable points will be dealt with in this section. Take the gradient of Equation 4.3 and set it equal to zero, to get the necessary conditions for a minimum.

$$\begin{aligned} 0 &= \sum_{i=1}^{q} \left(\|A_i x - b_i\| + \eta_i \|x\| + \eta_{b,i}\right) \left(\frac{A_i^T (A_i x - b_i)}{\|A_i x - b_i\|} + \eta_i \frac{x}{\|x\|} \right) \\ &= \sum_{i=1}^{q} \left(\left(1 + \frac{\eta_i \|x\| + \eta_{b,i}}{\|A_i x - b_i\|} \right) A_i^T (A_i x - b_i) + \eta_i^2 \frac{1 + \frac{\eta_i \|x\| + \eta_{b,i}}{\|A_i x - b_i\|}}{\frac{\eta_i \|x\|}{\|A_i x - b_i\|}} x \right) \end{aligned}$$

Make the following definitions:

$$\begin{aligned} \zeta_i &= \frac{\eta_i \|x\|}{\|A_i x - b_i\|} \\ \delta_i &= 1 + \zeta_i + \frac{\eta_{b,i}}{\|A_i x - b_i\|} \\ \Delta &= \operatorname{diag}(\delta_1 I, \delta_2 I, \ldots, \delta_q I) \\ \psi &= \sum_{i=1}^{q} \eta_i^2 \frac{\delta_i}{\zeta_i}. \end{aligned}$$

Then the gradient expression becomes

$$\begin{aligned}
0 &= \sum_{i=1}^{q}\left(\delta_i A_i^T(A_i x - b_i) + \eta_i^2 \frac{\delta_i}{\zeta_i} x\right) \\
&= A^T \Delta (Ax - b) + \left(\sum_{i=1}^{q} \eta_i^2 \frac{\delta_i}{\zeta_i}\right) x \\
&= A^T \Delta (Ax - b) + \psi x.
\end{aligned}$$

Solving for x yields

$$x = (A^T \Delta A + \psi I)^{-1} A^T \Delta b.$$

This can be simplified once more by defining $\phi_i = \frac{\delta_i}{\psi}$ and $\Phi = \frac{\Delta}{\psi}$ yielding

$$x = (A^T \Phi A + I)^{-1} A^T \Phi b \tag{4.4}$$

$$\Phi = \operatorname{diag}(\phi_1 I, \phi_2 I, \ldots, \phi_q I) \tag{4.5}$$

$$\phi_i = \frac{1 + \zeta_i + \frac{\eta_{b,i}}{\|A_i x - b_i\|}}{\sum_{j=1}^{q} \eta_j^2 \frac{1+\zeta_j+\frac{\eta_{b,j}}{\|A_j x - b_j\|}}{\zeta_j}} \tag{4.6}$$

$$\zeta_i = \frac{\eta_i \|x\|}{\|A_i x - b_i\|}. \tag{4.7}$$

4.3 Secular Equation

In this section secular equations are obtained for the ζ_i. When all the $\eta_{b,i} = 0$ this is all that needs to be done. When $\eta_{b,i} \neq 0$ the value of $\|A_i x - b_i\|$ is also needed. In this case two methods are suggested, to handle the added difficulty. The first way is to use a continuation method, starting with $\hat{\eta}_{b,i} = 0$ and progressing to $\hat{\eta}_{b,i} = \eta_{b,i}$. The second way is to note that $\|A_i x - b_i\| = \frac{\eta_i}{\zeta_i}\|x\|$ and then the only additional information needed is the value of $\|x\|$. This is shown in Appendix C to be bounded by

$$0 \leq \frac{\sqrt{a_1^2 + 4a_2 r_{ls}} - a_1}{2a_2} \leq \|x\| \leq \frac{\sqrt{a_1^2 + 4a_2 a_0} - a_1}{2a_2}$$

with

$$
\begin{aligned}
a_2 &= \sum_{i=1}^{q} \eta_i^2 \left(1 + \frac{1}{\zeta_i}\right)^2 \\
a_1 &= \sum_{i=1}^{q} 2\eta_{b,i}\eta_i \left(1 + \frac{1}{\zeta_i}\right) \\
a_0 &= \sum_{i=1}^{q} \left(\|b_i\|^2 + 2\eta_{b,i}\|b_i\|\right)
\end{aligned}
$$

and r_{ls} is the square of the LS residual, i.e. $\|(I - AA^\dagger)b\|^2$. The lower bound for $\|x\|$ is the more important one for the estimation of $\|A_i x - b_i\|$, as the greatest influence of that term is when it is small, since it appears in the denominator.

Returning to the secular equation, note that Equation 4.7 can be rewritten as

$$
\frac{\zeta_i^2}{\eta_i^2}\|A_i x - b_i\|^2 = \|x\|^2.
$$

Define the secular equations to be

$$
g_i(\zeta_1, \zeta_2, \ldots, \zeta_q) = \frac{\zeta_i^2}{\eta_i^2}\|A_i x - b_i\|^2 - \|x\|^2, \tag{4.8}
$$

for $i = 1, 2, \ldots, q$. For ease of writing the norms in the secular equations first use the Sherman-Morrison-Woodbury formula to rewrite Equation 4.4,

$$
\begin{aligned}
x &= \left(A^T \Phi A + I\right)^{-1} A^T \Phi b \\
&= \left(I - A^T \left(AA^T + \Phi^{-1}\right)^{-1} A\right) A^T \Phi b \\
&= A^T \left(I - \left(AA^T + \Phi^{-1}\right)^{-1} AA^T\right) \Phi b \\
&= A^T \left(AA^T + \Phi^{-1}\right)^{-1} \left(\left(AA^T + \Phi^{-1}\right) - AA^T\right) \Phi b \\
&= A^T \left(AA^T + \Phi^{-1}\right)^{-1} \Phi^{-1} \Phi b \\
&= A^T \left(AA^T + \Phi^{-1}\right)^{-1} b.
\end{aligned}
$$

Using this form the residual can be written as

$$\begin{aligned} Ax - b &= AA^T \left(AA^T + \Phi^{-1}\right)^{-1} b - b \\ &= \left(AA^T - \left(AA^T + \Phi^{-1}\right)\right) \left(AA^T + \Phi^{-1}\right)^{-1} b \\ &= -\Phi^{-1} \left(AA^T + \Phi^{-1}\right)^{-1} b. \end{aligned}$$

Thus

$$\|\hat{x}\|^2 = b^T \left(AA^T + \Phi^{-1}\right)^{-1} AA^T \left(AA^T + \Phi^{-1}\right)^{-1} b. \tag{4.9}$$

Row partition an $m \times m$ identity matrix compatibly with how A was partitioned

$$I_{m \times m} = \begin{bmatrix} \mathcal{I}_1 \\ \vdots \\ \mathcal{I}_q \end{bmatrix}, \tag{4.10}$$

thus $A_i = \mathcal{I}_i A$ and

$$\|A_i x - b_i\|^2 = \frac{1}{\phi_i^2} b^T \left(AA^T + \Phi^{-1}\right)^{-1} \mathcal{I}_i^T \mathcal{I}_i \left(AA^T + \Phi^{-1}\right)^{-1} b. \tag{4.11}$$

Using all the above, the secular equations can be written as

$$g_i(\zeta_1, \zeta_2, \ldots, \zeta_q) = \frac{\zeta_i^2}{\eta_i^2} \|A_i x - b_i\|^2 - \|x\|^2 \tag{4.12}$$

$$= b^T \left(AA^T + \Phi^{-1}\right)^{-1} \left(\left(\frac{\zeta_i}{\eta_i \phi_i}\right)^2 \mathcal{I}_i^T \mathcal{I}_i - AA^T \right) \left(AA^T + \Phi^{-1}\right)^{-1} b. \tag{4.13}$$

Before moving on, a couple of implementation related points will be examined. First, the equations, g_i, can be rewritten so that the method of Steepest Descent can be used. To see this, define M_i to be the Choleski factor of

$$\left(\frac{\zeta_i}{\eta_i \phi_i}\right)^2 \mathcal{I}_i^T \mathcal{I}_i - AA^T,$$

and let

$$y = \left(AA^T + \Phi^{-1}\right)^{-1} b.$$

The secular equation is thus

$$g_i(\zeta_1, \zeta_2, \ldots, \zeta_q) \quad = \quad y^T M_i^T M_i y.$$

Thus when $g_i = 0$, $M_i y = 0$. Using this, note that also $M_i^T M_i y$, which means

$$f_i(\zeta_1, \zeta_2, \ldots, \zeta_q) \quad = \quad y^T \left(\left(\frac{\zeta_i}{\eta_i \phi_i} \right)^2 \mathcal{I}_i^T \mathcal{I}_i - AA^T \right)^2 y$$

is zero, when $g_i = 0$, and $f_i \geq 0$. The quantity,

$$F(\zeta_1, \zeta_2, \ldots, \zeta_q) \quad = \quad \sum_{i=1}^{q} f_i(\zeta_1, \zeta_2, \ldots, \zeta_q),$$

has its minimum at $F = 0$ and this minimum occurs only where $g_i = 0$ for all $i = 1, 2, \ldots, q$. While Steepest Descent is only linear in its convergence, it is useful in finding a starting point for a faster method, like Newton's or a quasi-Newton method.

Second, both Newton's method and quasi-Newton methods, like Broyden's method, require at least an initial calculation of the Jacobian of the g_i. This can be done numerically, but note that by using the Choleski factorization defined above, the derivative of g_i with respect to ζ_k at the solution is

$$\begin{aligned}
\frac{\partial g_i}{\partial \zeta_k} \quad &= \quad 2 \frac{\partial y}{\partial \zeta_k}^T M_i^T M_i y + y^T \left(2 \frac{\zeta_i}{\eta_i \phi_i} \frac{2 \frac{\partial \zeta_i}{\partial \zeta_k} \eta_i \phi_i - \zeta_i \eta_i \frac{\partial \phi_i}{\partial \zeta_k}}{\eta_i^2 \phi_i^2} \right) \mathcal{I}_i^T \mathcal{I}_i y \\
&= \quad y^T \left(2 \frac{\zeta_i}{\eta_i \phi_i} \frac{2 \frac{\partial \zeta_i}{\partial \zeta_k} \eta_i \phi_i - \zeta_i \eta_i \frac{\partial \phi_i}{\partial \zeta_k}}{\eta_i^2 \phi_i^2} \right) \mathcal{I}_i^T \mathcal{I}_i y,
\end{aligned}$$

and $\frac{\partial \zeta_i}{\partial \zeta_k}$ is one if $i = k$ and zero otherwise (i.e. the Kronecker delta). Additionally, if $\eta_{b,j} = 0$ for all $j = 1, 2, \ldots, q$, the term $\frac{\partial \phi_i}{\partial \zeta_k}$, is

$$\frac{\partial \phi_i}{\partial \zeta_k} \quad = \quad \begin{cases} \frac{\psi \zeta_k^2 + \delta_k}{\psi^2 \zeta_k^2} & i = k \\ \frac{\delta_k}{\psi^2 \zeta_k^2} & i \neq k \end{cases}.$$

Note that these yield that $\frac{\partial g_i}{\partial \varsigma_k} \geq 0$. Using these terms, the Jacobian can be directly calculated.

4.4 Non-Differentiable Points

As has been mentioned, the non-differentiable points for the multiple row case are $x = 0$ and $\|A_i x - b_i\| = 0$ for all $i = 1, 2, \ldots, q$. No simple condition on the parameters has been found to specify in advance when the solution will be at a differentiable point or at a non-differentiable point. The cost function is convex, however, so if a minimum (and thus a solution) exists at a differentiable point then it is the global minimum. This gives one way of checking if the problem has a solution at the non-differentiable points, namely does the secular equation have a root. If the secular equation has a root (by checking for a sign change between $g_i(0)$ and $g_i(\infty)$), then the solution is at a differentiable point else it is at a non-differentiable point. If no solution exists at a differentiable point then the problem could be perturbed so it has a solution at a differentiable point, and then the original problem's solution can be found by a continuation method.

The cases of $\|A_i x - b_i\| = 0$ are more complicated than the case of $\|x\| = 0$ because A_i could have more columns than rows even when A has more rows than columns. Also, the fewer rows there are, the more likely $b_i \in \mathcal{R}(A_i)$. Finally note that A_i could have a non-empty null space even when A is full column rank. While $\|x\| = 0$ is easier, it is by no means trivial as the analysis below shows.

Denote the global minimum of $C_{row}(x)$ as x^*. Note that

$$\begin{aligned} C_{row}(x^*) &\leq C_{row}(0) \\ &= \frac{1}{2}\sum_{i=1}^{q}(\|b_i\| + \eta_{b,i})^2. \end{aligned}$$

Now let $\sigma_{i,max}$ be the largest singular value of A_i. It then follows that

$$\|A_i x - b_i\| \geq \begin{cases} \|b_i\| - \sigma_{i,max}\|x\| & \|x\| \leq \frac{\|b_i\|}{\sigma_{i,max}} \\ 0 & else. \end{cases}$$

Now define for $i = 1, 2, \ldots, q$,

$$\begin{aligned} \Gamma_1 &= \left\{ i : \|x\| \leq \frac{\|b_i\|}{\sigma_{i,max}} \right\} \\ \Gamma_2 &= \left\{ i : \|x\| > \frac{\|b_i\|}{\sigma_{i,max}} \right\}. \end{aligned}$$

Then in Γ_1

$$\begin{aligned} \|A_i x - b_i\| + \eta_i\|x\| + \eta_{b,i} &\geq \|b_i\| - \sigma_{i,max}\|x\| + \eta_i\|x\| + \eta_{b,i} \\ &= \|b_i\| + \eta_{b,i} + (\eta_i - \sigma_{i,max})\|x\|. \end{aligned}$$

In Γ_2

$$\begin{aligned} \|A_i x - b_i\| + \eta_i\|x\| + \eta_{b,i} &\geq \eta_i\|x\| + \eta_{b,i} \\ &> \eta_i \frac{\|b_i\|}{\sigma_{i,max}} + \eta_{b,i} \\ &= \frac{\eta_i}{\sigma_{i,max}}\|b_i\| + \eta_{b,i}. \end{aligned}$$

Combining these

$$C_{row}(x) \geq \frac{1}{2}\sum_{i\in\Gamma_1}(\|b_i\| + \eta_{b,i} + (\eta_i - \sigma_{i,max})\|x\|)^2 + \frac{1}{2}\sum_{i\in\Gamma_2}\left(\frac{\eta_i}{\sigma_{i,max}}\|b_i\| + \eta_{b,i}\right)^2.$$

Now if $\eta_i \geq \sigma_{i,max}$ for all i then

$$\begin{aligned} C_{row}(x) &\geq \frac{1}{2}\sum_{i\in\Gamma_1}(\|b_i\| + \eta_{b,i})^2 + \frac{1}{2}\sum_{i\in\Gamma_2}(\|b_i\| + \eta_{b,i})^2 \\ &\geq C_{row}(0). \end{aligned}$$

Thus $C_{row}(0) \geq C_{row}(x^*) \geq C_{row}(0)$ and so $x^* = 0$ if $\eta_i \geq \sigma_{i,max}$ for all i. The condition is sufficient but not necessary.

Practically the best way to see if the solution is at a differentiable point is to see if the secular equation has a root. If the secular equation has a root the solution is at a differentiable point. If $g_i < 0$ then $\|A_i x - b_i\| = 0$, else if $g_i > 0$ then $\|x\| = 0$. These can easily be seen from Equation 4.12. Since $\frac{\partial g_i}{\partial \zeta_k} \geq 0$ this means that $g_i > 0$ if $g_i(0, 0, \ldots, 0) > 0$ and $g_i < 0$ if $g_i(\zeta_i = \infty, \zeta_k = 0) < 0$. These become the checks that will be used to see if the solution is at a non-differentiable point.

4.5 Solution Algorithm

Should the reader need to implement the methods of this chapter, this section presents pseudo-code for the algorithm. It is assumed that Newton's or a quasi-Newton's method is being used, for more information on the benefits and implementation of particular methods see [47, 85, 2], or other numerical texts. An optional Steepest Descent section is suggested to refine the guess before the algorithm begins. The term y has been calculated to speed the algorithm, as it is used frequently. Let ∞ be approximated by $\frac{1}{\epsilon}$, i.e.: one over machine precision. The value *tol* is a user specified tolerance, which is used to determine when the calculation is close enough to the actual solution. Note that if $\|A_k x - b_k\| = 0$, it is possible for this to not completely specify x. In that case, x can be written as $A_k^\dagger b_k + (I - A_k A_k^\dagger)z$ and the algorithm can be run to find z on the reduced problem.

1. if $g_i(0, 0, \ldots, 0) > 0$ then $\|A_i x - b_i\| = 0$

2. else if $g_i(\zeta_i = \infty, \zeta_k = 0) < 0$ then $\|x\| = 0$

3. else differentiable, do below

4. set $\zeta_i = \frac{\eta_i}{\|A_i\|}$ for all $i = 1, 2, \ldots, q$

5. (optional) use Steepest Descent to refine ζ_i for all $i = 1, 2, \ldots, q$

6. calculate Jacobian;

7. repeat

   (a) $y = \left(AA^T + \Phi^{-1}\right)^{-1} b$

   (b) update ζ_i for all $i = 1, 2, \ldots, q$

   (c) update or calculate new Jacobian

   until($g_i \leq tol$ for all $i = 1, 2, \ldots, q$)

8. $x = A^T y$

4.6 Conclusions

A secular equation has been obtained to find the solution, but no simple necessary and sufficient condition(s) have been obtained to show beforehand when the solution is at a differentiable point versus a non-differentiable point. Methods to handle the difficulties have been suggested and a sufficient condition for when $x = 0$ has been shown. Additionally the secular equation can be used to see if the solution is at a differentiable point or not. While the secular equation technique works, it does not lend itself to simple condition(s) on A_i, η_i, b_i, and $\eta_{b,i}$. The method is thus usable, though it would be improved by having the simple conditions. An outline of a general algorithm to solve the problem has been included for reference.

Chapter 5

General Block Min Max Criterion

The final remaining area is to have arbitrary matrix perturbations. As has been seen there is a lot of similarity in the form of the answers derived which suggests that there should be a nice expression for generic perturbations. The general (block) perturbation min max problem can be stated as

$$\min_x \max_{\substack{\|E_{i,j}\|_2 \leq \eta_{i,j} \\ \|E_{b,i}\|_2 \leq \eta_{b,i}}} \frac{1}{2} \left\| \begin{bmatrix} A_{1,1} + E_{1,1} & \dots & A_{1,p} + E_{1,p} \\ \vdots & \ddots & \vdots \\ A_{q,1} + E_{q,1} & \dots & A_{q,p} + E_{q,p} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_p \end{bmatrix} - \begin{bmatrix} b_1 + E_{b,1} \\ \vdots \\ b_q + E_{b,q} \end{bmatrix} \right\|^2 \tag{5.1}$$

where $i = 1, 2, \dots, q$, $j = 1, 2, \dots, p$ and the norm has been squared, and a factor of $\frac{1}{2}$ has been inserted to make later formulas neater.

The flow of the chapter is as follows. In Section 5.1, the maximization of the perturbations is solved. In Section 5.2 the form of solution is shown. In Section 5.3 the secular is found and basic properties are shown. An algorithm is not presented as was done in Chapter 4 due to similarities with what was presented there, rather in Section 5.4 an alternative fixed point algorithm is developed.

5.1 An Alternate Formulation

In all the preceding problems the maximization was performed by finding an upper bound and a perturbation that caused the cost to reach the upper bound. The same method will be used here, but as the method is familiar it will only be covered briefly. For simplicity let a block row be specified by $A_{i,*} = [A_{i,1} \quad \ldots \quad A_{i,p}]$, and let the expression that is being minimized over x and maximized over the perturbations in Equation 5.1 be denoted by J_{gen}.

Begin by finding an upper bound for the cost function,

$$\begin{aligned}
J_{gen} &= \frac{1}{2}\left\| \begin{bmatrix} A_{1,1}+E_{1,1} & \ldots & A_{1,p}+E_{1,p} \\ \vdots & \ddots & \vdots \\ A_{q,1}+E_{q,1} & \ldots & A_{q,p}+E_{q,p} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_p \end{bmatrix} - \begin{bmatrix} b_1+E_{b,1} \\ \vdots \\ b_q+E_{b,q} \end{bmatrix} \right\|^2 \\
&= \frac{1}{2}\sum_{i=1}^{q} \left\| \begin{bmatrix} A_{i,1} & \ldots & A_{i,p} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_p \end{bmatrix} - b_i + \begin{bmatrix} E_{i,1} & \ldots & E_{i,p} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_p \end{bmatrix} - E_{b,i} \right\|^2 \\
&= \frac{1}{2}\sum_{i=1}^{q} \left(\|A_{i,*}x - b_i\| + \sum_{j=1}^{p} \|E_{i,j}x_j\| + \|E_{b,i}\| \right)^2 \\
&= \frac{1}{2}\sum_{i=1}^{q} \left(\|A_{i,*}x - b_i\| + \sum_{j=1}^{p} \eta_{i,j}\|x_j\| + \eta_{b,i} \right)^2 .
\end{aligned}$$

Now consider the perturbations

$$\begin{aligned}
\bar{E}_{i,j} &= \eta_{i,j}\frac{(A_{i,*}x - b_i)x_j^T}{\|A_{i,*}x - b_i\|\|x_j\|} \\
\bar{E}_{b,i} &= -\eta_{b,i}\frac{A_{i,*}x - b_i}{\|A_{i,*}x - b_i\|},
\end{aligned}$$

and note that

$$\begin{aligned} \|\bar{E}_{i,j}\| &= \eta_{i,j} \\ \|\bar{E}_{b,i}\| &= \eta_{b,i}. \end{aligned}$$

When these perturbations are put in J_{gen}, they yield a potential maximization,

$$\begin{aligned} \bar{J}_{gen} &= \frac{1}{2}\sum_{i=1}^{q}\left\|A_{i,*}x - b_i + \sum_{j=1}^{p}\eta_{i,j}\frac{(A_{i,*}x-b_i)x_j^T}{\|A_{i,*}x-b_i\|\|x_j\|}x_j + \eta_{b,i}\frac{A_{i,*}x-b_i}{\|A_{i,*}x-b_i\|}\right\|^2 \\ &= \frac{1}{2}\sum_{i=1}^{q}\left\|(A_{i,*}x-b_i)\left(1+\sum_{j=1}^{p}\frac{\eta_{i,j}\|x_j\|}{\|A_{i,*}x-b_i\|}+\frac{\eta_{b,i}}{\|A_{i,*}x-b_i\|}\right)\right\|^2 \\ &= \frac{1}{2}\sum_{i=1}^{q}\left(\|A_{i,*}x-b_i\|\left(1+\sum_{j=1}^{p}\frac{\eta_{i,j}\|x_j\|}{\|A_{i,*}x-b_i\|}+\frac{\eta_{b,i}}{\|A_{i,*}x-b_i\|}\right)\right)^2 \\ &= \frac{1}{2}\sum_{i=1}^{q}\left(\|A_{i,*}x-b_i\|+\sum_{j=1}^{p}\eta_{i,j}\|x_j\|+\eta_{b,i}\right)^2. \end{aligned}$$

Thus the perturbations maximize the cost function and thus Equation 5.1 is equivalent to

$$C_{gen} = \min_x \frac{1}{2}\sum_{i=1}^{q}\left(\|A_{i,*}x-b_i\|+\sum_{j=1}^{p}\eta_{i,j}\|x_j\|+\eta_{b,i}\right)^2. \tag{5.2}$$

Note that the cost function is convex. As in the block row case, the cost function is not in a sum of Euclidean norms form, so Overton's method cannot be used as written. No method to rewrite this into a form compatible with Overton's method has been found, but methods that work directly on convex cost functions could be used.

5.2 The Form of the Solution

The non-differentiable points for the general block case are $\|A_{i,*}x - b_i\| = 0$ for all $i = 1, 2, \ldots, q$ and $x = 0$. In this section the differentiable points will be dealt with. As with the multiple (block) row min max, the non-differentiable points can be ruled out by checking if the secular equations, which will be developed in Section 5.3, have a solution.

Define L_i as $L_i = \|A_{i,*}x - b_i\| + \sum_{j=1}^{p} \eta_{i,j} \|x_j\| + \eta_{b,i}$ to simplify the notation. Take the gradient of Equation 5.2 and set it equal to zero, to get the necessary conditions for a minimum,

$$0 = \sum_{i=1}^{q} L_i \left(\frac{A_{i,*}^T (A_{i,*}x - b_i)}{\|A_{i,*}x - b_i\|} + \begin{bmatrix} \frac{\eta_{i,1}}{\|x_1\|} I & & 0 \\ & \ddots & \\ 0 & & \frac{\eta_{i,p}}{\|x_p\|} I \end{bmatrix} x \right).$$

Now define

$$\begin{aligned}
\zeta_{i,j} &= \frac{\eta_{i,j}\|x_j\|}{\|A_{i,*}x - b_i\|} && (5.3)\\
\phi_i &= \frac{L_i}{\|A_{i,*}x - b_i\|} \\
&= 1 + \sum_{k=1}^{p} + \frac{\eta_{b,i}}{\|A_{i,*}x - b_i\|} && (5.4)\\
\Phi &= \operatorname{diag}(\phi_1 I, \ldots, phi_q I) && (5.5)\\
\delta_{i,j} &= \frac{\eta_{i,j} L_i}{\|x_j\|} \\
&= \frac{\eta_{i,j}^2 \phi_i}{\zeta_{i,j}} && (5.6)\\
\psi_j &= \frac{\eta_{i,j} L_i}{\|x_j\|} && (5.7)\\
\Psi &= \operatorname{diag}(\psi_1 I, \ldots, psi_p I). && (5.8)
\end{aligned}$$

This yields

$$\begin{aligned}
0 &= \sum_{i=1}^{q} \left(\phi_i A_{i,*}^T (A_{i,*}x - b_i) + \begin{bmatrix} \delta_{i,1} I & & 0 \\ & \ddots & \\ 0 & & \delta_{i,p} I \end{bmatrix} x \right) \\
&= \sum_{i=1}^{q} \left(A^T \Phi (Ax - b) + \Psi x \right).
\end{aligned}$$

And the solution has the form

$$x = (A^T \Phi A + \Psi)^{-1} A^T \Phi b \tag{5.9}$$

if it occurs at a differentiable point.

5.3 The Secular Equation

Similar to the row case, the expressions for $\zeta_{i,j}$ can be used to develop clean secular equations to find the solution. Note that this will require pq secular equations to be solved. Alternately, the expressions for ϕ_i and ψ_j can be used. The second option only requires $p + q$ secular equations, but these equations require $\|x_i\|$ and $\|A_{i,*}x - b_i\|$ to be calculated each time and complicated rational expressions to be used. For these reasons the secular equations in terms of $\zeta_{i,j}$ will be used.

Note the solution x in Equation 5.9 can be rewritten using the matrix inversion lemma,

$$\begin{aligned}
x &= (A^T \Phi A + \Psi)^{-1} A^T \Phi b \\
&= (\Psi^{-1} - \Psi^{-1} A^T (A \Psi^{-1} A^T + \Phi^{-1})^{-1} A \Psi^{-1}) A^T \Phi b \\
&= \Psi^{-1} A^T (I - (A \Psi^{-1} A^T + \Phi^{-1})^{-1} A \Psi^{-1} A^T) \Phi b \\
&= \Psi^{-1} A^T (A \Psi^{-1} A^T + \Phi^{-1})^{-1} \Phi^{-1} \Phi b \\
&= \Psi^{-1} A^T (A \Psi^{-1} A^T + \Phi^{-1})^{-1} b.
\end{aligned}$$

This allows a nice expression for x_j

$$x_j = \psi_j^{-1} A_{*,j}^T (A \Psi^{-1} A^T + \Phi^{-1})^{-1} b,$$

with

$$A_{*,j}^T = [A_{1,j}^T \quad \ldots \quad A_{q,j}^T],$$

i.e. a block column. Thus the square of the norm of x_j is given by

$$\|x_j\| = \psi_j^{-2} b^T (A\Psi^{-1}A^T + \Phi^{-1})^{-1} A_{*,j} A_{*,j}^T (A\Psi^{-1}A^T + \Phi^{-1})^{-1} b. \quad (5.10)$$

An expression for $A_{i,*}x - b_i$ can also be developed,

$$\begin{aligned} A_{i,*}x - b_i &= A_{i,*}\Psi^{-1}A^T(A\Psi^{-1}A^T + \Phi^{-1})^{-1}b - b_i \\ &= -\phi_i^{-1}\mathcal{I}_i(A\Psi^{-1}A^T + \Phi^{-1})^{-1}b, \end{aligned}$$

with $\mathcal{I}_i$ defined by Equation 4.10. Thus

$$\|A_{i,*}x - b_i\|^2 = \phi_i^{-2} b^T (A\Psi^{-1}A^T + \Phi^{-1})^{-1} \mathcal{I}_i^T \mathcal{I}_i (A\Psi^{-1}A^T + \Phi^{-1})^{-1} b.$$

Using this, the expression for $\zeta_{i,j}$ in Equation 5.3 can be squared and rewritten to

$$\begin{aligned} 0 &= \frac{\zeta_{i,j}^2}{\eta_{i,j}^2}\|A_{i,*}x - b_i\|^2 - \|x_j\|^2 \\ &= b^T(A\Psi^{-1}A^T + \Phi^{-1})^{-1}\left(\frac{\zeta_{i,j}^2}{\eta_{i,j}^2\phi_i^2}\mathcal{I}_i^T\mathcal{I}_i - A_{*,j}A_{*,j}^T\right)(A\Psi^{-1}A^T + \Phi^{-1})^{-1}b. \end{aligned}$$

Thus define the secular equations to be

$$g_{i,j} = b^T(A\Psi^{-1}A^T + \Phi^{-1})^{-1}\left(\frac{\zeta_{i,j}^2}{\eta_{i,j}^2\phi_i^2}\mathcal{I}_i^T\mathcal{I}_i - A_{*,j}A_{*,j}^T\right)(A\Psi^{-1}A^T + \Phi^{-1})^{-1}b.$$

5.4 Fixed Point Method

While a direct calculation method exists, it is worthwhile to examine another method to solve the problem. The basic equation for the solution is given by

$$\begin{aligned} \left(A^T\Phi A + \Psi I\right)x &= A^T\Phi b \\ A^T\Phi Ax + \Psi x &= A^T\Phi b. \end{aligned} \quad (5.11)$$

Equation 5.11 can be rewritten in one of two forms.

$$\begin{aligned} x &= \Psi^{-1} A^T \Phi \left(b - Ax\right) \\ x &= \left(A^T \Phi A\right)^{-1} \left((A^T \Phi b - \Psi x\right) \end{aligned}$$

These forms can easily be rewritten as recursive equation,

$$x_i = \Psi^{-1} A^T \Phi \left(b - Ax_{i-1}\right) \tag{5.12}$$

$$x_i = \left(A^T \Phi A\right)^{-1} \left(A^T \Phi b - \Psi x_{i-1}\right) \tag{5.13}$$

Two forms are used because when Ψ is very small, Equation 5.12 can provide numerical challenges, while at the same time Equation 5.13 behaves nicely. In most cases we will use Equation 5.12, as often Ψ is not so small as to require the switch. The resulting method works nicely, and while the convergence appears to be linear, the constant multiplier is usually small. Ten thousand runs of random data were tried with matrix dimensions varying from $m = 8$ and $n = 4$ to $m = 32$ and $n = 16$. Two row and two column partitions were used. The perturbations for each block were different, and varied from 1% to 20% of the maximum singular value of the block. The stopping condition used was $\frac{\|x_i - x_{i-1}\|}{\|x_i\|} \leq \delta$ for δ between 10^{-4} and 10^{-8}. Every case took less than 10 iterations, and the size of A did not matter. The size independence is comparable to observations noted in [62] for convergence of the SOCP and SDP solvers used in their LMI problems. The value of δ was the primary influence of convergence, and for values of $\delta < 10^{-12}$ caused a virtual infinite loop requiring a maximum iteration condition to terminate. The value of x appeared to bounce around near the solution, with numerical difficulties suspected as a primary cause.

This is a fixed point algorithm, and it can be shown under reasonable assumptions that the method will converge for small or large values of ψ. A small region where it has not been proven exists, and fixed point methods can be slow in some cases, but the method performs well in practice.

Lemma 5.1 (Convergence) *The fixed point method specified by Equation 5.12 and Equation 5.13 will converge if the starting point is close to the solution and either*

1. $\psi_{min} > \sigma_{max}^2 \phi_{max}$

2. $\psi_{max} < \sigma_n^2$

Proof:

First note that both Ψ and Φ are rational functions with no poles in the first quadrant. Thus if two succeeding values of x_i are close then the values of Ψ and Φ will be approximately the same. The values will be close near the solution. Given this condition, write, using Equation 5.12

$$\begin{aligned} x_i - x_{i-1} &= \Psi^{-1} A^T \Phi \left(b - A x_{i-1}\right) - \Psi^{-1} A^T \Phi \left(b - A x_{i-2}\right) \\ &= \Psi^{-1} A^T \Phi A \left(x_{i-2} - x_{i-1}\right). \end{aligned}$$

Taking norms

$$\begin{aligned} \|x_i - x_{i-1}\| &= \left\|\Psi^{-1} A^T \Phi A \left(x_{i-2} - x_{i-1}\right)\right\| \\ &\leq \left\|\Psi^{-1} A^T \Phi A\right\| \|x_{i-1} - x_{i-2}\| \\ &\leq \frac{\sigma_{max}^2 \phi_{max}}{\psi_{min}} \|x_{i-1} - x_{i-2}\|. \end{aligned}$$

The fraction must be less than 1 for convergence, which yields the first alternative in the lemma. Now consider Equation 5.13

$$\begin{aligned} x_i - x_{i-1} &= \left(A^T \Phi A\right)^{-1} \left(A^T \Phi b - \Psi x_{i-1}\right) - \left(A^T \Phi A\right)^{-1} \left(\left(A^T \Phi b - \Psi x_{i-1}\right)\right. \\ &= \left(A^T \Phi A\right)^{-1} \Psi \left(x_{i-2} - x_{i-1}\right). \end{aligned}$$

Taking norms

$$
\begin{aligned}
\|x_i - x_{i-1}\| &= \left\| \left(A^T \Phi A\right)^{-1} \Psi \left(x_{i-2} - x_{i-1}\right) \right\| \\
&\leq \left\| \left(A^T \Phi A\right)^{-1} \Psi \right\| \|x_{i-1} - x_{i-2}\| \\
&\leq \frac{\psi_{max}}{\sigma^2_{max}} \|x_{i-1} - x_{i-2}\| .
\end{aligned}
$$

The fraction must be less than 1 for convergence, which yields the second alternative in the lemma and completes the proof.

$\diamondsuit$ SDG $\diamondsuit$

At each step, values are needed for ψ_j and ϕ_i. The values of ψ_j and ϕ_i for each iteration can be estimated by using the previous x, the only challenge is finding the starting values. Some general bounds exist, which establish where the solutions lie and thus give the starting values.

Lemma 5.2 (General bounds)

$$
\begin{aligned}
\sum_{i=1}^{q} \eta^2_{i,j} &\leq \psi_j \\
1 &\leq \phi_i < \frac{\sqrt{\sum_{k=1}^{p} \left(\|b_k\| + \eta_{b_k}\right)^2}}{\|P_{A^{\perp}_{i,*}} b\|}
\end{aligned}
$$

Proof:

The condition on ψ_j can be easily seen by noting that it is the sum of non-negative terms of the form

$$
\psi_j = \sum_{i=1}^{q} \frac{\eta_{i,j} L_i}{\|x_j\|},
$$

with

$$
L_i = \|A_{i,*} x - b_i\| + \sum_{k=1}^{p} \eta_{i,k} \|x_k\| + \eta_{b_i}.
$$

Now let

$$L_{i/j} \;=\; \|A_{i,*}x - b_i\| + \sum_{k \neq j} \eta_{i,k}\|x_k\| + \eta_{b_i},$$

then

$$\begin{aligned}
\psi_j &= \sum_{i=1}^{q} \left(\frac{\eta_{i,j} L_{i/j}}{\|x_j\|} + \frac{\eta_{i,j}^2 \|x_j\|}{\|x_j\|} \right) \\
&= \sum_{i=1}^{q} \left(\frac{\eta_{i,j} L_{i/j}}{\|x_j\|} + \eta_{i,j}^2 \right) \\
&= \sum_{i=1}^{q} \frac{\eta_{i,j} L_{i/j}}{\|x_j\|} + \sum_{i=1}^{q} \eta_{i,j}^2.
\end{aligned}$$

A trivial implication of this is that ψ_j is not zero unless $\eta_{i,j} = 0$ for all i. No upper bound exists, because if $x_j \to 0$ then $\psi_j \to \infty$.

The lower bound on ϕ_i can be seen by noting the terms are all non-negative

$$\begin{aligned}
L_i &= \|A_{i,*}x - b_i\| + \sum_{j=1}^{p} \eta_{i,j} \|x_j\| + \eta_{b,i} \\
&\geq \|A_{i,*}x - b_i\|.
\end{aligned}$$

The result follows from the fact that $\phi_i = \frac{L_i}{\|A_{i,*}x - b_i\|}$.

The upper bound on ϕ_i can be established as follows. First note the optimal value of the cost function must be finite and in particular it must cost less than $x = 0$:

$$\begin{aligned}
C_{gen} &= \sum_{i=1}^{p} L_i^2 \\
&< \sum_{k=1}^{p} \left(\|b_k\| + \eta_{b_k}\right)^2.
\end{aligned}$$

This means $L_i^2 < \sum_{k=1}^{p} \left(\|b_k\| + \eta_{b_k}\right)^2$ or $L_i < \sqrt{\sum_{k=1}^{p} \left(\|b_k\| + \eta_{b_k}\right)^2}$, since $L_i^2 < C_{gen}$. The denominator of the bound is found by noting that the smallest the norm of the residual can be is the norm of b_i projected onto the orthogonal complement of the range of $A_{i,*}$.

5.5 Numerical Example

To motivate the nice performance of the general block min max problem, consider how it handles random ill-conditioned A matrices of size 8×4. The condition number of A was set to be around 100, and the uncertainty bounds of the A matrix to be small (1% to 20%). A hundred random systems were produced and the least squares (LS), total least squares (TLS), and general block min max (GBMM) solutions were calculated. The quality of the solution was calculated by taking the norm of the difference of the true and estimated x for each method. The performance of each method was then plotted against the system number. This test was performed many times with two samples provided as Figure 5.1 and Figure 5.2. The good performance of GBMM is easily seen. The LS solution performance is worse by a factor of 2 in most runs (i.e.: $\|x_{true} - x_{LS}\| \approx 2\|x_{true} - x_{GBMM}\|$). The TLS solution does not do well, which is to be expected in an ill-conditioned system with modeling errors. Note the deliberate choice of the case (ill-conditioned system with modeling errors) that is best for GBMM, as should the system lack modeling errors or if it has a small condition number a less conservative method would better fit the problem.

5.6 Results And Directions for General Partitioning

Two secular equation techniques were presented to find the solution if it is at a differentiable point. As with the row case, the non-differentiable points have not been ruled out so the secular equation must be checked, to see if there is a solution. When the secular equation has a solution, the problem's solution is at a differentiable point, else it is at a non-differentiable point. The perturbed problem can be used to get the solution by a continuation method and the convexity of the problem. A second method of solution which uses the form of solution to create a fixed point algorithm was also presented. Finally, a numeric example was used to demonstrate the benefits of the formulation.

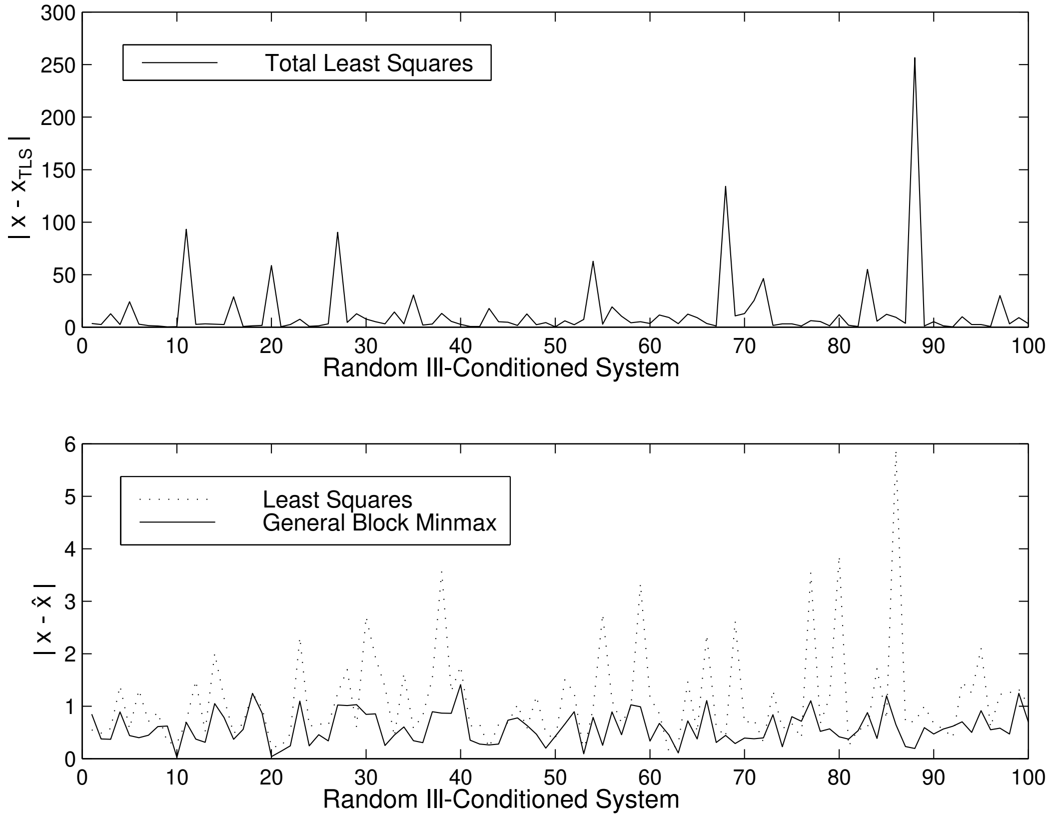

Figure 5.1. First Comparison of Least Squares to General Block Min Max

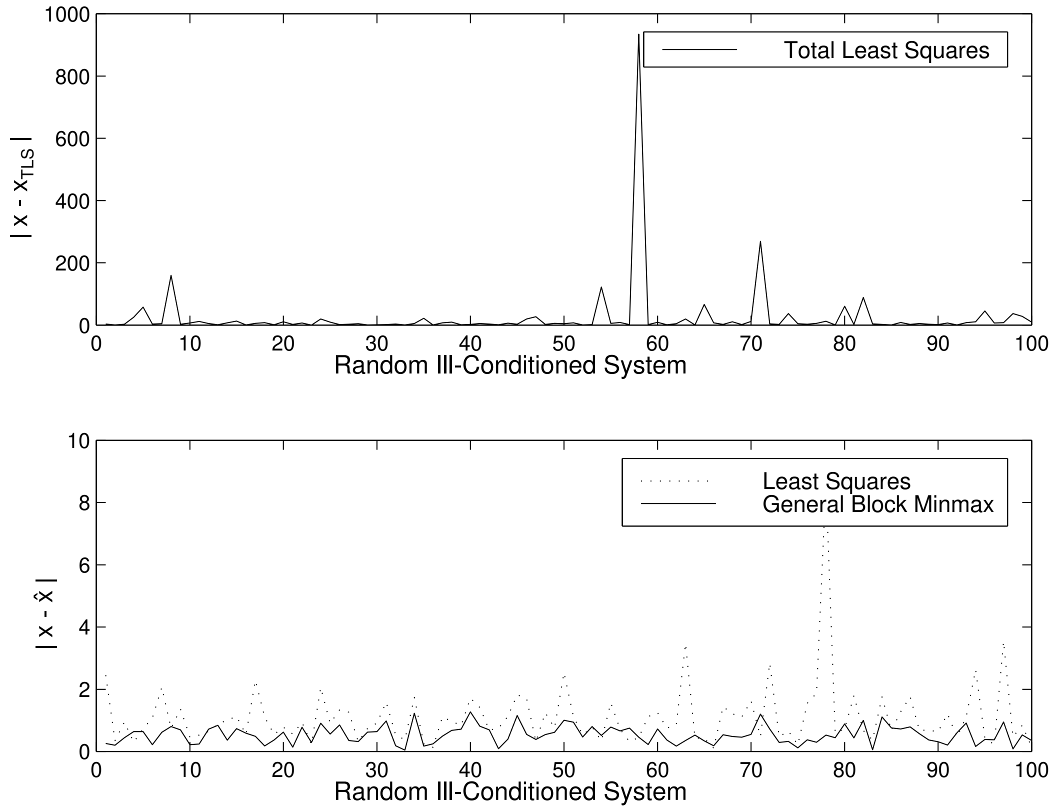

Figure 5.2. Second Comparison of Least Squares to General Block Min Max

Chapter 6

Degenerate Min Min Criterion

This chapter[1] is concerned with the following problem

$$\min_{x \in \mathcal{R}^n} \min_{\|E\| \leq \eta} \|(A+E)x - b\| \tag{6.1}$$

where A is an $m \times n$ real matrix and b is a real n-vector. This problem is a special case of the errors-in-variables problem, which has been given the formal name of the degenerate bounded errors-in-variables problem. For ease of reference the problem is usually called the degenerate min min problem. This problem can be viewed as a total least squares problem [67, 82] with bounds on the uncertainty in the coefficient matrix, which will be explained in more detail in Section 6.1. In this chapter frequent use is made of the terms degenerate and non-degenerate. Simply put, a degenerate problem is one where multiple solutions exist. The non-degenerate (unique solution) case of this problem occurs when η is small and b is in some sense far from the range of A. That η should be small is intuitive, since for $\eta = 0$ the problem reduces to the least squares problem, which is non-degenerate when A has full column rank. Conversely when η is larger than the

[1]Most of the material in this chapter originally appeared in [26]. SIAM holds the copyright, and this chapter appears in compliance with SIAM's requirements that the material can appear in later work provided their copyright is stated.

smallest singular value of A, one would anticipate degeneracy (multiple solutions) as the perturbed matrix $A + E$ is not guaranteed to be full column rank. The intuition behind b needing to be far from the range of A for non-degeneracy, comes from the fact that if b were close enough that multiple perturbations E existed such that b was in the range of $A+E$, then multiple solutions (degeneracy) would exist. In [25] and Section 2.19 the non-degenerate case of this problem was considered and how to compute its unique solution in $O(mn^2)$ flops was shown. This chapter considers the problem when it is degenerate; that is, when it has multiple solutions. In particular an $O(mn^2)$ algorithm is presented to find the solution with the minimum Euclidean norm. Note that the degenerate case is actually the generic case for this problem, and hence is more important than the non-degenerate case. This can be seen from the simple discussion above, since the non-degenerate case holds only for certain combinations of b and A when η is smaller than the smallest singular value of A. This is very restrictive, and hence the claim.

The basic structure of the chapter is as follows. First the geometric understanding of the problem is considered in Section 6.1. Then a brief outline of the proof is provided in Section 6.2 to make the full proof easier to follow. Section 6.3 provides pseudocode for the solution algorithm. Section 6.4 explains how the minimization over E is performed, which then allows Section 6.5 to cover computable conditions for degeneracy. Next, Section 6.6 shows that the feasibility constraint, $\|Ax - b\| \leq \eta\|x\|$, is actually an equality, $\|Ax - b\| = \eta\|x\|$. Section 6.7 explains how the secular equation is derived, and then in Section 6.8 the zero of secular equation which specifies the solution is identified. The majority of the rest of the chapter is concerned with proving the zero identified in Section 6.8 does indeed correspond to the solution. The chapter finishes with a summary in Section 6.18

and an extension of the min min problem to multiple columns with restricted perturbations in Section 6.19.

6.1 Geometric Understanding

Probably the easiest way to understand the problem at hand is to look at it geometrically. For ease of drawing consider A and b to be vectors of length 2. Note that while this is useful for getting a basic understanding, some of the key features of the problem do not appear in this case. For instance, when A has multiple columns the problem can be degenerate for small values of η. In such a case the degenerate min min problem has several advantages over other formulations, such as total least squares (TLS). One such advantage is the perturbation on A is much smaller in the degenerate min min problem than in the TLS problem.

In the general min min problem (degenerate or not), A and b are projected into a plane between the two similar to the TLS problem, but with a bound on how far A can be perturbed (see Figure 6.1). Note that the cone around A shows the boundary of possible perturbations to A. In essence, the min min formulation is solving the problem $\min \|[E \quad f]\|$ such that $(b+f) \in \mathcal{R}(A+E)$ and $\|E\| \leq \eta$. The problem at hand can thus be thought of as a TLS problem with bounds on the errors in A.

A better understanding of the degenerate problem can be obtained by considering one of the ways the problem can become degenerate. The easiest to visualize, and the only one that can be drawn in two dimensions, is the case when b lies in the cone of possible perturbations of A. In this case note that any $\hat{x}$ such that $x_l \leq \hat{x} \leq x_u$ is a solution to the problem. The perturbations $E(\hat{x})$ change, but each $\hat{x}$ in the range still solves the problem. The problem is which

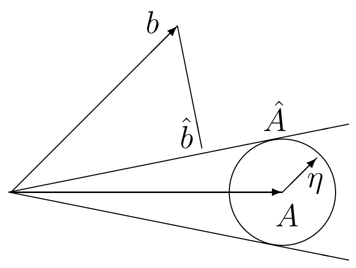

Figure 6.1. Min Min Problem

$\hat{x}$ should be chosen. The most conservative choice is to pick the smallest one, which is what shall be done. This choice has a lot to recommend it, but a full discussion is outside the bounds of this dissertation. In Section 6.4, this basic insight (picking the smallest solution) is exploited to reformulate the problem into a unique problem.

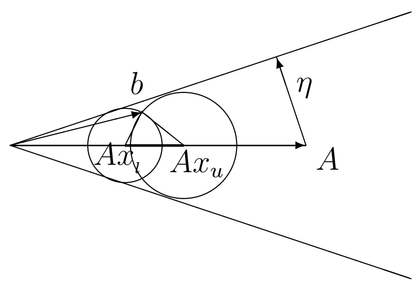

Figure 6.2. Degenerate Min Min Problem

6.2 Proof Outline

The cost function presented is useful for seeing how this problem handles the uncertainty in the matrix A, but it is not immediately useful in solving the problem. For instance checking if a problem is degenerate in the original form of the problem is tedious. The problem must be rewritten in a simpler form, and then a computable condition for degeneracy found. The proof begins in Section 6.4 with the cost function to solve

$$\min_{x \in \mathcal{R}^n} \min_{\|E\| \leq \eta} \|(A+E)x - b\|$$

and the degeneracy condition which was found in [25]

$$\eta\|x\| \geq \|Ax - b\|.$$

Since the non-degenerate case is already solved, the proof proceeds by assuming the degeneracy condition holds. The first step is to minimize the cost function over $\|E\| \leq \eta$, and find that the optimal cost is zero. Since the problem is degenerate and the cost function is zero, the solution with the smallest norm is chosen to obtain the problem

$$\min_{\|Ax-b\| \leq \eta\|x\|} \|x\|.$$

The condition $\eta\|x\| \geq \|Ax - b\|$, is not practical for checking for degeneracy in a problem, as mentioned above, since it requires checking multiple values of x to hopefully find one that holds and thus showing the problem is degenerate. The second step is thus to find a computable condition for degeneracy, which is done in Section 6.5. The proof proceeds by squaring the condition for degeneracy and using the singular value decomposition (SVD) of A to find the two cases in which the problem is degenerate. The first case is when η is larger than the smallest singular value of A. The first case is always degenerate. The second case is when η is not larger than the smallest singular value of A. The second case is degenerate only when

$$b^T(I - A(A^TA - \eta^2 I)^{-1}A^T)b \leq 0.$$

It still remains to show how to get the solution. Lagrange multiplier techniques are used to find the solution, so the inequality $\eta\|x\| \geq \|Ax - b\|$ must be reduced to an equality if possible. The third step of the proof is a proof that the solution, $\hat{x}$ is actually on the boundary of the inequality, and thus $\eta\|\hat{x}\| = \|A\hat{x} - b\|$. This is covered in Section 6.6.

The fourth step is to use Lagrange multiplier techniques to parameterize the solution, $\hat{x} = x(\alpha)$, in terms of a single variable, α, thus reducing the problem to finding the zeros a secular equation. This is covered in Section 6.7. A secular equation is a rational expression of one variable, which is constructed so that all the critical points of the original problem occur at zeros of the secular equation. The secular equation reduces the n-dimensional search for the solution, $\hat{x}$, to a 1-dimensional search. The solution to the original problem is denoted as $x(\alpha^o)$, and note that it will occur at one of the $2n$ zeros of the secular equation. The zero of the secular equation which corresponds to $x(\alpha^o)$ is denoted α^o.

The remainder of the chapter is concerned with showing which zero is α^o, and it starts with an assertion of the answer in Section 6.8. The unique zero of the secular equation in the interval $[\max(-\sigma_n^2, -\eta^2), \eta\sigma_1]$ is α^o, where σ_1 is the largest singular value of A and σ_n is the smallest. This is proven by a process of elimination.

Lagrange techniques are used (first and second order conditions on the Lagrangian) in Section 6.9 to narrow down the search area. By employing these techniques, it is found that α^o must lie in the interval $[\max(-\sigma_{n-1}^2, -\eta^2), \eta\sigma_1]$. This still admits several possibilities, see Figure 6.3. First, there are two critical points ($\alpha = -\sigma_n^2$ and $\alpha = -\sigma_{n-1}^2$) which could be α^o. Second, there are two intervals ($(-\sigma_n^2, \eta\sigma_1)$ and $(-\sigma_{n-1}^2, -\sigma_n^2)$) which could have α^o in one of them.

In particular, note that the interval $(-\sigma_{n-1}^2, -\sigma_n^2)$ can have multiple zeros in it, so that must also be dealt with. In Section 6.10 the second order condition

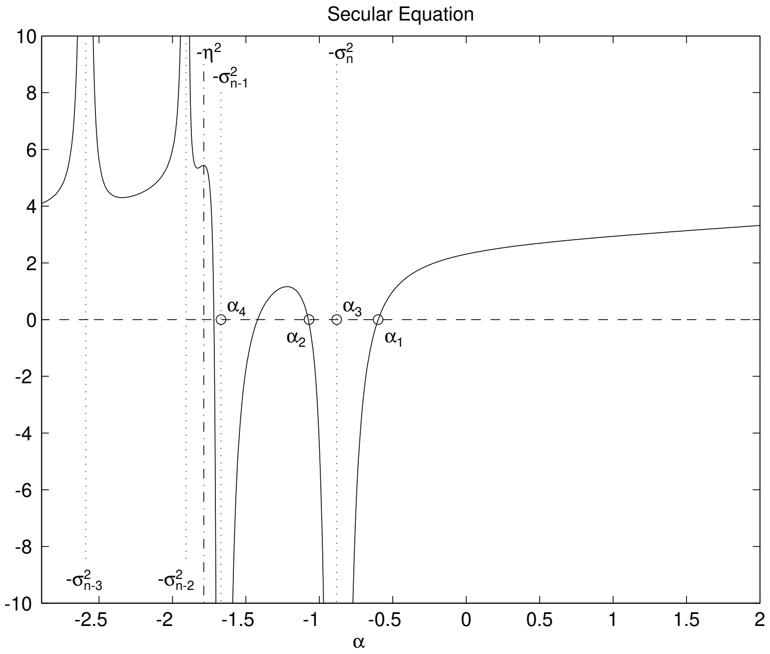

Figure 6.3. Secular Equation

is used to rule out half of the zeros in the second interval. The arguments that only the rightmost root in the interval $(-\sigma_{n-1}^2, -\sigma_n^2)$ is a candidate to be α^o is in appendix E. With this dealt with there are only four candidates zeros of $g(\alpha)$ to handle, which are denoted by α_1 through α_4. Section 6.11 introduces the four candidates: $\alpha_1 \in (-\sigma_n^2, \eta\sigma_1]$, α_2 is the rightmost root in $(-\sigma_{n-1}^2, -\sigma_{n-1}^2)$, $\alpha_3 = -\sigma_n^2$, and $\alpha_4 = -\sigma_{n-1}^2$. To show that α^o is the unique root in $[-\sigma_n^2, \eta\sigma_1]$, six cases are examined. Most of the work is involved at this stage, and hence most of the mathematical difficulties occur here. The basic idea is to eliminate the possibility that any root except the one that occurs in the interval $[\max(-\sigma_n^2, -\eta^2), \eta\sigma_1]$ can be α^o. Additionally, the existence and uniqueness of the zero must be shown. With this, established bisection or Newton's method can be used to find the root in the algorithm.

It is reasonable to wonder why six cases are needed to prove the assertion that α^o lies in the interval $[\max(-\sigma_n^2, -\eta^2), \eta\sigma_1]$. The reason lies in three basic factors which affect the shape of the secular equation. The first and most obvious is the size of η. Note for instance that if $\eta < \sigma_n$ then only one of the zeros α_1 is a candidate for α^o since an earlier condition (1^{st} order condition on the Lagrangian) states that $\alpha^o > -\eta^2$. Obviously to consider some of the candidates, such as α_4, it must be assumed that η is large enough to admit the possibility. The cases just allowed for an organization of the assumptions into convenient groups to handle. See Figure 6.4. The dotted vertical lines mark where the singular values are,

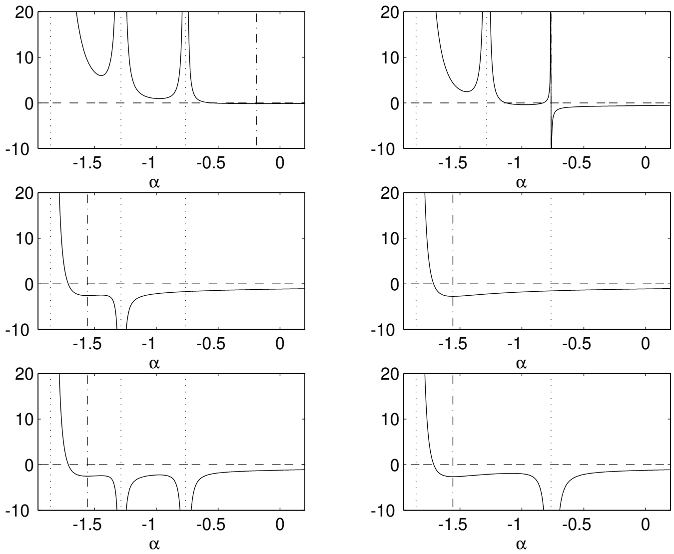

Figure 6.4. Six Cases of Proof. (UL) Case 1:$\eta < \sigma_n$; (UR) Case 2:$\eta = \sigma_n$; (ML) Case 3:$\eta > \sigma_n$, b is orthogonal to the left singular vector of σ_n, and $\sigma_n < \sigma_{n-1}$; (MR) Case 4:$\eta > \sigma_n$, b is orthogonal to the left singular vectors of σ_n, and σ_n has multiplicity k; (LL) Case 5:$\eta > \sigma_n$, b is not orthogonal to the left singular vector of σ_n, and $\sigma_n < \sigma_{n-1}$; (LR) Case 6:$\eta > \sigma_n$, b is not orthogonal to the left singular vectors of σ_n, and σ_n has multiplicity k.

and the dash-dotted vertical line indicates where $-\eta^2$ is. Note that in case 1 of Figure 6.4, it looks like the secular equation becomes flat to the right of $\alpha = -0.5$

but it does not. The scale makes the graph hard to read, so an expanded view of the region is provided in Figure 6.5. Case 1 examines η small ($\eta < \sigma_n$), case 2

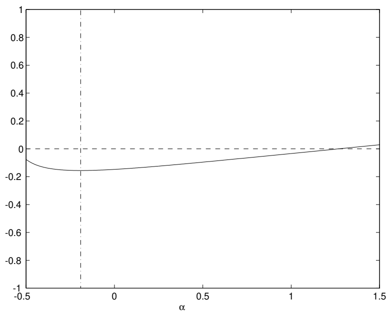

Figure 6.5. Expanded View of Case 1 Zero

considers the special case of $\eta = \sigma_n$, and finally cases 3-6 cover when η is large ($\eta > \sigma_n$).

When η is large there are more possibilities. The first is that the smallest singular value might have multiplicity of two or more. This can be exploited to simplify the problem. In particular, α_2 does not exist in this case, and $\alpha_3 = \alpha_4$. The cases where $\sigma_n < \sigma_{n-1}$ are the more difficult ones. The second is that b might be orthogonal to the left singular vector(s) of A, which correspond to smallest singular value. This drastically changes the shape of the graph of the secular equation in the region around $\alpha = -\sigma_n^2$. See, for instance, the middle left graph in Figure 6.4. The pole which normally appears at $-\sigma_n^2$ is not present. In fact, the only time α_3 can be α^o is when b is orthogonal to the left singular vector(s) of A, which corresponds to smallest singular value (σ_n). Similarly the only time α_4 can be α^o is when b is orthogonal to the left singular vector(s) of A, which correspond to the second smallest singular value (σ_{n-1}). Note that if the smallest singular value has multiplicity of at least two, then $\sigma_n = \sigma_{n-1}$. This case is shown on the middle right graph of Figure 6.4. The last four cases cover all the combinations of singular value multiplicity and b vector orthogonality which occurs when η is large.

6.3 Algorithm

This Section presents pseudo-code for the algorithm. The syntax has been designed to be Matlab[2]-like. Three lines deserve particular attention, though. The first one to appear states 'solve non-degenerate problem'. In this case the problem is not degenerate so a call to the code for the non-degenerate case as outlined in Section 2.9 and fully covered in [25] is used. The next line that could be confusing starts with 'pick any Θ'. In this case any unit vector, Θ, will solve the problem. An additional condition could be placed on the solution, $\hat{x}$, to select a specific Θ or to meet special requirements of the specific problem, so it is left unspecified in the pseudo-code. The final line that requires clarification starts with $\alpha \in [\max(-\sigma_n^2, -\eta^2), \eta\sigma_1]$. In this case, find the root of $g(\alpha)$ in the specified range, so any root finder (for instance bisection or Newton's method) can be used.

$[U, \Sigma, V] = svd(A);$

$b_1 = U^T b;$

$cond = 0;$

if $(\eta < \sigma_n)$ or $(\eta = \sigma_n$ and $b_1(n) = 0)$

    if $(b^T(I - A(A^TA - \eta^2 I)^{-1}A^T)b > 0)$

        solve non-degenerate problem

    else

        $cond = 1;$

    end

[2]Matlab is a registered trademark of The Mathworks, Inc.

else

if ($\eta = \sigma_n$)

$cond = 1;$

else

if ($\sigma_n < \sigma_{n-1}$) and ($b_1(n) = 0$) and ($g(-\sigma_n^2) \geq 0$)

$\bar{\Sigma}_1 = \Sigma(1:n-1, 1:n-1);$

$\bar{b}_1 = b_1(1:n-1);$

$$\hat{x} = V \begin{bmatrix} (\bar{\Sigma}_1^2 - \sigma_n^2 I)^{-1}\bar{\Sigma}_1 \bar{b}_1 \\ \pm\sqrt{\frac{g(-\sigma_n^2)}{\eta^2-\sigma_n^2}} \end{bmatrix};$$

elseif ($\sigma_n = \sigma_{n-k+1} < \sigma_{n-k}$) and ($\|b_1(n-k+1:n)\| = 0$)

and ($g(-\sigma_n^2) \geq 0$)

$\bar{\Sigma}_1 = \Sigma(1:n-k+1, 1:n-k+1);$

$\bar{b}_1 = b_1(1:n-k+1);$

$r = \sqrt{\frac{g(-\sigma_n^2}{\eta^2-\sigma_n^2}};$

Pick any $\Theta \in \mathbf{R}^k$ such that $\|\Theta\| = 1$;

$$\hat{x} = V \begin{bmatrix} (\bar{\Sigma}_1^2 - \sigma_n^2 I)^{-1}\bar{\Sigma}_1 \bar{b}_1 \\ r\Theta \end{bmatrix};$$

else

$cond = 1;$

end

end

end

if $cond == 1$

$\alpha \in [\max(-\sigma_n^2, -\eta^2), \eta\sigma_1]$ such that $g(\alpha) = 0$

$\hat{x} = (A^T A + \alpha I)^{\dagger} A^T b;$

end

Where $g(\alpha)$ is given by

$$g(\alpha) = b_2^T b_2 + b_1^T \left(\Sigma_1^2 + \alpha I\right)^{-2} \left(\alpha^2 I - \eta^2 \Sigma_1^2\right) b_1$$

and

$$\begin{aligned} A &= \begin{bmatrix} U_1 & U_2 \end{bmatrix} \begin{bmatrix} \Sigma_1 \\ 0 \end{bmatrix} V^T \\ b_1 &= U_1^T b \\ b_2 &= U_2^T b. \end{aligned}$$

6.4 Minimization over E

For the reader's convenience major milestones will be placed in boxes at the end of the sections where the milestone occurs. The problem is assumed degenerate and in particular that there exists an x such that $\eta\|x\| \geq \|Ax - b\|$. An equivalent computable criteria for degeneracy will be provided, however this formulation is more useful for the present. The goal in this Section is to reduce the problem to an equivalent formulation that does not involve E. The goal is accomplished by showing the degenerate problem is equivalent to requiring the solution to be in the set $\{x|\eta\|x\| \geq \|Ax - b\|\}$. The first step is to show that the problem requires that the solution be in the set, then show that any $\hat{x}$ in the set solves the problem. Note that the method used to get E is related to the formulation in [151], though the full argument is provided for completeness. Under the assumption that the problem is degenerate it follows that

$$\min_x \min_{\|E\| \leq \eta} \|Ax - b + Ex\| = 0,$$

since for any x such that $\eta\|x\| \geq \|Ax - b\|$, E can be chosen to be

$$E = -\gamma\eta \frac{(Ax - b)x^T}{\|Ax - b\|\|x\|}, \quad 0 \leq \gamma \leq 1,$$

and thus obtain

$$0 \leq \min_{x} \min_{\|E\| \leq \eta} \|Ax - b + Ex\| \leq \|Ax - b\| \left|1 - \gamma \frac{\eta\|x\|}{\|Ax - b\|}\right|.$$

Note that for the choice

$$\gamma = \frac{\|Ax - b\|}{\eta\|x\|} \leq 1,$$

the upper bound is zero. Since there exists an E which makes the minimum zero, the minimum value of the norm is zero. Therefore only the equation

$$Ax - b + Ex = 0 \tag{6.2}$$

with the constraint $\|E\| \leq \eta$, needs considering. This constrained equation is equivalent to being on the set defined by

$$\|Ax - b\| \leq \eta\|x\|. \tag{6.3}$$

To prove this, it is first shown that if the constrained equation (6.2) is met by any x then that x is in the set (6.3).

$$Ax - b + Ex = 0$$

$$Ax - b = -Ex$$

Taking the norm of both sides obtains

$$\|Ax - b\| = \|Ex\|.$$

It is noted that this implies

$$\|Ax - b\| \leq \|E\|\|x\|.$$

Then using the constraint on the perturbation size, $\|E\| \leq \eta$, yields

$$\|Ax - b\| \leq \eta\|x\|.$$

And the desired result is obtained. Now it must be shown that if some x is in the set (6.3) then the constraint Equation 6.2 is met by that x. This is accomplished by showing that for any x in the set, there exists a perturbation, E_0, such that the constraint equation is satisfied. To do this consider

$$E_0 = -\frac{(Ax - b)x^T}{\|x\|^2}.$$

First note that this perturbation satisfies the constraint on the size of the perturbations ($\|E\| \leq \eta$).

$$\|E_0\| \leq \frac{\|Ax - b\|}{\|x\|}$$

Since x is on the set $\|Ax - b\| \leq \eta\|x\|$, this reduces to

$$\|E_0\| \leq \eta.$$

Now consider the equation given by $Ax - b + E_0x$ and observe that this is

$$Ax - b + E_0x = Ax - b - (Ax - b).$$

Thus trivially this yields $Ax - b + E_0 = 0$ and the assertion is proven.

By assumption there are multiple solutions which will solve the problem as stated. Since any will solve the original problem, additional constraints may be added, which will simplify the solution and ensure the solution meets other requirements. A reasonable choice is to pick the solution with the minimum norm. Other nice properties of this choice also recommend it. For instance it is possible under certain conditions for the min max solution (from [24]) to also solve the degenerate min min problem. When this occurs the min max solution is the solution to the degenerate problem with minimum norm. This is not proven for reasons of space, but it does provide good motivation for the choice. Using the choice of the minimum norm solution, the problem can be written as

$$\min_{\|Ax-b\|\leq\eta\|x\|} \|x\|.$$

The degenerate problem can be reformulated as a unique problem by considering $$\min_{\|Ax-b\|\leq\eta\|x\|} \|x\|.$$

6.5 Computable Conditions for Degeneracy

The constraint, $\|Ax - b\| \leq \eta\|x\|$, defines the set on which the solution lies and is thus referred to as the feasibility constraint. The feasibility constraint can be squared and expanded to obtain

$$x^T A^T Ax - 2x^T A^T b + b^T b \leq \eta^2 x^T x. \tag{6.4}$$

Let $A = U\Sigma V^T$ be the SVD of A conformally partitioned as follows

$$U = \begin{pmatrix} U_1 & U_2 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_1 \\ 0 \end{pmatrix},$$

and define both $b_i = U_i^T b$ for $i = 1, 2$, and $z = V^T x$. These definitions are made solely to simplify the expressions under consideration, and provide a convenient shorthand for the rest of the problem. Then inequality (6.4) can be simplified to obtain

$$z^T \Sigma_1^2 z - 2z^T \Sigma_1 b_1 + b_1^T b_1 + b_2^T b_2 \leq \eta^2 z^T z. \tag{6.5}$$

Now assuming that the singular values are in decreasing order partition Σ_1 as follows

$$\Sigma_1 = \begin{pmatrix} \Sigma_+ & 0 \\ 0 & \Sigma_- \end{pmatrix},$$

where $\Sigma_+^2 - \eta^2 I \geq 0$ and $\Sigma_-^2 - \eta^2 I < 0$. Also conformally partition z and b_1

$$z = \begin{pmatrix} z_+ \\ z_- \end{pmatrix} \quad b_1 = \begin{pmatrix} b_{1+} \\ b_{1-} \end{pmatrix}.$$

Then inequality (6.5) can be expanded into

$$\begin{aligned} 0 \geq \\ & z_+^T(\Sigma_+^2 - \eta^2 I)z_+ - 2z_+^T\Sigma_+ b_{1+} + b_{1+}^T b_{1+} + \\ & z_-^T(\Sigma_-^2 - \eta^2 I)z_- - 2z_-^T\Sigma_- b_{1-} + b_{1-}^T b_{1-} + \\ & b_2^T b_2. \end{aligned}$$

Now observe that if Σ_- is nonempty then the inequality always has at least one z which makes it true. In other words if $A^TA - \eta^2 I$ is indefinite then the problem is always degenerate. On the other hand, if $A^TA - \eta^2 I$ is positive-semidefinite then degeneracy depends on the vector b. To get a computable condition for degeneracy, first note that when $x = 0$ the constraint is non-negative. Proceed by minimizing the expression

$$x^T(A^TA - \eta^2 I)x - 2x^TA^Tb + b^Tb$$

and when $\eta \neq \sigma_i$ obtain

$$x_o = (A^TA - \eta^2 I)^{-1}A^Tb.$$

Now, when $A^TA - \eta^2 I$ is positive, the constraint is non-positive at this point. By plugging this back into the expression being minimized, obtain

$$b^T(I - A(A^TA - \eta^2 I)^{-1}A^T)b \leq 0 \tag{6.6}$$

as the required computable condition for the problem to be degenerate when $\eta < \sigma_n$.

The problem is degenerate if either

$$\eta > \sigma_n$$

or

$$b^T(I - A(A^TA - \eta^2 I)^{-1}A^T)b \leq 0.$$

6.6 Solution is on the Boundary

This Section proves that the optimal solution is obtained at the boundary of the feasible set; that is, at the minimum norm solution the inequality is actually an equality. Mathematically this means the feasibility constraint, which is given by the inequality $\|Ax - b\| \leq \eta\|x\|$, is actually an equality, $\|Ax - b\| = \eta\|x\|$. To prove this, use the shorthand developed in the last Section that given the SVD of A then $b_i = U_i^T b$ for $i = 1, 2$, and $z = V^T x$. The problem of finding the solution with the smallest norm to the degenerate problem can now be recast as minimizing $z^T z$ subject to the inequality constraint (6.6).

Now if $b = 0$, then clearly the minimum norm solution is $z = 0$ which does lie on the boundary ($\|\Sigma z - 0\| = 0 = 0 = \eta\|z\|$). So consider the case when $b \neq 0$. Denote by $f(z)$ the expression on the left-hand side of the inequality in Equation (6.6). Then it is clear that $f(0) > 0$, and therefore $z = 0$ is not a feasible point. Now suppose that contrary to the hypothesis that the optimal solution occurs at an interior point. Denote that optimal solution by z_0. Since it is an interior point, $0 > f(z_0)$. Let γ denote a scalar and consider the function

$f(\gamma z_0)$ as γ varies. Since $f(\cdot)$ is a continuous function it follows that as γ is decreased from 1 towards 0, the value of $f(\gamma z_0)$ must at sometime become equal to 0. But now there is a contradiction as $\|\gamma z_0\| < \|z_0\|$ for $0 < \gamma < 1$. Hence, the optimal solution must lie on the boundary of the feasible set.

Therefore the requirement becomes

$$\min_{\|Ax-b\|=\eta\|x\|} \|x\|.$$

Note that the problem is unaffected by squaring, thus to simplify the algebra we will work with the problem

$$\min_{\|Ax-b\|^2=\eta^2\|x\|^2} \|x\|^2.$$

The problem is equivalently stated as:

$$\min_{\|Ax-b\|^2=\eta^2\|x\|^2} \|x\|^2.$$

6.7 Reduction to Secular Equation

Since the problem has been reduced to an equality constrained minimization problem, the method of Lagrange multipliers can be used. Letting λ denote the Lagrange multiplier obtain the following set of equations that characterize the critical points

$$x + \lambda \left(A^T(Ax-b) - \eta^2 x \right) = 0$$

Simplifying, obtain

$$(A^TA + \frac{1-\lambda\eta^2}{\lambda} I)x = A^Tb.$$

Make the definition $(1 - \lambda\eta^2)/\lambda = \alpha$. Then

$$x = (A^T A + \alpha I)^{-1} A^T b.$$

Plugging this into $\|Ax - b\|^2 = \eta^2 \|x\|^2$ and using the SVD of A obtain

$$\begin{aligned} &b_2^T b_2 + b_1^T \Sigma_1^4 (\Sigma_1^2 + \alpha I)^{-2} b_1 - 2 b_1^T \Sigma_1^2 (\Sigma_1^2 + \alpha I)^{-1} b_1 + b_1^T b_1 \\ &\quad = \eta^2 b_1^T \Sigma_1^2 (\Sigma_1^2 + \alpha I)^{-2} b_1. \end{aligned}$$

Simplifying

$$b_2^T b_2 + b_1^T (\Sigma_1^2 + \alpha I)^{-2} (\alpha^2 I - \eta^2 \Sigma_1^2) b_1 = 0.$$

Since the goal is to find the values of α for which the right hand side of the above equation is zero, define the function $g(\alpha)$ as

$$g(\alpha) = b_2^T b_2 + b_1^T (\Sigma_1^2 + \alpha I)^{-2} (\alpha^2 I - \eta^2 \Sigma_1^2) b_1$$

and then study the zeros of this function. The function $g(\alpha)$ is called the "secular equation", since it is rational function of one variable. If σ_i denotes the ith singular value of A, then the above secular equation has poles at $-\sigma_i^2$.

This secular equation can have up to $2n$ real zeros. One of them will be the minimum norm solution to the problem, $x(\alpha^o)$. Note that if $\alpha > \eta\sigma_1$ in the secular equation, then $b = 0$, which as was stated earlier requires $z = 0$, and thus $x = 0$. Since by assumption $b \neq 0$, $\alpha \leq \eta\sigma_1$.

The secular equation, $g(\alpha)$ is given by:

$$g(\alpha) = b_2^T b_2 + b_1^T (\Sigma_1^2 + \alpha I)^{-2} (\alpha^2 I - \eta^2 \Sigma_1^2) b_1$$

6.8 Main Theorem

In all cases where a degenerate solution exists, the solution is determined by the unique root of the secular equation in the interval $[\max(-\sigma_n^2, -\eta^2), \eta\sigma_1]$.

The rest of the chapter is devoted to establishing this claim. This is a difficult task due to the non-convex nature of the problem and the presence of multiple local minima.

The solution to the problem, $\hat{x}$, is given by $\hat{x} = x(\alpha^o)$ with α^o the unique zero of $$g(\alpha) = b_2^T b_2 + b_1^T(\Sigma_1^2 + \alpha I)^{-2}(\alpha^2 I - \eta^2\Sigma_1^2)b_1$$ in the interval $[\max(-\sigma_n^2, -\eta^2), \eta\sigma_1]$.

6.9 First and Second Order Conditions

Since the Lagrange multiplier must be non-negative at a local minimum and $\lambda = 1/(\alpha + \eta^2)$ conclude that

$$\alpha \geq -\eta^2. \tag{6.7}$$

To narrow down the interesting zeros look at the second order conditions for a local minimum. The Lagrangian is

$$L(x, \lambda) = \|x\|^2 + \lambda(\|Ax - b\|^2 - \eta^2\|x\|^2).$$

The second order condition for a local minimum is that the Hessian of $L(x, \lambda)$ with respect to x be positive-semidefinite when restricted to the tangent subspace of the constraint. Differentiating once

$$\nabla_x L(x, \lambda) = 2x + \lambda\left(2A^T(Ax - b) - 2\eta^2 x\right).$$

Differentiating once more

$$\nabla_x^2 L(x, \lambda) = 2I + \lambda \left(2A^T A - 2\eta^2 I\right),$$

which on simplifying yields

$$\nabla_x^2 L(x, \lambda) = 2\lambda \left(\alpha I + A^T A\right).$$

The constraint is

$$c(x) = \|Ax - b\|^2 - \eta^2 \|x\|^2 .$$

The gradient of the constraint is

$$\nabla_x c(x) = 2A^T(Ax - b) - 2\eta^2 x,$$

which can be simplified by noting that

$$A^T(Ax - b) = -\alpha x$$

thus

$$\nabla_x c(x) = -(\alpha + \eta^2)x.$$

The tangent subspace of the constraint has $n-1$ dimensions (even when $\eta = \sigma_i$). Now construct a basis for this subspace. Using the SVD notation developed in Section 6.5

$$V^T \nabla_x c(x) = -(\alpha + \eta^2)z.$$

Similarly change the basis for the Hessian of the Lagrangian

$$V^T \nabla_x^2 L(x, \lambda) V = 2\lambda \left(\Sigma_1^2 + \alpha I\right),$$

Partition z as

$$z = \begin{pmatrix} z_1 \\ z_2 \end{pmatrix},$$

where z_1 is a scalar. Let

$$H = \begin{pmatrix} z_2^T \\ -z_1 I \end{pmatrix}.$$

Then $H^T z = 0$. Therefore the restricted Hessian is

$$H^T V^T \nabla_x^2 L(x, \lambda) V H = 2\lambda \left(H^T \Sigma_1^2 H + \alpha H^T H \right).$$

Note that the second order condition requires that the restricted Hessian be positive-semidefinite, and Cauchy's interlacing theorem gives the subsequent requirements on the non-restricted Hessian. Cauchy's interlacing theorem states that the smallest eigenvalue for the restricted Hessian must lie between the smallest and second smallest eigenvalues for the non-restricted Hessian. Thus for a local minimum the second smallest eigenvalue of the non-restricted Hessian must be greater than zero. For the condition on the second smallest eigenvalue to be met α must satisfy the constraint $\alpha \geq -\sigma_{n-1}^2$, where σ_{n-1} is the second smallest singular value of A.

This raises the question of how many zeros of the secular equation are larger than $\max(-\eta^2, -\sigma_{n-1}^2)$ **and which of them corresponds to the global minimum.** The proof proceeds by systematically eliminating zeros in this range. There are two critical points (where the secular equation becomes infinite), which correspond to $\alpha = -\sigma_{n-1}^2$ and $\alpha = -\sigma_n^2$. There also are two intervals to worry about, namely $(-\sigma_n^2, \eta\sigma_1)$ and $(-\sigma_{n-1}^2, -\sigma_n^2)$. In the first interval it will be shown that there is only one zero, but this is not true for the second interval. In Section 6.10 the second order condition is used to rule out half of the zeros in the second interval. Appendix E shows that only the rightmost root in the second interval is actually a candidate. Four candidates remain, two in the intervals and two critical points, and six cases are used to prove which one corresponds to the global minimum.

$$\alpha^o > \max(-\eta^2, -\sigma_{n-1}^2)$$

6.10 Squeezing the Second-Order Conditions

The goal of this section is to use the second-order conditions to discard some zeros in the interval $(-\sigma_{n-1}^2, -\sigma_n^2)$. Recall that the restricted Hessian is

$$H^T V^T \nabla_x^2 L(x,\lambda) V H = 2\lambda \left(H^T \Sigma_1^2 H + \alpha H^T H \right).$$

This can be expanded to obtain

$$H^T V^T \nabla_x^2 L(x,\lambda) V H = 2\lambda \left(\sigma_1^2 z_2 z_2^T + z_1^2 \Sigma_2^2 + \alpha z_2 z_2^T + \alpha z_1^2 I \right),$$

where

$$\Sigma_1 = \begin{pmatrix} \sigma_1 & 0 \\ 0 & \Sigma_2 \end{pmatrix}.$$

Then make the conformal partition

$$b_1 = \begin{pmatrix} b_{11} \\ b_{12} \end{pmatrix},$$

and use the representation $z = (\Sigma_1^2 + \alpha I)^{-1} \Sigma_1 b_1$ to simplify the expansion. Additionally, make the definition $M = H^T V^T \nabla_x^2 L(x,\lambda) V H$ for ease of reading and thus obtain the simplified expansion

$$\begin{aligned} M &= 2\lambda \frac{b_{11}^2 \sigma_1^2}{(\sigma_1^2 + \alpha)^2} (\Sigma_2^2 + \alpha I) \\ &\quad \left(I + \frac{(\sigma_1^2 + \alpha)^3}{b_{11}^2 \sigma_1^2} (\Sigma_2^2 + \alpha I)^{-2} \Sigma_2 b_{12} b_{12}^T \Sigma_2 (\Sigma_2^2 + \alpha I)^{-1} \right). \end{aligned}$$

Now compute the determinant,

$$\det(M) = \left(\frac{2\lambda b_{11}^2\sigma_1^2}{(\sigma_1^2+\alpha)^2}\right)^n \det(\Sigma_2^2+\alpha I)$$
$$\left(1+\frac{(\sigma_1^2+\alpha)^3}{b_{11}^2\sigma_1^2} b_{12}^T\Sigma_2^2(\Sigma_2^2+\alpha I)^{-3}b_{12}\right),$$

which can be further simplified to obtain

$$\det(M) = \left(\frac{2\lambda b_{11}^2\sigma_1^2}{(\sigma_1^2+\alpha)^2}\right)^n \frac{(\sigma_1^2+\alpha)^3}{b_{11}^2\sigma_1^2}$$
$$\det(\Sigma_2^2+\alpha I)\left(b_1^T\Sigma_1^2(\Sigma_1^2+\alpha I)^{-3}b_1\right). \quad (6.8)$$

Recall the definition of the secular equation, $g(\alpha)$, given in Section 6.7:

$$g(\alpha) = b_2^T b_2 + b_1^T(\Sigma_1^2+\alpha I)^{-2}(\alpha^2 I - \eta^2\Sigma_1^2)b_1.$$

Then differentiating once obtain

$$g'(\alpha) = 2(\alpha+\eta^2)b_1^T\Sigma_1^2(\Sigma_1^2+\alpha I)^{-3}b_1. \quad (6.9)$$

Using this rewrite Equation (6.8) as

$$\det(M) = \left(\frac{2\lambda b_{11}^2\sigma_1^2}{(\sigma_1^2+\alpha)^2}\right)^n \frac{(\sigma_1^2+\alpha)^3}{2(\alpha+\eta^2)b_{11}^2\sigma_1^2}\det(\Sigma_2^2+\alpha I)g'(\alpha).$$

Therefore, if a root of the secular equation lies in the interval $(-\sigma_{n-1}^2, -\sigma_n^2)$ then it can correspond to a local minimum only if $g'(\alpha)$ is non-positive. **This essentially means that only half of the zeros in the interval correspond to local minima.**

> A zero, α_k, of $g(\alpha)$ in the interval $(-\sigma_{n-1}^2, -\sigma_n^2)$ can correspond to a local minimum of the Lagrangian (and thus have a chance of being the global minimum α^o) only if
>
> $$g'(\alpha_k) \leq 0.$$

6.11 Four Candidate Zeros

At this point there are several potential candidates for α. First there is the possibility of a root in the interval $[-\sigma_n^2, \eta\sigma_1]$ designated α_1. The uniqueness and conditions for existence of α_1 will be shown later. Second, there are potentially many roots in the interval $(-\sigma_{n-1}^2, -\sigma_n^2)$, but only the rightmost one matters as will be shown later and it is thus designated α_2. Finally, there are two critical points, $\alpha_3 = -\sigma_n^2$ and $\alpha_4 = -\sigma_{n-1}^2$. The candidates are summarized in table 6.1.

$\alpha_1 \in [-\sigma_n^2, \eta\sigma_1]$
$\alpha_2 \in (-\sigma_{n-1}^2, -\sigma_n^2)$
$\alpha_3 = -\sigma_n^2$
$\alpha_4 = -\sigma_{n-1}^2$

Table 6.1. Candidate Zeros

The proof involves six cases, which cover special conditions for the problem. See Table 6.2. While the material is complicated it is very useful to understand the intricacies of the problem. The first two cases involve small values of η. The second two cases cover when b is orthogonal to the left singular vector(s) of the smallest singular value. The last two cases cover when b is not orthogonal to the left singular vector(s) of the smallest singular value. Now proceed to prove this and show which candidate root will yield the solution to the problem, $\hat{x}$.

Case 1: $\eta < \sigma_n$
Case 2: $\eta = \sigma_n$
Case 3: $\eta > \sigma_n$, $b_{1,n} = 0$, $\sigma_n < \sigma_{n-1}$
Case 4: $\eta > \sigma_n$, $\|b_{1,(n-k+1,n)}\| = 0$, $\sigma_n = \sigma_{n-k+1}$
Case 5: $\eta > \sigma_n$, $b_{1,n} \neq 0$, $\sigma_n < \sigma_{n-1}$
Case 6: $\eta > \sigma_n$, $\|b_{1,n-k+1}\| \neq 0$, $\sigma_n = \sigma_{n-k+1}$

Table 6.2. Six Cases of the Proof

6.12 Case 1: $\eta < \sigma_n$

This is the easiest case to handle. This is because there is only one root in the interval $[-\eta^2, \eta\sigma_1]$ and this must correspond to the global minimum, as there are no other local minima to worry about. The only candidate zero is α_1 because of the first order condition, Equation 6.7. Only the existence and uniqueness of α_1 must be proven.

Since $\alpha + \eta^2 \geq 0$ from Equation 6.7, it follows by using Equation 6.9 that $g'(\alpha)$ is positive in the interval $(-\eta^2, \infty)$ when $\eta \leq \sigma_n$. Therefore there can be at most one root in the the interval $[-\eta^2, \eta\sigma_1]$.

It will now be shown that there is at least one root in the interval $[-\eta^2, \eta\sigma_1]$. Simplifying the degeneracy condition in Equation 6.6 by using the SVD of A

obtain

$$b_2^T b_2 - \eta^2 b_1^T (\Sigma_1^2 - \eta^2 I)^{-1} b_1 \leq 0,$$

which is identical to $g(-\eta^2) \leq 0$. Furthermore,

$$\lim_{\alpha \to \eta\sigma_1} g(\alpha) > 0.$$

Therefore there must be a zero of $g(\alpha)$ in the interval $[-\eta^2, \eta\sigma_1]$.

6.13 Case 2: $\eta = \sigma_n$

It is claimed that there is a unique root of $g(\alpha)$ in $[-\sigma_n^2, \eta\sigma_1]$, and this is the global minimum. Uniqueness is established by the same method as in Section 6.12, and thus if a root exists in the interval $[-\sigma_n^2, \eta\sigma_1]$ it is unique. Only two candidates, the zero α_1 and the critical point α_3, are possible because of the first order condition, Equation 6.7. The claim will be proven in two steps. Before starting, note that if σ_n is multiple with multiplicity k then $\tilde{b}_1 = b_{1,(n-k+1:n)}$ is the partitioning of b_1 corresponding to the multiple singular values of σ_n.

The first case is when $b_{1,n} \neq 0$ or $\|\tilde{b}_1\| \neq 0$. First note that in this case the candidate zero α_3 is not possible. To see this, partition Σ_1 as

$$\Sigma_1 = \begin{pmatrix} \bar{\Sigma}_1 & 0 \\ 0 & \sigma_n \end{pmatrix}.$$

Similarly partition z into $\bar{z}$ and z_n, and b_1 into $\bar{b}_1$ and $b_{1,n}$. Use these to rewrite the Lagrange condition, $(A^T A + \alpha I)x = A^T b$ as

$$\begin{pmatrix} \bar{\Sigma}_1^2 + \alpha_3 I & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \bar{z} \\ z_n \end{pmatrix} = \begin{pmatrix} \bar{\Sigma}_1 \bar{b}_1 \\ b_{1,n} \end{pmatrix}$$

Since $b_{1,n} \neq 0$, α_3 cannot be α^o. The existence of a root in the interval $(-\sigma_n^2, \eta\sigma_1)$ follows from the observation that

$$\begin{aligned}\lim_{\alpha \to -\sigma_n^2+} g(\alpha) &= -\infty \\ \lim_{\alpha \to \eta\sigma_1} g(\alpha) &\geq 0.\end{aligned}$$

Thus when $b_{1,n} \neq 0$ or $\|\tilde{b}_1\| \neq 0$, $\alpha^o = \alpha_1$.

The second case is $b_{1,n} = 0$ when $\sigma_n < \sigma_{n-1}$ or $\|\tilde{b}_1\| = 0$ when σ_n is multiple. In this case note that there is no longer a pole in $g(\alpha)$ at $\alpha = -\sigma_n^2$. By observing the degeneracy condition given by Equation 6.6 that the degeneracy in this case is determined by b so for degeneracy, Equation 6.6 must hold for a smaller problem. Simplifying the Equation 6.6 using the SVD of A obtain

$$b_2^T b_2 - \eta^2 b_1^T (\Sigma_1^2 - \eta^2 I)^{-1} b_1 \leq 0,$$

which is identical to $g(-\eta^2) \leq 0$. Furthermore,

$$\lim_{\alpha \to \eta\sigma_1} g(\alpha) \geq 0.$$

Therefore there must be a root in the interval $[-\eta^2, \eta\sigma_1]$, so α_1 exists. It will now be shown when α_3 is α^o. To satisfy the equation

$$\begin{pmatrix} \bar{\Sigma}_1^2 + \alpha_3 I & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \bar{z} \\ z_n \end{pmatrix} = \begin{pmatrix} \bar{\Sigma}_1 \bar{b}_1 \\ 0 \end{pmatrix},$$

the following must hold

$$\bar{z} = \left(\bar{\Sigma}_1^2 + \alpha_3 I\right)^{-1} \bar{\Sigma}_1 \bar{b}_1.$$

The constraint equation can be written in z and simplified to

$$\alpha_3 \bar{b}_1^T \left(\bar{\Sigma}_1^2 + \alpha_3 I\right)^{-1} \bar{b}_1 + b_2^T b_2 = 0.$$

Note that this is exactly $g(\alpha_3) = 0$. Thus for α_3 to be a candidate it must also be the unique root in the interval $[-\eta^2, \eta\sigma_1]$. The condition for α_3 to be α^o is that $\alpha_1 = \alpha_3$, and thus we can easily see that in all cases the unique zero which corresponds to the problem solution, $x(\alpha^o)$, is given by α_1.

6.14 Case 3: $\eta > \sigma_n$, $b_{1,n} = 0$, $\sigma_n < \sigma_{n-1}$

Again it is claimed that there is a unique root in $[-\sigma_n^2, \eta\sigma_1]$ and this is the global minimum, which this section proves. Two cases arise when $b_{1,n} = 0$ by observing the equation

$$\begin{pmatrix} \bar{\Sigma}_1^2 + \alpha I & 0 \\ 0 & \sigma_n^2 + \alpha \end{pmatrix} \begin{pmatrix} \bar{z} \\ z_n \end{pmatrix} = \begin{pmatrix} \bar{\Sigma}_1 \bar{b}_1 \\ 0 \end{pmatrix}. \tag{6.10}$$

First, it could be that $\alpha = \alpha_3 = -\sigma_n^2$, which can only happen when $b_{1,n} = 0$. The second case is $z_n = 0$. Note that it is still true that

$$\lim_{\alpha \to \eta\sigma_1} g(\alpha) \geq 0.$$

It is also true that $g'(\alpha) > 0$ on the interval $(-\sigma_{n-1}^2, \infty)$, thus if a root exists it is unique. Start by finding the form of the solution $\hat{x}$ when $\alpha = \alpha_3$ and then show the conditions for determining which candidate zero yields the global minimum.

When $\alpha = -\sigma_n^2$ the solution is found in two steps. First solve for $\bar{z}$ from Equation 6.10. Obtain

$$\bar{z} = \left(\bar{\Sigma}_1^2 - \sigma_n^2 I\right)^{-1} \bar{\Sigma}_1 \bar{b}_1.$$

Note that the constraint can be written in z as

$$\left\| \begin{matrix} \Sigma_1 z - b_1 \\ b_2 \end{matrix} \right\|^2 - \eta^2 \|z\|^2 = 0.$$

Now separate z_n in the constraint and obtain

$$\bar{b}_1^T \left(\bar{\Sigma}_1^2 - \sigma_n^2 I\right)^{-2} \left(\sigma_n^4 I - \eta^2 \bar{\Sigma}_1^2\right) \bar{b}_1 + b_2^T b_2 + \left(\sigma_n^2 - \eta^2\right) z_n^2 = 0.$$

Note that this can be rewritten in terms of $g(-\sigma_n^2)$ as

$$g\left(-\sigma_n^2\right) + \left(\sigma_n^2 - \eta^2\right) z_n^2 = 0.$$

Thus see there are two answers (positive and negative squares) for z_n. The answers for z_n are given by

$$z_n^2 = \frac{g\left(-\sigma_n^2\right)}{\eta^2 - \sigma_n^2}. \tag{6.11}$$

Note that for a solution for z_n to exist $g(-\sigma_n^2) \geq 0$. The solution is then given by

$$\hat{x} = V \begin{bmatrix} \left(\bar{\Sigma}_1^2 - \sigma_n^2 I\right)^{-1} \bar{\Sigma}_1 \bar{b}_1 \\ \pm\sqrt{\frac{g(-\sigma_n^2)}{\eta^2 - \sigma_n^2}} \end{bmatrix}.$$

Which of the potential roots is the actual solution still must be shown. This is broken into two steps. The first is when $g(-\sigma_n^2) \leq 0$, and the second is $g(-\sigma_n^2) > 0$. If $g(-\sigma_n^2) \leq 0$ then trivially a unique root in $[-\sigma_n^2, \eta\sigma_1]$ exists. Moreover, no root exists in the interval $(-\sigma_{n-1}^2, -\sigma_n^2)$ so α_2 is not a candidate. Note that for $\alpha_4 = -\sigma_{n-1}^2$ to be a candidate, it must be true that $b_{1,n-1} = 0$. When $b_{1,n-1} = 0$, $g'(\alpha) > 0$ on the interval $(-\sigma_{n-2}^2, \infty)$, which means $g(-\sigma_{n-1}^2) < 0$. If it is assumed that $\alpha = \alpha_4$ and proceed similarly to Section 6.14 it is seen that $g(-\sigma_{n-1}^2) \geq 0$ is required and thus α_4 cannot be α^o. Note that when $g(-\sigma_n^2) < 0$, it is impossible for $\alpha = -\sigma_n^2$. When $g(-\sigma_n^2) = 0$, the unique root is $\alpha = -\sigma_n^2$ and thus the two remaining candidate zeros can easily be seen to coincide. Thus when $g(-\sigma_n^2) \leq 0$, the unique zero is given by α_1.

When $g(-\sigma_n^2) > 0$ no root exists in $(-\sigma_n^2, \eta\sigma_1]$ so α_1 is not α^o but as was seen in Section 6.14 this is the condition for $\alpha = \alpha_3 = -\sigma_n^2$. Note that when $g(-\sigma_n^2) > 0$, there can be a root in the interval $(-\sigma_{n-1}^2, -\sigma_n^2)$, but the slope is positive in this interval and by the results of Section 6.10 it cannot be a minimum. The only remaining question in this case is if $\alpha_4 = -\sigma_{n-1}^2$ is a candidate when $g(-\sigma_n^2) > 0$. Again recall that for $-\sigma_{n-1}^2$ to be a candidate, it must be true that $b_{1,n-1} = 0$ and $g(-\sigma_{n-1}^2) \geq 0$. The equation

$$\begin{pmatrix} \tilde{\Sigma}_1^2 + \alpha I & 0 & 0 \\ 0 & \sigma_{n-1}^2 + \alpha & 0 \\ 0 & 0 & \sigma_n^2 + \alpha \end{pmatrix} \begin{pmatrix} \tilde{z} \\ z_{n-1} \\ z_n \end{pmatrix} = \begin{pmatrix} \tilde{\Sigma}_1 \tilde{b}_1 \\ 0 \\ 0 \end{pmatrix} \tag{6.12}$$

must be satisfied. The following argument shows that $-\sigma_{n-1}^2$ is not a candidate when $g(-\sigma_{n-1}^2) \geq 0$. Note that since $b_{1,n-1} = 0 = b_{1,n}$, $g'(\alpha) > 0$ on the interval $(-\sigma_{n-2}^2, \infty)$. Now introduce the parameter $\gamma = \|b_{1,n-1}\|^2$ and consider a continuity argument on γ similar to the continuity argument in Section 6.16. Since the argument is very similar to the one to be constructed, only a sketch of the details will be provided here. Note that for $\gamma \neq 0$ there is a root in the interval $(-\sigma_{n-1}^2, -\sigma_n^2)$ which is not the global minimum. As γ goes to zero, this root moves to the left, and it reaches $-\sigma_{n-1}^2$ when $\gamma = 0$, since $g(-\sigma_{n-1}^2) \geq 0$. The derivative of the cost with respect to γ can be seen to be negative in the interval $(-\sigma_{n-1}^2, -\sigma_n^2)$ by the following method. First take the derivative and note there appears the term $d\alpha(\gamma)/d\gamma$, which is solved for by taking the derivative of $g(\alpha(\gamma)) = 0$ with respect to γ. Substituting back in and simplifying observe that as γ increases, the cost decreases in the interval $(-\sigma_{n-1}^2, -\sigma_n^2)$ and thus the x corresponding to the root which appears in the interval when $\gamma \neq 0$ has a lower cost than the x which corresponds to $-\sigma_{n-1}^2$. The root is not a global minimum however and so neither can be α^o at $-\sigma_{n-1}^2$. The only possibility when $g(-\sigma_n^2) \geq 0$ is thus $\alpha^o = -\sigma_n^2$.

6.15 Case 4: $\eta > \sigma_n$, $\|b_{1,(n-k+1,n)}\| = 0$, $\sigma_n = \sigma_{n-k+1}$

Again the claim is made that there is a unique root in $[-\sigma_n^2, \eta\sigma_1]$, and this is the global minimum. For simplicity partition Σ_1 as

$$\Sigma_1 = \begin{pmatrix} \bar{\Sigma}_1 & 0 \\ 0 & \sigma_n I \end{pmatrix},$$

where $\bar{\Sigma}_1$ corresponds to the singular values that are strictly greater than σ_n. Similarly partition z into $\bar{z}$ and $\tilde{z}$, and b_1 into $\bar{b}_1$ and $\tilde{b}_1$. Two cases arise when $\tilde{b}_1 = 0$ by observing the equation

$$\begin{pmatrix} \bar{\Sigma}_1^2 + \alpha I & 0 \\ 0 & (\sigma^2 + \alpha) I \end{pmatrix} \begin{pmatrix} \bar{z} \\ \tilde{z} \end{pmatrix} = \begin{pmatrix} \bar{\Sigma}_1 \bar{b}_1 \\ 0 \end{pmatrix}. \tag{6.13}$$

First it could be that $\alpha = -\sigma_n^2$, which note can only happen when $b_{1,n} = 0$. The second case is $\tilde{z} = 0$. Note that

$$\lim_{\alpha \to \eta\sigma_1} g(\alpha) \geq 0$$

is still true, and that $g'(\alpha) > 0$ on the interval $(-\sigma_{n-1}^2, \infty)$, thus if a root exists it is unique.

When $\alpha = -\sigma_n^2$, the solution is found in two steps. First solve for $\bar{z}$ from Equation 6.13. Obtain

$$\bar{z} = \left(\bar{\Sigma}_1^2 - \sigma_n^2 I\right)^{-1} \bar{\Sigma}_1 \bar{b}_1.$$

Note that the constraint can be written in z as

$$\left\| \begin{matrix} \Sigma_1 z - b_1 \\ b_2 \end{matrix} \right\|^2 - \eta^2 \|z\|^2 = 0.$$

Now separate $\tilde{z}$ in the constraint and obtain

$$\bar{b}_1^T \left(\bar{\Sigma}_1^2 - \sigma_n^2 I\right)^{-2} \left(\sigma_n^4 I - \eta^2 \bar{\Sigma}_1^2\right) \bar{b}_1 b_2^T b_2 + \left(\sigma_n^2 - \eta^2\right) \tilde{z}^T \tilde{z} = 0.$$

Similar to what was seen in the last Section, note that the above equation can be written in terms of $g(-\sigma_n^2)$. Doing so, obtain

$$g\left(-\sigma_n^2\right) + \left(\sigma_n^2 - \eta^2\right)\tilde{z}^T\tilde{z} = 0.$$

Note that this defines a hypersphere with radius

$$r = \sqrt{\frac{g\left(-\sigma_n^2\right)}{\eta^2 - \sigma_n^2}}.$$

To be able to solve for the radius, $g(-\sigma_n^2) \geq 0$, and thus this is a condition on the when $\alpha = -\sigma_n^2$. Let Θ be any vector with unit Euclidean norm. The solutions for $\tilde{z}$ are given by

$$\tilde{z} = r\Theta.$$

The solution is then given by

$$\hat{x} = V\begin{bmatrix} \left(\bar{\Sigma}_1^2 - \sigma_n^2 I\right)^{-1}\bar{\Sigma}_1\bar{b}_1 \\ r\Theta \end{bmatrix}.$$

Note that the second order condition requires that $\alpha \leq -\sigma_n^2$ and thus the only candidates are α_1 and α_3. If $g(-\sigma_n^2) \leq 0$ then trivially there is a unique root in $[-\sigma_n^2, \eta\sigma_1]$, and it is impossible for $\alpha = -\sigma_n^2$. If $g(-\sigma_n^2) > 0$ no root exists in $(-\sigma_n^2, \eta\sigma_1]$ but as was seen above this is the condition for $\alpha = -\sigma_n^2$. When $g(-\sigma_n^2) = 0$ the two zeros can easily be seen to coincide.

6.16 Case 5: $\eta > \sigma_n$, $b_{1,n} \neq 0$, $\sigma_n < \sigma_{n-1}$

The claim is again made that there is a unique root in $(-\sigma_n^2, \eta\sigma_1]$ and it is the global minimum. Note that since $b_{1,n} \neq 0$, $\alpha = -\sigma_n^2$ cannot be a solution.

The existence of a root in the interval $[-\sigma_n^2, \eta\sigma_1]$ follows from the observation that

$$\begin{aligned}
\lim_{\alpha \to -\sigma_n^2+} g(\alpha) &= -\infty \\
\lim_{\alpha \to \eta\sigma_1} g(\alpha) &\geq 0.
\end{aligned}$$

Uniqueness is established by the same method as in Section 6.12.

Now proceed to show that of the three candidate roots only the one in the interval $(-\sigma_n^2, \eta\sigma_1]$ can be the global minimum. The argument proceeds by continuation on $\beta = b_{1,n}^2$. Begin by defining

$$\bar{g}(\alpha) = b_2^T b_2 + \bar{b}_1^T \left(\bar{\Sigma}_1^2 + \alpha I\right)^{-2} \left(\alpha^2 I - \eta^2 \bar{\Sigma}_1^2\right) \bar{b}_1.$$

Rewrite the secular equation $g(\alpha)$ in terms of α and β as

$$g(\alpha, \beta) = \bar{g}(\alpha) + \beta \frac{\alpha^2 - \eta^2 \sigma_n^2}{(\sigma_n^2 + \alpha)^2}.$$

Note that when $\beta = 0$, $g(\alpha, 0) = \bar{g}(\alpha)$. Also note that $\bar{g}'(\alpha, 0) > 0$ when α lies in the interval $(\max(-\sigma_{n-1}^2, -\eta^2), \infty)$. Let $\alpha_1(\beta)$ denote the unique root in the interval $(-\sigma_n^2, \eta\sigma_1]$, and $\alpha_2(\beta)$ denote the rightmost root in the interval $(-\sigma_{n-1}^2, -\sigma_n^2)$ of $g(\alpha, \beta)$. Also let $y_1(\beta)$ denote the stationary point $V^T\hat{x}$ corresponding to $\alpha_1(\beta)$, and similarly for $y_2(\beta)$ corresponding to $\alpha_2(\beta)$.

When $\bar{g}(-\sigma_n^2) < 0$, note that neither $\alpha_1(\beta)$ nor $\alpha_2(\beta)$ converges to $-\sigma_n^2$ as β goes to zero. As already observed, at $\beta = 0$, $g'(\alpha, 0) > 0$ when α lies in the interval $(\max(-\sigma_{n-1}^2, -\eta^2), \infty)$, and since $\bar{g}(-\sigma_n^2) < 0$ this implies that $g'(\alpha, 0) > 0$ when α lies in the interval $(\max(-\sigma_{n-1}^2, -\eta^2), -\sigma_n^2)$. Thus $y_2(\beta)$ does not exist at $\beta = 0$ and so it must not exist for some open neighborhood around $\beta = 0$. For $y_2(\beta)$ to be a candidate there must exist some value of β, say β_2 for which $y_2(\beta)$ first exists. At the point β_2, $\alpha_2(\beta_2)$ must be at least a double root, and thus the slope of $g(\alpha_2(\beta))$ must be zero at β_2. From Section 6.10, note that $\alpha_2(\beta_2)$ cannot be the α^o, so it must be that $\|y_2(\beta_2)\|^2 \geq \|y_1(\beta_2)\|^2$.

Now proceed with the case when $\bar{g}(-\sigma_n^2) \geq 0$, and it will then be shown that in both cases $\|y_2(\beta)\|^2$ gets larger as β increases, while $\|y_1(\beta)\|^2$ decreases. It is easy to note from the form of $g(\alpha)$ that

$$\lim_{\beta \to 0+} \alpha_1(\beta) = -\sigma_n^2 = \lim_{\beta \to 0+} \alpha_2(\beta)$$

when $\bar{g}(-\sigma_n^2) \geq 0$. Now proceed to show that

$$\lim_{\beta \to 0+} |y_{1,i}(\beta)| = |y_{1,i}(0)| = |y_{2,i}(0)| = \lim_{\beta \to 0+} |y_{2,i}(\beta)|, \quad 1 \leq i \leq n.$$

First observe that this is trivially true for $i \neq n$. Next note that $\bar{g}(\alpha)$ is continuous at $\alpha = -\sigma_n^2$, thus

$$\begin{aligned}
\lim_{\beta \to 0+} \left(y_{2,n}(\beta)^2 - y_{1,n}(\beta)^2\right) &= \lim_{\beta \to 0+} \left(\frac{\sigma_n^2}{\alpha_2(\beta)^2 - \eta^2\sigma_n^2} \frac{\alpha_2(\beta)^2 - \eta^2\sigma_n^2}{(\alpha(\beta) + \sigma_n^2)^2}\beta \right. \\
&\quad \left. - \frac{\sigma_n^2}{\alpha_1(\beta)^2 - \eta^2\sigma_n^2} \frac{\alpha_1(\beta)^2 - \eta^2\sigma_n^2}{(\alpha(\beta) + \sigma_n^2)^2}\beta \right) \\
&= \frac{1}{\sigma_n^2 - \eta^2} \\
&\quad \lim_{\beta \to 0+} \left(\frac{\alpha_2(\beta)^2 - \eta^2\sigma_n^2}{(\alpha_2(\beta) + \sigma_n^2)^2}\beta + \bar{g}(\alpha_2(\beta), \beta) \right. \\
&\quad \left. - \frac{\alpha_1(\beta)^2 - \eta^2\sigma_n^2}{(\alpha_1(\beta) + \sigma_n^2)^2}\beta - \bar{g}(\alpha_1(\beta), \beta) \right) \\
&= \frac{1}{\sigma_n^2 - \eta^2} \lim_{\beta \to 0+} \left(g(\alpha_2(\beta), \beta) - g(\alpha_1(\beta), \beta)\right) \\
&= 0.
\end{aligned}$$

Note that this shows that $\|y_1(\beta)\|$ and $\|y_2(\beta)\|$ are continuous for $\beta \geq 0$, with $\|y_1(0)\| = \|y_2(0)\|$.

Now examine the derivative of the cost function, $\|x\|^2$ with respect to β. Use this to show that in both cases $\|y_1(\beta)\|$ is less than $\|y_2(\beta)\|$ for all $\beta \geq 0$. The derivative is

$$\frac{d\left\|x(\alpha(\beta))\right\|^2}{d\beta} = \frac{\sigma_n^2}{\left(\sigma_n^2+\alpha(\beta)\right)^2} - 2\frac{d\alpha(\beta)}{d\beta} b_1^T \left(\Sigma_1^2+\alpha(\beta)I\right)^{-3} \Sigma_1^2 b_1.$$

The derivative of $\alpha(\beta)$ with respect to β must be calculated, so take the derivative of $g(\alpha(\beta))=0$

$$\begin{aligned} 0 &= \frac{dg(\alpha(\beta))}{d\beta} \\ &= \frac{\alpha(\beta)^2-\eta^2\sigma_n^2}{\left(\alpha(\beta)+\sigma_n^2\right)^2} + 2\left(\alpha(\beta)+\eta^2\right)\frac{d\alpha(\beta)}{d\beta}\left(b_1^T\left(\Sigma_1^2+\alpha(\beta)I\right)^{-3}\Sigma_1^2 b_1\right). \end{aligned}$$

Solving for the derivative of $\alpha(\beta)$ with respect to β yields

$$\frac{d\alpha(\beta)}{d\beta} = -\frac{\alpha(\beta)^2-\eta^2\sigma_n^2}{2\left(\alpha(\beta)+\eta^2\right)\left(\sigma_n^2+\alpha(\beta)\right)^2\left(b_1^T\left(\Sigma_1^2+\alpha(\beta)I\right)^{-3}\Sigma_1^2 b_1\right)}.$$

Substituting this into the derivative of $\|x\|^2$ with respect to β obtain

$$\frac{d\left\|x(\alpha(\beta))\right\|^2}{d\beta} = \frac{\sigma_n^2}{\left(\sigma_n^2+\alpha(\beta)\right)^2} + \frac{\alpha(\beta)^2-\eta^2\sigma_n^2}{\left(\alpha(\beta)+\eta^2\right)\left(\sigma_n^2+\alpha(\beta)\right)^2}.$$

Simplifying this yields

$$\frac{d\left\|x(\alpha(\beta))\right\|^2}{d\beta} = \frac{\alpha(\beta)}{\left(\alpha(\beta)+\eta^2\right)\left(\alpha(\beta)+\sigma_n^2\right)}.$$

Clearly for increasing β it can be seen that $\alpha_1(\beta)$ decreases the cost function when $\alpha_1(\beta)<0$, while $\alpha_2(\beta)$ increases the cost function for all β. When $0 \le \alpha_1(\beta) \le \eta\sigma_n$, $d\alpha(\beta)/d\beta \ge 0$ and note that the cost is increasing for both $y_1(\beta)$ and $y_2(\beta)$. Since the cost is increasing for $y_1(\beta)$ when $0 \le \alpha_1(\beta) \le \eta\sigma_n$, $\|y_1(\beta)\|^2 \le \|y_1(\eta\sigma_n)\|^2$ on this interval. Additionally, note that for $\alpha_1(\beta)$ in the interval $[\eta\sigma_n, \eta\sigma_1]$, $d\alpha(\beta)/d\beta \le 0$ and the cost increases with increasing β. Note that while these observations are true for $[\eta\sigma_n, \infty]$, the interval $[\eta\sigma_n, \eta\sigma_1]$ is specified because the root cannot lie in $[\eta\sigma_1, \infty]$. Observe that it has been shown that $\|y_1(\beta)\|^2 \le \|y_1(\eta\sigma_n)\|^2$ when $\alpha_1(\beta)$ is in the interval $[\eta\sigma_n, \eta\sigma_1]$. Thus the maximum value of the cost, when $\alpha_1(\beta)$ is in the interval $[\eta\sigma_n, \eta\sigma_1]$, occurs at $\beta = \eta\sigma_n$. The maximum rate of change for the cost, when$\alpha_1(\beta)$ is in the interval $[0, \eta\sigma_1]$, can easily be found to be

$$\max \frac{d\,\|x(\alpha(\beta))\|^2}{d\beta} = \frac{\eta\sigma_n}{(\eta\sigma_n + \eta^2)\,(\eta\sigma_n + \sigma_n^2)}.$$

Simplifying, obtain

$$\max \frac{d\,\|x(\alpha(\beta))\|^2}{d\beta} = \frac{1}{(\eta + \sigma_n)^2}.$$

A similar calculation can be done for the interval $(\max(-\eta^2, -\sigma_{n-1}^2), -\sigma_{n-1}^2)$ and thus find that the minimum increase in the cost occurs at $\beta = -\eta\sigma_n$ and is given by

$$\min \frac{d\,\|x(\alpha(\beta))\|^2}{d\beta} = \frac{1}{(\eta - \sigma_n)^2}.$$

Now note that the maximum rate of increase for $y_1(\beta)$ is less than the minimum rate of increase for $y_2(\beta)$, and for β sufficiently small, $\|y_1(\beta)\| \le \|y_2(\beta)\|$. It can now be easily seen that $\|y_1(\beta)\| \le \|y_2(\beta)\|$ for all β thus α_2 cannot be the global minimum.

Now consider the third candidate zero, namely $-\sigma_{n-1}^2$. Note that for it to be a candidate, $b_{1,n-1} = 0$ and $g(-\sigma_{n-1}^2) \ge 0$. Observe that similar to what was seen in Appendix E, the minimum on the interval $(-\sigma_{n-2}^2, -\sigma_n^2)$ must occur between the second to the rightmost and the rightmost roots of the secular equation on the interval. In Appendix E the only options were the roots themselves, but in this case there is also the possibility of $-\sigma_{n-1}^2$. Note that if $-\sigma_{n-1}^2$ is not one of the two rightmost roots on the interval $(-\sigma_{n-2}^2, -\sigma_n^2)$ then it cannot be the global minimum. It is already known that the rightmost root, designated α_2 is not the global minimum, and additionally the second most right root cannot be the global minimum since the slope of $g(\alpha)$ is not negative at this point.

Re-introduce the parameter $\gamma = \|b_{1,n-1}\|^2$ and consider a continuity argument on γ similar to the continuity argument presented in Section 6.16. Since the argument is very similar to the one constructed, it will only be sketched here. Note that for $\gamma \neq 0$ there are multiple roots in the interval $(-\sigma_{n-1}^2, -\sigma_n^2)$, none of which is the global minimum. As γ goes to zero all of the roots are made to move to the left, and all but the rightmost either reaches $-\sigma_{n-1}^2$ or becomes complex valued as $\gamma \to 0$, since $g(-\sigma_{n-1}^2) \geq 0$. The derivative of the cost with respect to γ can be seen to be negative in the interval $(-\sigma_{n-1}^2, -\sigma_n^2)$ by the following method. First take the derivative and note there appears the term $d\alpha(\gamma)/d\gamma$, which is solved for by taking the derivative of $g(\alpha(\gamma)) = 0$ with respect to γ. Substituting back in and simplifying it is seen that as γ increases the cost decreases in the interval $(-\sigma_{n-1}^2, -\sigma_n^2)$ and thus the x which corresponds to the root which appears in the interval when $\gamma \neq 0$ has a lower cost than the x which corresponds to $-\sigma_{n-1}^2$. That root is not a global minimum however and so neither can the root at $-\sigma_{n-1}^2$. The possibility that $-\sigma_{n-1}^2$ is α^o is thus excluded, and the case is finished.

6.17 Case 6: $\eta > \sigma_n$, $\|b_{1,n-k+1}\| \neq 0$, $\sigma_n = \sigma_{n-k+1}$

Once more it is claimed that there is a unique root, α_1, in the interval $(-\sigma_n^2, \eta\sigma_1]$ and it is the global minimum.

The existence of a root , α_1, in the interval $(-\sigma_n^2, \eta\sigma_1]$ follows from the observation that

$$\begin{aligned} \lim_{\alpha \to -\sigma_n^2+} g(\alpha) &= -\infty \\ \lim_{\alpha \to \eta\sigma_1} g(\alpha) &\geq 0. \end{aligned}$$

$\eta < \sigma_n$ and $b^T(I - A(A^TA - \eta^2 I)^{-1}A^T)b \leq 0$
$\eta = \sigma_n$, $b_{1,n} = 0$, and $\bar{b}_1^T(I - \bar{\Sigma}_1^2(\bar{\Sigma}_1 - \eta^2 I)^{-1})\bar{b}_1 \leq 0$
$\eta = \sigma_n$, $b_{1,n} \neq 0$
$\eta > \sigma_n$

Table 6.3. Degeneracy Conditions

Uniqueness is established by the same method as in Section 6.12.

Since $\|b_{1,n-k+1}\| \neq 0$, $\alpha = \alpha_3 = -\sigma_n^2$ is not possible. Note that the second order condition gives the additional requirement that $\alpha \geq -\sigma_n^2$. Since $\alpha \geq -\sigma_n^2$ then trivially there are no additional roots to worry about. The only candidate is thus the unique root, α_1, in the interval $(-\sigma_n^2, \eta\sigma_1]$.

6.18 Summary of Results

The problem under consideration is

$$\min_{x \in \mathcal{R}^n} \min_{\|E\| \leq \eta} \|(A + E)x - b\|$$

where A is an $m \times n$ real matrix and b is an n-dimensional real column vector. The assumption is made that the problem is degenerate and in particular that there exists an x such that $\eta\|x\| \geq \|Ax - b\|$. Degeneracy can be easily checked as outlined in Table 6.3. To obtain a solution to the degenerate problem the optimization problem

$$\min_{\|Ax-b\|\le\eta\|x\|} \|x\|$$

is considered. The SVD of A is given by

$$A = \begin{bmatrix} U_1 & U_2 \end{bmatrix} \begin{bmatrix} \Sigma_1 \\ 0 \end{bmatrix} V^T.$$

And define $b_1 = U_1^T b$, and $b_2 = U_2^T b$. When $b_{1,n} = 0$ if σ_n is unique or $\|b_{1,n-k+1,n}\| = 0$ if σ_n is of multiplicity k, partition Σ_1 as

$$\Sigma_1 = \begin{pmatrix} \bar{\Sigma}_1 & 0 \\ 0 & \sigma_n I \end{pmatrix}.$$

Similarly partition b_1 into $\bar{b}_1$ and $b_{1,n} = 0$. The secular equation is given by

$$g(\alpha) = b_2^T b_2 + b_1^T(\Sigma_1^2 + \alpha I)^{-2}(\alpha^2 I - \eta^2\Sigma_1^2)b_1.$$

Given these definitions, the solution to the problem is given in Table 6.4. Note that to find the unique root of the secular equation, $g(\alpha)$ in the interval specified can be easily and quickly done by a method such as bisection or Newton's method.

6.19 Restricted Perturbations

So far, the case in which all the columns of the A matrix are subject to perturbations has been considered. It may happen in practice, however, that only selected columns are uncertain, while the remaining columns are known precisely. This situation can be handled by the approach of this chapter as is now clarified.

Given $A \in \mathbb{R}^{m\times n}$, partition it into block columns,

$$A = \begin{bmatrix} A_1 & A_2 \end{bmatrix},$$

Condition	Solution
$\eta > \sigma_n$, $\sigma_n < \sigma_{n-1}$, $b_{1,n} = 0$, $g(-\sigma_n^2) \geq 0$	$x = V \begin{bmatrix} \left(\bar{\Sigma}_1^2 - \sigma_n^2 I\right)^{-1} \bar{\Sigma}_1 \bar{b}_1 \\ \pm\sqrt{\frac{g(-\sigma_n^2)}{\eta^2-\sigma_n^2}} \end{bmatrix}$
$\eta > \sigma_n$, $\sigma_n = \sigma_{n-k+1}$, $\lVert b_{1,(n-k+1,n)} \rVert = 0$, $g(-\sigma_n^2) \geq 0$	$\hat{x} = V \begin{bmatrix} \left(\bar{\Sigma}_1^2 - \sigma_n^2 I\right)^{-1} \bar{\Sigma}_1 \bar{b}_1 \\ r\Theta \end{bmatrix}$ $r = \sqrt{\frac{g(-\sigma_n^2)}{\eta^2-\sigma_n^2}}$ $\lVert \Theta \rVert = 1$
else	$x = (A^T A + \alpha I)^{\dagger} A^T b$ $\alpha_1 \in [\max(-\sigma_n^2, -\eta^2), \eta\sigma_1]$ such that $g(\alpha_1) = 0$

Table 6.4. Solution to the Problem

and assume, without loss of generality, that only the columns of A_2 are subject to perturbations while the columns of A_1 are known exactly. Then pose the following problem:

Given $A \in \mathbb{R}^{m \times n}$, with $m \geq n$ and A full rank, $b \in \mathbb{R}^m$, and nonnegative real number η_2, determine $\hat{x}$ such that

$$\min_{\hat{x}} \min_{\lVert E_{A_2} \rVert \leq \eta_2} \left\{ \left\lVert \begin{bmatrix} A_1 & A_2 + E_{A_2} \end{bmatrix} \hat{x} - b \right\rVert \right\}. \tag{6.14}$$

Partition $\hat{x}$ accordingly with A_1 and A_2, say

$$\hat{x} = \begin{bmatrix} \hat{x}_1 \\ \hat{x}_2 \end{bmatrix}$$

then write

$$\left\lVert \begin{bmatrix} A_1 & A_2 + E_{A_2} \end{bmatrix} \hat{x} - b \right\rVert = \lVert A\hat{x} - b + E_{A_2}\hat{x}_2 \rVert .$$

Assuming the fundamental condition for this case, which is

$$\eta_2 \|\hat{x}_2\| \geq \|Ax - b\|,$$

and following the development of Section 6.4 conclude the problem is equivalent to

$$\min_{\|Ax-b\|^2 = \eta_2^2 \|x_2\|^2} \|x\|^2.$$

Note that the constraint can be rewritten as

$$\|Ax - b\|^2 + \eta_2^2 \|x_1\|^2 = \eta_2^2 \|x_2\|^2 + \eta_2^2 \|x_1\|^2,$$

which becomes

$$\left\| \begin{bmatrix} A_1 & A_2 \\ \eta_2 I & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - \begin{bmatrix} b \\ 0 \end{bmatrix} \right\|^2 = \eta_2^2 \|x\|^2.$$

Now define the following

$$\tilde{A} = \begin{bmatrix} A_1 & A_2 \\ \eta_2 I & 0 \end{bmatrix},$$

and

$$\tilde{b} = \begin{bmatrix} b \\ 0 \end{bmatrix}.$$

The problem thus becomes

$$\min_{\|\tilde{A}x - \tilde{b}\|^2 = \eta_2^2 \|x\|^2} \|x\|^2,$$

which is easily seen to be of the same form as the original problem, though of slightly larger dimension. This can thus be solved by the method developed in this chapter.

Chapter 7

Min Max Backward Error Criterion

This chapter examines four related problems, which cover special cases of the min max backward error criterion. The cost functions are rational, and provide a natural sequence of solutions to examine the problem. First the basic motivation and formulations of the problems are examined in Section 7.1. The third and fourth formulations are solved in Section 7.2. A relation to the TLS problem is then presented in Section 7.3. The second cost function formulation is solved in Section 7.4. The form of solution, and current results for the first and primary formulation is presented in Section 7.5.

7.1 Motivation and Formulation

A common way of seeking a good answer in numerical analysis is to find a method of calculation that minimizes the backward error of the problem. The basic idea of the backward error analysis technique is to show the solution obtained

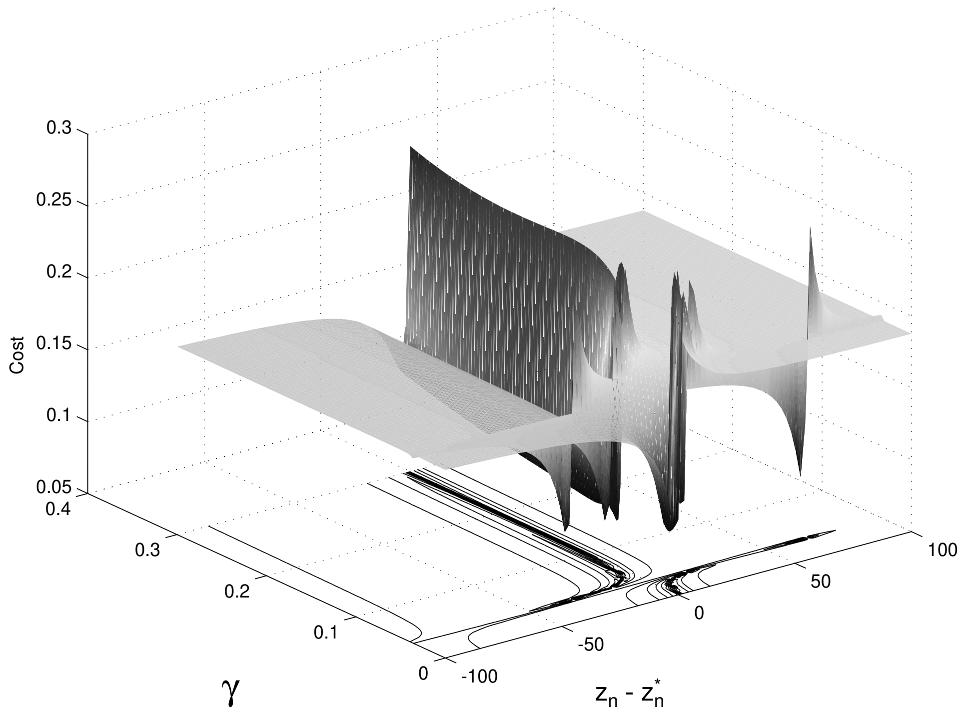

Figure 7.1. Cost function, showing the many singularities and relative minima.

is the exact solution of a nearby problem, by use of floating-point arithmetic error bounds. A method that does this is said to be a stable one. A method with a small backward error is not guaranteed accurate answers, but rather that the method used is a good one in the numerical sense. See [88, 135] for a more complete treatment. It seems reasonable to ask then what the regression problem would be for a backward error criterion. The problem is thus stated

$$\min_x \max_{\|E\| \leq \eta} \frac{\|(A+E)x - b\|}{\|A\|\|x\| + \|b\|}.$$

It is important to note that this cost function is not convex, as most other cost functions are. This can be seen in Figure 7.1, which shows a region around the global minimum for a randomly generated matrix with 8 rows and 4 columns. Note in particular that a number of the minimum are close to the global minimum in cost (one within 2.5% of the cost). This problem is thus more difficult in nature than most, but it holds forth a potential of a numerically superior way of approaching the problem. To help understand the backward error problem better it will be considered as part of a family of four related costs functions. The problems are all rational cost functions and are generated by the presence or lack of the uncertainty in A (E) and the $\|b\|$ on the bottom. The costs functions are thus

1. $\min_x \max_{\|E\|\leq\eta} \frac{\|(A+E)x-b\|}{\|A\|\|x\|+\|b\|}$,
2. $\min_x \frac{\|Ax-b\|}{\|A\|\|x\|+\|b\|}$,
3. $\min_x \max_{\|E\|\leq\eta} \frac{\|(A+E)x-b\|}{\|A\|\|x\|}$,
4. $\min_x \frac{\|Ax-b\|}{\|A\|\|x\|}$.

The three additional problems are thus nominal models in some sense. First, consider how to handle the maximization of the uncertainty (for problems 1 and 3). Since E appears only in the numerator, the maximization is accomplished by maximizing the numerator. Similar to what is done in [24], the original problem is identical to

$$\min_x \frac{\|Ax-b\|+\eta\|x\|}{\|A\|\|x\|+\|b\|}$$

and the third problem is identical to

$$\min_x \frac{\|Ax-b\|+\eta\|x\|}{\|A\|\|x\|}.$$

By a little rearranging, the third problem can be written as

$$\min_x \left(\frac{\|Ax-b\|}{\|A\|\|x\|}+\frac{\eta}{\|A\|}\right).$$

Essentially, problem three and problem four have the same solution! The four problems are thus

1. $\min_x \frac{\|Ax-b\|+\eta\|x\|}{\|A\|\|x\|+\|b\|} = \min_x C_1(x)$,

2. $\min_x \frac{\|Ax-b\|}{\|A\|\|x\|+\|b\|} = \min_x C_2(x)$,

3. $\min_x \frac{\|Ax-b\|+\eta\|x\|}{\|A\|\|x\|} = \min_x C_3(x)$,

4. $\min_x \frac{\|Ax-b\|}{\|A\|\|x\|} = \min_x C_4(x)$.

The problems will be examined in reverse order (simple to complex).

7.2 Formulations Three and Four

Since the third and fourth cost functions will give the same answer, consider the simpler model (fourth cost function) to get both. The model is

$$\bar{C}_4(x) = \frac{\|Ax-b\|^2}{\|A\|^2\|x\|^2}.$$

The square of the fourth formula is being used since it has an identical solution to the non-squared version and it only has a non-differentiable point at $x = 0$. Begin by noting that for this function, $x = 0$ is not a possible answer since the cost for $x = 0$ is always more costly than say the least squares solution x_{LS}. The solution is thus always at a differentiable point. Assume that $b \notin \mathcal{R}(A)$, since if it is not, the solution is trivially identical to the least squares solution. For all $x \neq 0$, the gradient of $\bar{C}_4(x)$ with respect to x is

$$\begin{aligned}
\nabla_x \bar{C}_4(x) &= 2\frac{A^T(Ax-b) - \frac{\|Ax-b\|^2}{\|x\|^2}x}{\|A\|^2\|x\|^2} \\
&= 2\frac{\left(A^TA - \frac{\|Ax-b\|^2}{\|x\|^2}I\right)x - A^Tb}{\|A\|^2\|x\|^2}.
\end{aligned}$$

The candidate solutions, x_{opt}, are then found by setting $\nabla_x \bar{C}_4(x) = 0$ and solving to find

$$x_{opt} = \left(A^T A - \frac{\|Ax_{opt} - b\|^2}{\|x_{opt}\|^2} I \right)^{-1} A^T b.$$

Thus consider the parameterized family $x(\gamma)$ given by

$$x(\gamma) = \left(A^T A - \gamma I\right)^{-1} A^T b.$$

Define the regression parameter γ such that γ_{opt} is given by

$$\gamma_{opt} = \frac{\|Ax_{opt} - b\|^2}{\|x_{opt}\|^2} > 0. \tag{7.1}$$

An expression to calculate γ_{opt} is needed. Rewrite the definition of γ_{opt} and substitute the expression for x_{opt} found above.

$$\gamma_{opt} \| \left(A^T A - \gamma_{opt} I\right)^{-1} A^T b\|^2 \;=\; \|A \left(A^T A - \gamma_{opt} I\right)^{-1} A^T b - b\|^2.$$

To simplify the expression above, introduce the singular value decomposition (SVD) of A as $A = \begin{bmatrix} U_1 & U_2 \end{bmatrix} \begin{bmatrix} \Sigma & 0 \end{bmatrix}^T V^T$. Also introduce the notation $b_1 = U_1^T b$ and $b_2 = U_2^T b$. The expression becomes

$$\begin{aligned}
\gamma_{opt} \| \left(\Sigma^2 - \gamma_{opt} I\right)^{-1} \Sigma b_1 \|^2 &= \left\| \begin{bmatrix} \left(\Sigma^2 - \gamma_{opt} I\right)^{-1} \Sigma^2 b_1 - b_1 \\ -b_2 \end{bmatrix} \right\|^2 \\
\gamma_{opt} b_1^T \left(\Sigma^2 - \gamma_{opt} I\right)^{-2} \Sigma^2 b_1 &= \left\| \begin{bmatrix} -\gamma_{opt} \left(\Sigma^2 - \gamma_{opt} I\right)^{-1} b_1 \\ -b_2 \end{bmatrix} \right\|^2 \\
\gamma_{opt} b_1^T \left(\Sigma^2 - \gamma_{opt} I\right)^{-2} \Sigma^2 b_1 &= \gamma_{opt}^2 b_1^T \left(\Sigma^2 - \gamma_{opt} I\right)^{-2} b_1 + b_2^T b_2 \\
\gamma_{opt} b_1^T \left(\Sigma^2 - \gamma_{opt} I\right)^{-2} \left(\Sigma^2 - \gamma_{opt} I\right) b_1 - b_2^T b_2 &= 0 \\
\gamma_{opt} b_1^T \left(\Sigma^2 - \gamma_{opt} I\right)^{-1} b_1 - b_2^T b_2 &= 0.
\end{aligned}$$

Call the expression,

$$g_1(\gamma) = \gamma b_1^T \left(\Sigma^2 - \gamma I\right)^{-1} b_1 - b_2^T b_2, \tag{7.2}$$

the secular equation in keeping with the literature [24, 25]. Thus the value of γ_{opt} is specified by the roots of the secular equation. To find the root several questions need to be answered. Does the root exist? Is the root unique? Is there an interval where the root occurs? Immediately note from the expression for γ_{opt}, Equation 7.1, that it is greater than zero. The basic outline is to find an upper bound on the size of the γ_{opt}, which is a local minimum. When the upper bound is found, establish uniqueness and finally existence.

Before proceeding further, note a simple relation that can be derived from $x(\gamma) = \left(A^T A - \gamma I\right)^{-1} A^T b$ and will prove useful in our development. Note that this relation holds for all values of γ.

$$\begin{aligned} x(\gamma) = \left(A^T A - \gamma I\right)^{-1} A^T b \\ \left(A^T A - \gamma I\right) x(\gamma) = A^T b \\ A^T A x(\gamma) - \gamma x(\gamma) = A^T b \\ A^T (Ax(\gamma) - b) &= \gamma x(\gamma) \end{aligned}$$

Now proceed with taking the second derivative of the cost,

$$\begin{aligned} \nabla_x^2 \bar{C}_4(x) &= 2\frac{A^T A - \gamma I - 2x\left(\frac{(Ax-b)^T A}{\|x\|^2} - \frac{x^T \|Ax-b\|^2}{\|x\|^4}\right)}{\|A\|^2\|x\|^2} \\ &\quad -4\frac{\left(A^T A - \frac{\|Ax-b\|^2}{\|x\|^2} I\right) x - A^T b}{\|A\|^2\|x\|^4} x^T \\ &= 2\frac{A^T A - \gamma I - 2x\left(\frac{\gamma x^T}{\|x\|^2} - \frac{x^T \|Ax-b\|^2}{\|x\|^4}\right)}{\|A\|^2\|x\|^2} \\ &\quad -2\frac{\nabla_x \bar{C}_4(x) x^T}{\|x\|^2} \\ &= 2\frac{A^T A - \gamma I - 2P_x\left(\gamma - \frac{\|Ax-b\|^2}{\|x\|^2}\right)}{\|A\|^2\|x\|^2} \\ &\quad -2\frac{\nabla_x \bar{C}_4(x) x^T}{\|x\|^2}. \end{aligned}$$

Note that P_x is the projection onto x and is given by $\frac{xx^T}{\|x\|^2}$. Since only the slope at the roots of the secular equation are of concern, that means $\nabla_x \bar{C}_4(x) = 0$ and $\gamma = \frac{\|Ax-b\|^2}{\|x\|^2}$. The second derivative becomes $2(A^TA - \gamma I)$, which is positive definite for $\gamma < \sigma_n^2$, where σ_n is the smallest singular value of A. This means

$$\boxed{0 \leq \gamma \leq \sigma_n^2.}$$

To show uniqueness, take the derivative of the secular equation,

$$\begin{aligned} g_1'(\gamma) &= b_1^T \left(\Sigma^2 - \gamma I\right)^{-1} b_1 + \gamma b_1^T \left(\Sigma^2 - \gamma I\right)^{-2} b_1 \\ &= b_1^T \left(\Sigma^2 - \gamma I\right) \left(\Sigma^2 - \gamma I\right)^{-2} b_1 + \gamma b_1^T \left(\Sigma^2 - \gamma I\right)^{-2} b_1 \\ &= b_1^T \Sigma^2 \left(\Sigma^2 - \gamma I\right)^{-2} b_1 \\ &> 0. \end{aligned}$$

The derivative is positive, so the root will be unique. Note that discontinuities exist, but not in the interval where the solution must lie. All that remains is then to show existence.

Begin by observing that $g_1(0) = -b_2^T b_2 \leq 0$. For simplicity, assume that the smallest singular value of A is unique, the extension is obvious. Now if the n^{th} element of b_1, denoted $b_{1,n}$ is not zero then $\lim_{\gamma \to \sigma_n^2} g_1(\gamma) = \infty$. If $b_{1,n} \neq 0$ then trivially the root exists. If $b_{1,n} = 0$ then note that

$$g_1(\sigma_n^2) = \sigma_n^2 \bar{b_1}^T (\bar{\Sigma}^2 - \gamma I)^{-1} \bar{b_1} - b_2^T b_2,$$

with

$$b_1 = \begin{bmatrix} \bar{b_1} \\ 0 \end{bmatrix} \quad \Sigma = \begin{bmatrix} \bar{\Sigma} & 0 \\ 0 & \sigma_n \end{bmatrix}.$$

If this number is non-negative, again the root exists. The question remaining is what happens when the number is negative? In this case look at the gradient of the cost function

$$\begin{bmatrix} \bar{\Sigma}^2 - \gamma I & 0 \\ 0 & \sigma_n^2 - \gamma \end{bmatrix} \begin{bmatrix} \bar{z} \\ z_n \end{bmatrix} = \begin{bmatrix} \bar{\Sigma}\bar{b_1} \\ 0 \end{bmatrix}$$

where,

$$\begin{bmatrix} \bar{x} \\ x_n \end{bmatrix} = v \begin{bmatrix} \bar{z} \\ z_n \end{bmatrix}.$$

Since $g_1(\sigma_n^2) < 0$ and $g_1'(\gamma) > 0$, there is no root of $g_1(\gamma)$ for $\gamma < \sigma_n^2$. The global minimum root is less than or equal to σ_n^2, so $\gamma = \sigma_n^2$. This means that $\bar{z} = (\bar{\Sigma}^2 - \sigma_n^2 I)^{-1}\bar{\Sigma}\bar{b_1}$. To find the value of z_n , substitute the values of γ and $\bar{z}$ back into the cost function and find that

$$\bar{C}_4\left(v\begin{bmatrix} \bar{z} \\ z_n \end{bmatrix}\right) = \frac{\sigma_n^2}{\sigma_1^2}\frac{\sigma_n^4 \bar{b_1}^T\left(\bar{\Sigma}^2 - \gamma I\right)^{-2}\bar{b_1} + b_2^T b_2 + \sigma_n^2 z_n^2}{\sigma_n^2 \bar{b_1}^T \bar{\Sigma}^2\left(\bar{\Sigma}^2 - \gamma I\right)^{-2}\bar{b_1} + \sigma_n^2 z_n^2}.$$

Only three possibilities exist, $z_n = 0$, $z_n = \pm\infty$, or z_n can be anything. For a rational function of the form

$$\frac{\alpha + \sigma_n^2 z_n^2}{\beta + \sigma_n^2 z_n^2},$$

the value of z_n is given by

$$z_n = \begin{cases} 0 & \beta - \alpha > 0 \\ \pm\infty & \beta - \alpha < 0 \\ * & \beta - \alpha = 0. \end{cases}$$

For this problem $\beta - \alpha = g_1(\sigma_n^2) < 0$ and thus $z_n = \pm\infty$. The entire solution is thus characterized.

7.3 Relation to TLS

Two regression problems hold predominance in estimation, least squares and total least squares. For formulations three and four, the regression parameter, γ is greater than zero, which is the least squares regression parameter. The parameter γ also has a relation to the total least squares regression parameter σ_{n+1}.

Start by noting an interesting bound on the size of the total least squares regression parameter. To see the bound, consider the TLS problem written as

$$\begin{bmatrix} A^T A & A^T b \\ b^T A & b^T b \end{bmatrix} \begin{bmatrix} x_{TLS} \\ -1 \end{bmatrix} = \sigma_{n+1}^2 \begin{bmatrix} x_{TLS} \\ -1 \end{bmatrix}.$$

The top line specifies the form of the solution as

$$x_{TLS} = \left(A^T A - \sigma_{n+1}^2 I\right)^{-1} A^T b.$$

The bottom line gives the secular equation for the TLS problem. This can be seen by inserting the form of solution into the bottom line.

$$\begin{aligned} b^T A x_{TLS} - b^T b &= -\sigma_{n+1}^2 \\ b^T A \left(A^T A - \sigma_{n+1}^2 I\right)^{-1} A^T b - b^T b &= -\sigma_{n+1}^2. \end{aligned}$$

Recall the SVD of A and the definitions of b_1 and b_2 that have been used in earlier sections,

$$\begin{aligned} A &= \begin{bmatrix} U_1 & U_2 \end{bmatrix} \begin{bmatrix} \Sigma \\ 0 \end{bmatrix} V^T \\ b_1 &= U_1^T b \\ b_2 &= U_2^T b. \end{aligned}$$

This yields

$$\begin{aligned}
b_1^T\Sigma^2(\Sigma^2-\sigma_{n+1}^2 I)^{-1}b_1 - b_1^T b_1 - b_2^T b_2 &= -\sigma_{n+1}^2\\
\sigma_{n+1}^2 b_1^T(\Sigma^2-\sigma_{n+1}^2 I)^{-1}b_1 - b_2^T b_2 &= -\sigma_{n+1}^2\\
\sigma_{n+1}^2 b_1^T(\Sigma^2-\sigma_{n+1}^2 I)^{-1}b_1 + \sigma_{n+1}^2 &= b_2^T b_2\\
\sigma_{n+1}^2\left(b_1^T(\Sigma^2-\sigma_{n+1}^2 I)^{-1}b_1 + 1\right) &= b_2^T b_2. \qquad (7.3)
\end{aligned}$$

Recalling that $\sigma_{n+1}\sigma_n$ it can be seen that

$$b_1^T(\Sigma^2-\sigma_{n+1}^2 I)^{-1}b_1 > 0,$$

and thus

$$b_1^T(\Sigma^2-\sigma_{n+1}^2 I)^{-1}b_1 + 1 > 1.$$

Finally arrive at the desired upper bound of the TLS parameter,

$$\boxed{\sigma_{n+1}^2 < b_2^T b_2}$$

Now proceed to see how the regression parameter from Section 7.2 compares with the TLS parameter, σ_{n+1}. To do so we subtract Equation 7.3 from Equation 7.2 and obtain

$$\begin{aligned}
\gamma b_1^T\left(\Sigma^2-\gamma I\right)^{-1}b_1 - \sigma_{n+1}^2\left(b_1^T(\Sigma^2-\sigma_{n+1}^2 I)^{-1}b_1 + 1\right) &= 0\\
\gamma b_1^T\left(\Sigma^2-\gamma I\right)^{-1}b_1 - \sigma_{n+1}^2 b_1^T(\Sigma^2-\sigma_{n+1}^2 I)^{-1}b_1 &= \sigma_{n+1}^2\\
b_1^T\left(\gamma\left(\Sigma^2-\gamma I\right)^{-1} - \sigma_{n+1}^2(\Sigma^2-\sigma_{n+1}^2 I)^{-1}\right)b_1 &= \sigma_{n+1}^2\\
b_1^T(\Sigma^2-\gamma I)^{-1}(\Sigma^2-\sigma_{n+1}^2 I)^{-1}\left(\gamma(\Sigma^2-\sigma_{n+1}^2 I) - \sigma_{n+1}^2(\Sigma^2-\gamma I)\right)b_1 &= \sigma_{n+1}^2.
\end{aligned}$$

Recall that $0 \leq \sigma_{n+1} \leq \sigma_n$ and $0 \leq \gamma \leq \sigma_n$ so

$$\begin{aligned} \Sigma^2 - \gamma I &\geq 0 \\ \Sigma^2 - \sigma_{n+1}^2 I &\geq 0. \end{aligned}$$

Thus the following results

$$\begin{aligned} \gamma(\Sigma^2 - \sigma_{n+1}^2 I) - \sigma_{n+1}^2(\Sigma^2 - \gamma I) &\geq 0 \\ \gamma\Sigma^2 - \gamma\sigma_{n+1}^2 I - \sigma_{n+1}^2\Sigma^2 + \gamma\sigma_{n+1}^2 I &\geq 0 \\ \gamma\Sigma^2 - \sigma_{n+1}^2\Sigma^2 &\geq 0 \\ \gamma - \sigma_{n+1}^2 &\geq 0. \end{aligned}$$

The final result is

$$\gamma \geq \sigma_{n+1}^2.$$

Formulations three and four always deregularize more than TLS. This gives the even better bounds for γ of

$$\boxed{\sigma_{n+1}^2 \leq \gamma \leq \sigma_n^2.}$$

7.4 Formulation Two

The main difficulty in solving the general problem (problem 1) is the denominator. In particular the addition of the $\|b\|$ in the denominator adds considerable complexity. It is reasonable to ask why to bother with it, after all a solution exists without it. The main problem with formulations three and four are that they are always too optimistic. As x approaches zero from sufficiently close, the denominator goes to zero and thus the cost rises. It is never possible for this regression technique to return $x = 0$ even though there are times when that is the best choice from a physical standpoint. The constant addition of $\|b\|$ in the denominator of the cost prevents the cost function from going to infinity as x is near zero. Formulations one and two are thus more realistic than formulations three and four. It is important to note that for formulation two, the only time $x = 0$ is when $A^\dagger b = 0$, as this is the only way to have $\|Ax - b\| = \|A\|\|x\| + \|b\|$. In practical terms this will not happen so it will be assumed that $A^\dagger b \neq 0$ for this section and thus $x = 0$ is not a candidate. When $b = AA^\dagger b$, the choice of $x = A^\dagger b$ yields a cost of zero, so it is the solution. When $b \neq AA^\dagger b$, the only non-differentiable point has been ruled out so the solution is at a differentiable point.

The next step in solving the general problem (formulation) is to consider the next most difficult problem (formulation two). Formulation two is defined by the cost function

$$\min_x C_2(x) = \min_x \frac{\|Ax - b\|}{\|A\|\|x\| + \|b\|}.$$

As with the other problems, take the gradient with respect to x and by rearranging terms, find that

$$\begin{aligned}
\nabla_x C_2(x) &= \frac{A^T(Ax - b) - \frac{\|Ax-b\|^2\|A\|}{\|x\|(\|A\|\|x\|+\|b\|)}x}{\|Ax - b\|(\|A\|\|x\| + \|b\|)} \\
&= \frac{A^T(Ax - b) - \gamma_2 x}{\|Ax - b\|(\|A\|\|x\| + \|b\|)}
\end{aligned}$$

with $\gamma_2 = \frac{\|Ax-b\|^2\|A\|}{\|x\|(\|A\|\|x\|+\|b\|)}$. By setting $\nabla_x C_2(x)$ equal to zero, this yields

$$x(\gamma_2) = (A^T A - \gamma_2 I)^{-1} A^T b.$$

The Hessian is given by

$$\nabla_x^2 C_2(x) = \frac{1}{\|Ax-b\|(\|A\|\|x\|+\|b\|)} \left(A^T A - \gamma_2 I + \gamma_2 \frac{\|b\| x x^T}{(\|A\|\|x\|+\|b\|)\|x\|^2} \right).$$

Denote the singular value decomposition of A as before, with the smallest singular value of A given by σ_n. First note that when $\gamma_2 \leq \sigma_n^2$ then the Hessian is positive semidefinite.

The Hessian will be positive semidefinite if $A^T A - \gamma_2 I + \gamma_2 \frac{\|b\| x x^T}{(\|A\|\|x\|+\|b\|)\|x\|^2}$ is positive semidefinite. Using the SVD and denoting $z = v^T x$ the condition becomes $\Sigma^2 - \gamma_2 I + \gamma_2 \frac{\|b\| z z^T}{(\|A\|\|z\|+\|b\|)\|z\|^2}$ must be non-negative. By partitioning z into $z = [\bar{z}^T \quad z_n]^T$ and partitioning the remaining matrices similarly, the Hessian condition can be written as:

$$\begin{bmatrix} \bar{\Sigma}^2 - \gamma I & 0 \\ 0 & \sigma_n^2 - \gamma \end{bmatrix} + \gamma_2 \frac{\|b\|}{(\|A\|\|z\|+\|b\|)\|z\|^2} \begin{bmatrix} \bar{z} \\ z_n \end{bmatrix} \begin{bmatrix} \bar{z} \\ z_n \end{bmatrix}^T$$

and the form of solution becomes

$$(\Sigma^2 - \gamma_2 I) z = \Sigma^T b_1$$

with $b_1 = U_1 b$ as before. When $b_{1,n} = 0$ then either $z_n = 0$ or $\gamma_2 = \sigma_n^2$. If $\gamma_2 = \sigma_n^2$, trivially $\gamma_2 \leq \sigma_n^2$. But note that if $z_n = 0$ then by the Hessian, $\gamma_2 \leq \sigma_n^2$.

$$\boxed{b_{1,n} = 0 \quad \Rightarrow \quad \gamma_2 \leq \sigma_n^2}$$

7.4.1 Perturbation Analysis

It has already been shown that when $b_{1,n} = 0$ that $\gamma_2 \leq \sigma_n^2$, so it remains to be shown that this remains true when $b_{1,n} \neq 0$. Note from perturbation theory (for example Section 8.6.1 of [68]) that for the Hessian condition to be non-negative it must have either $\gamma_2 \leq \sigma_{n-1}^2$ or $\gamma_2 \leq \sigma_n^2 \frac{\|A\|\|z\|+\|b\|}{\|A\|\|z\|}$. In particular if the smallest singular value is a multiple singular value, so $\sigma_n = \sigma_{n-1}$, then trivially $\gamma_2 \leq \sigma_n^2$. As a final observation, $\gamma_2 = \sigma_n^2$ only when $b_{1,n} = 0$ so by continuity of γ_2, in order to have $\gamma_2 > \sigma_n^2$ it must first have $\gamma_2 = \sigma_n^2$ when $b_{1,n} = 0$. It only remains to show that when $b_{1,n} \neq 0$ and σ_n is a unique singular value that $\gamma_2 \leq \sigma_n^2$.

This is done by first performing a perturbation analysis on $b_{1,n}$. When $b_{1,n} = 0$, $\gamma_2 = \sigma_n^2$, so when $b_{1,n} = \delta b_{1,n} \ll 1$, $\gamma_2 = \sigma_n^2 \pm \delta\gamma_2$. Note that by continuity there is always found a small value of $\delta b_{1,n}$ such that $\delta\gamma_2 \ll \sigma_{n-1}^2 - \sigma_n^2$. To show that $\gamma_2 < \sigma_n^2$ it needs to be shown that the cost for $\gamma_2 = \sigma_n^2 + \delta\gamma_2$ is greater than the cost for $\gamma_2 = \sigma_n^2 - \delta\gamma_2$. Proceed by examining the cost function and the first order condition of the cost function.

Note that for the partitioning

$$z(\gamma_2) = \begin{bmatrix} \bar{z} \\ z_n \end{bmatrix}$$

the cost function can be rewritten as

$$C_2 = \frac{\left\| \begin{bmatrix} \bar{\Sigma} & 0 \\ 0 & \sigma_n \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \bar{z} \\ z_n \end{bmatrix} - \begin{bmatrix} \bar{b}_1 \\ b_{1,n} \\ b_2 \end{bmatrix} \right\|}{\|A\| \left\| \begin{bmatrix} \bar{z} \\ z_n \end{bmatrix} \right\| + \|b\|}.$$

The first order condition can likewise be written as

$$\begin{bmatrix} \bar{\Sigma}^2 - \gamma_2 I & 0 \\ 0 & \sigma_n^2 - \gamma_2 \end{bmatrix} \begin{bmatrix} \bar{z} \\ z_n \end{bmatrix} = \begin{bmatrix} \bar{\Sigma} & 0 \\ 0 & \sigma_n \end{bmatrix} \begin{bmatrix} \bar{b}_1 \\ b_{1,n} \end{bmatrix}.$$

Using the fact that $\gamma_2 = \sigma_n^2 \pm \delta\gamma_2$ it can be seen that $\mp\delta\gamma_2 = \sigma_n^2 - \gamma_2$. Define $\bar{D} = \bar{\Sigma} - \sigma_n^2 I$ then rewrite the first order condition as

$$\begin{bmatrix} \bar{D} \mp \delta\gamma_2 I & 0 \\ 0 & \mp\delta\gamma_2 \end{bmatrix} \begin{bmatrix} \bar{z} \\ z_n \end{bmatrix} = \begin{bmatrix} \bar{\Sigma} & 0 \\ 0 & \sigma_n \end{bmatrix} \begin{bmatrix} \bar{b}_1 \\ b_{1,n} \end{bmatrix}.$$

Since it was noted above that $\delta\gamma_2 \ll \sigma_{n-1}^2 - \sigma_n^2$ and $\sigma_{n-1}^2 - \sigma_n^2$ is the smallest element of the diagonal matrix $\bar{D}$, approximate $\bar{D} \mp \delta\gamma_2 I$ by $\bar{D}$. This results in

$$\begin{aligned} \begin{bmatrix} \bar{D} & 0 \\ 0 & \mp\delta\gamma_2 \end{bmatrix} \begin{bmatrix} \bar{z} \\ z_n \end{bmatrix} &= \begin{bmatrix} \bar{\Sigma} & 0 \\ 0 & \sigma_n \end{bmatrix} \begin{bmatrix} \bar{b}_1 \\ b_{1,n} \end{bmatrix} \\ \begin{bmatrix} \bar{z} \\ z_n \end{bmatrix} &= \begin{bmatrix} \bar{D}^{-1}\bar{\Sigma}\bar{b}_1 \\ \frac{\sigma_n b_{1,n}}{\mp\delta\gamma_2} \end{bmatrix}. \end{aligned}$$

From this note that the norm of z will not be affected by the the sign of $\mp\delta\gamma_2$, thus only the numerator of the cost matters. Substituting the result into the cost function find

$$\begin{aligned} C_2 &= \frac{\left\| \begin{bmatrix} \bar{\Sigma} & 0 \\ 0 & \sigma_n \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \bar{z} \\ z_n \end{bmatrix} - \begin{bmatrix} \bar{b}_1 \\ b_{1,n} \\ b_2 \end{bmatrix} \right\|}{\|A\|\|z\| + \|b\|} \\ &= \frac{\left\| \begin{bmatrix} \bar{\Sigma} & 0 \\ 0 & \sigma_n \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \bar{D}^{-1}\bar{\Sigma}\bar{b}_1 \\ \frac{\sigma_n b_{1,n}}{\mp\delta\gamma_2} \end{bmatrix} - \begin{bmatrix} \bar{b}_1 \\ b_{1,n} \\ b_2 \end{bmatrix} \right\|}{\|A\|\|z\| + \|b\|} \\ &= \frac{\left\| \begin{bmatrix} (\bar{\Sigma} - \sigma_n^2 I)^{-1}\bar{\Sigma}^2\bar{b}_1 \\ \frac{\sigma_n^2 b_{1,n}}{\mp\delta\gamma_2} \\ 0 \end{bmatrix} - \begin{bmatrix} \bar{b}_1 \\ b_{1,n} \\ b_2 \end{bmatrix} \right\|}{\|A\|\|z\| + \|b\|} \end{aligned}$$

$$C_2 = \frac{\left\| \begin{bmatrix} \sigma_n^2(\bar{\Sigma} - \sigma_n^2 I)^{-1}\bar{b}_1 \\ \frac{\sigma_n^2 \pm \delta\gamma_2}{\mp\delta\gamma_2} b_{1,n} \\ -b_2 \end{bmatrix} \right\|}{\|A\|\|z\| + \|b\|}.$$

The only term where the sign of $\pm\delta\gamma_2$ is significant is the second row of the numerator. When the sign is positive the second row is clearly larger than when the sign is negative. Thus the norm of the numerator will be larger when the sign is positive, so the cost is less when $\gamma_2 < \sigma_n^2$ then when $\gamma_2 > \sigma_n^2$. This eliminates the possibility that γ_2 lies in the range $(\sigma_n^2, \sigma_{n-1}^2)$.

7.4.2 Final Case

There only remains the possible alternative of $\gamma_2 = \sigma_{n-1}^2$, which will be disproven now. Consider the first derivative of the cost, with $z = \begin{bmatrix} \bar{z}^T & z_{n-1} & z_n \end{bmatrix}^T$. Partition the other matrices similarly to obtain

$$\begin{bmatrix} \bar{\Sigma}^2 - \gamma_2 I & 0 & 0 \\ 0 & \sigma_{n-1}^2 - \gamma_2 & 0 \\ 0 & 0 & \sigma_{n-1}^2 - \gamma_2 \end{bmatrix} \begin{bmatrix} \bar{z} \\ z_{n-1} \\ z_n \end{bmatrix} = \begin{bmatrix} \bar{\Sigma}\bar{b} \\ \sigma_{n-1} b_{1,n-1} \\ \sigma_n b_{1,n} \end{bmatrix}.$$

For $\gamma_2 = \sigma_{n-1}^2$ it is required that $b_{1,n-1} = 0$ so

$$\begin{bmatrix} \bar{\Sigma}^2 \sigma_{n-1}^2 I & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \sigma_{n-1}^2 - \sigma_{n-1}^2 \end{bmatrix} \begin{bmatrix} \bar{z} \\ z_{n-1} \\ z_n \end{bmatrix} = \begin{bmatrix} \bar{\Sigma}\bar{b} \\ 0 \\ \sigma_n b_{1,n} \end{bmatrix}.$$

It has been shown that if $b_{1,n} = 0$ then $\gamma_2 \leq \sigma_n^2$, so if $\gamma_2 = \sigma_{n-1}^2$ then the determinant of the second derivative of the cost must be positive for all values of $b_{1,n} \neq 0$. Consider the second derivative of the cost partitioned as was done for the first derivative and with $\gamma_2 = \sigma_{n-1}^2$.

$$\begin{bmatrix} \bar{\Sigma}^2 - \sigma_{n-1}^2 I & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \sigma_n^2 - \sigma_{n-1}^2 \end{bmatrix} + \frac{\sigma_{n-1}^2 \|b\|}{(\|A\|\|z\| + \|b\|)\|z\|^2} \begin{bmatrix} \bar{z} \\ z_{n-1} \\ z_n \end{bmatrix} \begin{bmatrix} \bar{z} \\ z_{n-1} \\ z_n \end{bmatrix}^T$$

The determinant of this is always less than or equal to zero (proof of this is in Appendix G) in some neighborhood of $b_{1,n} = 0$, which means that it can never be the minimum at any point, since in order to be the minimum it must be the minimum at some point around $b_{1,n} = 0$. It is thus impossible for $\gamma_2 = \sigma_{n-1}^2$.

7.4.3 Final Result

From what has been shown, the solution to problem 2 is given by

$$x(\gamma_2) = (A^T A - \gamma_2 I)^{-1} A^T b$$

where

$$\gamma_2 \in [0, \sigma_n^2].$$

The value of γ_2 can be found as the root of

$$\gamma_2 - \frac{\|Ax(\gamma_2) - b\|^2 \|A\|}{\|x(\gamma_2)\|(\|A\|\|x(\gamma_2)\| + \|b\|)}$$

in the given range. Since the Hessian is strictly positive in this range and for $\gamma_2 = 0$ the equation is trivially less than zero, the root is unique in this range. The root can be found in n^2 time by using a root finding method like bisection or Newton's method.

7.5 Formulation One

The ultimate formulation has not been completely solved, but several important results exist. First, consider when it is possible for $x = 0$ to be a candidate. The cost for $x = 0$ is 1, so for $x = 0$ to be the solution it must be true that for all $x \neq 0$ that

$$\begin{aligned}
\|Ax - b\| + \eta\|x\| &\geq \|A\|\|x\| + \|b\| \\
\eta\|x\| &\geq \sigma_1\|x\| + \|b\| - \|Ax - b\| \\
\eta &\geq \sigma_1 + \frac{\|b\| - \|Ax - b\|}{\|x\|} \\
\eta &\geq \sigma_1 + \frac{\|b\| - |\|Ax\| - \|b\||}{\|x\|}.
\end{aligned}$$

If $\|b\| > \|Ax\|$ then

$$\begin{aligned}
\eta &\geq \sigma_1 + \frac{\|b\| - (\|b\| - \|Ax\|)}{\|x\|} \\
\eta &\geq \sigma_1 + \frac{\|Ax\|}{\|x\|} \\
\eta &\geq 2\sigma_1.
\end{aligned}$$

If $\|b\| \leq \|Ax\|$ then

$$\begin{aligned}
\eta &\geq \sigma_1 + \frac{\|b\| - (\|Ax\| - \|b\|)}{\|x\|} \\
\eta &\geq \sigma_1 + \frac{2\|b\| - \|Ax\|}{\|x\|} \\
\eta &\geq \sigma_1 + \frac{\|Ax\|}{\|x\|} \\
\eta &\geq 2\sigma_1.
\end{aligned}$$

Thus no matter what, for $x = 0$ to be the solution, $\eta \geq 2\sigma_1$. The point at which $x = 0$ is a candidate solution can be adjusted by changing the term, $\|A\|\|x\|$, in the denominator to $\alpha\|x\|$ for $0 < \alpha \leq \|A\|$. This does not alter the analysis but does permit practical values of η to yield $x = 0$. Note that $\alpha = 0$ is excluded as this is the min max problem discussed in Section 2.7.

The second result of interest is the form of solution and secular equation(s) to find it when the solution is not at a singular point. The cost function is

$$C_1(x) = \frac{\|Ax - b\| + \eta\|x\|}{\|A\|\|x\| + \|b\|}.$$

Taking the gradient and setting it equal to zero yields

$$x(\gamma_1) = (A^T A + \gamma_1 I)^{-1} A^T b,$$

with

$$\begin{aligned}
\gamma_1 &= \frac{\|Ax - b\|}{\|x\|}\left(\eta - \frac{\|Ax - b\| + \eta\|x\|}{\|A\|\|x\| + \|b\|}\|A\|\right) \\
&= \frac{\|Ax - b\|}{\|x\|}\left(\eta - C_1(x)\|A\|\right).
\end{aligned}$$

Technically this is a solution, and it can be calculated by running down all the roots of $g_1(\gamma_1) = \gamma_1 - \frac{\|Ax-b\|}{\|x\|}\left(\eta - C_1(x)\|A\|\right)$. Multiplying by the denominator and then repeatedly squaring the expression and combining terms gives another expression that does not involve norms to odd powers. By substituting in expressions for $\|Ax - b\|$ and $\|x\|$, an alternate secular equation which does not require the calculation of x at each step is found:

$$\begin{aligned}
g_2(\gamma_1) &= \left(\|A\|b^T(A^T A + \gamma_1 I)^{-1}b\right)^4 - 2\left(\|A\|\|b\|b^T(A^T A + \gamma_1 I)^{-1}b\right)^2 \\
&\quad \times b^T(A^T A + \gamma_1 I)^{-1}(A^T A + \eta^2 I)(A^T A + \gamma_1 I)^{-1}b \\
&\quad + \left(\|b\|^2 b^T(A^T A + \gamma_1 I)^{-1}(A^T A - \eta^2 I)(A^T A + \gamma_1 I)^{-1}b\right)^2 \\
&= \left(\|A\|\alpha_1(\gamma_1)\right)^4 - 2\left(\|A\|\|b\|\alpha_1(\gamma_1)\right)^2\alpha_2(\gamma_1) + \left(\|b\|^2\alpha_3(\gamma_1)\right)^2
\end{aligned}$$

with

$$
\begin{aligned}
\alpha_1(\gamma_1) &= b^T(A^TA + \gamma_1 I)^{-1}b \\
\alpha_2(\gamma_1) &= b^T(A^TA + \gamma_1 I)^{-1}(A^TA + \eta^2 I)(A^TA + \gamma_1 I)^{-1}b \\
\alpha_3(\gamma_1) &= b^T(A^TA + \gamma_1 I)^{-1}(A^TA - \eta^2 I)(A^TA + \gamma_1 I)^{-1}b.
\end{aligned}
$$

The only disadvantage to $g_2(\gamma_1)$ is that it has four times the roots as the original (three fictitious for each true). In all the trials conducted the desired root for $g_1(\gamma_1)$ was the only one in $(\infty, -\sigma_n^2]$ if it existed or $-\sigma_n^2]$ if it did not. In the case of $g_2(\gamma_1)$ the root in $(\infty, -\sigma_n^2]$ which was also a root of $g_1(\gamma_1)$ was the solution or $-\sigma_n^2$ was the solution if there were no roots in the interval. Practically this means that all the roots in the interval had to be run down and the costs compared. The root that yielded the lowest cost was the "real" one and the solution. While compelling, this is not a proof so it is listed as incomplete.

7.6 Conclusions

This is arguably the most challenging and interesting problem in the dissertation. Four sub-problems are presented and three completely solved. The final sub-problem has two secular equations derived and thus a solution exists, but some work remains to narrow which root is actually the desired solution. The method has much to recommend it, despite this limitation, as can be seen in Chapter 8 and Chapter 9.

Chapter 8

System Identification Example

There are a great number of books and articles covering how system identification is to be done, for instance see [68, 78, 91, 92, 93, 99, 101, 109, 149, 160]. A basic, linear state-space model will be assumed for the system to be identified, in which the input and output are subject to white noise. Mathematically the system can be written as

$$\begin{aligned} \dot{x} &= Ax + Bu \\ y &= Cx + Du \end{aligned}$$

The state is x and the input, u, and the output, y have additive white noise. The goal in system ID is to find the matrices, A, B, C, and D. The subspace identification problem as outlined in [101] and [92] will be used to find the matrices. This method has become quite popular due to its numerical properties and its ability to directly identify a state space model from input-output data. Note that an infinite number of state space models exist for a given system due to similarity transforms ($T^{-1}AT$, $T^{-1}B$, CT, D for a given transform T). The subspace identification method begins by forming two Hankel matrices of the input-output data,

$$H_1 = \begin{bmatrix} u[k] & u[k+1] & \dots & u[k+j-1] \\ y[k] & y[k+1] & \dots & y[k+j-1] \\ u[k+1] & u[k+2] & \dots & u[k+j] \\ y[k+1] & y[k+2] & \dots & y[k+j] \\ \vdots & \vdots & & \vdots \\ u[k+i-1] & u[k+i] & \dots & u[k+i+j-2] \\ y[k+i-1] & y[k+i] & \dots & y[k+i+j-2] \end{bmatrix}$$

$$H_2 = \begin{bmatrix} u[k+i] & u[k+i+1] & \dots & u[k+i+j-1] \\ y[k+i] & y[k+i+1] & \dots & y[k+i+j-1] \\ u[k+i+1] & u[k+i+2] & \dots & u[k+i+j] \\ y[k+i+1] & y[k+i+2] & \dots & y[k+i+j] \\ \vdots & \vdots & & \vdots \\ u[k+2i-1] & u[k+2i] & \dots & u[k+2i+j-2] \\ y[k+2i-1] & y[k+2i] & \dots & y[k+2i+j-2] \end{bmatrix}.$$

Note that $j \gg \max(pi, mi)$ where p is the number of inputs and m is the number of outputs. It can be shown (ref [101]) that any basis for the intersection of the row space of these two Hankel matrices constitutes a sequence of state vectors for the system. A simple way of calculating a basis for the state vectors is as follows. First, take the singular value decomposition (SVD) of the concatenation of H_1 and H_2,

$$\begin{aligned} \begin{bmatrix} H_1 \\ H_2 \end{bmatrix} &= USV^T \\ &= \begin{bmatrix} U_{11} & U_{12} \\ U_{21} & U_{22} \end{bmatrix} \begin{bmatrix} S_{11} & 0 \\ 0 & 0 \end{bmatrix} V^T \end{aligned}$$

where the dimensions are

$$\begin{aligned}
\dim(U_{11}) &= i(p+m) \times (2mi+n) \\
\dim(U_{12}) &= i(p+m) \times (2pi-n) \\
\dim(U_{21}) &= i(p+m) \times (2mi+n) \\
\dim(U_{22}) &= i(p+m) \times (2pi-n) \\
\dim(S_{11}) &= (2mi+n) \times (2mi+n).
\end{aligned}$$

Then take a SVD of the following product

$$U_{12}^T U_{11} S_{11} = \begin{bmatrix} U_q & U_q^{\perp} \end{bmatrix} \begin{bmatrix} S_q & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} V_q^T \\ V_q^{\perp T} \end{bmatrix}.$$

Thus a valid state vector sequence, say X, is given by

$$X = U_q^T U_{12}^T H_1.$$

Now write the identification problem as

$$\begin{bmatrix} U_q^T U_{12}^T U(m+p+1 : (i+1)(m+p), :)S \\ U(mi+pi+m+1 : (i+1)(m+p), :)S \end{bmatrix} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} U_q^T U_{12}^T U(1 : i(m+p), :)S \\ U(mi+pi+1 : mi+pi+m, :)S \end{bmatrix},$$

where "j:k" specifies all the rows or columns in the range $\{j, j+1, \ldots, k\}$ (rows or columns are specified by context of the comma), and ":" means all rows or columns in the matrix. The notation is taken from Matlab[1]. This problem is typically not consistent and thus it is usually solved by posing it as a least

[1]Matlab is a registered trademark of The Mathworks, Inc.

squares problem. In this case, the problem is $\min_Y \|FY - G\|$ with

$$\begin{aligned}
F &= \begin{bmatrix} U_q^T U_{12}^T U(1 : i(m+p), :)S \\ U(mi + pi + 1 : mi + pi + m, :)S \end{bmatrix}^T \\
G &= \begin{bmatrix} U_q^T U_{12}^T U(m + p + 1 : (i+1)(m+p), :)S \\ U(mi + pi + m + 1 : (i+1)(m+p), :)S \end{bmatrix}^T \\
Y &= \begin{bmatrix} A & B \\ C & D \end{bmatrix}^T
\end{aligned}$$

thus yielding a solution of $Y^T = G^T F^{\dagger T}$ or in our original problem

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} = \begin{bmatrix} U_q^T U_{12}^T U(m + p + 1 : (i+1)(m+p), :)S \\ U(mi + pi + m + 1 : (i+1)(m+p), :)S \end{bmatrix}
\begin{bmatrix} U_q^T U_{12}^T U(1 : i(m+p), :)S \\ U(mi + pi + 1 : mi + pi + m, :)S \end{bmatrix}^{\dagger}.$$

This works very well for small state dimensions. Note that if the original measurements were subject to noise, then there would be errors in both the matrices, F and G, that are being worked with. Moreover, Hankel matrices are notoriously ill-conditioned, so that if the original measurements had small noise in them, the noise could be greatly amplified particularly for larger state dimension. The regularization methods that have been discussed can be used to account for this possibility by recasting the problem either with favorable perturbations (degenerate min min problem) or with minimum backward error (for numerical considerations). Total least squares will also be considered as a comparative case. One other technique is not shown, is the LMI structured robust least squares technique of [61, 59, 62, 63]. The prime advantage of the LMI technique in this case is its ability to account for the Hankel structure. The LMI technique is not included as the problem required large amounts of memory to run and the only machine available at the time did not have LMI solvers on it. Given the structure, it is reasonable to assume that the LMI technique would do well, and should be considered for future problems as memory is becoming less and less expensive.

8.1 Problem Setup

Subspace identification is typically done on small state dimension problems, say around six to ten states. This is partially due to the fact that it requires a great deal of memory to perform the two SVD's required to set up the regression problem. A second reason has been alluded to already, that being numerical difficulties inherent in the regression problem. To demonstrate this, a 65 state subspace identification problem was set up, as this was the largest problem that could run on a computer that was accessible at the time (300 MB RAM required). The system to be identified was chosen to be (for n=65)

$$A = \begin{bmatrix} \frac{n}{3n} & 1 & 0 & \dots & 0 \\ 0 & \frac{n-1}{3n} & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \frac{2}{3n} & 1 \\ 0 & \dots & \dots & 0 & \frac{1}{3n} \end{bmatrix}$$

This state matrix was chosen because it is well known to have numerical problems, and thus it provided an excellent test case for the problem at hand. A pseudo-random binary sequence (PRBS) was used as input to the system to be identified, to guarantee persistency of excitation and since PRBS signals are known to be good for identification. The condition numbers of the Hankel matrices were consistently on the order of 10^{16}, amply demonstrating the ill-conditioned nature

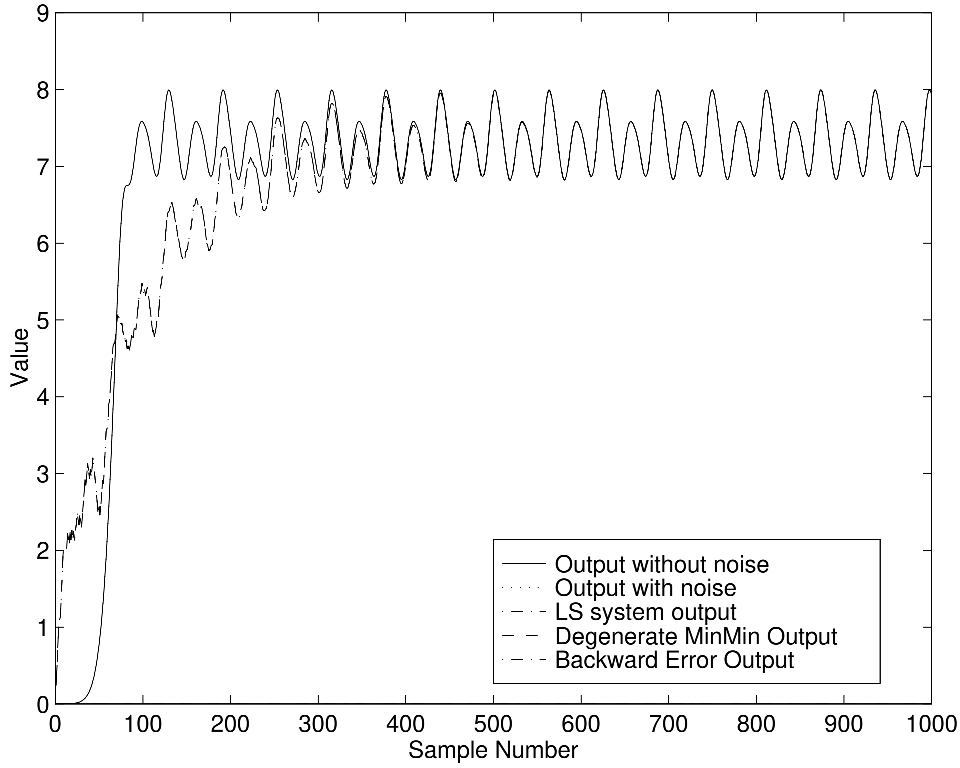

Figure 8.1. Solution Tracking for System ID Problem. The LS output is zero because $C = 0$. The remaining methods performed too similarly to distinguish on the graph.

of the problem. The noise was set at six orders of magnitude less than the signal, and thus cannot be seen when plotted out on a graph so no graph is included. It is easy to see that any problems in calculating a solution must be on numerical and not systems grounds, due to the tremendous signal to noise ratio in the problem. Several simulation runs were calculated, and the five methods (LS, TLS, TR, DMM, BE) were each used to fit a system to the output. The results are graphed in Figures 8.1, 8.2, 8.3, and 8.4. Note that in all cases the output and the output with noise are too close to distinguish, but are both included to show that noise is not a factor in the problem.

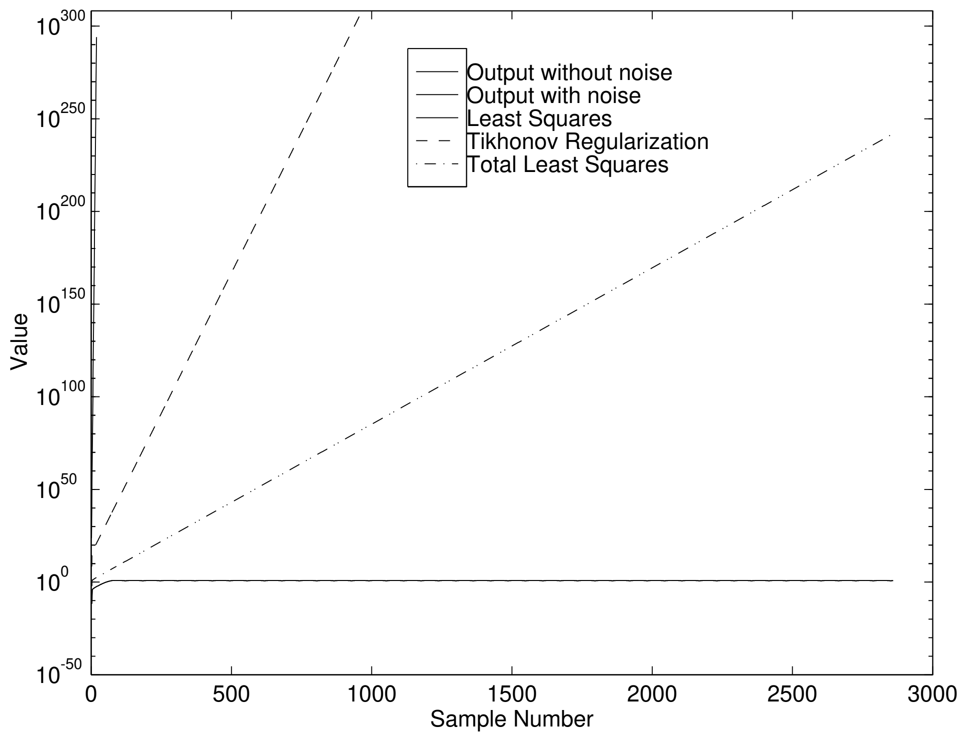

Figure 8.2. Solution Tracking for System ID Problem Showing Unstable LS, TLS, and TR Solutions on Log Scale. The almost flat line is the system output. The LS solution is almost on top of the y-axis, but can be distinguished.

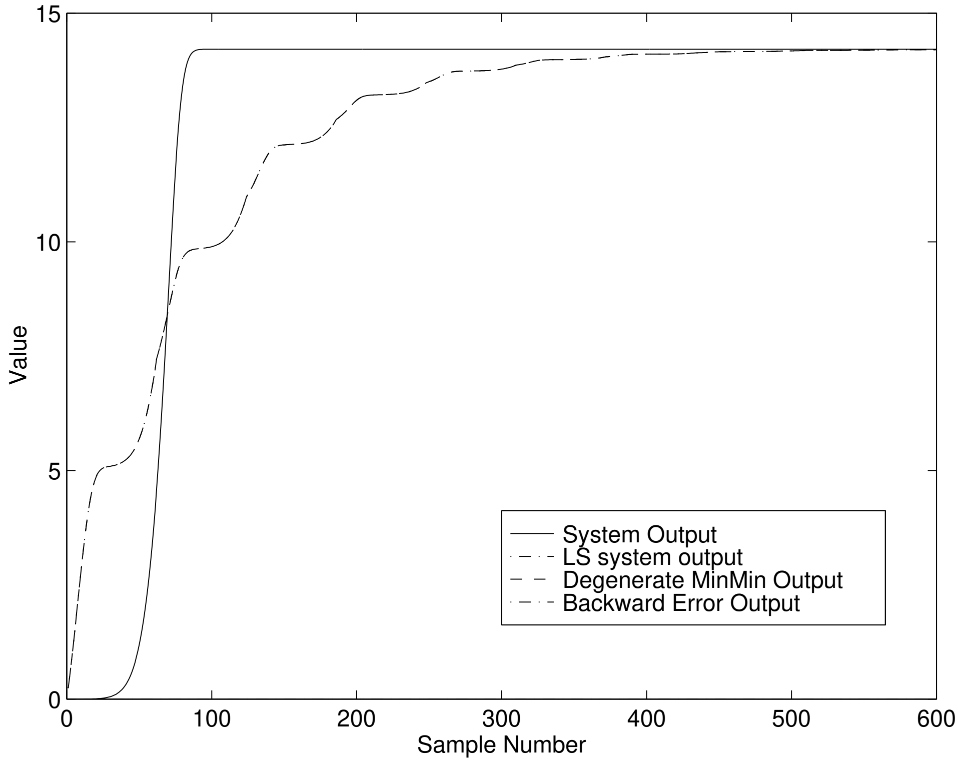

Figure 8.3. Step Response for System ID Problem Solutions. The LS output is zero because $C = 0$. The remaining methods performed too similarly to distinguish on the graph.

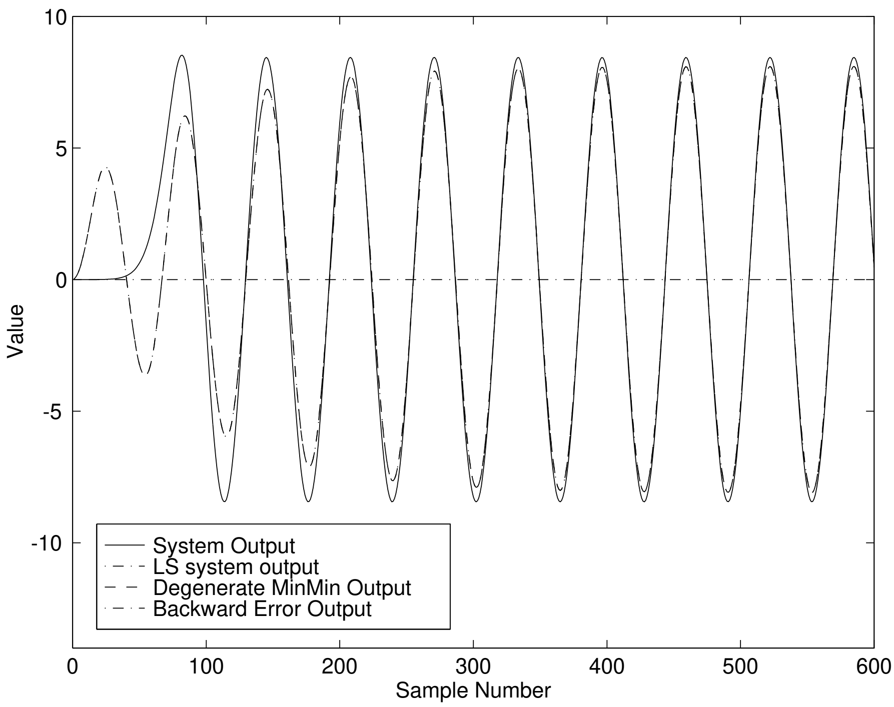

Figure 8.4. Sin Response for System ID Problem Solutions. The LS output is zero because $C = 0$. The remaining methods performed too similarly to distinguish on the graph.

It can be readily seen that both the proposed regression methods, degenerate min min (DMM) and backward error (BE) yielded good results, while the current methods, least squares (LS), total least squares (TLS), and Tikhonov (TR) do not. The LS solution always selected an unstable system and additionally it frequently found the state observation matrix, C to be the null matrix, which is why its output is zero on most of the graphs. Note that in Figure 8.2 the output of a LS solution that did not have the null matrix for C was graphed. The TLS solution generated several unstable poles in every run, which is why the graph is unbounded in Figure 8.2. It is possible to try to clean up the unstable poles by projecting the unstable poles into the stable region to see how a stabilized system would do, but this is against the basic goals of this analysis, which is to see how each method works on its own. A major overall goal of this dissertation is to find methods which provide good results without resorting to special, ad-hoc measures to clean up bad results. The interesting one is the Tikhonov solution, which in theory should be more robust. The Tikhonov solution does not generate good results because the standard choice of η^2 is actually almost machine precision and thus has almost no effect on the results. A larger value could obviously do better, as the techniques in this dissertation could be considered as special cases of Tikhonov, but again this is an ad-hoc measure.

A key interest in system ID is to get a model that behaves like the actual system and Figure 8.1 shows how well the DMM and BE track the output used in the system ID problem. For comparison purposes, compare how well the two methods do at tracking the step response of the actual system, Figure 8.3, and the actual system's response to a sine wave, Figure 8.4. Note that the two methods produce essentially identical results even though the regression parameters are not the same. This is because while the parameters are different (the BE parameter is roughly four times the size of the DMM parameter) the parameters are both very small. For all ten runs, the parameters were between zero and 0.001. The parameters were small enough not to really effect the solution accuracy much, but still large enough to keep the numerical errors from destroying the solution.

Chapter 9

Image Processing Example

The second example is a simple two-dimensional image processing application. A small picture with the grey-scale words, 'HELLO WORLD' of early programming fame, has been blurred. The image is 20x35 and the blur is done by multiplying each column of the image by a Gaussian blur matrix of size 20. The component of a Gaussian blur, G, with standard deviation, σ, in position, (i,j), is given by

$$G_{i,j} = e^{-\left(\frac{i-j}{\sigma}\right)^2}.$$

The true blur is not known exactly but is corrupted by noise. The blur is not so strong that some of the features cannot be seen, and in particular one can see that there is writing but the specifics are hard to make out. See the first two images in Figure 9.1. The image does not look that bad and the actual matrix is known to within 3%, meaning the ratio of the norms of the perturbation to the matrix is under 3%. The perturbation is small and the matrix is small, two key aspects of an 'easy to solve' problem. The condition is on the order of 1000, which is large but not unreasonable. The two most popular techniques, least squares (LS) and

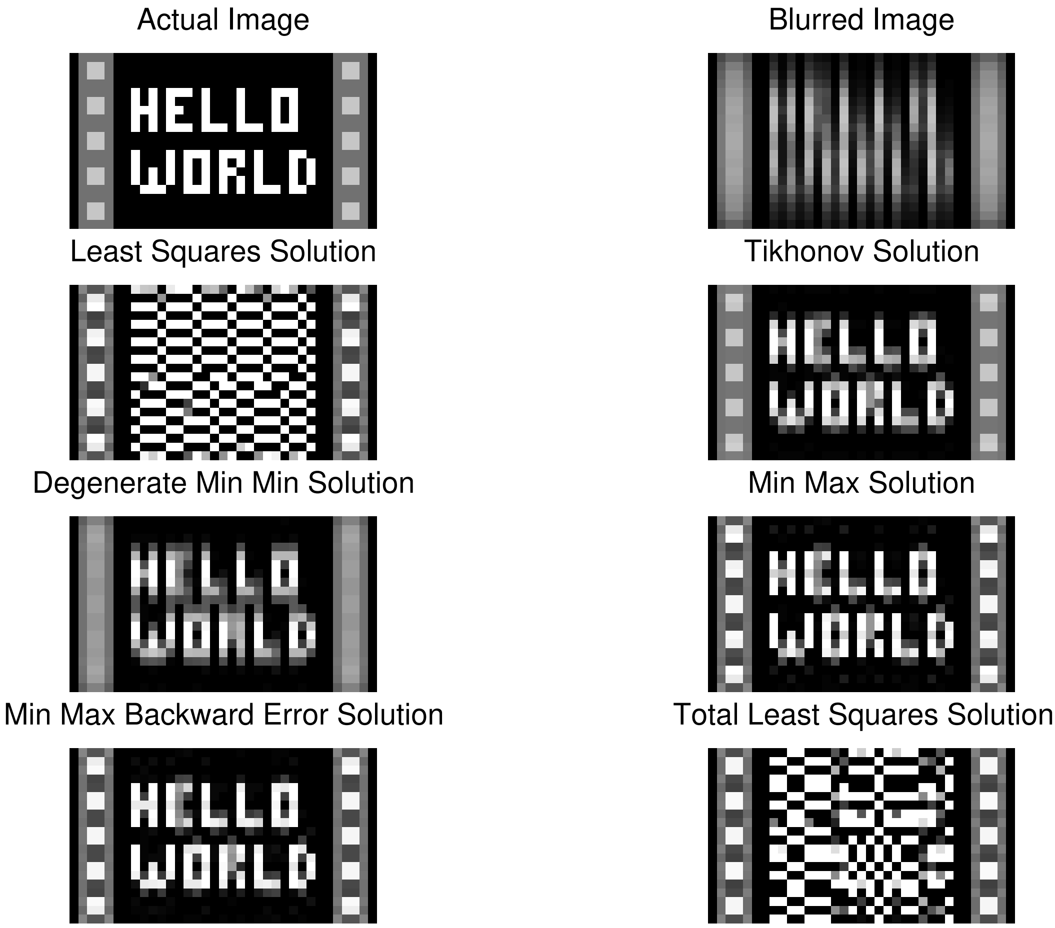

Figure 9.1. Hello World Problem

total least squares (TLS), both fail badly in recovering the text. It is interesting to note that both recover the border acceptably though. It is important to keep in mind that just because a technique fails in one area it does not necessarily fail everywhere.

All of the remaining techniques do a good enough job on the text to allow it to be figured out. The worst is the degenerate min min but interestingly it does the best job on recovering the middle prong on the 'E' in hello. The remaining techniques do arguably similar jobs. The backward error (BE) and min max (or BDU - bounded data uncertainty) estimators do a little better job on getting edge distinction, but because of this there is a slight degradation in the quality of the border which is not as sharp as these techniques would like. The Tikhonov regulator does the best job on the border, and a fairly good job on the text. Note that some fading is apparent on the end and corners of the letters but all in all the quality is good.

A key aspect of all of these regression techniques is selection of the regression parameter. In the suggested techniques, this is done semi-automatically. Semi-automatically because the error bound on the matrix must still be supplied, and it is not guaranteed to be known accurately. This becomes critical as the selection of the regression parameter is mostly influenced by this error bound. Select an error that is too large and data losses can result, select one too small and there will not be enough regulation or deregulation to improve the regression. The 'HELLO WORLD' picture will be used as an example of selecting an error bound and as a basis for comparing how the methods look next to each other.

First consider setting the error for all methods to be the 2-norm of the perturbation matrix, i.e.: assume $A = A_{True} + E$ and select $\eta = \|E\|_2$. The results can be seen in Figure 9.2. Note that this is the error bound used in the original

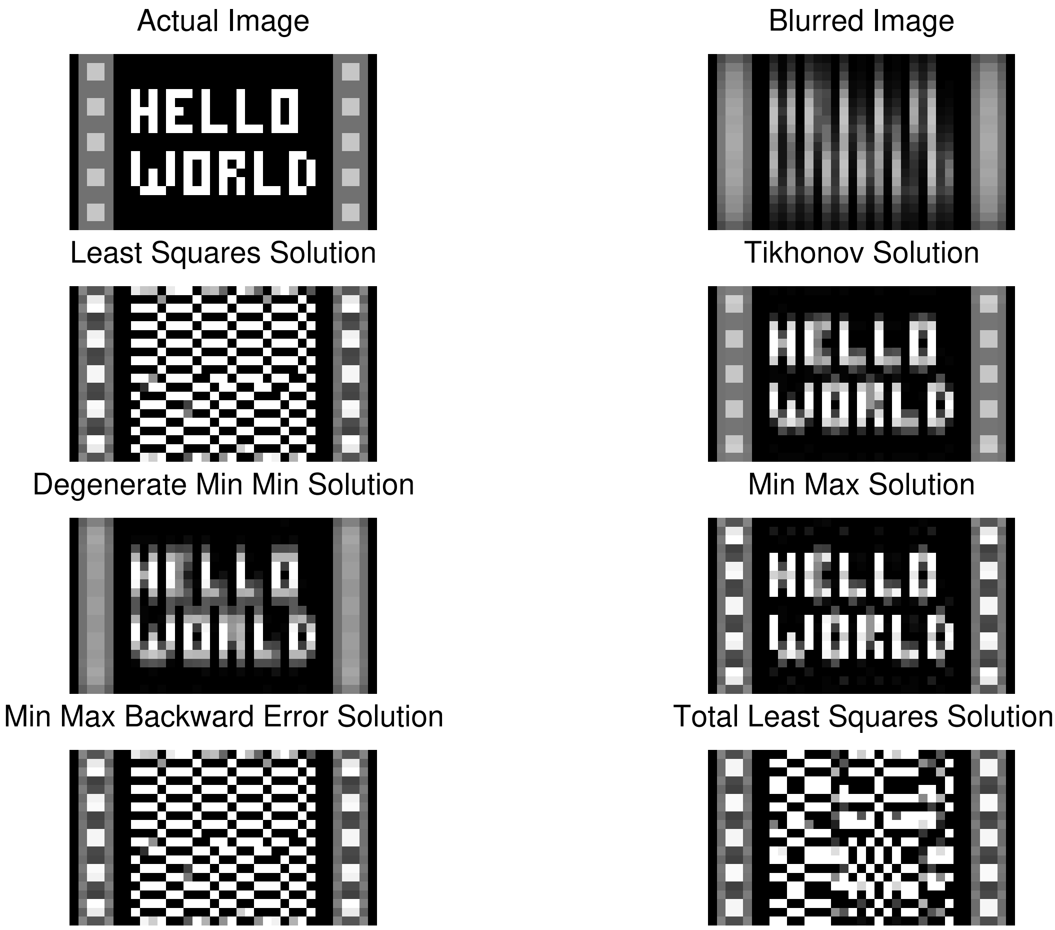

Figure 9.2. Hello World with $\eta = \|E_A\|_2$

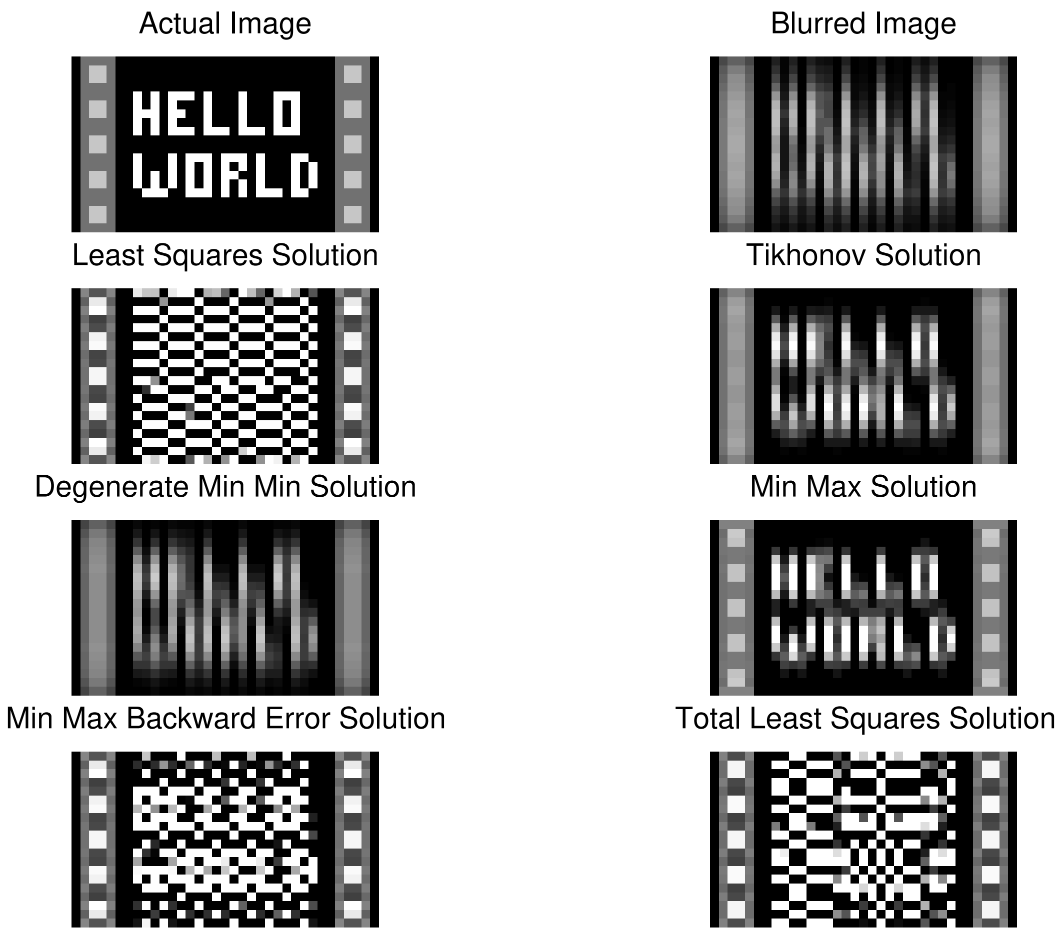

Figure 9.3. Hello World with $\eta = \|E\|_F$

method with the exception of the BE estimator, which is too optimistic and fails similar to the LS and TLS techniques. This underscores the comment about the techniques being semi-automatic, as in this case the technique which usually performs best does not with a bad choice of error bound.

Next consider the error bound that gives the best indicator of the effects of the unknown matrix perturbation, the Frobenius norm. Since the Frobenius norm takes into account all of the energy in the perturbation (all singular values), it gives a good feel for what a perturbation matrix can do. The results can be seen in Figure 9.3. Notice that the BE technique does well, but the others, which

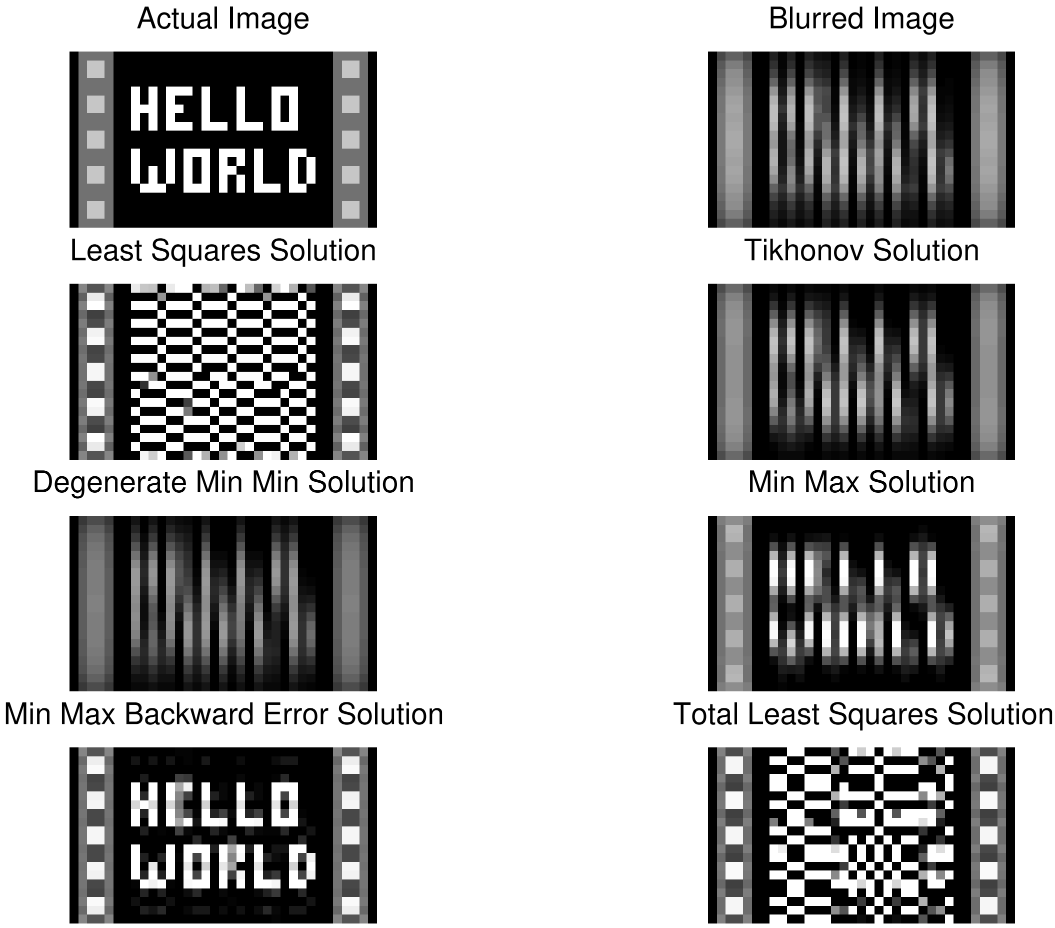

Figure 9.4. Hello World with $\eta = 2\|E\|_F$

are more conservative have varying degrees of success. It is interesting to note that the BDU (min max) technique does surprisingly well given that it is more conservative by design than the degenerate min min technique. The degenerate min min technique can become conservative under the right circumstances and that is what is seen here.

It is reasonable to consider error bounds in the range of

$$\left\{\frac{\|E\|_2}{n}, n\|E\|_2\right\}.$$

The range was picked as a slightly larger interval than that which is necessary

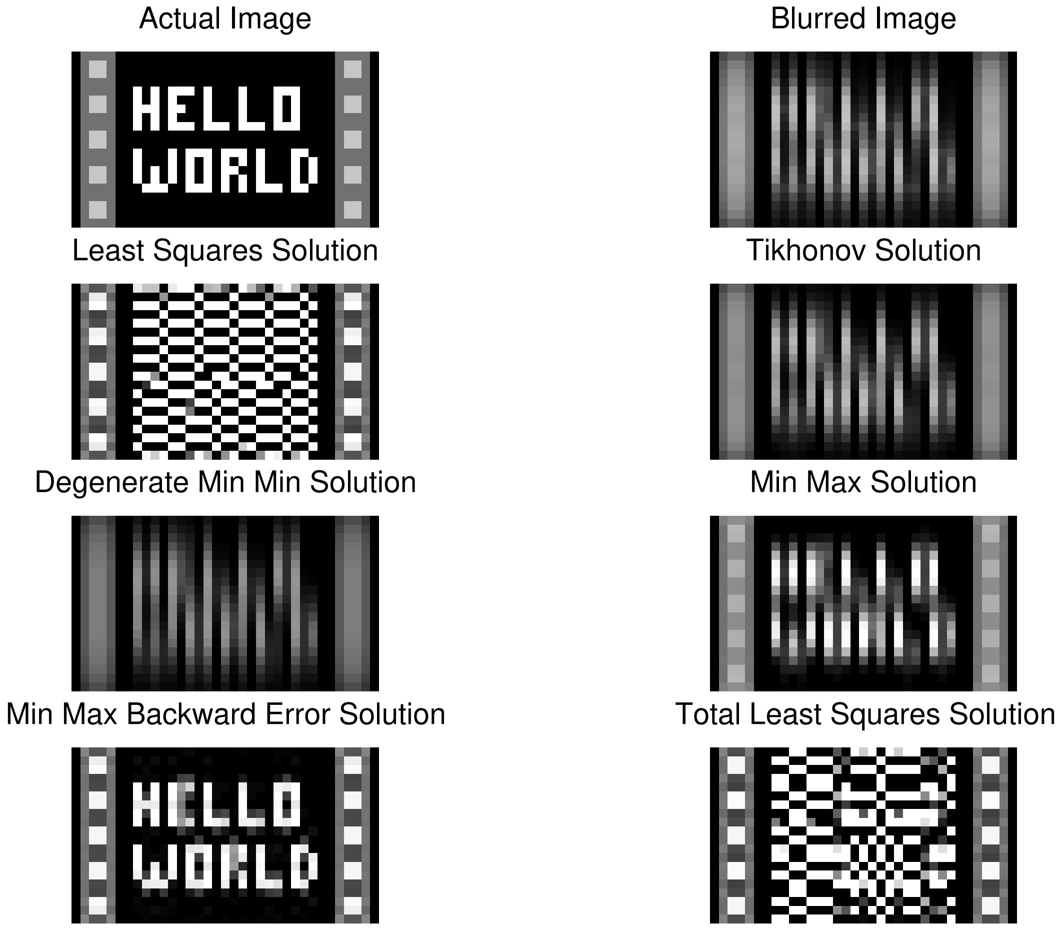

Figure 9.5. Hello World with $\eta = \frac{n}{2}\|E\|_2$

to contain all the norms of the perturbation matrix. The range was increased to allow for uncertainty in the bound. In this case, the upper level was too conservative. A slightly larger scaling of the Frobenius norm, namely twice it, was picked for the next point to evaluate. It can be seen that the BE technique continues to improve while the others continue to get worse. By this point only the BDU still produces output from which letters can be recognized.

As a final point, half the upper limit was chosen. Improvement can still be seen in the BE technique, while the others are so conservative that they are useless. It takes until nearly the upper limit until the BE technique gives as bad a result as the others generate by this half the upper limit. The BE technique ends up giving reasonable answers for a longer region, but again it must be noted that fine tuning of the perturbation error bound is very helpful even for self-tuning systems.

Chapter 10

Conclusion

All models contain uncertainty, and thus it should always be considered when working with estimation and identification problems. When conditioning is good and uncertainty is small, the standard techniques perform well and it is reasonable to examine techniques which can improve the solution, such as total least squares or min min solvers. When a system becomes ill-conditioned or the uncertainty is large a robust solver is needed, such as Tikhonov techniques, min max techniques, constrained problems, or Ridge Regression. Contemporary techniques were discussed in Chapter 2.

Five problems were formulated in this dissertation.

1. Three problems were unstructured, partitioned min max problems. The unstructured, block column partitioned min max problem was solved in Chapter 3. The unstructured, block row partitioned min max problem was solved in Chapter 4. The unstructured, general block partitioned min max problem was solved in Chapter 5. In each case the form of solution and the secular equation was provided to find the solution when it is at a

differentiable point. Non-differentiable points were discussed as were other techniques to reach the solution.

2. One problem was the degenerate case of the min min problem. This solution was totally solved and characterized. An algorithm for solving the problem was presented, and a first step to the partitioned case was presented.

3. The final problem was the min max Backward Error problem. This was broken into four sub-problems, three of which were completely solved and one of which has a solution but would benefit from additional characterization.

Two examples were covered to examine the usefulness of the formulations presented. One example was from system identification and one example from image processing. The performance advantages of the methods presented in this dissertation were discussed.

10.1 Future Directions

While much has been done, several areas deserve further attention. The areas for further investigation are:

1. Prove which zero of the backward error secular equation is the optimal one, and develop an algorithm for finding it, for the remaining sub-problem.

2. Test the secular equation based algorithm for the multi-column partitioning of the min max problem against the quadratically convergent method to determine more precisely when to use each.

3. Find secular equation techniques for the structured min max problem and compare it to the LMI technique.

4. Find a LMI formulation for the partitioned cases and compare it to the secular equation techniques.
5. Solve the partitioned cases of the degenerate and non-degenerate min min problems.

Appendix A

Existence and Form of Slope

Lemma A.1 (Existence and Form of the Slope) *Let $b \notin R\{A\}$. The following limits exist for $i \in \{1, 2, \ldots, p\}$:*

$$\lim_{x_i \to 0^+} \frac{\partial J(x_1, x_2, \ldots, x_p)}{\partial x_i} = \frac{A_i^T \left(A_{/i} x_{/i} - b\right)}{\left\|A_{/i} x_{/i} - b\right\|} + \eta_i \tag{A.1}$$

and

$$\lim_{x_i \to 0^-} \frac{\partial J(x_1, x_2, \ldots, x_p)}{\partial x_i} = \frac{A_i^T \left(A_{/i} x_{/i} - b\right)}{\left\|A_{/i} x_{/i} - b\right\|} - \eta_i. \tag{A.2}$$

Note that $A_{/i}$ denotes the A matrix with the i^{th} column removed (similar for $x_{/i}$).

Proof:

Note first that the cost function J is defined by

$$J(x_1, x_2, \ldots, x_p) = \|Ax - b\| + \sum_{j=1}^{p} \eta_j \|x_j\|.$$

It is thus obvious that the only points at which the derivative with respect to x_i will not exist are $\{x | Ax - b = 0\}$ which cannot happen given that $b \notin R\{A\}$ and

$x_i = 0$. Thus trivially the limits exist. Now consider the cost function given that $x_i > 0$,

$$J\left(x_1, x_2, \ldots, x_p | x_i > 0\right) \;=\; \|Ax - b\| + \sum_{j \neq i} \eta_j \|x_j\| + \eta_i x_i.$$

The derivative with respect to x_i is given by

$$\begin{aligned}
\frac{\partial J\left(x_1, x_2, \ldots, x_p | x_i > 0\right)}{\partial x_i} &= \frac{\partial \|Ax - b\|}{\partial x_i} + \frac{\partial \sum_{j \neq i} \eta_j \|x_j\|}{\partial x_i} + \frac{\partial \eta_i x_i}{\partial x_i} \\
&= \frac{A_i^T \left(Ax - b\right)}{\|Ax - b\|} + \eta_i.
\end{aligned}$$

Now take the limit as $x_i \to 0^+$,

$$\begin{aligned}
\lim_{x_i \to 0^+} \frac{\partial J\left(x_1, x_2, \ldots, x_p | x_i > 0\right)}{\partial x_i} &= \lim_{x_i \to 0^+} \left(\frac{A_i^T \left(Ax - b\right)}{\|Ax - b\|} + \eta_i \right) \\
&= \lim_{x_i \to 0^+} \left(\frac{A_i^T \left(Ax - b\right)}{\|Ax - b\|} \right) + \eta_i \\
&= \lim_{x_i \to 0^+} \left(\frac{A_i^T \left(A_{/i} x_{/i} - b\right)}{\left\|A_{/i} x_{/i} - b + A_i x_i\right\|} \right) \\
&\quad + \lim_{x_i \to 0^+} \left(\frac{A_i^T A_i x_i}{\left\|A_{/i} x_{/i} - b + A_i x_i\right\|} \right) + \eta_i \\
&= \frac{A_i^T \left(A_{/i} x_{/i} - b\right)}{\left\|A_{/i} x_{/i} - b\right\|} + \eta_i.
\end{aligned}$$

The last line is easily seen by noting that

$$\begin{aligned}
\lim_{x_i \to 0^+} x_i &= 0 \\
\lim_{x_i \to 0^+} \left\|A_{/i} x_{/i} - b + A_i x_i\right\| &= \left\|A_{/i} x_{/i} - b\right\|,
\end{aligned}$$

and then referencing a book on Analysis, with regard to quotients of limits, such as [122] page 49, Theorem 3.3. Now consider the cost function given that $x_i < 0$

$$J\left(x_1, x_2, \ldots, x_p | x_i > 0\right) \;=\; \|Ax - b\| + \sum_{j \neq i} \eta_j \|x_j\| - \eta_i x_i.$$

The derivative with respect to x_i is given by

$$\begin{aligned}\frac{\partial J\left(x_1, x_2, \ldots, x_p | x_i > 0\right)}{\partial x_i} &= \frac{\partial \left\|Ax - b\right\|}{\partial x_i} + \frac{\partial \sum_{j \neq i} \eta_j \left\|x_j\right\|}{\partial x_i} - \frac{\partial \eta_i x_i}{\partial x_i} \\ &= \frac{A_i^T \left(Ax - b\right)}{\left\|Ax - b\right\|} - \eta_i.\end{aligned}$$

Now take the limit as $x_i \to 0^-$,

$$\begin{aligned}\lim_{x_i \to 0^-} \frac{\partial J\left(x_1, x_2, \ldots, x_p | x_i < 0\right)}{\partial x_i} &= \lim_{x_i \to 0^-} \left(\frac{A_i^T \left(Ax - b\right)}{\left\|Ax - b\right\|} - \eta_i\right) \\ &= \lim_{x_i \to 0^-} \left(\frac{A_i^T \left(Ax - b\right)}{\left\|Ax - b\right\|}\right) - \eta_i \\ &= \lim_{x_i \to 0^-} \left(\frac{A_i^T \left(A_{/i} x_{/i} - b\right)}{\left\|A_{/i} x_{/i} - b + A_i x_i\right\|}\right) \\ &\qquad + \lim_{x_i \to 0^-} \left(\frac{A_i^T A_i x_i}{\left\|A_{/i} x_{/i} - b + A_i x_i\right\|}\right) - \eta_i \\ &= \frac{A_i^T \left(A_{/i} x_{/i} - b\right)}{\left\|A_{/i} x_{/i} - b\right\|} - \eta_i.\end{aligned}$$

The last line is easily seen by noting that

$$\begin{aligned}\lim_{x_i \to 0^-} x_i &= 0 \\ \lim_{x_i \to 0^-} \left\|A_{/i} x_{/i} - b + A_i x_i\right\| &= \left\|A_{/i} x_{/i} - b\right\|\end{aligned}$$

and again referencing a book on Analysis with regard to quotients of limits, such as has been noted can be found in [122] page 49, Theorem 3.3. Thus the lemma is proved.

♢ SDG ♢

Appendix B

Sign of x_i in Multi-Column Min Max

Lemma B.1 (Sign of x_i in Multi-Column Min Max) *Let $b \notin \mathcal{R}\{A\}$ and let $x^*_{j/i}$ denote the optimal solution for x_j given that $x_i = 0$ and $x_k = x^*_k$ for all $k \notin \{j, i\}$. Then one and only one of the following must hold.*

1.
$$\lim_{x_i \to 0^+} \frac{\partial J\left(x^*_{\{j|j \neq i\}}, x_i\right)}{\partial x_i} < 0 \tag{B.1}$$

   Thus, the solution lies in the half plane defined by $x_i > 0$.

2.
$$\lim_{x_i \to 0^-} \frac{\partial J\left(x^*_{\{j|j \neq i\}}, x_i\right)}{\partial x_i} > 0 \tag{B.2}$$

   Thus, the solution lies in the half plane defined by $x_i < 0$.

3. *The solution lies on the hyperplane defined by $x_i = 0$ and moreover is given by $x^*_{j/i}$ and $x_i = 0$.*

Proof:

Begin by noting that the cost function on the hyperplane $x_i = 0$ is at a minimum if $x_j = x^*_{j/i}$ for all $j \notin \{i\}$. This is easily observed because $x^*_{j/i}$ is defined as the solution to the reduced order problem of

$$\min_{x_j} \left(\left\| A_{/i} x_{/i} - b \right\| + \sum_{j \neq i} \eta_j \left\| x_j \right\| \right).$$

Thus by definition it is the point of minimum cost on the hyperplane $x_i = 0$. If the cost is smaller for the general problem, say at some $x_i > 0$, than the minimum cost on the hyperplane $x_i = 0$, then the minimum for the general problem must be on the same side of the hyperplane (in this case $x_i > 0$.). To prove this, assume that the minimum for the general problem is on the other side of the hyperplane $x_i = 0$. Without loss of generalization, let $x^*_i < 0$, but assume that there exists a $x^o_i > 0$ such that $J(x^o_j, x^o_i) < J(x^*_{j/i}, x_i = 0)$. Let x^*_j be the optimal value for x_j when $x_i = x^*_i$. Now after all that work the proof is simple. Consider the line between (x^o_j, x^o_i) and (x^*_j, x^*_i) defined by

$$\begin{aligned} &\{(x_j(\alpha), x_i(\alpha))| \\ &\quad x_j(\alpha) = \left(\alpha x^o_j + (1-\alpha)\, x^*_j\right), x_i(\alpha) = \left(\alpha x^o_i + (1-\alpha)\, x^*_i\right), 0 \leq \alpha \leq 1\big\}. \end{aligned}$$

Since the cost function is convex, for all α between zero and one

$$\begin{aligned} J\left(x_j(\alpha), x_i(\alpha)\right) &\leq \alpha J\left(x^o_j, x^o_i\right) + (1-\alpha) J\left(x^*_j, x^*_i\right) \\ &< \alpha J\left(x^o_j, x^o_i\right) + (1-\alpha) J\left(x^o_j, x^o_i\right) \\ &= J\left(x^o_j, x^o_i\right) \\ &< J\left(x^*_{j/i}, x_i = 0\right) \\ &\leq J\left(x^*_j\left(\frac{-x^*_i}{x^o_i - x^*_i}\right), x_i\left(\frac{-x^*_i}{x^o_i - x^*_i}\right) = 0\right). \end{aligned}$$

Note that the last line is true since $\alpha = \frac{-x_i^*}{(x_i^o - x_i^*)}$ is the value of α for which $x_i(\alpha) = 0$, and thus this is the point on the line at which the line intersects the hyperplane $x_i = 0$. Since the optimal cost on the hyperplane is given by $J\left(x_{j/i}^*, x_i = 0\right)$, the cost at this point must be greater than or equal to the minimum. Note however, that the last line creates a contradiction, for it says that the cost function is not convex. Thus the minimum cost of the general problem must be on the same side as any point found to be less than the minimum cost on the hyperplane $x_i = 0$. Now to use this fact note that if

$$\lim_{x_i \to 0^+} \frac{\partial J\left(x_{\{j|j \neq i\}}^*, x_i\right)}{\partial x_i} < 0.$$

Then, there exists a point $(x_{j/i}^*, x_i > 0)$ such that $J(x_{j/i}^*, x_i > 0) < J(x_{j/i}^*, x_i = 0)$. Thus from the fact shown above, the optimal x_i must also be greater than zero. This is simply case (1). Similarly, if

$$\lim_{x_i \to 0^-} \frac{\partial J\left(x_{\{j|j \neq i\}}^*, x_i\right)}{\partial x_i} > 0.$$

Then, there exists a point $(x_{j/i}^*, x_i < 0)$ such that $J(x_{j/i}^*, x_i < 0) < J(x_{j/i}^*, x_i = 0)$. Thus from the fact shown above, the optimal x_i must also be less than zero. This is simply case (2), and moreover note that both conditions on the limits in case (1) and case (2) cannot hold at the same time, since it would violate convexity by having a saddle point at $(x_{j/i}^*, x_i = 0)$. This simply points out that case (1) and case (2) are mutually exclusive.

Note that trivially, if the minimum cost for the general problem does not happen when $x_i > 0$ or $x_i < 0$ then trivially it must happen on the hyperplane $x_i = 0$, and the minimum cost on that hyperplane is then the solution to the general problem. Clearly this indicates that if the solution is on the hyperplane it is located at $(x_{j/i}^*, x_i = 0)$. This is simply case (3), which has been noted cannot happen if either case (1) or case (2)happens.

Thus either case (1), case (2), or case (3) can happen, and the cases are mutually exclusive. It still must shown that if the optimal $x_i > 0$ then the limit in case (1) must hold. Analogously it must be shown that if the optimal $x_i < 0$ then the limit in case (2) must hold.

Let the optimal $x_i > 0$ but assume that the condition on the derivative in case (1) does not hold, thus

$$\lim_{x_i \to 0^+} \frac{\partial J\left(x^*_{\{j|j\neq i\}}, x_i\right)}{\partial x_i} \geq 0.$$

Note that since the minimum cost on the hyperplane is given by the point $(x^*_{j/i}, x_i = 0)$, the slope is non-negative as $x^*_{j/i}$ increases (or at least the limit of the slope from above). Also the slope is non-positive as $x^*_{j/i}$ decreases (or at least the limit of the slope from below). Denote the point $(x^*_{j/i}, x_i = 0)$ by $\bar{x}^*$.

Since the global minimum is not in the hyperplane $x_i = 0$, then the cost function must decrease in some direction off of the hyperplane, which will be denoted as u. From Analysis, for example see [122, pps. 215 – 219], that the directional derivative can be expressed as

$$\begin{aligned}
\lim_{t \to 0} \frac{J(\bar{x}^* + tu) - J(\bar{x}^* + tu)}{t} &= (\nabla J)(\bar{x}^*)\,\dot{u} \\
&= (D_u J)(\bar{x}^*) \\
&= \sum_{k=1}^{p} (D_{x_k} J)(\bar{x}^*)\, u_k.
\end{aligned}$$

Extend this to consider the case of left and right derivatives by taking the limit from one side only. For example the right derivative is

$$\lim_{t \to 0^+} \frac{J(\bar{x}^* + tu) - J(\bar{x}^* + tu)}{t} = \sum_{k=1}^{p} \left(\lim_{x_k \to \bar{x}^*_k} \frac{\partial J(\bar{x}^*)}{\partial x_k} \right) |u_k|.$$

Note that as long as there are no discontinuities in some neighborhood of $\bar{x}^*$ to one side of the hyperplane $x_i = 0$, then for the cost function to decrease off the

the point $\bar{x}^*$, case (1) must hold. Since this is against the assumption, there must be a discontinuity in every neighborhood. This can only happen if one of the components $x_{k\neq i} = 0$. From what has already been shown the minimum on the hyperplane $x_{k\neq i} = 0$ must be on the hyperplane $x_i = 0$ or on the same side of the hyperplane $x_i = 0$ as the global minimum, so that convexity is not violated. Since the assumption does not allow for the slope of the cost function to decrease in the hyperplane $x_{k\neq i} = 0$ as the hyperplane $x_i = 0$ is moved off, the minimum must occur on the two (or more) hyperplanes.

Now consider the quadrant where the solution lies. Without loss of generality assume that the quadrant in which the solution lies is such that $x_j > 0$ for all j. The cost function in this quadrant is then identical to the function

$$\|Ax - b\| + \sum_{j=1}^{p} \eta_j x_j. \tag{B.3}$$

Note that since by assumption $b \notin \mathcal{R}(A)$ this new cost function is $\mathcal{C}^\infty$. From the analysis results above, trivially

$$\frac{\|Ax - b\| + \sum_{j=1}^{p} \eta_j x_j}{\partial x_{i\neq k}} < 0. \tag{B.4}$$

Further note that

$$\frac{\partial J(x_{j\neq k} > 0, x_k > 0)}{\partial x_{i\neq k}} = \frac{\partial J(x_{j\neq k} > 0, x_k < 0)}{\partial x_{i\neq k}}. \tag{B.5}$$

Thus

$$\frac{\partial J(x_{j\neq k} > 0, x_k = 0)}{\partial x_{i\neq k}} = \frac{\partial \|Ax - b\| + \sum_{j=1}^{p} \eta_j x_j}{\partial x_{i\neq k}}. \tag{B.6}$$

Which then contradicts the assumption that the limit in case (1) does not hold. The argument follows directly for case (2) and the proof is complete.

◇ SDG ◇

Appendix C

Bounds on $\|x\|$ in Multiple Row Min Max

This appendix develops bounds on the size of $\|x\|$ for the multiple (block) row partitioned min max formulation. The cost function is

$$C = \sum_{i=1}^{q} \left(\|A_i x - b_i\| + \eta_i \|x\| + \eta_{b,i}\right)^2 .$$

Use the fact that at the solution $\|A_i x - b_i\| = \frac{\eta_i}{\zeta_i}\|x\|$ to see that

$$\begin{aligned} C &= \sum_{i=1}^{q} \left(\frac{\eta_i}{\zeta_i}\|x\| + \eta_i\|x\| + \eta_{b,i}\right)^2 \\ &= \sum_{i=1}^{q} \left(\eta_i \left(1 + \frac{1}{\zeta_i}\right)\|x\| + \eta_{b,i}\right)^2 \\ &= \sum_{i=1}^{q} \left(\eta_i^2 \left(1 + \frac{1}{\zeta_i}\right)^2 \|x\|^2 + 2\eta_{b,i}\eta_i \left(1 + \frac{1}{\zeta_i}\right)\|x\| + \eta_{b,i}^2\right). \end{aligned}$$

Now note that at $x = 0$ the cost is $\sum_{i=1}^{q}(\|b_i\| + \eta_{b,i})^2$, so the optimal cost must be less than or equal to this, so

$$\sum_{i=1}^{q}(\|b_i\| + \eta_{b,i})^2 \geq \sum_{i=1}^{q} \left(\eta_i^2 \left(1 + \frac{1}{\zeta_i}\right)^2 \|x\|^2 + 2\eta_{b,i}\eta_i \left(1 + \frac{1}{\zeta_i}\right)\|x\| + \eta_{b,i}^2\right).$$

Define

$$\begin{aligned}
a_2 &= \sum_{i=1}^{q} \eta_i^2 \left(1 + \frac{1}{\zeta_i}\right)^2 \\
a_1 &= \sum_{i=1}^{q} 2\eta_{b,i}\eta_i \left(1 + \frac{1}{\zeta_i}\right) \\
a_0 &= \sum_{i=1}^{q} \left(\|b_i\|^2 + 2\eta_{b,i}\|b_i\|\right).
\end{aligned}$$

Which simplifies the cost requirement to $0 \geq a_2\|x\|^2 + a_1\|x\| - a_0$. Using the quadratic formula and noting that since $a_1 \geq 0$ the parabola is concave up and the value of $\|x\|$ must be between the two roots. Additionally, $\|x\| \geq 0$, which replaces the negative root with 0, thus

$$0 \leq \|x\| \leq \frac{\sqrt{a_1^2 + 4a_2a_0} - a_1}{2a_2}.$$

To find the lower bound note that by dropping the $\eta_i\|x\|$ terms and noting that for positive terms the sum of the squares is less than the square of the sum and then using this to minimize, yields

$$\begin{aligned}
C &\geq \sum_{i=1}^{q}(\|A_i x - b_i\| + \eta_{b,i})^2 \\
&> \|Ax - b\|^2 + \sum_{i=1}^{q}(\eta_{b,i})^2 \\
&\geq \|(I - AA^{\dagger})b\|^2 + \sum_{i=1}^{q}(\eta_{b,i})^2
\end{aligned}$$

with $A^{\dagger}$ the pseudo-inverse of A. Thus

$$\begin{aligned}
&\|(I - AA^{\dagger})b\|^2 \sum_{i=1}^{q}(\eta_{b,i})^2 \\
&\quad \leq \sum_{i=1}^{q}\left(\eta_i^2\left(1 + \frac{1}{\zeta_i}\right)^2\|x\|^2 + 2\eta_{b,i}\eta_i\left(1 + \frac{1}{\zeta_i}\right)\|x\| + \eta_{b,i}^2\right).
\end{aligned}$$

A similar parabola results, namely $a_2\|x\|^2 + a_1\|x\| - \|(I - AA^{\dagger})b\|^2 \geq 0$. Again only the positive root is needed because $\|x\| \geq 0$ which gives the lower bound of

$$0 \leq \frac{\sqrt{a_1^2 + 4a_2 r_{ls}} - a_1}{2a_2} \leq \|x\|.$$

Appendix D

Piecewise Convexity of $\|x(\alpha)\|$

This appendix deals only with the degenerate min-min problem. In particular, it shows that $\|x(\alpha)\|^2$ is strictly convex in the interval $(-\sigma_{n-1}^2, -\sigma_n^2)$, which is a key step in showing that only the zero closest to $-\sigma_n^2$ can correspond to a potential candidate for the global minimum.

Start by noting

$$\|x(\alpha)\|^2 = b_1^T \Sigma_1^2 (\Sigma_1^2 + \alpha I)^{-2} b_1.$$

Differentiating once with respect to α yields

$$\frac{d}{d\alpha}\|x(\alpha)\|^2 = -2b_1^T \Sigma_1^2 (\Sigma_1^2 + \alpha I)^{-3} b_1.$$

Differentiating once more gives

$$\frac{d^2}{d\alpha^2}\|x(\alpha)\|^2 = 6b_1^T \Sigma_1^2 (\Sigma_1^2 + \alpha I)^{-4} b_1,$$

which shows that $\|x(\alpha)\|^2$ is strictly convex on the interval $(-\sigma_{n-1}^2, -\sigma_n^2)$ and hence that it has a unique minimum on that interval.

Appendix E

Rightmost Root

This appendix deals only with the degenerate min-min problem. In particular, this appendix proves that of all the roots in the interval $(-\sigma_{n-1}^2, -\sigma_n^2)$ only the right-most one can possibly correspond to the global minimum.

Let $\alpha_0, \ldots, \alpha_l$ denote the zeros of the secular equation $g(\alpha)$ in the interval $(-\sigma_{n-1}^2, -\sigma_n^2)$, in increasing order; that is

$$-\sigma_{n-1}^2 < \alpha_0 < \alpha_1 < \cdots < \alpha_l < -\sigma_n^2.$$

From the result in Section 6.10 it is known that only the roots corresponding to negative slopes of the secular equation can correspond to local minima. Since

$$\lim_{\alpha \to -\sigma_n^2 -} g(\alpha) = -\infty,$$

it follows that

$$g'(\alpha_l) < 0 \quad \text{and} \quad g'(\alpha_{l-1}) > 0.$$

(The degenerate multiple root cases are ignored for now as the argument can be extended to them by continuity.)

Now there are two possibilities. Either $\|x(\alpha_l)\| \leq \|x(\alpha_{l-1})\|$ or not. The first case implies that $\|x(\alpha_{i+1})\| < \|x(\alpha_i)\|$ due to the convexity of $\|x(\alpha)\|$ on $(-\sigma_{n-1}^2, -\sigma_n^2)$.

For the second case, $\|x(\alpha_{l-1})\| < \|x(\alpha_l)\|$. It must be shown that this implies $\|x(\alpha_{l-1})\| < \|x(\alpha)\|$ for $-\sigma_{n-1}^2 < \alpha < -\alpha_{l-1}$, and that this is not the global minimum. Toward this end take the derivative of $x(\alpha)$ with respect to α and get

$$\frac{dx(\alpha)}{d\alpha} = -\left(A^TA + \alpha I\right)^{-1} x(\alpha).$$

It has already been shown that $\|x(\alpha)\|$ is convex on this interval, and thus it suffices to find if the derivative of $\|x(\alpha)\|^2$ with respect to α is negative at α_{l-1}, which shows that $x(\alpha)$ is then decreasing. Note that the derivative of $\|x(\alpha)\|^2$ is obtained by premultiplying the derivative of $x(\alpha)$ by $x(\alpha)^T$. To do the analysis use the SVD of A and thus have

$$\frac{d\|x(\alpha)\|^2}{d\alpha} = -b_1^T\Sigma_1^2\left(\Sigma_1^2 + \alpha I\right)^{-3} b_1.$$

Note that the matrix in parenthesis is indefinite and thus it must be determined if the expression is negative or not at $\alpha = \alpha_{l-1}$. To do this consider another function whose derivative has already been examined. Consider the constraint function, $\|Ax - b\|^2 - \eta^2\|x\|^2$, and since at $\alpha = \alpha_{l-1}$ the infeasible region is being entered as α increases, the derivative of the constraint must be positive. This condition can be expressed as

$$2(\alpha_{l-1} + \eta^2)b_1^T\Sigma_1^2\left(\Sigma_1^2 + \alpha I\right)^{-3} b_1 > 0.$$

We note that $2(\alpha_{l-1} + \eta^2) > 0$ thus the condition is

$$b_1^T\Sigma_1^2\left(\Sigma_1^2 + \alpha I\right)^{-3} b_1 > 0.$$

This trivially gives

$$\frac{d\|x(\alpha)\|^2}{d\alpha} < 0,$$

and thus $x(\alpha)$ must be decreasing at $\alpha = \alpha_{l-1}$ for increasing α. Applying convexity to this result gives $\|x(\alpha_{l-1})\| < \|x(\alpha)\|$ for $-\sigma_{n-1}^2 < \alpha < -\alpha_{l-1}$, and thus the minimum feasible value for $x(\alpha)$ on $-\sigma_{n-1}^2 < \alpha < -\sigma_n^2$ is $x(\alpha_{l-1})$.

Now since $x(\alpha_{l-1})$ does not correspond to a local minimum it follows that there is a neighborhood of $x(\alpha_{l-1})$ of the constraint surface such that in this neighborhood $\|x\| < \|x(\alpha_{l-1})\|$. Thus since $x(\alpha_{l-1})$ does not correspond to a local minimum, it can be discarded from further consideration, since it is not the global minimum. Either way only the rightmost in the interval $(-\sigma_{n-1}^2, -\sigma_n^2)$ remains as a candidate.

Appendix F

Symmetry of $A^T(Ax - b)x^T$

First recall that the expression for $x(\psi)$ is given by

$$\begin{aligned} x &= (A^TA + \psi I)^{-1}A^Tb \\ &= A^T(AA^T + \psi I)^{-1}b \end{aligned}$$

and the expression for $Ax(\psi) - b$ is given by

$$\begin{aligned} Ax(\psi) - b &= AA^T(AA^T + \psi I)^{-1}b - b \\ &= -\psi(AA^T + \psi I)^{-1}b. \end{aligned}$$

Now note that

$$\begin{aligned} A^T(Ax(\psi) - b) &= -\psi A^T(AA^T + \psi I)^{-1}b \\ &= -\psi x(\psi), \end{aligned}$$

thus

$$\begin{aligned} A^T(Ax(\psi) - b)x(\psi)^T &= -\psi A^T(AA^T + \psi I)^{-1}bx(\psi)^T \\ &= -\psi x(\psi)x(\psi)^T. \end{aligned}$$

Trivially, the matrix is symmetric.

Appendix G

Determinant of Second Derivative Is Negative

Begin by recalling what the second derivative is for case 2 of the Backward Error.

$$\begin{bmatrix} \bar{\Sigma}^2 - \sigma_{n-1}^2 I & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \sigma_n^2 - \sigma_{n-1}^2 \end{bmatrix} + \frac{\sigma_{n-1}^2 \|b\|}{(\|A\|\|z\| + \|b\|)\|z\|^2} \begin{bmatrix} \bar{z} \\ z_{n-1} \\ z_n \end{bmatrix} \begin{bmatrix} \bar{z} \\ z_{n-1} \\ z_n \end{bmatrix}^T$$

Since the first matrix is singular, re-write this slightly for convenience later.

$$\begin{bmatrix} \bar{\Sigma}^2 - \sigma_{n-1}^2 I & 0 & 0 \\ 0 & \delta & 0 \\ 0 & 0 & \sigma_n^2 - \sigma_{n-1}^2 \end{bmatrix} + \frac{\sigma_{n-1}^2 \|b\|}{(\|A\|\|z\| + \|b\|)\|z\|^2} \begin{bmatrix} \bar{z} \\ z_{n-1} \\ z_n \end{bmatrix} \begin{bmatrix} \bar{z} \\ z_{n-1} \\ z_n \end{bmatrix}^T$$
$$- \frac{1}{\delta} \begin{bmatrix} 0 \\ \delta \\ 0 \end{bmatrix} \begin{bmatrix} 0 \\ \delta \\ 0 \end{bmatrix}^T$$

Notice that these equations are the same. Define the following

$$\begin{aligned}
A &= -\frac{\|z\|^2(\|A\|\|z\| + \|b\|)}{\sigma_{n-1}^2\|b\|} \\
B &= z^T \\
C &= z \\
D &= \begin{bmatrix} \bar{\Sigma}^2 - \sigma_{n-1}^2 I & 0 & 0 \\ 0 & \delta & 0 \\ 0 & 0 & \sigma_n^2 - \sigma_{n-1}^2 \end{bmatrix}.
\end{aligned}$$

Now take the determinant.

$$\begin{aligned}
\det(\nabla^2 C_2) &= \det\left((D - CA^{-1}B) - \frac{1}{\delta}\begin{bmatrix} 0 \\ \delta \\ 0 \end{bmatrix}\begin{bmatrix} 0 \\ \delta \\ 0 \end{bmatrix}^T \right) \\
&= \det\left(\begin{bmatrix} \delta & \begin{bmatrix} 0 & \delta & 0 \end{bmatrix} \\ \begin{bmatrix} 0 \\ \delta \\ 0 \end{bmatrix} & \begin{bmatrix} \\ D - CA^{-1}B \\ \\ \end{bmatrix} \end{bmatrix} \right)
\end{aligned}$$

Use Pascal's expansion on the first row (or first column as it is symmetric), and denote the location of the δ in $\begin{bmatrix} 0 & \delta & 0 \end{bmatrix}$ as the p^{th} place (note that the first zero is a vector of zeros not a scalar).

$$
\begin{aligned}
\det(\nabla^2 C_2) &= \delta \det\left(D - CA^{-1}B\right) + (-1)^p \delta \\
&\quad \det\left(\begin{bmatrix} 0 & \delta & 0 \\ \bar{\Sigma} - \sigma_{n-1}^2 I & 0 & 0 \\ 0 & 0 & \sigma_n^2 - \sigma_{n-1}^2 \end{bmatrix} - A^{-1} \begin{bmatrix} 0 \\ \bar{z} \\ z_n \end{bmatrix} \begin{bmatrix} 0 \\ \bar{z} \\ z_n \end{bmatrix}^T\right) \\
&= \delta \det\left(D - CA^{-1}B\right) + (-1)^p (-1)^{p-1} \delta \\
&\quad \det\left(\begin{bmatrix} \delta & 0 & 0 \\ 0 & \bar{\Sigma} - \sigma_{n-1}^2 I & 0 \\ 0 & 0 & \sigma_n^2 - \sigma_{n-1}^2 \end{bmatrix} - A^{-1} \begin{bmatrix} 0 \\ \bar{z} \\ z_n \end{bmatrix} \begin{bmatrix} 0 \\ \bar{z} \\ z_n \end{bmatrix}^T\right) \\
&= \delta \det\left(D - CA^{-1}B\right) \\
&\quad -\delta \det\left(\begin{bmatrix} \delta & 0 & 0 \\ 0 & \bar{\Sigma} - \sigma_{n-1}^2 I & 0 \\ 0 & 0 & \sigma_n^2 - \sigma_{n-1}^2 \end{bmatrix} - A^{-1} \begin{bmatrix} 0 \\ \bar{z} \\ z_n \end{bmatrix} \begin{bmatrix} 0 \\ \bar{z} \\ z_n \end{bmatrix}^T\right)
\end{aligned}
$$

Now use Pascal's expansion on the first row (or first column as it is symmetric) one more time and note there are two terms to be solved for.

$$
\begin{aligned}
\det(\nabla^2 C_2) &= \delta \det\left(D - CA^{-1}B\right) \\
&\quad -\delta^2 \det\left(\begin{bmatrix} \bar{\Sigma} - \sigma_{n-1}^2 I & 0 \\ 0 & \sigma_n^2 - \sigma_{n-1}^2 \end{bmatrix} - A^{-1} \begin{bmatrix} \bar{z} \\ z_n \end{bmatrix} \begin{bmatrix} \bar{z} \\ z_n \end{bmatrix}^T\right) \\
&= \delta\,(\text{Term1}) - \delta^2\,(\text{Term2}) \qquad \text{(G.1)}
\end{aligned}
$$

G.1 Term 1

The basic idea here is to rewrite Term 1 using properties of the determinant and then show that no matter what the matrix Σ is there exists a b that will make this term negative. Only one b is needed for the perturbation analysis, and it turns out a "bad" b always exists that has small $b_{1,n}$. For the following development, note that A is a scalar thus $\det(A) = A$.

$$\begin{aligned}
\text{Term1} &= \det\left(D - CA^{-1}B\right) \\
&= \det\left(\begin{bmatrix} A & B \\ C & D \end{bmatrix}\right) / \det(A) \\
&= \frac{\det(D)}{A} \det\left(A - BD^{-1}C\right) \\
&= \frac{\det(D)}{A} \left(A - BD^{-1}C\right)
\end{aligned}$$

Note that D is a non-singular diagonal matrix with only 1 negative term, so its determinant is negative. The term A is also negative, so the fraction is positive. All that must be shown is that the second part is negative. Before doing this, note that from the first order condition on the cost the following holds

$$z_n^2 = \left(\frac{\sigma_n b_{1,n}}{\sigma_n^2 - \sigma_{n-1}^2}\right)^2.$$

Now

$$\begin{aligned}
\left(A - BD^{-1}C\right) &= -\frac{\|z\|^2 \left(\|A\|\|z\| + \|b\|\right)}{\|b\|\sigma_{n-1}^2} \\
&\quad - \left(\bar{z}^T \left(\bar{\Sigma}^2 - \sigma_{n-1}^2 I\right) \bar{z} + \frac{z_{n-1}^2}{\delta} + \frac{z_n^2}{\sigma_n^2 - \sigma_{n-1}^2}\right).
\end{aligned}$$

Substitute in the value of z_n^2 found above to obtain

$$\begin{aligned}
\left(A - BD^{-1}C\right) &= -\frac{\|z\|^2\left(\|A\|\|z\| + \|b\|\right)}{\|b\|\sigma_{n-1}^2} \\
&\quad - \left(\bar{z}^T\left(\bar{\Sigma}^2 - \sigma_{n-1}^2 I\right)\bar{z} + \frac{z_{n-1}^2}{\delta} + \frac{(\sigma_n b_{1,n})^2}{(\sigma_n^2 - \sigma_{n-1}^2)^3}\right) \\
&= -\left(\frac{\|z\|^2\left(\|A\|\|z\| + \|b\|\right)}{\|b\|\sigma_{n-1}^2} + \bar{z}^T\left(\bar{\Sigma}^2 - \sigma_{n-1}^2 I\right)\bar{z} + \frac{z_{n-1}^2}{\delta}\right) \\
&\quad - \frac{(\sigma_n b_{1,n})^2}{(\sigma_n^2 - \sigma_{n-1}^2)^3} \\
&= -\left(\frac{\|z\|^2\left(\|A\|\|z\| + \|b\|\right)}{\|b\|\sigma_{n-1}^2} + \bar{z}^T\left(\bar{\Sigma}^2 - \sigma_{n-1}^2 I\right)\bar{z} + \frac{z_{n-1}^2}{\delta}\right) \\
&\quad + \frac{(\sigma_n b_{1,n})^2}{(\sigma_{n-1}^2 - \sigma_n^2)^3} \\
&= -\alpha_1 + \beta b_{1,n}^2
\end{aligned}$$

with

$$\begin{aligned}
\alpha_1 &= \frac{\|z\|^2\left(\|A\|\|z\| + \|b\|\right)}{\|b\|\sigma_{n-1}^2} + \bar{z}^T\left(\bar{\Sigma}^2 - \sigma_{n-1}^2 I\right)\bar{z} + \frac{z_{n-1}^2}{\delta} \\
\beta &= \frac{(\sigma_n)^2}{(\sigma_{n-1}^2 - \sigma_n^2)^3}.
\end{aligned}$$

Now when $b_{1,n} < \sqrt{\frac{\alpha_1}{\beta}}$, Term 1 will be negative. Consider in particular,

$$b_{1,n} = \sqrt{\frac{\alpha_1}{2\beta}}$$

so that

$$b_{1,n}^2 = \frac{\alpha_1}{2\beta}.$$

Thus,

$$
\begin{aligned}
\text{Term1} &= \frac{\det(D)}{A}\left(A - BD^{-1}C\right) \\
&= \frac{\det(D)}{A}\left(-\alpha_1 + \beta b_{1,n})^2\right) \\
&= \frac{\det(D)}{A}\left(-\alpha_1 + \beta\frac{\alpha_1}{2\beta}\right) \\
&= \frac{\det(D)}{A}\left(-\alpha_1 + \frac{\alpha_1}{2}\right) \\
&= -\frac{\alpha_1}{2}\frac{\det(D)}{A}.
\end{aligned}
$$

G.2 Term 2

Now look at Term 2,

$$
\begin{aligned}
\text{Term2} &= \det\left(\begin{bmatrix}\bar{D} & 0 \\ 0 & d_n\end{bmatrix} - A^{-1}\begin{bmatrix}\bar{z} \\ z_n\end{bmatrix}\begin{bmatrix}\bar{z} \\ z_n\end{bmatrix}^T\right) \\
&= \frac{\det\left(\begin{bmatrix}A & \bar{z}^T & z_n \\ \bar{z}^T & \bar{D} & 0 \\ z_n & 0 & d_n\end{bmatrix}\right)}{A} \\
&= \frac{\det\left(\begin{bmatrix}\bar{D} & 0 \\ 0 & d_n\end{bmatrix}\right)}{A}\det\left(A - \bar{z}^T\bar{D}^{-1}\bar{z} - \frac{z_n^2}{d_n}\right).
\end{aligned}
$$

Notice that the first term is similar to the first term of the last section. The two are related by

$$
\frac{\det(D)}{A} = \delta\frac{\det\left(\begin{bmatrix}\bar{D} & 0 \\ 0 & d_n\end{bmatrix}\right)}{A}
$$

or

$$\frac{1}{\delta}\frac{\det(D)}{A} = \frac{\det\left(\begin{bmatrix}\bar{D} & 0 \\ 0 & d_n\end{bmatrix}\right)}{A}.$$

Since the two terms are related by δ, which is positive, they must have the same sign. This means the first term is positive. The second term is the determinant of a scalar. The expression for z_n must be the same, and the same value of $b_{1,n}$ will be used, as they are going to be used in the same equation. Keeping all this in mind,

$$\begin{aligned}
A - \bar{z}^T\bar{D}^{-1}\bar{z} - \frac{z_n^2}{d_n} &= -\alpha + \frac{z_{n-1}^2}{\delta} + \beta b_{1,n}^2 \\
&= -\alpha + \frac{z_{n-1}^2}{\delta} + \beta\frac{\alpha}{2\beta} \\
&= -\alpha + \frac{z_{n-1}^2}{\delta} + \frac{\alpha}{2} \\
&= \frac{z_{n-1}^2}{\delta} - \frac{\alpha}{2}.
\end{aligned}$$

The second term is thus given by

$$\text{Term2} = \frac{1}{\delta}\frac{\det(D)}{A}\left(\frac{z_{n-1}^2}{\delta} - \frac{\alpha}{2}\right).$$

G.3 Putting It All Together

Take the results of the last two sections and substitute them into Equation G.1.

$$\begin{aligned}
\det(\nabla^2 C_2) &= \delta\,(\text{Term1}) - \delta^2\,(\text{Term2}) \\
&= \delta\left(-\frac{\alpha_1}{2}\frac{\det(D)}{A}\right) - \delta^2\left(\frac{1}{\delta}\frac{\det\,(D)}{A}\left(\frac{z_{n-1}^2}{\delta} - \frac{\alpha}{2}\right)\right) \\
&= \delta\left(-\frac{\alpha_1}{2}\frac{\det(D)}{A}\right) - \delta\left(\frac{\det\,(D)}{A}\left(\frac{z_{n-1}^2}{\delta} - \frac{\alpha}{2}\right)\right) \\
&= \delta\frac{\det(D)}{A}\left(-\frac{\alpha_1}{2} - \left(\frac{z_{n-1}^2}{\delta} + \frac{\alpha}{2}\right)\right) \\
&= -\delta\frac{\det(D)}{A}\frac{z_{n-1}^2}{\delta}
\end{aligned}$$

From the last line it is easy to see that second derivative is negative except when $z_{n-1} = 0$ when it is zero. In all cases the second derivative is not positive for values of $b_{1,n} < \sqrt{\frac{\alpha_1}{\beta}}$.

Bibliography

[1] K. D. Andersen. An Efficient Newton Barrier Method for Minimizing a Sum of Euclidean Norms. *SIAM J. Optim.*, 6:74–95, 1996.

[2] K. Atkinson. *Elementary Numerical Analysis, Second Ed.* Wiley, New York, 1993.

[3] S. J. Benbow. Solving Generalized Least-Squares Problems With LSQR. *SIMAX*, 21:165–177, 1999.

[4] Å. Björck. Iterative Refinement of Linear Least Squares Solutions I. *BIT*, 7:257–278, 1967.

[5] Å. Björck. Iterative Refinement of Linear Least Squares Solutions II. *BIT*, 8:8–30, 1968.

[6] Å. Björck. A General Updating Algorithm for Constrained Linear Least Squares Problems. *SISSC*, 5:394–402, 1984.

[7] Å. Björck. Stability Analysis of the Method of Seminormal Equations for Linear Least Squares Problems. *Linear Alg. and Its Applic.*, 88/89:31–48, 1987.

[8] Å. Björck. *Least Squares Methods: Handbook of Numerical Analysis*, volume 1. Elsiver, Holland, 1988.

[9] Å. Björck. Component-wise Perturbation Analysis and Error Bounds for Linear Least Squares Solutions. *BIT*, 31:238–244, 1991.

[10] Å. Björck. *Numerical Methods for Least Squares Problems*. SIAM, Philadelphia, PA, 1996.

[11] Å. Björck and L. Eldén. Methods in Numerical Algebra for Ill-posed Problems. Technical Report LiTH-MAT-R-1979-33, Department of Mathematics, Linköping University, 1979.

[12] Å. Björck and T. Elfving. Accelerated Projection Methods for Computing Pseudoinverse Solutions of Systems of Linear Equations. *BIT*, 19:145–163, 1979.

[13] Å. Björck and G. H. Golub. Iterative Refinement of Linear Least Squares Solutions by Householder Transformation. *BIT*, 7:322–337, 1967.

[14] S. Boyd and L. El Ghaoui. Method of Centers for Minimizing Generalized Eigenvalues. *Linear Algebra and its Applications*, 188:63–111, July 1993. Special Issue on Systems and Control.

[15] S. Boyd, L. El Ghaoui, E. Feron, and V. Balakrishnan. *Linear Matrix Inequalities in System and Control Theory*. SIAM, Philadelphia, PA, 1994.

[16] S. Boyd and L. Vandenberghe. *Semidefinite Programming Relaxations of Non-convex Problems in Control and Combinatorial Optimization*, chapter 15, pages 279–288. Kluwer, 1997.

[17] S. Boyd and L. Vandenberghe. Advances in Convex Optimization: Theory, Algorithms, and Applications. In *IEEE International Symposium on Information Theory*. IEEE, July 2002. Talk.

[18] S. Boyd and L. Vandenberghe. Convex Optimization. `http://www.stanford.edu/~boyd/cvxbook.html`, December 2002.

[19] S. Boyd, L. Vandenberghe, and M. Grant. Efficient Convex Optimization for Engineering Design. In *Proceedings IFAC Symposium on Robust Control Design*, pages 14–23, September 1994.

[20] P.H. Calamai and A.R. Conn. A Stable Algorithm for Solving the Multifacility Location Problem Involving Euclidean Distances. *SIAM Journal on Scientific and Statistical Computing*, 1:512–526, 1980.

[21] S. Chandrasekaran, G. H. Golub, M. Gu, and A. H. Sayed. Best-fit Parameter Estimation for a Bounded Errors-in-Variables Model. In *Proc. American Control Conference*, Albuquerque, NM, 1997.

[22] S. Chandrasekaran, G. H. Golub, M. Gu, and A. H. Sayed. Efficient Algorithms for Least Squares Type Problems with Bounded Uncertainties. In S. Van Huffel, editor, *Recent Advances in Total Least Squares Techniques and Errors-in-Variables Modeling*, pages 171–180. SIAM, Philadelphia, PA, 1997.

[23] S. Chandrasekaran, G. H. Golub, M. Gu, and A. H. Sayed. Parameter Estimation in the Presence of Bounded Modeling Errors. *IEEE Signal Process. Lett.*, 4:195–197, 1997.

[24] S. Chandrasekaran, G. H. Golub, M. Gu, and A. H. Sayed. Parameter Estimation in the Presence of Bounded Data Uncertainties. *SIMAX*, 19(1):235–252, 1998.

[25] S. Chandrasekaran, G. H. Golub, M. Gu, and A. H. Sayed. An Efficient

Algorithm for a Bounded Errors-in-Variables Model. *SIMAX*, 20(4):839–859, 1999.

[26] S. Chandrasekaran, M. Gu, A. H. Sayed, and K. E. Schubert. The Degenerate Bounded Errors-In-Variables Model. *SIMAX*, 23(1):138–166, 2001.

[27] S. Chandrasekaran and K. E. Schubert. Models for Robust Estimation and Identification. In S. Van Huffel and P. Lemmerling, editors, *Total Least Squares and Errors-In-Variables Modeling*, pages 199–208, Dordrecht, 2001. Kluwer Academic Publishers.

[28] A. K. Cline. An Elimination Method for the Solution of Linear Least Squares Problems. *SINUM*, 10:283–289, 1973.

[29] F. Cordellier and J.Ch. Giorot. On the Fermat-Weber Problem with Convex Cost Functionals. *Mathematical Programming*, 14:295–311, 1978.

[30] M. G. Cox. The Least Squares Solution of Overdetermined Linear Equations Having Gand or Augmented Band Structure. *IMA J. Numer. Anal.*, 1:3–22, 1981.

[31] J. Cullum. Ill-posed Deconvolutions: Regularization and Singular Value Decompositions. In *Proceedings 19th IEEE Conference on Decision and Control*, volume 1, pages 29–35, 1980.

[32] G. Cybenko. The Numerical Stability of the Lattice Algorithm for Least Squares Linear Prediction Problems. *BIT*, 24:441–455, 1984.

[33] C. Daniel and F. S. Wood. *Fitting Equations to Data*. Wiley, New York, 1971.

[34] L.M. Delves and I. Barrodale. A Fast Direct Method for the Least Squares Solution of Slightly Overdetermined Sets of Linear Equations. *J. Inst. Maths. Applics.*, 24:149–156, 1979.

[35] G. Demoment. Image Reconstruction and Restoration: Overview of Common Estimation Problems. *IEEE Trans. Acoustic Speech and Signal Processing*, 37:2024–2036, 1989.

[36] I. S. Duff. Pivot Selection and Row Ordering in Givens Reduction on Sparse Matrices. *Computing*, 13:239–248, 1974.

[37] U. Eckhardt. On an Optimization Problem Related to Minimal Surfaces with Obstacles. In R. Bulirsch, W. Oettli, and J. Stoer, editors, *Optimization and Optimal Control*. Springer-Verlag, 1975.

[38] U. Eckhardt. Weber's Problem and Weiszfeld's Algorithm in General Spaces. *Mathematical Programming*, 18:186–196, 1980.

[39] M. P. Ekstrom and R. L. Rhoads. On the Application of Eigenvector Expansions to Numerical Deconvolutions. *J. of Comp. Phys.*, 14:395–417, 1974.

[40] L. Eldèn. Algorithms for the Regularization of Ill-Conditioned Least Squares Problems. *BIT*, 17:134–145, 1977.

[41] L. Eldèn. Perturbation Theory for the Least Squares Problem with Linear Equality Constraints. *SINUM*, 17:338–350, 1980.

[42] L. Eldèn. A Weighted Pseudoinverse, Generalize Singular Values, and Constrained Least Squares Problems. *BIT*, 22:487–502, 1983.

[43] L. Eldèn. An Algorithm for the Regularization of Ill-Conditioned, Banded Least Squares Problems. *SISSC*, 5:237–254, 1984.

[44] L. Eldèn. A Note on the Computation of the Generalized Cross-Validation Function for Ill-Conditioned Least Squares Problems. *BIT*, 24:467–472, 1985.

[45] H. W. Engl and H. Gfrerer. A Posteriori Parameter Choice for General Regulariztion Methods for Solving Linear Ill-posed Problems. *Appl. Numer. Math.*, 4:395–417, 1988.

[46] J.W. Eyster, J.A. White, and W.W. Wierwille. On Solving Multifacility Location Problems Using a Hyperboloid Approximation Procedure. *AIIE Transactions*, 5:1–6, 1973.

[47] J. D. Faires and R. L. Burden. *Numerical Methods*. PWS-Kent, Boston, 1993.

[48] M. K. H. Fan, A. L. Tits, and J. C. Doyle. Robustness in the Presence of Mixed Parabetric Uncertainty and Unmodeled Dynamics. *ITAC*, 36:25–38, 1991.

[49] R. D. Fierro and J. R. Bunch. Colinearity and Total Least Squares. *SIMAX*, 15:1167–1181, 1994.

[50] R. D. Fierro, G. H. Golub, P. C. Hansen, and D. P. O'Leary. Regularization by Truncated Total Least Squares. *SISC*, 18:1223–1241, 1997.

[51] A. Forsgren and G. Sporre. On Weighted Linear Least-Squares Problems Related to Interior Methods for Convex Quadratic Programming. *SIMAX*, 23:42–56, 2001.

[52] G. E. Forsythe and G. H. Golub. On the Stationary Values of a Second-Degree Polynomial on the Unit Sphere. *SIAM J. App. Math.*, 14:1050–1068, 1965.

[53] J. N. Franklin. Minimum Principles for Ill-posed Problems. *SIAM J. Math. Anal.*, 9:638–650, 1978.

[54] M. Furuya, H. Ohmori, and A. Sano. Optimization of Weighting Constant for Regularization in Least Squares System Identification. *Grans. Inst Elec. Inform. Comm. Eng. A*, J72A:1012–1015, 1989.

[55] W. Gander. Least Squares with a Quadratic Constraint. *Numer. Math.*, 36:291–307, 1981.

[56] C. F. Gauss and G. W. Stewart. *Theory of the Combination of Observations Least Subject to Errors.* SIAM, Philadelphia, PA, 1995.

[57] J. A. George and M. T. Heath. Solution of Sparse Linear Least Squares Problems Using Givens Rotations. *Lin. Alg. and Its Applic.*, 34:69–83, 1980.

[58] H. Gfrerer. An A Posteriori Parameter Choice for Ordinary and Iterated Tikhonov Regularization of Ill-Posed Problems Leading to Optimal Convergence Rates. *Math. Comp.*, 49:507–522, 1987.

[59] L. El Ghaoui. and G. Calafiore. Worst-case Prediction Under Structured Uncertainty. In *Proc. Amer. Control Conf.*, pages 3402–3406, San Diego, CA, 1999.

[60] L. El Ghaoui and G. Galafiore. Robust Filtering for Discrete-Time Systems with Bounded Noise and Parametric Uncertainty. *ITAC*, 46:1084–1089, 2001.

[61] L. El Ghaoui and H. Lebret. Robust Least Squares and Applications. In *Proceedings of the 35th Conference on Decision and Control*, 1996.

[62] L. El Ghaoui. and H. Lebret. Robust Solutions to Least-Squares Problems with Uncertain Data. *SIMAX*, 18(4):1035–1064, 1997.

[63] L. El Ghaoui. and H. Lebret. Robust Solutions to Least-Squares Problems with Uncertain Data. In S. Van Huffel, editor, *Recent Advances in Total Least Squares Techniques and Errors-in-Variables Modeling*, pages 161–170. SIAM, Philadelphia, PA, 1997.

[64] G. H. Golub, P. C. Hansen, and D. P. O'Leary. Tikhonov Regularization and Total Least Squares. *SIMAX*, 30(1):185–194, 1999.

[65] G. H. Golub, M. Heath, and G. Wahba. Generalized Cross-Validation as a Method for Chosing a Good Ridge Parameter. *Technometrics*, 21:215–223, 1979.

[66] G. H. Golub and C. F. Van Loan. Total Least Squares. In T. Gasser and M. Rosenblatt, editors, *Smoothing Techniques for Curve Estimation*, pages 69–76, New York, 1979. Springer-Verlag.

[67] G. H. Golub and C. F. Van Loan. An Analysis of the Total Least Squares Problem. *SIMAX*, 17:883–893, 1980.

[68] G. H. Golub and C. F. Van Loan. *Matrix Computations*. The Johns Hopkins University Press, Baltimore, Md, 1996.

[69] G. H. Golub and U. von Matt. Quadratically Constrained Least Squares and Quadratic Problems. *Numer. Math.*, 59:561–580, 1991.

[70] G. H. Golub and J. H. Wilkinson. Note on Iterative Refinement of Least Squares Solutions. *Numerical Math.*, 9:139–148, 1966.

[71] Ming Gu. Backward Perturbation Bounds for Linear Least Squares Prolems. *SIMAX*, 20:363–372, 1998.

[72] M. L. Hambaba. The Robust Generalized Least-squares Estimator. *Signal Processing*, 26:359–368, 1992.

[73] M. Hanke and P. C. Hansen. Regularization Methods for Large-scale Problems. *Surveys on Mathematics for Industry*, 3:253–315, 1993.

[74] M. Hanke and T. Raus. A General Heuristic for Choosing the Regularization Parameter in Ill-Posed Problems. *SIAM J. Sci. Comput.*, 7:956–972, 1996.

[75] P. C. Hansen. Analysis of Discrete Ill-posed Problems by Means of the L-curve. *Siam Review*, 34:561–580, 1992.

[76] P. C. Hansen. *Rank-Deficient and Discrete Ill-Posed Problems.* SIAM, Philadelphia, PA, 1998.

[77] K. H. Haskell and R. J. Hanson. Selected Algorithms for the Linearly Constrained Least Squares Problem: A User's Guide. Technical Report SAND78-1290, Sandia National Laboratories, Albuquerque, NM., 1979.

[78] S. Haykin. *Adaptive Filter Theory.* Prentice Hall, Englewood Cliffs, NJ, 1991.

[79] D. J. Higham and N. J. Higham. Backward Error and Condition of Structured Linear Systems. *SIMAX*, 13:162–175, 1992.

[80] A. E. Hoerl and R. W. Kennard. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, 12:55–67, 1970.

[81] P. D. Hough and S. A. Vavasis. Complete Orthogonal Decomposition For Weighted Least Squares. *SIMAX*, 18:369–392, 1997.

[82] S. Van Huffel and J. Vandewalle. *The Total Least Squares Problem: Computational Aspects and Analysis*. SIAM, Philadelphia, PA, 1991.

[83] B. R. Hunt. The Application of Constrained Least-squares Estimation to Image Restoration by Digital Computer. *IEEE Trans. Comput.*, C-22:805–812, 1973.

[84] M. E. Kilmer and D. P. O'Leary. Choosing Regularization Parameters in Iterative Methods for Ill-posed Problems. *SIMAX*, 22:1204–1221, 2001.

[85] D. Kincaid and W. Cheney. *Numerical Analysis, Second Ed.* Brooks/Cole, Boston, 1996.

[86] S. Kourouklis and C. C. Paige. A Constrained Least Squares Approach to the General Gauss-Markov Linear Model. *J. Amer Stat. Assoc.*, 76:620–625, 1981.

[87] H.W. Kuhn. A Note on Fermat's Problem. *Mathematical Programming*, 4:98–107, 1973.

[88] A.J. Laub. Computational Matrix Analysis. Class notes for ECE 234 and ECE 231A.

[89] C. L. Lawson and R. J. Hanson. *Solving Least Squares Problems*. SIAM, Philadelphia, PA, 1995.

[90] H. Lebret and S. Boyd. Antenna Array Pattern Synthesis Via Convex Optimization. *IEEE Transactions on Signal Processing*, 45(3):526–532, March 1997.

[91] L. Ljung. *System Identification Toolbox User's Guide*. The Math Works, Inc., Natick, MA, 1991.

[92] L. Ljung. System Identification. In *The Control Handbook*. CRC Press, 1996.

[93] Lenart Ljung. *System Identification: Theory for the User*. Prentice Hall, Englewood Cliffs, NJ, 1987.

[94] M. Lobo, L. Vandenberghe, S. Boyd, and H. Lebret. Applications of Second-Order Cone Programming. *Linear Algebra and its Applications*, 284:193–228, November 1998. Special Issue on Linear Algebra in Control, Signals and Image Processing.

[95] R. Lorenz and S. Boyd. Robust Minimum Variance Beamforming. Submitted to IEEE Transactions on Signal Processing, October 2001.

[96] R.F. Love. Locating Facilities in Three-dimensional Space by Convex Programming. *Naval Research Logistics Quarterly*, 16:503–516, 1969.

[97] N. Mastronardi, P. Lemmerling, and S. Van Huffel. Fast Structured Total Least Squares Algorithm for Solving the Basic Deconvolution Problem. *SIMAX*, 22:533–553, 2000.

[98] D. A. McQuarrie and P. A. Rock. *General Chemistry*. W. H. Freeman and Company, New York, NY, 1987.

[99] J. M. Mendel. *Lessons in Estimation Theory for Signal Processing, Communications, and Control.* Prentice Hall, Englewood Cliffs, NJ, 1995.

[100] K. Miller. Least Squares Methods for Ill-posed Problems with a Prescribed Bound. *SIAM J. Math. Anal.*, 1:52–74, 1970.

[101] M. Moonen, B. De Moor, L. Vandenberghe, and J. Vandewalle. On- and Off-line Identifiation of Linear State-Space Models. In R. V. Patel, A. J. Laub, and P. M. Van Dooren, editors, *Numerical Linear Algebra Techniques for Systems and Control.* IEEE Press, 1994.

[102] B. De Moor. Structured Total Least Squares and L_2 Approximation Problems. *Linear Algebra Appl.*, 188-189:163–207, 1993.

[103] V. A. Morozov. On the Solution of Functional Equations by the Method of Regularization. *Soviet Math. Dokl.*, 7:414–417, 1966. cited in [76].

[104] Y. Nesterov and A. Nemirovski. *Interior-point Polynomial Algorithms in Convex Programming.* SIAM, 1994.

[105] A. Neubauer and O. Scherzer. Regularization for Curve Representations: Uniform Convergence for Discontinuous Solutions of Ill-posed Problems. *SIAM J. Appl. Math.*, 58:1891–1900, 1998.

[106] Y. Nievergelt. Total Least Squares: State-of-the-Art Regression in Numerical Analysis. *SIAM Rev.*, 36:258–264, 1994.

[107] D. P. O'Leary. Near-Optimal Parameters for Tikhonov and Other Regularization Methods. Technical report, CS Dept., University of Maryland, 1999. CS-TR-4004.

[108] D. P. O'Leary and J. A. Simmons. A Bidiaonalization-Regularization Procedure for Large Scale Discreizations of Ill-Posed Problems. *SISSC*, 2:474–489, 1981.

[109] W. J. Ostrander. *Reflection Seismology.* UCSB Bookstore Custom Publishing, Goleta, CA, 1999.

[110] M. L. Overton. A Quadratically Convergent Method for Minimizing a Sum of Euclidean Norms. *Mathematical Programming*, 27:34–63, 1983.

[111] C. C. Paige. Computer Solution and Perturbation Analysis of Generalized Least Squares Problems. *Math. Comp.*, 33:171–184, 1979.

[112] C. C. Paige. Fast Numerically Stable Computations for Generalized Linear Least Squares Problems. *SINUM*, 16:165–171, 1979.

[113] C. C. Paige. The General Limit Model and the Generalized Singular Value Decomposition. *Lin. Alg. and Its Aplic.*, 70:269–284, 1985.

[114] C. C. Paige and S. Zdeněk. Bounds for the Least Squares Residual using Scaled Total Least Squares. In S. van Huffel and P. Lemmerling, editors, *Total Least Squares and Errors-in-Variables Modelling: Analysis, Algorithms and Applications*, pages 35–43, Dordrecht, 2002. Kluwer Academic Publishers.

[115] C. C. Paige and S. Zdeněk. Unifying Least Squares, Total Least squares and Data Least Squares. In S. van Huffel and P. Lemmerling, editors, *Total Least Squares and Errors-in-Variables Modelling: Analysis, Algorithms and Applications*, pages 25–34, Dordrecht, 2002. Kluwer Academic Publishers.

[116] R.L. Parker. *Geophysical Inverse Theory.* Princeton University Press, Princeton, NJ, 1994.

[117] J. R. Partington and P. M. Mäkilä. Worst-case Analysis of the Least-squares Method and Related Identification Methods. *Systems Control Lett.*, 24:193–200, 1995.

[118] G. Peters and J. H. Wilkinson. The Least Squares Problem and Pseudo-Inverses. *Comp. J.*, 13:309–316, 1970.

[119] J. E. Pierce and B. W. Rust. Constrained Least Squares Interval Estimation. *SIAM Jour. Sci. Stat. Comput.*, 6:670–683, 1985.

[120] J. G. Proakis and D. G. Manolakis. *Introduction to Digital Signal Processing.* Macmillan Publishing Co., New York, NY, 1988.

[121] T. Raus. The Principle of the Residual in the Solution of Ill-posed Problems with Nonselfadjoint Operator. *Uchen. Zap. Tartu Gos. Univ.*, 715:12–20, 1985. In Russian. Referenced in [74].

[122] W. Rudin. *Principles of Mathematical Analysis, Third Edition.* McGraw-Hill, New York, NY, 1976.

[123] B. W. Rust. Truncating the Singular Value Decomposition for Ill-posed Problems. Technical report, National Institute of Standards and Technology, U.S. Dept. of Commerce, 1998. Tech. Report NISTIR 6131.

[124] M. A. Saunders. Sparse Least Squares by Conjugate Gradients: a Comparison of Preconditioning Methods. In *Proceedings of Computer Science and Statistics: Twelfth Annual Conference on the Interface*, Waterloo, Canada, 1979.

[125] A. H. Sayed. A Framework for State-Space Estimation with Uncertain Models. *ITAC*, 46:998–1013, 2001.

[126] A. H. Sayed and S. Chandrasekaran. Estimation in the Presence of Multiple Sources of Uncertainties with Applications. In *Proceedings of the Asilomar Conference*, pages 1811–1815, Pacific Grove, CA, 1998.

[127] A. H. Sayed, A. Garulli, and S. Chandrasekaran. A Fast Iterative Solution for Worst-case Parameter Estimation with Bounded Model Uncertainties. In *Proc. American Control Conference*, Albuquerque, NM, 1997.

[128] A. H. Sayed and V. H. Nascimento. Design Criteria for Uncertain Models with Structured and Unstructered Uncertainties. *Linear Alg. Appl.*, 284:259–306, 1998.

[129] S. Schechter. Minimization of a Convex Function by Relaxation. In J. Abadie, editor, *Integer and Nonlinear Programming*, pages 177–190. North-Holland, 1970.

[130] K. Schittkowski and J. Stoer. A Factorization Method for the Solution of Constrained Linear Least Squares Problems Allowing for Subsequent Data Changes. *Numer. Math.*, 31:431–463, 1979.

[131] K. Schittkowski and J. Stoer. A Factorization Method for the Solution of Constrained Linear Least Squares Problems Allowing Subsequent Data Changes. *Numer. Math.*, 31:431–463, 1979.

[132] B. L. Shader. Least Squares Sign-solvability. *SIMAX*, 16:1056–1073, 1995.

[133] P.M. Shearer. *Introduction To Seismology.* Cambridge University Press, New York, NY, 1999.

[134] H. Stark and J. W. Woods. *Probability, Random Processes and Estimation Theory for Engineers.* Prentice Hall, Englewood Cliffs, NJ, 1994.

[135] G. W. Stewart. *Introduction to Matrix Computations.* Academic Press, New York, NY, 1973.

[136] G. W. Stewart. A Note on the Perturbation of Singular Values. *Lin. Alg. Appl.*, 28:213–216, 1979.

[137] G. W. Stewart. The Effects of Rounding Error on an Algorithm for Downdating a Cholesky Factorization. *Journal of the Institute for Mathematics and Applications*, 23:203–213, 1979.

[138] G. W. Stewart. On the Asymptotic Behavior of Scaled Singular Value and QR Decompositions. *Math. Comp.*, 43:483–490, 1984.

[139] G. W. Stewart. On Scaled Projections and Pseudoinverses. *Linear Algebra Appl.*, 112:189–193, 1989.

[140] G. Strang. A Framework for Equilibrium Equations. *SIREV*, 30:283–297, 1988.

[141] A. Ben Tal and A. Nemirovski. *Lectures on Modern Convex Optimization.* SIAM, 2001.

[142] R. Tempo. Worst-case Optimality of Smothing Algorithms for Parametric System Identification. *Automatica*, 31:759–764, 1995.

[143] A. Tikhonov and V. Arsenin. *Solutions of Ill-Posed Problems.* Wiley, New York, 1977.

[144] M. J. Todd. A Dantzig-Wolfe-Like Variant of Karmarkar's Interior-Point Linear Programming Algorithm. *Oper. Res.*, 38:1006–1018, 1990.

[145] L. Vandenberghe and S. Boyd. Semidefinite Programming. *SIAM Review*, 38(1):49–95, March 1996.

[146] L. Vandenberghe, S. Boyd, and S.-P. Wu. Determinant maximization with linear matrix inequality constraints. *SIMAX*, 19(2):499–533, 1998.

[147] J. M. Varah. A Practical Examination of some Numerical Methods for Linear Discrete Ill-posed Problems. *SIAM Review*, 21:100–111, 1979.

[148] S. A. Vavasis. Stable Numerical Algorithms For Equilibrium Systems. *SIMAX*, 15:1108–1131, 1994.

[149] M. Verhaegen. Identification of Continuous-time MIMO State-Space Models from Sampled Data, in the Presence of Process and Measurement Noise. In *Proceedings of the Workshop Mathematics for Systems*, 1997.

[150] H. Voss and U. Eckhardt. Linear Convergence of Generalized Weiszfeld's Method. *Computing*, 25:243–251, 1980.

[151] B. Walden, R. Karlson, and J. Sun. Optimal Backward Perturbation Bounds for the Linear Least Squares Problem. *Numerical Linear Algebra with Applications*, pages 271–286, 1995.

[152] R. H. Wampler. L2A and L2B, Weighted Least Squares Solutions by Modified Gram–Schmidt with Iterative Refinement. *ACM Trans. Math. Software*, 5:494–99, 1979.

[153] R. H. Wampler. Solutions to Weighted Least Squares Problems by Modified Gram–Schmidt with Iterative Refinement. *ACM Trans. Math. Software*, 5:457–465, 1979.

[154] G. A. Watson. Solving Data Fitting Problems in l_p norms with Bounded Uncertainties in the Data. In D. F. Griffiths and G. A. Watson, editors, *Proceedings of the Dundee Conference, Numerical Analysis*, pages 249–265. Chapman and Hall / CRC, 2000.

[155] G. A. Watson. Data Fitting Problems with Bounded Uncertainties in the Data. *SIMAX*, 22(4):1274–1293, 2001.

[156] R. Webster. *Convexity*. Oxford, New York, NY, 1994.

[157] P.-Å. Wedin. Notes on the Constrained Linear Least Squares Problem. A New Approach Based on Generalized Inverses. Technical Report UMINF–75.79, Institute of Information Processing, University of Umeå, 1979.

[158] E. Weiszfeld. Sur le point par lequel la somme des distances de n points donnés est minimum. *Tohoku Mathematics Journal*, 43:355–386, 1937.

[159] B. Widrow and S. D. Stearns. *Adaptive Signal Processing*. Prentice Hall, Englewood Cliffs, NJ, 1985.

[160] A. S. Willsky. *Digital Signal Processing and Control and Estimation Theory: Points of Tangency, Areas of Intersection, and Parallel Directions*. MIT Press, Cambridge, Ma, 1979.

[161] G. H. Yang and J. L. Wang. Robust Nonfragile Kalman Filtering for Uncertain Linear Systems with Estimator Gain Uncertainty. *ITAC*, 46:343–348, 2001.

[162] M. E. Zervakis and T. M. Kwon. Robust Estimation Techniques in Regularized Image Restoration. *Op. Eng.*, 31:2174–2190, 1992.