Lawrence Berkeley National Laboratory report LBNL-52434

Optimal Sensitivity Analysis of Linear Least Squares

Joseph F. Grcar

Lawrence Berkeley National Laboratory
Mail Stop 50A-1148
One Cyclotron Road
Berkeley, CA 94720-8142 USA
e-mail: jfgrcar@lbl.gov

Abstract

Results from the many years of work on linear least squares problems are combined with a new approach to perturbation analysis to explain in a definitive way the sensitivity of these problems to perturbation. Simple expressions are found for the asymptotic size of optimal backward errors for least squares problems. It is shown that such formulas can be used to evaluate condition numbers. For full rank problems, Frobenius norm condition numbers are determined exactly, and spectral norm condition numbers are determined within a factor of square-root-two. As a result, the necessary and sufficient criteria for well conditioning are established. A source of ill conditioning is found that helps explain the failure of simple iterative refinement. Some textbook discussions of ill conditioning are found to be fallacious, and some error bounds in the literature are found to unnecessarily overestimate the error. Finally, several open questions are described.

Keywords: asymptotic estimates,
backward errors,
condition numbers,
linear least squares,
optimal backward errors.

2000 MSC: primary 65F99 (numerical linear algebra),
secondary 62J99 (linear inference, regression),
65G99 (error analysis and interval analysis).

Pages 5, 26, 35, 36, and 50 should be reproduced in color.

Contents

1 Introduction **4**

2 Numerical Analysis Background **8**

3 Pertinent Real Analysis **10**
3.1 Sensitivity Analysis of Metric Projections 11
3.2 Alternate Formula for Condition Numbers 12
3.3 Example of Evaluating Condition Numbers 16

4 Optimal Backward Errors for Least Squares **18**
4.1 Literature on Optimal Backward Error 18
4.2 Asymptotic Size of Optimal Backward Error 23
4.3 Calculable Asymptotic Estimate 25

5 Condition Numbers for Least Squares **27**
5.1 Literature on Error Bounds 27
5.2 Condition Numbers 37
5.3 Dependence on κ_2^2 39
5.4 Examination of the Problem's Conditioning 42

6 Applications **44**
6.1 Failure of Simple Iterative Refinement 44
6.2 Ill-Conditioned Without κ_2^2 48
6.3 Error Bounds that Overestimate the Error 49

7 Conclusion **49**
7.1 Narrative 49
7.2 Summary 51
7.3 Open Questions 52

Nomenclature **55**

References **58**

List of Figures

1 Function that lists entries of matrices as column vectors 17
2 Jacobian matrix for the residual of problem LS 23
3 Simple iterative improvement 44

List of Tables

1 Timeline for least squares bibliography 5
2 Operation counts for bounds on optimal backward error 22
3 Bounds on optimal backward error 26
4 Leading terms of error bounds 36
5 Values in the example of Golub and Wilkinson 46
6 Two instances of Example 6.1. 50
7 Ratios of bounds to errors for Example 6.1. 50

This paper is dedicated to John von Neumann at the centennial of his birth on 28th December 1903.

1 Introduction

On the Shoulders of Giants Alan Turing [58] introduced the sensitivity of a numerical problem's solution to changes in its data as a way to measure the difficulty of solving the problem accurately. "Condition numbers" are now regarded as fundamental to understanding numerical calculations. Yet textbooks exhibit surprisingly few. For every problem whose condition number is given, it is easy to find another whose optimal (minimal) condition number is unknown.

The year before Turing's contribution, John von Neumann [41] mentioned the converse sensitivity of a problem's data to changes in the solution. He observed, an inaccurate solution may solve the problem for some perturbed data, and he suggested, the size of perturbation (which now is called backward error) is the more appropriate measure of numerical accuracy. Oettli and Prager [42] rediscovered von Neumann's point of view years later, and showed by example that formulas could be derived for optimal (minimal) backward errors.

This paper uses linear least squares problems to illustrate an unanticipated combination of von Neumann's and Turing's ideas. It is proved that optimal condition numbers in general depend on the size of optimal backward errors. Since the differential theory of metric projections [11] [25] provides asymptotic expressions for the latter, in principle it is possible to derive the condition numbers. Indeed, all the formulas are easy to evaluate for linear least squares problems. This results in simple expressions for the size of optimal backward errors and for condition numbers of the problems.

The optimal formulas describe the perturbational properties of linear least squares problems in a definitive manner. Thus they clarify the years of research into these problems. Conversely, the previous work guides the interpretation of the optimal formulas by suggesting a computable asymptotic estimate for the optimal backward error, and by helping identify the exact spectral condition numbers. In this way a survey of the archival literature is an integral part of this paper, see Table 1.

Although this paper is about numerical linear algebra and error analysis, it is reasonable to say its conclusions are broadly relevant. Among all the calculations studied in numerical analysis, none are more frequently performed in engineering, the sciences, and statistics than least squares estimation.

Optimal Backward Errors The size of optimal backward errors for linear least squares problems was an open question for many years [31, p. 198] [49, pp. 6–7] [51, p. 163] until it was answered by Waldén, Karlson, and Sun [61] in 1995. The discovery stimulated much additional work, in particular because the exact size is difficult to evaluate, so bounds and estimates have been constructed for it [27] [35] [40] [61].

Table 1: *Timeline for sensitivity analysis of linear least squares problems, and for historical perspective, of selected contributions to solution algorithms. This specialized list does not reflect the composition of the large literature about least squares and related topics, for which see [10].*

Algorithms, Error Bounds, Optimal Backward Errors

circa 1800	Gauss
1924	Bénoit (Cholesky) [5]
1938	Banachiewicz [4]
1944	Dwyer [17]
1958	Householder [33]
1965	Golub [21]
	Businger, Golub [13]
1966	Golub, Wilkinson [24]
1967	Björck [7]
	Björck [6]
1969	Hanson, Lawson [29]
	Pereyra [44]
1972	Stoer [52]
1973	Wedin [62]
1974	Abdelmalek [1]
	Lawson, Hanson [36]
1975	van der Sluis [59]
1977	Stewart [48]
	Stewart [49]
1979	Paige [43]
1980	Eldén [19]
1983	Golub, Van Loan [22]
	Lötstedt [37]
1984	Lötstedt [38]
1985	Wedin [63]
1989	Arioli, Duff, de Rijk [3]
1989	Björck [8]
1990	Higham [31]
1990	Wei [64]
1991	Björck [9]
1995	Waldén, Karlson, Sun [61]
1996	Higham [32]
	Sun [55]
1997	Karlson, Waldén [35]
	Sun [56]
1999	Cox, Higham [14]
	Cox, Higham [15]
	Gu [27]
	Gu [28]
2001	Malyshev [39]
2002	Malyshev, Sadkane [40]

Regarding Gauss, an engaging account of the least squares problems that he actually solved can be found in [34, pp. 212–214], formal histories and sources are cited in [50, p. 323–324], and translations of Gauss's relevant writings are in [20]. As for Banachiewicz, several papers on related topics appear in Bulletin International de l'Académie Polonaise des Sciences et des Lettres, *series A, but specifically [4] seems not to have formally appeared before publication ceased in 1939.*

This paper finds two simple formulas that asymptotically equal the size of the Frobenius norm optimal backward error as the approximate solution becomes more accurate. If x_0 solves the problem

$$\min_u \|b - Au\|_2 \,,$$

and if $x = x_0 + \delta x$ solves the similar problem for the perturbed matrix $A + \delta A$, then as x nears x_0, the size of the optimal backward error asymptotically is

$$\min_{\delta A} \|\delta A\|_F \;\simeq\; \left\|(\|r_0\|_2^2\, I + \|x_0\|_2^2\, A^t A)^{-1/2} A^t r\right\|_2 \,, \tag{1}$$

provided the matrix in parentheses is invertible (usually the case), where $r_0 = b - Ax_0$ is the least squares residual, and $r = b - Ax$ is the approximate residual. The meanings of "asymptotic" are explained in Section 3.1 but include that the approximation error is $o(\|\delta x\|_2)$.

Equation (1) cannot be evaluated in practice because it depends on an exact solution and residual, r_0 and x_0. The following, calculable expression also is proved to asymptotically equal the size of the optimal backward error.

$$\min_{\delta A} \|\delta A\|_F \;\simeq\; \left\|(\|r\|_2^2\, I + \|x\|_2^2\, A^t A)^{-1/2} A^t r\right\|_2 \tag{2}$$

Gu [27] showed that this quantity is boundedly near the optimal size when A has full column rank. Karlson and Walden [35] established a lower bound that is within half of this expression. These bounds and this paper's asymptotic results suggest that equation (2) is an accurate, robust estimate for the size the optimal backward error of linear least squares problems.

Condition Numbers The condition number of linear least squares problems has been an open question since 1966 when Golub and Wilkinson [24] found an error bound that contains the square of the coefficient matrix's condition number. Many error bounds were subsequently derived, [1] [3] [7] [8] [9] [22] [29] [31] [32] [36] [44] [48] [52] [59] [62] [64], some of which have been used to study the conditioning of these problems.

If A has full rank, then this paper proves that the Frobenius norm relative condition number is,

$$\chi_F^{(\text{LS, rel})}(A) \;=\; \left(\frac{\|r_0\|_2^2}{\|x_0\|_2^2\, \sigma_{\min}^2} + 1\right)^{1/2} \frac{\|A\|_F}{\sigma_{\min}} \,.$$

where $\sigma_{\min}$ is the smallest nonzero singular value of A. Moreover, the following expression overestimates the spectral norm relative condition number by at most the factor $\sqrt{2}$,

$$\chi_2^{(\text{LS, rel})}(A) \;\approx\; \left(\frac{\|r_0\|_2}{\|x_0\|_2\, \sigma_{\min}} + 1\right) \kappa_2 \,, \tag{3}$$

where $\kappa_2 = \sigma_{\max}/\sigma_{\min}$ is the spectral matrix condition number, and $\sigma_{\max}$ is the

largest singular value of A. This quantity has previously appeared in several upper bounds for the error [7] [32] [36] [52] [59] [62] [64]. The only lower bound, by van der Sluis, showed that $\sigma_{\max}/\sigma_{\min}^2$ must be part of the condition number [59, p. 250, rem. 5.2].

The tight limits for the spectral condition number permit definitive statements to be made about the conditioning of the problem. In particular, the full rank problem is ill conditioned with respect to perturbations of the coefficient matrix if and only if *either:*

1. $\|r_0\|_2$ is substantially larger than $\|x_0\|_2\,\sigma_{\min}$, or
2. A is ill conditioned.

Stoer [52] has previously noted that these criteria imply large error bounds. However, most of the literature interprets conditions like (a) not as a separate source of ill conditioning but rather as causing the condition number to depend on κ_2^2. The sharp bounds on the condition number imply necessary and sufficient criteria for the problem's condition to be governed by the square of the matrix condition number (tangent theorem), and sufficient, geometric criteria for the problem to be ill conditioned (secant theorem).

Applications The results of this paper are applied to investigate some questions about linear least squares problems.

1. The reason for the failure of simple iterative improvement in the famous example of Golub and Wilkinson [24] is explained.
2. A simple example shows linear least squares problems can be ill conditioned even though the coefficient matrix is well conditioned.
3. The error bounds in the literature are examined numerically. It is found that some, including the bound used by LAPACK [2], systematically overestimate the error by a factor of κ_2

Plan of This Paper This is the plan of the paper. Section 2 introduces notation and supplies background information on the sensitivity analysis of numerical problems. Section 3 discusses real analysis, and proves that optimal backward errors give optimal condition numbers. Sections 4 and 5 apply results from the previous section to derive the optimal backward error formulas and the condition number formulas, respectively. Section 6 contains applications. Section 7 gives a brief historical narrative, a summary of results, and a list of open questions. A nomenclature precedes the references.

The research methodology employed in Sections 4 and 5 always begins with a thorough review of the literature (Sections 4.1 and 5.1). Application of Section 3's theories is kept succinct (Sections 4.2 and 5.2). The literature's results and this paper's new results are then combined to clarify the understanding of least squares problems (Sections 4.3, 5.2, 5.3, 5.4).

2 Numerical Analysis Background

Optimal backward errors and condition numbers are intrinsic to all numerical problems, so they can be defined quite generally. A numerical problem consists of data $y \in \mathbb{R}^m$, solutions $x \in \mathbb{R}^n$, and a residual function $F : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}^p$. The solutions for some data y_0 are those x_0 at which the residual vanishes, $F(y_0, x_0) = 0$.

With this notation it is possible to express the optimal backward errors, for the approximate solution $x \approx x_0$, as the solution of a minimization problem,

$$\mu(x) = \min_{y \,:\, F(y,x)\,=\,0} \|y - y_0\| \,. \tag{4}$$

If y attains equation (4)'s minimum, then $y - y_0$ is an optimal backward error, and $\mu(x)$ is its size. The size of the optimal backward error is a function of x that depends on y_0 and on the norm chosen for $\mathbb{R}^m$.

The same notation can be used to define condition numbers. The current approach owes as much to Wilkinson as to Turing, whose statistical reasoning [58, p. 298] has not been adopted. Although Wilkinson knew of more formal definitions, he pragmatically viewed condition numbers as the coefficients of data perturbations in bounds for solution errors [65, p. 29]. Thus a condition number, for data y_0 and solution x_0, bounds the ratio of changes in the solution to perturbations in the data. An optimal (minimal) ratio exists in a limiting sense for arbitrarily small data perturbations.

This can be made precise in the following way.[1] Suppose there is a neighborhood N of y_0 for which every $y \in N$ has a solution. With this assumption and notation, an optimal condition number can be expressed as,

$$\chi^{(\mathrm{abs})}(y_0) = \lim_{y \,\to\, y_0} \sup_{x \,:\, F(y,x)\,=\,0} \frac{\|x - x_0\|}{\|y - y_0\|} \,.$$

If a distinguished branch of solutions is of interest, then it is given by a function $f : N \to \mathbb{R}^n$ where $f(y_0) = x_0$ and $f(y) = x$ is the desired solution of the problem for the data y, in which case,

$$\chi^{(\mathrm{abs})}(y_0) = \limsup_{y \,\to\, y_0} \frac{\|f(y) - x_0\|}{\|y - y_0\|} \,. \tag{5}$$

This number is a function of y_0 that depends on the norms for the data and solution (the norms chosen for $\mathbb{R}^m$ and $\mathbb{R}^n$, respectively) as well as on the chosen branch of solutions, f. Often the solution is uniquely determined by the data, so the last qualification may be unnecessary.

Usually condition numbers are wanted that are invariant with respect to scaling the data and the solution. The simplest way to achieve this is to make

[1] The use of limits to define condition numbers originated with Rice [45, p. 288, def. 2]. The limit superior version was developed by Skeel [47] and has been used by Demmel [16].

equation (5)'s numerator and denominator relative to $\|x_0\|$ and $\|y_0\|$, respectively. Equivalently,

$$\chi^{(\mathrm{rel})}(y_0) = \frac{\|y_0\|}{\|x_0\|}\, \chi^{(\mathrm{abs})}(y_0)\,, \tag{6}$$

which is called a norm-wise relative condition number. With some justification equation (6) might be called *the* condition number because for linear equations it gives the universally recognized value of the matrix condition number. See the example in Section 3.3.

Although it will not be used in this paper, another way to define scale-invariant condition numbers uses norms that are relative to the selected data, y_0, and to the selected solution, x_0. Examples of such norms are,

$$\|x\|_{x_0} \;=\; \left\| \frac{x}{|x_0|} \right\|_\infty \quad \text{and} \quad \|y\|_{y_0} \;=\; \left\| \frac{y}{|y_0|} \right\|_\infty ,$$

where the notation $x/|x_0|$ means each entry of x is divided by the magnitude of the corresponding entry of the reference vector x_0, and similarly for y. These norms commonly are based on the infinity norm though any is acceptable. When they are used in equation (5), then the result is called a component-wise relative condition number.

Returning to Wilkinson's point of view, condition numbers should provide a bound on the first-order variation in the solution with respect to perturbations of the data. The importance of equations (5) and (6) is that they give condition numbers in Wilkinson's sense that are optimally small.

Theorem 2.1 (Optimal Error Bounds) *Suppose $f : N \to \mathbb{R}^n$ where N is a neighborhood of $y_0 \in \mathbb{R}^m$, and $f(y_0) = x_0$. If f is Fréchet differentiable at y_0, then equation (5)'s condition number is well defined and is the smallest possible coefficient in any error bound of the form,*

$$\|f(y) - x_0\| \;\le\; \chi^{(\mathrm{abs})}(y_0)\, \|y - y_0\| \;+\; o\left(\|y - y_0\|\right) . \tag{7}$$

Proof. (Part 1.) Differentiability implies $\chi^{(\mathrm{abs})}(y_0) = \|\mathcal{D}f(y_0)\|$ where the operator norm is the one induced from the norms for $\mathbb{R}^m$ and $\mathbb{R}^n$. This has been remarked by Demmel [16, p. 253] and for completeness it is proved here.

The Fréchet derivative of f at y_0 is the unique linear operator $\mathcal{L} = \mathcal{D}f(y_0)$ for which the following (deleted) limit vanishes,

$$\lim_{y \to y_0} \frac{\|f(y) - f(y_0) - \mathcal{L}(y - y_0)\|}{\|y - y_0\|} \;=\; 0\,. \tag{8}$$

The triangle inequality can replace the norm in the numerator with the smaller difference of norms.

$$\lim_{y \to y_0} \left| \frac{\|f(y) - f(y_0)\|}{\|y - y_0\|} - \frac{\|\mathcal{L}(y - y_0)\|}{\|y - y_0\|} \right| \;=\; 0$$

This means, for every $\epsilon > 0$ there is some $\delta > 0$ so $0 < \|y - y_0\| \leq \delta$ implies,

$$-\epsilon \;\leq\; \frac{\|f(y) - f(y_0)\|}{\|y - y_0\|} - \frac{\|\mathcal{L}(y - y_0)\|}{\|y - y_0\|} \;\leq\; \epsilon\,.$$

Since $\mathcal{L}$ is a linear operator, the second term's supremum with respect to all such y is $\|\mathcal{L}\|$. Thus for each ϵ the ratio in equation (5) has a supremum too, which in the limit of vanishingly small ϵ must converge to $\|\mathcal{L}\|$.

(Part 2.) Equation (8) is equivalent to

$$f(y) - x_0 \;=\; \mathcal{L}(y - y_0) \;+\; o\left(\|y - y_0\|\right)\,,$$

therefore

$$\|f(y) - x_0\| \;\leq\; \|\mathcal{L}\|\,\|(y - y_0)\| \;+\; o\left(\|y - y_0\|\right)\,.$$

This shows that $\|\mathcal{L}\| = \|\mathcal{D}f(y_0)\| = \chi^{(\mathrm{abs})}(y_0)$ does appear in an error bound of the kind in the theorem.

(Part 3.) Conversely, if

$$\|f(y) - x_0\| \;\leq\; c\,\|y - y_0\| \;+\; o\left(\|y - y_0\|\right)\,,$$

then

$$\frac{\|f(y) - x_0\|}{\|y - y_0\|} \;\leq\; c \;+\; \frac{o\left(\|y - y_0\|\right)}{\|y - y_0\|}\,,$$

so passing to the limit superior as $y \to y_0$ gives,

$$\chi^{(\mathrm{abs})}(y_0) \;\leq\; c \;+\; 0\,,$$

which completes the theorem. ∎

For component-wise relative condition numbers, equation (7)'s bound is scale-invariant owing to the use of relative norms. The scale-invariant error bound for norm-wise relative condition numbers is

$$\frac{\|f(y) - x_0\|}{\|x_0\|} \;\leq\; \chi^{(\mathrm{rel})}(y_0)\,\frac{\|y - y_0\|}{\|y_0\|} \;+\; o\left(\frac{\|y - y_0\|}{\|y_0\|}\right)\,, \tag{9}$$

where equation (6) gives the smallest possible coefficient, $\chi^{(\mathrm{rel})}(y_0)$, in such a bound.

3 Pertinent Real Analysis

This section discusses some real analysis that pertains to sensitivity questions in numerical analysis. The rest of the paper depends on this material only through Theorems 3.3 and 3.5. Readers who are primarily interested in least squares may begin at Section 4.

3.1 Sensitivity Analysis of Metric Projections

Equation (4) reveals that the determination of optimal backward errors is a metric projection from the point y_0 to the set $\mathcal{S}(x) = \{y : F(y, x) = 0\}$. As x varies so does the set $\mathcal{S}(x)$. In particular, the distance, $\mu(x)$, between y_0 and its metric projection (the nearest point in the set) varies near 0 as the approximate solution, x, varies near x_0. Thus, a metric projection's first order sensitivity to its set's deformations and translations gives estimates for the size of optimal backward errors.

It might be thought that sensitivity to perturbation is best studied in terms of derivatives. In many cases of interest these are difficult if not impossible to evaluate directly. Instead, asymptotic analysis suggests the following concept that indirectly leads to more tractable expressions.

Definition 3.1 (Asymptotic Equality, Rational Equivalence) *Suppose the functions f and g are defined on a neighborhood of $x_0 \in \mathbb{R}^n$ and have values in $\mathbb{R}$. The functions are asymptotically equal at x_0 in a rational sense,*

$$f \simeq g\,,$$

when every $\epsilon > 0$ has a neighborhood of x_0 where

$$(1 - \epsilon)g(x) \leq f(x) \leq (1 + \epsilon)g(x)\,.$$

This paper uses three consequences of Definition 3.1 that are easily verified. First, if either of two asymptotically equal functions does not vanish in a deleted neighborhood of the point, x_0, then the other also does not vanish there, and

$$\lim_{x \to x_0} \frac{f(x)}{g(x)} = 1\,. \tag{10}$$

Second, asymptotic equality is an equivalence relation among functions. Third, any function asymptotically equal to equation (4)'s function μ has the same differential properties at x_0 as μ, in the following sense.

Lemma 3.2 (Differential Equivalence) *Suppose F is continuously differentiable and its Jacobian matrix with respect to the first block of variables, evaluated at y_0 and x_0, has full row rank. If equation (4)'s function μ asymptotically equals a function f at x_0, then $\mu - f$ has a vanishing Fréchet derivative there, and consequently*

$$\mu(x) = f(x) + o(\|x - x_0\|)\,.$$

Proof. For the proof see [25, cor. 2.8]. ∎

With this preparation, the following theorem shows that equation (4)'s optimal backward error can be estimated in an asymptotic sense by solving optimization problems with simpler, linear constraints.

Theorem 3.3 (Asymptotic Size of Optimal Backward Errors) *Suppose a numerical problem is defined by a residual function $F(y, x)$ of the data y and solutions x, with the following properties.*

1. *$F : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}^p$ is continuously differentiable.*
2. *There are a y_0 and x_0 with $F(y_0, x_0) = 0$.*
3. *The $p \times m$ partial Jacobian matrix $\mathcal{J}_1 F(y_0, x_0)$ has full row rank.*

   (The notation $\mathcal{J}_1 F(y_0, x_0)$ means the matrix of derivatives with respect to F's first block of variables, y, evaluated at (y_0, x_0).)

Choose any norms for $\mathbb{R}^m$, $\mathbb{R}^n$, and $\mathbb{R}^p$.

$\Rightarrow$ *In the limit as $x \to x_0$, equation (4)'s size of the optimal backward error, $\mu(x)$, asymptotically equals*

$$\begin{aligned} \mu^{(0)}(x) &= \min_{\delta y \,:\, \mathcal{J}_1 F(y_0, x_0)\delta y = F(y_0, x)} \|\delta y\| \\ &= \max_{f \,:\, \|\mathcal{J}_1 F(y_0, x_0)^* f\| \le 1} f(F(y_0, x)) \,. \end{aligned} \tag{11}$$

Definition 3.1 describes what is meant by the asymptotic equality, $\mu \simeq \mu^{(0)}$ at x_0.

Proof. For the proof see [25, thm. 6.1]. The result for 2-norms also can be obtained by specializing from the case of the more general constraints treated by Bonnans and Shapiro [11, p. 434]. ∎

Equation (11)'s optimization problems have an explicit solution for 2-norms.

Corollary 3.4 (2-norm Asymptotic Size) *With the hypotheses and notation of Theorem 3.3, if the norm chosen for $\mathbb{R}^m$ is the Euclidean norm, then the optimal backward error asymptotically equals*

$$\mu^{(0)}(x) = \left\| \left[\mathcal{J}_1 F(y_0, x_0)\right]^\dagger F(y_0, x) \right\|_2 \,, \tag{12}$$

where † is the pseudoinverse.

3.2 Alternate Formula for Condition Numbers

There is a relationship between the size of the optimal backward error and the condition number. Equation (5)'s limit with respect to data, y, can be replaced by a limit with respect to solutions, x.

Theorem 3.5 (Condition Number Formula) *Suppose a numerical problem is defined by a residual function F with the following properties.*

1. *$F : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}^p$ is continuously differentiable.*

2. *There are a y_0 and x_0 with $F(y_0, x_0) = 0$.*

3. *The $p \times m$ partial Jacobian matrix $\mathcal{J}_1 F(y_0, x_0)$ has full row rank.*

   (The first three hypotheses are those of Theorem 3.3.)

4. *The $p \times n$ partial Jacobian matrix $\mathcal{J}_2 F(y_0, x_0)$ has full row rank.*

5. *There is a neighborhood $\mathcal{N}$ of y_0 where each $y \in \mathcal{N}$ has one and only one $x \in \mathbb{R}^n$ that solves $F(y, x) = 0$.*

Choose any norms for $\mathbb{R}^m$, $\mathbb{R}^n$, and $\mathbb{R}^p$.

$\Rightarrow$ *The condition number, in equation (5), is well defined on a neighborhood of y_0.*

(This is stronger than necessary since the remainder of the theorem only evaluates $\chi^{(\text{abs})}$ at y_0.)

$\Rightarrow$ *The size of the optimal backward error, $\mu(x)$ in equation (4), is well defined on a neighborhood of x_0.*

$\Rightarrow$ *Equation (5)'s condition number is given by the size of the optimal backward error,*

$$\chi^{(\text{abs})}(y_0) = \limsup_{x \to x_0} \frac{\|x - x_0\|}{\mu(x)} \,. \tag{13}$$

Proof. The proof depends on finding:

1. a function f of data y so $x = f(y)$ is a solution for y,

2. a function g of x so $y = g(x)$ is data for which x is a solution, and

3. f and g are inverses in the sense that $f(g(x)) = x$.

Each of these establishes one piece of the theorem's three-part conclusion.

(Part 1.) Hypotheses 1, 2, and 4 suffice to invoke the implicit function theorem to obtain a differentiable solution function $f : N_{y_0} \to \mathbb{R}^n$ where N_{y_0} is a neighborhood of y_0 and $f(y_0) = x_0$. Note that the implicit function theorem usually is stated for a nonsingular matrix, so when $\mathcal{J}_2 F(y_0, x_0)$ only has full row rank, then the theorem must be applied to a subset of linearly independent columns. (The columns select a subset of solution variables. Fix the others at their values in x_0, then obtain the selected variables from the implicit function of y, and finally extend the function's range to $\mathbb{R}^n$ by using the fixed values in the remaining coordinates, thereby obtaining f.)

Since $\mathbb{R}^m$ has finite dimension, y_0 has a compact, convex neighborhood in N_{y_0}, and since f is continuously differentiable, it satisfies a Lipschitz condition there. The Lipschitz constant bounds equation (5)'s suprema, which suffices to define χ throughout the subneighborhood's interior. This establishes the theorem's first conclusion.

(Part 2.) Similarly, hypotheses 1, 2, and 3 suffice to invoke the implicit function theorem for a differentiable function $g : N_{x_0} \to \mathbb{R}^m$ where N_{x_0} is a neighborhood of x_0 with $g(x_0) = y_0$ and $F(g(y), x) = 0$ for all $x \in N_{x_0}$. Equation (4) therefore has a nonempty feasible set for each of these x. The sets are closed because F is continuous so the minimum distance from y_0 to each set is attained. This establishes the theorem's second conclusion.

(Part 3.) In whatever space is indicated, let $\mathcal{B}_c(r)$ be the open ball with center c and radius r. The immediate use is to define the limit superior in terms of suprema over balls collapsing to the limit point,

$$\begin{aligned}
\chi^{(\text{abs})}(y_0) &= \limsup_{y \to y_0} \frac{\|f(y) - x_0\|}{\|y - y_0\|} \\
&= \lim_{\epsilon \to 0} \sup_{y \in \mathcal{B}_{y_0}(\epsilon)} \frac{\|f(y) - x_0\|}{\|y - y_0\|} \\
&= \lim_{\epsilon \to 0} \sup_{x \in f(\mathcal{B}_{y_0}(\epsilon))} \sup_{y \in f^{-1}(x) \cap \mathcal{B}_{y_0}(\epsilon)} \frac{\|x - x_0\|}{\|y - y_0\|} \\
&= \lim_{\epsilon \to 0} \sup_{x \in f(\mathcal{B}_{y_0}(\epsilon))} \frac{\|x - x_0\|}{\inf_{y \in f^{-1}(x) \cap \mathcal{B}_{y_0}(\epsilon)} \|y - y_0\|} \,. \qquad (14)
\end{aligned}$$

Without loss of generality these limits can be restricted to $\epsilon < \epsilon_0$ where ϵ_0 is sufficiently small that $\mathcal{B}_{y_0}(\epsilon) \subseteq N_{y_0} \cap \mathcal{N}$.

The next step changes the set from which y is chosen in equation (14). The choice $x \in f(\mathcal{B}_{y_0}(\epsilon))$ means that both the following infema are well defined and no larger than ϵ,

$$\inf_{y \in f^{-1}(x) \cap \mathcal{B}_{y_0}(\epsilon)} \|y - y_0\| \qquad \text{and} \qquad \mu(x) = \inf_{y \,:\, F(y,x) = 0} \|y - y_0\| \,.$$

The second infemum's feasible set contains the first's, so the second's value is a lower bound for the pair. When $\epsilon < \epsilon_0$ then $F(y, x) = 0$ and $\|y - y_0\| < \epsilon$ together imply $y \in \mathcal{B}_{y_0}(\epsilon) \subseteq \mathcal{N}$ so $f(y) = x$ by hypothesis 5 which means y also belongs to the first infemum's feasible set. Thus the two infema are equal for the x and ϵ so chosen. Equation (14) thus becomes,

$$\chi^{(\text{abs})}(y_0) = \lim_{\epsilon \to 0} \sup_{x \in f(\mathcal{B}_{y_0}(\epsilon))} \frac{\|x - x_0\|}{\mu(x)} \,. \qquad (15)$$

The final step changes the set from which x is chosen in equation (15). That Part 2's function g is continuous and $g(x_0) = y_0$ mean, for every $\epsilon_1 > 0$ but no larger than ϵ_0 there is some $\delta_1 > 0$ so $g(\mathcal{B}_{x_0}(\delta_1)) \subseteq \mathcal{B}_{y_0}(\epsilon_1)$. Consider any $x \in \mathcal{B}_{x_0}(\delta_1)$. Since x is a solution for the data $g(x)$ by the choice of g, and since $f(g(x))$ is a solution for the data $g(x)$ by the choice of f, therefore $f(g(x)) = x$ by the choice $g(x) \in \mathcal{B}_{y_0}(\epsilon_1) \subseteq \mathcal{B}_{y_0}(\epsilon_0) \subseteq N_{y_0} \cap \mathcal{N}$ and by hypothesis 5. Altogether,

$$\mathcal{B}_{x_0}(\delta_1) = f(g(\mathcal{B}_{x_0}(\delta_1))) \subseteq f(\mathcal{B}_{y_0}(\epsilon_1))\,,$$

from which follows the inequality,

$$\forall\ \epsilon_1 < \epsilon_0,\ \exists\ \delta_1 :\ \sup_{x\ \in\ f(\mathcal{B}_{y_0}(\epsilon_1))} \frac{\|x - x_0\|}{\mu(x)}\ \geq\ \sup_{x\ \in\ \mathcal{B}_{x_0}(\delta_1)} \frac{\|x - x_0\|}{\mu(x)}\,. \tag{16}$$

Conversely, that Part 1's function f is continuous and $f(y_0) = x_0$ mean, for every $\epsilon_2 > 0$ there is some $\delta_2 > 0$ so that $f(\mathcal{B}_{y_0}(\delta_2)) \subseteq \mathcal{B}_{x_0}(\epsilon_2)$. This gives the reverse inequality,

$$\forall\ \epsilon_2,\ \exists\ \delta_2 :\ \sup_{x\ \in\ f(\mathcal{B}_{y_0}(\delta_2))} \frac{\|x - x_0\|}{\mu(x)}\ \leq\ \sup_{x\ \in\ \mathcal{B}_{x_0}(\epsilon_2)} \frac{\|x - x_0\|}{\mu(x)}\,. \tag{17}$$

The complementary inequalities in equations (16) and (17) allow equation (15)'s x to be chosen from the set $\mathcal{B}_{x_0}(\epsilon)$ without affecting the value of the limit. Thus the derivation continues from equation (15) and concludes as follows,

$$\begin{aligned}
\chi^{(\text{abs})}(y_0) &= \lim_{\epsilon\ \to\ 0}\ \sup_{x\ \in\ \mathcal{B}_{x_0}(\epsilon)} \frac{\|x - x_0\|}{\mu(x)} \\
&= \limsup_{x\ \to\ x_0} \frac{\|x - x_0\|}{\mu(x)}\,.
\end{aligned}$$

∎

Theorems 3.3 and 3.5 combine to express the condition number in terms of the asymptotic estimates for the optimal backward error.

Corollary 3.6 *Under the hypotheses of Theorem 3.5, equation (5)'s condition number is given by equation (11)'s asymptotic expression for the size of the optimal backward error,*

$$\chi^{(\text{abs})}(y_0) = \limsup_{x\ \to\ x_0} \frac{\|x - x_0\|}{\mu^{(0)}(x)}\,. \tag{18}$$

Proof. For the data y_0 with exact solution x_0 and any approximate solution $x \approx x_0$, the size of the optimal backward error, $\mu(x)$, vanishes if and only if x solves the numerical problem for the data y_0. Theorem 3.5's fifth hypothesis says that x_0 is the unique solution for y_0. Thus $x \neq x_0$ implies $\mu(x) \neq 0$. It is therefore possible to invoke Definition 3.1's consequence in equation (10),

$$\lim_{x\ \to\ x_0} \frac{\mu(x)}{\mu^{(0)}(x)} = 1\,.$$

Multiplying this with equation (13) gives the Corollary's equation (18). ∎

3.3 Example of Evaluating Condition Numbers

This section verifies Theorem 3.5 by showing that equation (13) does give the well-known condition number for linear equations. The familiar condition number is the one that supposes only the matrix entries (and not also the right-side vector's entries) are perturbable data.

Problem 3.7 (LE) *Solve $Au = b$ for u, where A is a $p \times n$ matrix.*

Equation (13) requires an expression for the size of optimal backward errors. Choose some norms for $\mathbb{R}^n$ and $\mathbb{R}^p$, and use the corresponding operator norm on $p \times n$ matrices to measure data perturbations. By this measure, for any $x \neq 0$, the size of the optimal backward error is known to be

$$\mu^{(\mathrm{LE})}(x) = \frac{\|Ax - b\|}{\|x\|} \,. \tag{19}$$

This formula was originally derived for ∞-norms by Oettli and Prager [42], and then for arbitrary vector norms and the induced matrix norm by Rigal and Gaches [46]. The proofs in many textbooks are unnecessarily complicated because they simultaneously treat perturbations to b. It is interesting that this formula is valid even if the equations are inconsistent. Moreover, the formula for the spectral matrix norm also minimizes the Frobenius norm of the matrix perturbation because the optimal spectral δA has rank 1.

Before using equations (13) and (19) to evaluate condition numbers, it is necessary to check Theorem 3.5's hypotheses.

1. The residual function is

   $$F^{(\mathrm{LE})}(v(A), x) = Ax - b$$

   where $v(A)$ is Figure 1's function that lists matrix entries in column vector form.

2. The problem should have a solution for the given data. This is the assumption that the linear equations are consistent, so suppose for the data $y_0 = v(A)$ that $Ax_0 = b$ for some x_0.

3. With Figure 1's ordering of matrix entries it is easy to see that

   $$\mathcal{J}_1 F^{(\mathrm{LE})}(v(A), x) = \big[\mathrm{diag}\,(x_1), \mathrm{diag}\,(x_2), \ldots, \mathrm{diag}\,(x_n)\big]$$

   where $\mathrm{diag}\,(x_i)$ is the $n \times n$ diagonal matrix that replicates the i-th entry of x on its diagonal. Therefore the hypothesis that $\mathcal{J}_1 F^{(\mathrm{LE})}(y_0, x_0)$ has full row rank is simply the assumption that $x_0 \neq 0$.

4. It is easy to evaluate $\mathcal{J}_2 F^{(\mathrm{LE})}(y_0, x_0) = A$, so the fourth hypothesis is that A has full row rank.

$$v(A) = \begin{bmatrix} A_{1,1} \\ \vdots \\ A_{1,n} \\ A_{2,1} \\ \vdots \\ A_{2,n} \\ \cdots \\ A_{p,1} \\ \vdots \\ A_{p,n} \end{bmatrix} \in \mathbb{R}^{p \times n} ,$$

Figure 1: *The function v lists the entries of $p \times n$ matrices as column vectors by stacking the columns.*

---

5. Uniqueness for the solution of A's linear equations is the assumption that A has a trivial right null space. In combination with hypothesis 4 this means A must be invertible, which is then true of all sufficiently near matrices.

Thus for linear equations there is considerable overlap among Theorem 3.5's many hypotheses. They reduce to the simple assumptions that A is nonsingular and $b \neq 0$.

With Theorem 3.5's hypotheses satisfied, equations (13) and (19) can be combined to evaluate the condition number.

$$\begin{aligned} \chi^{(\text{LE, abs})}(A) &= \limsup_{x \to x_0} \frac{\|x - x_0\|}{\mu^{(\text{LE})}(x)} \\ &= \limsup_{x \to x_0} \frac{\|x - x_0\| \, \|x_0\|}{\|Ax - b\,\|} \\ &= \limsup_{\delta x \to 0} \frac{\|\delta x\| \, \|x_0\|}{\|A\,\delta x\|} \\ &= \|A^{-1}\| \, \|x_0\| \end{aligned}$$

The relative condition number is therefore the familiar quantity,

$$\chi^{(\text{LE, rel})}(A) = \frac{\|A\|}{\|x_0\|} \chi^{(\text{LE, abs})}(A) = \|A^{-1}\| \, \|A\| \, .$$

This verifies that equation (13) gives the universally recognized value. Theorem 3.5 can be used with confidence to evaluate condition numbers of problems more complicated than linear equations.

4 Optimal Backward Errors for Least Squares

4.1 Literature on Optimal Backward Error

Having completed the preparations of the previous sections, the subject of remainder of the paper is the linear least squares problem (LS).

Problem 4.1 (LS) *Solve*

$$\min_u \|b - Au\|_2$$

where A is a $m \times n$ matrix.

For a given vector x, a backward error is a perturbation matrix, δA, for which the given x exactly solves the perturbed problem,

$$\min_u \|b - (A + \delta A)u\|_2 \,.$$

Stewart [49, p. 6, thm. 3.2] gave formulas for two such δA's in 1977, but the smallest size of backward errors remained open for many years [31, p. 198].

Waldén, Karlson, and Sun [61, p. 273, thm. 2.2] solved the problem in 1995. The smallest Frobenius-norm perturbations, δA, that make a given *nonzero* x into a solution of the perturbed problem were found to have size,

$$\begin{aligned} \mu_F^{(\mathrm{LS})}(x) &= \left(\frac{\|r\|_2^2}{\|x\|_2^2} + \min\{0, \lambda\} \right)^{1/2} \\ \text{for } \lambda &= \lambda_{\min}\left(AA^t - \frac{rr^t}{\|x\|_2^2} \right) , \end{aligned} \tag{20}$$

where $r = b - Ax$ is the approximate least-square residual of the unperturbed problem, and λ is the smallest eigenvalue of the $m \times m$ matrix. Equation (20) is for the Frobenius norm, but the value for the spectral norm is nearby, because Waldén, Karlson, and Sun proved [61, p. 283],

$$\frac{1}{\sqrt{2}}\, \mu_F^{(\mathrm{LS})}(x) \;\le\; \mu_2^{(\mathrm{LS})}(x) \;\le\; \mu_F^{(\mathrm{LS})}(x) \tag{21}$$

(the lower inequality is nontrivial). From equation (20) Higham [32, p. 405] [61, p. 275] derived an equivalent formula,

$$\begin{aligned} \mu_F^{(\mathrm{LS})}(x) &= \min\left\{ \frac{\|r\|_2}{\|x\|_2}\,, \sigma \right\} \\ \text{for } \sigma &= \sigma_{\min}\left(\left[A\,,\, \frac{\|r\|_2}{\|x\|_2}\left(I - \frac{rr^t}{\|r\|_2^2} \right) \right] \right) , \end{aligned} \tag{22}$$

where σ is the smallest of the m singular values of the $m \times (n + m)$ matrix. Similar expressions hold when the vector b also is perturbable [32, p. 404, thm.

19.5] [61, p. 276, cor. 2.1]. The special case $x = 0$ also has a separate formula [61, p. 284, rem. 1].

Table 1 shows that many results followed Waldén, Karlson, and Sun's. Some addressed more complicated problems. Sun [56] showed that equation (20) may apply with the extra requirement that x be the unique, minimal 2-norm solution of the perturbed problem. He also considered problems with multiple right sides [55]. Cox and Higham studied problems with linear constraints. Malyshev [39] obtained a formula for the size of optimal backward errors of problems with spherical constraints

Since equations (20) and (22) evidently are expensive to evaluate, several estimates were proposed. The literature contains almost a dozen bounds which are surveyed here. The only hypothesis for many of them is that their denominators do not vanish.

1. Stewart (1977) first derived bounds of this kind. He stated them for spectral norms, but since they are realized by rank 1 perturbations, they also hold for Frobenius norms. Stewart's "first" bound is [49, p. 6, thm. 3.2],

$$\mu_F^{(\mathrm{LS})}(x) \;\le\; \|\delta A\|_F \;=\; \frac{\|A^t r\|_2}{\|r\|_2} \quad \text{where} \quad \delta A \;=\; -\,\frac{r r^t A}{\|r\|_2^2}\,. \tag{23}$$

2. Stewart's (1977) "second" bound is [49, p. 6, thm. 3.2],

$$\mu_F^{(\mathrm{LS})}(x) \;\le\; \|\delta A\|_F \;=\; \frac{\|r - r_0\|_2}{\|x\|_2} \quad \text{where} \quad \delta A \;=\; \frac{(r - r_0)x^t}{\|x\|_2^2}\,,$$

   in which $r_0 = b - Ax_0$ is the exact least squares residual and $r = b - Ax$ is the approximate residual. From this Stewart concluded, *if r is nearly correct, then the backward error is small* [51, p. 161].

   Stewart's second bound can be restated using $r - r_0 = \mathcal{P}r$, where $\mathcal{P}$ is the orthogonal projection into the column space of A,

$$\mu_F^{(\mathrm{LS})}(x) \;\le\; \|\delta A\|_F \;=\; \frac{\|\mathcal{P}r\|_2}{\|x\|_2} \;=\; \frac{\|Ax - \mathcal{P}b\,\|_2}{\|x\|_2}\,. \tag{24}$$

   In this form the bound and equation (19) show that *the size of the optimal backward error of the linear least squares problem is bounded by that of the consistent linear equations $Au = \mathcal{P}b$.*

3. Waldén, Karlson, and Sun's (1995) formula has been used to obtain a subtle bound. If $Au = b$ is inconsistent, then $r = b - Ax$ is not in the column space of A and it can be shown [61, p. 275] that $\lambda < 0$ in equation (20). In this case,

$$\mu_F^{(\mathrm{LS})}(x) \;=\; \frac{\|r\|_2}{\|x\|_2} - \lambda \;<\; \frac{\|r\|_2}{\|x\|_2} \;=\; \mu_F^{(\mathrm{LE})}(x)\,. \tag{25}$$

(Note the same follows from equation (24) using $\|\mathcal{P}r\|_2 < \|r\|_2$.) *Thus if $Au = b$ is inconsistent, then the nearest least squares problem that x does solve is too near A for x to solve the linear equations, so the problem has a nonzero residual and therefore is still inconsistent* [32, p. 405].

4. Waldén, Karlson, and Sun (1995) also showed [61, p. 279, thm. 3.2],

$$\left[\mu_2^{(\mathrm{LS})}(x)\right]^2 \;\geq\; c\,\frac{(r^tAx)^2}{\|x\|_2^2\left(\|Ax\|_2^2+\|r\|_2^2\right)}$$

where

$$c=\frac{2}{1+\sqrt{1-a}} \quad\text{and}\quad a \;=\; \frac{4\,(r^tAx)^2}{\left(\|Ax\|_2^2+\|r\|_2^2\right)^2}\,.$$

The bound is attained — meaning, it *is* the spectral norm of the optimal backward error — under conditions depending on the eigenvectors for equation (20)'s eigenvalue λ [61, p. 283, cor. 5.1]. As a lower bound it is "often quite good" [61, p. 280], but an example shows it can be zero when the backward error is not. Since $0 \leq a \leq 1$ hence $1 \leq c \leq 2$ so removing c simplifies the bound without much weakening it,

$$\mu_2^{(\mathrm{LS})}(x) \;\geq\; \frac{|\,r^tAx\,|}{\|x\|_2\left(\|Ax\|_2^2+\|r\|_2^2\right)^{1/2}}\,. \tag{26}$$

5. Karlson and Waldén (1997) showed for any v [35, p. 864, eqns. 2.5–6],

$$\mu_2^{(\mathrm{LS})}(x) \;\geq\; (\,2-\sqrt{2}\,)\,\frac{|\,v^tA^t\,r\,|}{\left\|\left(\|r\|_2^2\,I+\|x\|_2^2A^tA\right)^{1/2}v\right\|_2}\,.$$

They examined two cases of this formula: one with $v = A^tr$,

$$\mu_2^{(\mathrm{LS})}(x) \;\geq\; (\,2-\sqrt{2}\,)\,\frac{\|A^tr\|_2^2}{\left\|\left(\|r\|_2^2\,I+\|x\|_2^2A^tA\right)^{1/2}A^tr\right\|_2}\,, \tag{27}$$

and the other for the $v = (\|r\|_2^2\,I + \|x\|_2^2A^tA)^{-1}A^tr$ that maximizes the lower bound, resulting in

$$\mu_2^{(\mathrm{LS})}(x) \;\geq\; (\,2-\sqrt{2}\,)\,\left\|\left(\|r\|_2^2\,I+\|x\|_2^2A^tA\right)^{-1/2}A^tr\right\|_2\,. \tag{28}$$

Tests of these bounds are described in item 6.

6. Karlson and Waldén (1997) also showed how to use a QR factorization of A to convert equation (22) to a minimization over an arbitrary vector. Finding the exact minimum could be costly, so they chose a test vector to give an upper bound for the backward error [35, p. 868, eqn. 3.8].

   In tests Karlson and Waldén found that the ratio between their upper bound and equation (27)'s lower bound was "rather satisfying" [35, p.

868]. However, they also gave a simple example in which the ratio can be arbitrarily large [35, p. 869]. Equation (28) was not tested numerically, but it was very accurate at 1/2 of the optimal value for the simple example where the other bounds performed badly.

7. Gu (1999) derived an estimate that differs from the optimal size by a factor between 1 and the golden ratio [27, p. 365, thm. 2.1],

$$\frac{\sqrt{5}-1}{2}\ \mathcal{G}_1(x)\ \le\ \mu_F^{(\mathrm{LS})}(x)\ \le\ \mathcal{G}_1(x)\,.$$

The estimate $\mathcal{G}_1(x)$ is stated in terms of a singular value decomposition for A. Suppose

$$A = U \begin{bmatrix} \Sigma \\ 0 \end{bmatrix} V^t$$

where U and V are orthogonal matrices, and let

$$\begin{bmatrix} r_1 \\ r_2 \end{bmatrix} = U^t r \quad \text{and} \quad \rho = \mu_2^{(\mathrm{LE})}(x) = \frac{\|r\|_2}{\|x\|_2}\,.$$

In this notation the estimate is

$$\mathcal{G}_1(x)\ =\ \min\left\{\rho\,,\ \left(\frac{r_1^t\Sigma^2(\Sigma^2+\rho^2 I)^{-1}r_1}{\|r_2\|_2\,\rho^{-2}+\rho^2\,r_1^t(\Sigma^2+\rho^2 I)^{-2}\,r_1}\right)^{1/2}\right\}. \qquad (29)$$

Gu assumed that A has full column rank, but the derivation appears to have no restrictions other than A have more rows than columns and the denominators do not vanish. A special case for $x = 0$ is omitted for simplicity and is identical to the one not given for equation (20).

Equation (29) cannot be evaluated in practice because $\|r_2\|_2 = \|r_0\|_2$ is the size of the true least squares residual. An equivalent formula that is computable is given in Section 4.3's equation (35).

8. Gu (1999) derived another estimate that he called a special case of equation (29) but which on inspection is a simplification and weakening of $\mathcal{G}_1(x)$. This estimate is [27, p. 367, cor. 2.2],

$$\mathcal{G}_2(x)\ =\ \frac{\left(r_1^t\Sigma^2(\Sigma^2+\rho^2 I)^{-1}r_1\right)^{1/2}}{\|x\|_2}\,. \qquad (30)$$

where r_1 and ρ are as in equation (29). The bounds provided by this estimate are,

$$\frac{\sqrt{5}-1}{2}\ \mathcal{G}_2(x)\ \le\ \mu_F^{(\mathrm{LS})}(x)\ \le\ \frac{\|r\|_2}{\|r_0\|_2}\ \mathcal{G}_2(x)\,.$$

The upper bound is satisfactory when $r \approx r_0$, so the estimate is better applied to small perturbations, $x \approx x_0$.

9. Malyshev and Sadkane (2002) gave an algorithm [40, pp. 744-745] to evaluate equation (22) using Lanczos bidiagonalization. The calculation is not described here because it does not result in a closed form expression.

The intended use for most of these bounds and estimates is to assess the stability (meaning, the size of the backward error) for an approximate solution of a least squares problem. Cost of evaluation is a consideration in assessing their effectiveness. Table 2 compares their published operation counts with those of the exact formulas. Equations (20) and (22) are evaluated naïvely by forming and solving the dense, $m \times m$ or $m \times (n+m)$ eigenvalue or singular value problems, respectively.

Table 2's operation counts should be seen as preliminary. Waldén, Karlson, and Sun [61, p. 275] and Gu [27, p. 367] suggest (but do not explain how) equation (22) could be evaluated cheaply when a singular value decomposition is available from solving the least squares problem. In this case, Karlson and

Table 2: *Operation counts for evaluating exact formulas (=), upper bounds (↑), lower bounds (↓), and estimates (≈) for the size of the optimal backward error of an $m \times n$ linear least squares problem, $m > n$. The "reference" is to the source of the operation count. Dense matrices are assumed, except for the sparse method of Malyshev and Sadkane. For comparison, solving the least squares problem by either QR or SVD methods uses $\mathcal{O}(mn^2)$ operations [22, p. 177].*

	source	eqn.	type	operation count	reference
	Waldén, Karlson, Sun, 1995 [61]	(20)	=	$\mathcal{O}(m^2n^2) + \mathcal{O}(m^3)$	[22, p. 282]
	Higham, 1995 [61]	(22)	=	$\mathcal{O}(m^3) + \mathcal{O}(m^2n)$	[22, p. 175]
1.	Stewart, 1977 [49]	(23)	↑	$4mn + \mathcal{O}(m)$	
4.	Waldén, Karlson, Sun, 1995 [61]	(26)	↓	$2mn + \mathcal{O}(m)$	
5.	Karlson, Waldén, 1997 [35]	(27)	↓	$6mn + \mathcal{O}(m)$	
5.	Karlson, Waldén, 1997 [35]	(28)	↓	$\mathcal{O}(mn) + \mathcal{O}(n^3)$ given QR of A	[35, p. 864]
6.	Karlson, Waldén, 1997 [35]		↑	$\mathcal{O}(mn)$ given QR of A	[35, p. 865]
7.	Gu, 1999 [27]	(29)	≈	$\mathcal{O}(mn^2)$	[27, p. 367]
9.	Malyshev, Sadkane, 2002 [40]		↑	$\mathcal{O}(\ell\ nz)$ for ℓ iterations, nz nonzeroes	[40, p. 740]

Waldén's upper bound also might be simpler to evaluate. Gu's formulas are in terms of a singular value decomposition, but he did not consider reusing it because he studied fast algorithms for structured least squares problems.

4.2 Asymptotic Size of Optimal Backward Error

This section derives an asymptotic expression for the size of optimal backward errors for linear least squares problems.

The method of analysis is to apply Corollary 3.4's equation (12). It is therefore necessary to check Theorem 3.3's hypotheses before proceeding.

1. It is known that x solves problem (LS) if and only if $A^t(b - Ax) = 0$. Therefore an acceptable residual function is

$$F^{(\mathrm{LS})}(v(A), x) = A^t(b - Ax) \tag{31}$$

   where $v(A)$ is Figure 1's function that lists the matrix entries in column vector form.

2. Suppose that x_0 solves the least squares problem for the matrix A and the vector b. That is, $F^{(\mathrm{LS})}(y_0, x_0) = 0$ where $y_0 = v(A)$.

3. Figure 2 shows the matrix $J = \mathcal{J}_1 F^{(\mathrm{LS})}(v(A), x)$. Let J_0 be this matrix evaluated at A and x_0. In this case, Figure 2's vector r is the exact least squares residual, $r_0 = b - Ax_0$. Since $A^t r_0 = 0$, it is then easy to evaluate

$$J_0 J_0^t \;=\; (r_0^t r_0)\, I + (x_0^t x_0)\, A^t A\,. \tag{32}$$

   Thus J_0 has full row rank exactly when the matrix in equation (32) is nonsingular.

$$
\begin{aligned}
J &= \\
&= \begin{bmatrix}
r_1 - A_{1,1}x_1 & \cdots & r_m - A_{m,1}x_1 & \cdots & \cdots & -A_{1,1}x_n & \cdots & -A_{m,1}x_n \\
-A_{1,2}x_1 & \cdots & -A_{m,2}x_1 & \ddots & & \vdots & & \vdots \\
\vdots & & \vdots & & \ddots & -A_{1,n-1}x_n & \cdots & -A_{m,n-1}x_n \\
-A_{1,n}x_1 & \cdots & -A_{m,n}x_1 & \cdots & \cdots & r_1 - A_{1,n}x_n & \cdots & r_m - A_{m,n}x_n
\end{bmatrix} \\
&= \begin{bmatrix}
r^t & & & \\
& r^t & & \\
& & \ddots & \\
& & & r^t
\end{bmatrix} - \begin{bmatrix} x_1 A^t & x_2 A^t & \cdots & x_m A^t \end{bmatrix}
\end{aligned}
$$

Figure 2: *The Jacobian matrix $J = \mathcal{J}_1 F^{(\mathrm{LS})}(v(A), x) \in \mathbb{R}^{n \times mn}$ of partial derivatives for equation (31)'s residual function with respect to the matrix entries ordered as in Figure 1. The vector r is $b - Ax$.*

When these three hypotheses are satisfied (that is, when equation (32)'s matrix is nonsingular) and $x \approx x_0$, then equation (12) asymptotically equals the size of the optimal backward error for linear least squares problems. Equation (12)'s pseudoinverse is $J_0^t(J_0J_0^t)^{-1}$ of which $J_0^t(J_0J_0^t)^{-1/2}$ is irrelevant to the 2-norm because it has orthonormal columns. Since $\|v(\,\cdot\,)\|_2 = \|\cdot\|_F$, the matrix perturbations are actually measured in the Frobenius norm. All this proves the following theorem.

Theorem 4.2 (Asymptotic Size of Optimal Backward Errors for LS) *Suppose Problem 4.1 (LS) has a solution x_0 and a least squares residual $r_0 = b - Ax_0$ with the following property.*

1. *The matrix $\|r_0\|_2^2\, I + \|x_0\|_2^2\, A^tA$ is nonsingular.*

   (This hypothesis is usually true because most least squares problems are inconsistent, $r_0 \neq 0$.)

Let $\mu_F^{(\mathrm{LS})}(x)$ be the size of the smallest Frobenius-norm perturbations, δA, for which x solves the least squares problem with matrix $A + \delta A$.

$\Rightarrow$ *As x nears x_0, the size of optimal backward error asymptotically equals*

$$\begin{aligned} \mu_F^{(\mathrm{LS},\,0)}(x) &= \left\|(\|r_0\|_2^2\, I + \|x_0\|_2^2\, A^tA)^{-1/2}A^tr\right\|_2 \\ &= \left\|(\|r_0\|_2^2\, I + \|x_0\|_2^2\, A^tA)^{-1/2}A^tA(x_0 - x)\right\|_2 , \end{aligned} \tag{33}$$

where $r = b - Ax$ is the approximate least squares residual. The asymptotic equality is in the sense of Definition 3.1.

For purposes of analysis, equation (33) may be more usefully expressed in terms of singular value decompositions of A. The appropriate canonical decomposition is $A = U\Sigma V^t$ where U and V have orthonormal columns and Σ is a square diagonal matrix of the *nonzero* singular values. Such a decomposition of A does not immediately imply one for equation (32),

$$J_0J_0^t \;=\; \|r_0\|_2^2\, I + \|x_0\|_2^2\, V\Sigma^2V^t\,,$$

because VV^t is not necessarily equal to I. However, equation (33) applies $(J_0J_0^t)^{-1/2}$ only to the column space of V, which is invariant under the transformation. In the basis of V's columns, the transformation is represented by

$$V\left(\|r_0\|_2^2\, I + \|x_0\|_2^2\, \Sigma^2\right)^{-1/2} V^t\,.$$

This gives the following corollary.

Corollary 4.3 (Asymptotic Size of Optimal Backward Errors for LS) *With the hypotheses and notation of Theorem 4.2, let $A = U\Sigma V^t$ be a singular value decomposition where Σ is a square diagonal matrix of A's nonzero singular values. Then*

$$\begin{aligned} \mu_F^{(\mathrm{LS},\,0)}(x) &= \left\|(\|r_0\|_2^2\, I + \|x_0\|_2^2\, \Sigma^2)^{-1/2}\Sigma U^tr\right\|_2 \\ &= \left\|(\|r_0\|_2^2\, I + \|x_0\|_2^2\, \Sigma^2)^{-1/2}\Sigma^2V^t(x_0 - x)\right\|_2 . \end{aligned} \tag{34}$$

4.3 Calculable Asymptotic Estimate

Here, the previous Section's asymptotic formula combines with the literature's results to identify a computable asymptotic expression. Many of the formulas in Section 4.1's survey are restated in Table 3 to emphasize their resemblance to Theorem 4.2's equation (33) for small perturbations.

- Karlson and Waldén's second lower bound, equation (28), and equation (33) are within a constant multiple after the substitutions $r, x \leftrightarrow r_0, x_0$.
- After some difficult manipulations (which are not given here because they apply only to this expression), Gu's first estimate, equation (29), can be restated as

$$\mathcal{G}_1(x) \;=\; \min\left\{ \frac{\|r\|_2}{\|x\|_2},\; \frac{\left\| \left(\|r\|_2^2\, I + \|x\|_2^2 A^t A \right)^{-1/2} A^t r \right\|_2}{\left\| \left(\|r\|_2^2\, I + \|x\|_2^2\, AA^t \right)^{-1} r \right\|_2 \, \|r\|_2} \right\} \tag{35}$$

  If $r \approx r_0$, then $A^t r \approx 0$ so the denominator is

$$\left\| \left(\|r\|_2^2\, I + \|x\|_2^2\, AA^t \right)^{-1} r \right\|_2 \, \|r\|_2 \;\approx\; \left\| \left(\|r\|_2^2\, I \right)^{-1} r \right\|_2 \, \|r\|_2 \;=\; 1\,.$$

  In this way Gu's first estimate and equation (33) are nearly the same under the substitutions $r, x \leftrightarrow r_0, x_0$.
- Gu's second estimate, equation (30), and equation (33) are identical under these substitutions.

Thus the results of Karlson and Waldén, Gu, and Theorem 4.2 suggest that

$$\left\| (\|r\|_2^2\, I + \|x\|_2^2\, A^t A)^{-1/2} A^t r \right\|_2$$

may be an asymptotic estimate for $\mu_F^{(\mathrm{LS})}(x)$. This is proved in the following theorem. Note that this quantity is computable in practice because, unlike equation (33), it does not depend on the true residual and solution, r_0 and x_0.

Theorem 4.4 (Calculable Estimate) *With the hypotheses and notation of Theorem 4.2, and if A, r_0, and x_0 are not zero, then for $x \approx x_0$ the size of the optimal backward error asymptotically equals*

$$\widetilde{\mu}_F^{(\mathrm{LS})}(x) \;=\; \left\| (\|r\|_2^2\, I + \|x\|_2^2\, A^t A)^{-1/2} A^t r \right\|_2 \,,$$

where $r = b - Ax$ is the approximate least squares residual. The asymptotic equality is in the sense of Definition 3.1.

Proof. Let $x = x_0 + \delta x$, so that by the triangle inequality

$$\|x_0\|_2 - \|\delta x\|_2 \;\le\; \|x\|_2 \;\le\; \|x_0\|_2 + \|\delta x\|_2 \,,$$

hence

Table 3: *Bounds and estimates for the size of the optimal backward error in coefficient matrices of linear least squares problems. When a bound permits a choice of norms, then the Frobenius norm is chosen. Red indicates a subexpression common to some bounds.*

	source	eqn.	bound or estimate
1.	Stewart, 1977 [49]	(23)	$\mu_F^{(\mathrm{LS})}(x) \leq \dfrac{\|A^t r\|_2}{\|r\|_2}$
2.	Stewart, 1977 [49]	(24)	$\mu_F^{(\mathrm{LS})}(x) \leq \dfrac{\|\mathcal{P} r\|_2}{\|x\|_2} \leq \dfrac{\|Ax - \mathcal{P} b\|_2}{\|x\|_2}$
3.	Waldén, Karlson, Sun, 1995 [61]	(25)	$\mu_F^{(\mathrm{LS})}(x) < \mu_F^{(\mathrm{LE})}(x)$ when $b \notin \mathrm{col}(A)$
4.	Waldén, Karlson, Sun, 1995 [61]	(26)	$\mu_2^{(\mathrm{LS})}(x) \geq \dfrac{\| r^t A x \|}{\|x\|_2 \left(\|Ax\|_2^2 + \|r\|_2^2\right)^{1/2}}$
5.	Karlson, Waldén, 1997 [35]	(27)	$\mu_2^{(\mathrm{LS})}(x) \geq (2-\sqrt{2}) \dfrac{\|A^t r\|_2^2}{\left\|(\|r\|_2^2 I + \|x\|_2^2 A^t A)^{1/2} A^t r\right\|_2}$
5.	Karlson, Waldén, 1997 [35]	(28)	$\mu_2^{(\mathrm{LS})}(x) \geq (2-\sqrt{2}) \left\|(\|r\|_2^2 I + \|x\|_2^2 A^t A)^{-1/2} A^t r\right\|_2$
7.	Gu, 1999 [27]	(29)	$\dfrac{\sqrt{5}-1}{2} \mathcal{G}_1(x) \leq \mu_F^{(\mathrm{LS})}(x) \leq \mathcal{G}_1(x)$ where $\mathcal{G}_1(x) = \min\left\{ \dfrac{\|r\|_2}{\|x\|_2}, \dfrac{\left\|(\|r\|_2^2 I + \|x\|_2^2 A^t A)^{-1/2} A^t r\right\|_2}{\left\|(\|r\|_2^2 I + \|x\|_2^2 A A^t)^{-1} r\right\|_2 \|r\|_2} \right\}$
8.	Gu, 1999 [27]	(30)	$\dfrac{\sqrt{5}-1}{2} \mathcal{G}_2(x) \leq \mu_F^{(\mathrm{LS})}(x) \leq \dfrac{\|r\|_2}{\|r_0\|_2} \mathcal{G}_2(x)$ where $\mathcal{G}_2(x) = \left\|(\|r\|_2^2 I + \|x\|_2^2 A^t A)^{-1/2} A^t r\right\|_2$

$$(1-\epsilon)\,\|x_0\|_2 \;\le\; \|x\|_2 \;\le\; (1+\epsilon)\,\|x_0\|_2$$

whenever $\epsilon \ge \|\delta x\|_2/\|x_0\|_2$. Similarly $r = r_0 - A\,\delta x$, so

$$(1-\epsilon)\,\|r_0\|_2 \;\le\; \|r\|_2 \;\le\; (1+\epsilon)\,\|r_0\|_2$$

provided $\epsilon \ge \|A\,\delta x\|_2/\|r_0\|_2$. If additionally $\epsilon < 1$, then the inequalities are preserved by the following arithmetic steps: squaring them, multiplying the first by σ^2, adding them, taking the square root. This leaves,

$$\begin{aligned}(1-\epsilon)\,(\|r_0\|_2^2 + \|x_0\|_2^2\,\sigma^2)^{1/2} \;\le\; & \\ (\|r\|_2^2 + \|x\|_2^2\,\sigma^2)^{1/2} & \\ \le\; (1+\epsilon)\,(\|r_0\|_2^2 + \|x_0\|_2^2\,\sigma^2)^{1/2} &\end{aligned}$$

which rearranges to,

$$\begin{aligned}(1-\epsilon)\,(\|r\|_2^2 + \|x\|_2^2\,\sigma^2)^{-1/2} \;\le\; & \\ (\|r_0\|_2^2 + \|x_0\|_2^2\,\sigma^2)^{-1/2} & \\ \le\; (1+\epsilon)\,(\|r\|_2^2 + \|x\|_2^2\,\sigma^2)^{-1/2}\,. &\end{aligned} \tag{36}$$

Recall from equation (34) that,

$$\mu_F^{(\mathrm{LS},\,0)}(x) \;=\; \left\|(\|r_0\|_2^2\,I + \|x_0\|_2^2\,\Sigma^2)^{-1/2}\,v\right\|_2\,,$$

and for $\widetilde{\mu}^{(\mathrm{LS})}(x)$ similarly,

$$\widetilde{\mu}_F^{(\mathrm{LS})}(x) \;=\; \left\|(\|r\|_2^2\,I + \|x\|_2^2\,\Sigma^2)^{-1/2}\,v\right\|_2\,,$$

where $v = \Sigma U^t r$, and $A = U\Sigma V^t$ is Corollary 4.3's singular value decomposition. The operators in these equations are diagonal, so applying them to v and forming the 2-norm gives, from equation (36),

$$(1-\epsilon)\,\widetilde{\mu}_F^{(\mathrm{LS})}(x) \;\le\; \mu_F^{(\mathrm{LS},\,0)}(x) \;\le\; (1+\epsilon)\,\widetilde{\mu}_F^{(\mathrm{LS})}(x)\,.$$

Thus, if $1 > \epsilon > 0$, then Definition 3.1's inequality holds for $x = x_0 + \delta x$ with $\|\delta x\|_2 \le \epsilon \min\{\|x_0\|_2, \|r_0\|_2/\|A\|_2\}$. This proves that $\widetilde{\mu}_F^{(\mathrm{LS})} \simeq \mu_F^{(\mathrm{LS},\,0)}$ at x_0 in Definition 3.1's notation. Since asymptotic equality is an equivalence relation and $\mu_F^{(\mathrm{LS},\,0)} \simeq \mu_F^{(\mathrm{LS})}$ at x_0 by Theorem 3.3, therefore $\widetilde{\mu}_F^{(\mathrm{LS})} \simeq \mu_F^{(\mathrm{LS})}$. ∎

5 Condition Numbers for Least Squares

5.1 Literature on Error Bounds

The literature contains over one dozen error bounds for linear least squares problems. They and their hypotheses are stated here in roughly their original

forms. In this paper's notation, the bounds suppose that x_0 and $x = x_0 + \delta x$ solve, respectively,

$$\min_u \|b - Au\|_2 \quad \text{and} \quad \min_u \|(b + \delta b) - (A + \delta A)u\|_2 \,.$$

The bounds for δx usually are in terms of A, δA, b, δb, x, x_0, $r = b - Ax$, and $r_0 = b - Ax_0$. An unstated assumption is that $x_0 \neq 0$ so it is meaningful to discuss the relative error, $\|\delta x\|_2 / \|x_0\|_2$.

1. Golub and Wilkinson (1966) were the first to consider bounds of this kind. Their assumptions were,

   (a) A has full column rank,

   (b) $\|A\|_2 = 1$ and $\|b\|_2 = 1$,

   (c) $\|\delta A\|_2 = \varepsilon$ and $\|\delta b\|_2 = \varepsilon$,

   (d) ε is "arbitrarily small",

   from which they derived [24, p. 144, eqn. 43],

   $$\|\delta x\|_2 \;\leq\; \varepsilon\,\kappa_2 + \varepsilon\,\kappa_2\,\|x_0\|_2 + \varepsilon\,\kappa_2^2\,\|r_0\|_2 + \mathcal{O}(\varepsilon^2)\,. \tag{37}$$

   This error bound initiated a tradition whereby κ_2^2 explicitly appears in error bounds for least squares problems. The bound is inapplicable to most problems due to the very restrictive hypothesis (1b). This can be removed by transforming the error bound through the following steps, after which the dependence on κ_2^2 is not evident.

   - First, an examination of Golub and Wilkinson's derivation finds that their error bound's first term accounts for perturbations to b while the other terms account for perturbations to A. Grouping the terms thusly reveals that the actual error bound is

     $$\|\delta x\|_2 \;\leq\; \|\delta b\|_2\,\kappa_2 + \|\delta A\|_2\,\kappa_2\,(\|x_0\|_2 + \kappa_2\,\|r_0\|_2) + \mathcal{O}(\varepsilon^2)\,,$$

     where now $\varepsilon = \max\{\|\delta A\|_2, \|\delta b\|_2\}$.

   - Second, hypothesis (1b) means that κ_2 in Golub and Wilkinson's error bound is actually $1/\sigma_{\min}(A)$. With this substitution the bound becomes

     $$\|\delta x\|_2 \;\leq\; \frac{\|\delta b\|_2}{\sigma_{\min}(A)} + \frac{\|\delta A\|_2}{\sigma_{\min}(A)}\left(\|x_0\|_2 + \frac{\|r_0\|_2}{\sigma_{\min}(A)}\right) + \mathcal{O}(\varepsilon^2)\,. \tag{38}$$

   - Third, a general least squares problem

     $$\min_{\widetilde{x}} \|\widetilde{b} - \widetilde{A}\widetilde{x}\|_2$$

     can be scaled to satisfy hypotheses (1b) in just one way,

$$\min_{\tilde{x}/c_2} \left\| (c_1\tilde{b}) - (c_2 c_1 \widetilde{A})(\tilde{x}/c_2) \right\|_2 \quad \text{where} \quad c_1 = \frac{1}{\|\tilde{b}\|_2}\,, \quad c_2 = \frac{1}{\|c_1\widetilde{A}\|_2}\,.$$

Equation (38) applies to the scaled problem with the substitutions,

$$\begin{array}{c} \text{term in equation (38)} \mapsto \text{term in scaled problem} \\ \begin{array}{lll} A \;\mapsto\; c_1 c_2\, \widetilde{A} & r_0 \;\mapsto\; c_1\, \tilde{r}_0 & x_0 \;\mapsto\; \tilde{x}_0/c_2 \\ \delta A \;\mapsto\; c_1 c_2\, \delta\widetilde{A} & \delta b \;\mapsto\; c_1\, \delta\tilde{b} & \delta x \;\mapsto\; \delta\tilde{x}/c_2 \end{array} \end{array}$$

which result in the following bound,

$$\begin{aligned} \frac{\|\delta\tilde{x}\|_2}{c_2} \;&\leq\; \frac{\|c_1\,\delta\tilde{b}\|_2}{\sigma_{\min}(c_1 c_2\,\widetilde{A})} + \frac{\|c_1 c_2\,\delta\widetilde{A}\|_2}{\sigma_{\min}(c_1 c_2\,\widetilde{A})} \left(\frac{\|\tilde{x}_0\|_2}{c_2} + \frac{\|c_1\,\tilde{r}_0\|_2}{\sigma_{\min}(c_1 c_2\,\widetilde{A})} \right) \\ &+\; \mathcal{O}(\varepsilon^2)\,. \end{aligned}$$

- Finally, removing the common denominator c_2, cancelling c_1 from numerators and denominators, and discarding the tildes give,

$$\|\delta x\|_2 \;\leq\; \frac{\|\delta b\|_2}{\sigma_{\min}} + \frac{\|\delta A\|_2}{\sigma_{\min}} \left(\|x_0\|_2 + \frac{\|r_0\|_2}{\sigma_{\min}} \right) + \frac{\|b\|_2}{\|A\|_2}\,\mathcal{O}(\varepsilon^2)\,.$$

Thus Golub and Wilkinson's bound when applied to a general least squares problem is

$$\frac{\|\delta x\|_2}{\|x_0\|_2} \;\leq\; \frac{\|\delta b\|_2}{\|b\|_2}\,\frac{\|b\|_2}{\|x_0\|_2\,\sigma_{\min}} + \frac{\|\delta A\|_2}{\|A\|_2} \left(1 + \frac{\|r_0\|_2}{\|x_0\|_2\,\sigma_{\min}} \right) \kappa_2 + \mathcal{O}(\varepsilon^2)\,, \quad (39)$$

for which it is assumed,

(a) A has full column rank,
(b) $\|\delta A\|_2/\|A\|_2 \leq \varepsilon$ and $\|\delta b\|_2/\|b\|_2 \leq \varepsilon$, and
(c) ε is arbitrarily small.

2. Björck (1967) credits Golub for suggesting that a bound could be derived from the augmented system,

$$\begin{bmatrix} r_0 + Ax_0 \\ A^t r_0 \end{bmatrix} = \begin{bmatrix} I & A \\ A^t & \end{bmatrix} \begin{bmatrix} r_0 \\ x_0 \end{bmatrix} = \begin{bmatrix} b \\ \end{bmatrix}\,. \quad (40)$$

Björck assumed,

(a) A and $A + \delta A$ have full column rank,
(b) $\alpha = (\sqrt{2}+1)\,\|\delta A\|_2/\sigma_{\min} < 1$,

from which he obtained [7, p. 17],

$$\begin{aligned} \frac{\|\delta x\|_2}{\|x_0\|_2} \;\leq\; & \frac{\kappa_2}{\sqrt{1-\alpha}} \left(1 + \frac{\kappa_2}{\sqrt{1-\alpha}}\,\frac{\|r_0\|_2}{\|A\|_2\,\|x_0\|_2} \right) \frac{\|\delta A\|_2}{\|A\|_2} \\ & + \frac{\kappa_2}{\sqrt{1-\alpha}}\,\frac{\|\delta b\|_2}{\|A\|_2\,\|x_0\|_2}\,. \end{aligned} \quad (41)$$

3. Hanson and Lawson (1969) assumed

   (a) A has full column rank,

   (b) $\varepsilon\,\kappa_2 < 1$,

   where $\varepsilon = \|\delta A\|_2/\|A\|_2$. From this they derived that $A+\delta A$ has full column rank [29, p. 794, thm. 2.3.2], and then [29, p. 797, eqn. 2.4.10],

$$\begin{aligned}\frac{\|\delta x\|_2}{\|x_0\|_2} \;\le\; & \frac{\kappa_2}{1-\varepsilon\,\kappa_2}\frac{\|\delta b\|_2}{\|Ax_0\|_2} \\ & + \frac{\varepsilon\,\kappa_2}{1-\varepsilon\,\kappa_2}\left(1+\frac{\kappa_2}{1-\varepsilon\,\kappa_2}\,\frac{\|r_0\|_2}{\|Ax_0\|_2}\right).\end{aligned} \tag{42}$$

4. Pereyra (1969) assumed, in the notation of this paper,

   (a) A has full rank,

   (b) $\beta = \kappa_2\dfrac{\|\delta A\|}{\|A\|}(1+\kappa_2\dfrac{\|\delta A\|}{\|A\|}+\kappa_2) < 1$,

   (c) $\delta b = 0$.

   From assumption (4b) he derived that $A+\delta A$ has the same rank as A [44, p. 199, lem. 4.2], and then [44, p. 200, eqn. 4.8],

$$\frac{\|\delta x\|}{\|x_0\|} \;\le\; \frac{1}{1-\beta}\left(\beta+\kappa_2\,\frac{\|\delta A\|\,\|A^\dagger\|\,\|b\|}{\|A\|\,\|A^\dagger b\|}\right), \tag{43}$$

   in which all norms can be taken to be 2-norms.

5. Stoer (1972) assumed,

   (a) A and $A+\delta A$ have full column rank,

   (b) δA is small enough that $[(A+\delta A)^t(A+\delta A)]^{-1}$ can be approximated using $(I+E)^{-1} \approx I-E$ where $E = (A^tA)^{-1}(A^t\delta A+\delta A^tA)$,

   from which he derived [52, p. 176, eqn. 4.8.3.5] [53, p. 212, eqn. 4.8.3.5], in the notation of this paper,

$$\frac{\|\delta x\|_2}{\|x_0\|_2} \;\dot{\le}\; \kappa_2\left(1+\kappa_2\,\frac{\|r_0\|_2}{\sigma_{\max}\,\|x_0\|_2}\right)\frac{\|\delta A\|_2}{\|A\|_2}+\kappa_2\,\frac{\|\delta b\|_2}{\sigma_{\min}\,\|x_0\|_2}. \tag{44}$$

6. Wedin (1973) assumed,

   (a) A and $A+\delta A$ have equal rank,

   (b) $\|\delta A\|_2 \le \varepsilon\|A\|_2$,

   (c) $\varepsilon\,\kappa_2 < 1$,

   (d) $x_0 = A^\dagger b$ and $x = x_0+\delta x = (A+\delta A)^\dagger(b+\delta b)$,

where † is pseudoinverse, from which he obtained [62, p. 224, thm. 5.1],

$$\begin{aligned}\frac{\|\delta x\|_2}{\|x_0\|_2} \;\le\; & \frac{\kappa_2}{1-\varepsilon\,\kappa_2}\left\{\varepsilon + \frac{\|\delta b\|_2}{\|A\|_2\,\|x_0\|_2} + \frac{\varepsilon\,\kappa_2\,\|r_0\|_2}{\|A\|_2\,\|x_0\|_2}\right\} \\ & + \frac{\varepsilon\,\|(AA^t)^\dagger b\|_2\,\|A\|_2}{\|x_0\|_2}\,.\end{aligned} \tag{45}$$

The final term in equation (45) can be discarded when A has full column rank.

7. Abdelmalek (1974) assumed,

   (a) A and $A+\delta A$ have equal rank,
   (b) $\|A^\dagger\|_2\,\|\delta A\|_2 < 1$,

   from which he derived [1, p. 222, eqn. 42],

$$\begin{aligned}\frac{\|\delta x\|}{\|x_0\|} \;\le\; & \frac{1}{\|x_0\|}\left[\frac{\|A^\dagger\|\,\|\delta b\|}{1-\|A^\dagger\|\,\|\delta A\|}\right. \\ & + \frac{\|A^\dagger\|}{1-\|A^\dagger\|\,\|E_1\|}\left(\sqrt{2}\,\|E_1\|\,\|x_0\| + \frac{\|A^\dagger\|\,\|E_2\|\,\|r_0\|}{1-\|A^\dagger\|\,\|E_1\|}\right) \\ & \left. + \frac{\|A^\dagger\|^2\,\|E_2\|\,\|\delta A\|}{(1-\|A^\dagger\|\,\|\delta A\|)^2}\left(\sqrt{2}\,\|x_0\| + \frac{\|A^\dagger\|\,\|r_0\|}{1-\|A^\dagger\|\,\|\delta A\|}\right)\right]\,,\end{aligned} \tag{46}$$

   where all norms are 2-norms, $E_1 = \mathcal{P}\,\delta A$, $E_2 = (I-\mathcal{P})\,\delta A$, and $\mathcal{P}$ is the orthogonal projection into the column space of A.

8. Lawson and Hanson (1974) assumed,

   (a) A has full column rank,
   (b) $\varepsilon\,\kappa_2 < 1$,

   where $\varepsilon = \|\delta A\|_2/\|A\|_2$. From this they derived [36, p. 51, eqn. 9.14],

$$\frac{\|\delta x\|_2}{\|x_0\|_2} \;\le\; \frac{\kappa_2}{1-\varepsilon\,\kappa_2}\left[\left(1 + \frac{\kappa_2\,\|r_0\|_2}{\|A\|_2\,\|x_0\|_2}\right)\varepsilon + \frac{\|\delta b\|_2}{\|A\|_2\,\|x_0\|_2}\right]\,. \tag{47}$$

9. Van der Sluis (1975) assumed,

   (a) A has full column rank,
   (b) $\|\delta A\|_2 \le \varepsilon\|A\|_2$ and $\|\delta b\|_2 \le \varepsilon\|b\|_2$,
   (c) $\varepsilon\,\kappa_2 < 1$,

   from which he obtained [59, p, 251, eqn. 5.8],

$$\begin{aligned}\frac{\|\delta x\|_2}{\|x_0\|_2} \;\le\; \varepsilon \Bigg(& \frac{\sigma_{\max}}{\sigma_{\min}} \frac{R(x_0)}{\sigma_{\min}} \frac{\tan(\theta)}{1-(\varepsilon\,\kappa_2)^2} + \frac{\sigma_{\max}}{\sigma_{\min}} \frac{1}{1-\varepsilon\,\kappa_2} \\ & + \frac{R(x_0)}{\sigma_{\min}} \frac{1}{\cos(\theta)} \frac{1}{1-\varepsilon\,\kappa_2} \Bigg)\,,\end{aligned} \tag{48}$$

where $R(x_0) = \|Ax_0\|_2/\|x_0\|_2$, and θ is the angle between b and the column space of A.

10. Stewart (1977) assumed,

    (a) δA is an acute perturbation of A.

    This condition appears to be due to Wedin [62, p. 228]. For purposes of comparison with other bounds, it suffices to add the assumption,

    (b) A has full column rank,

    which gives a simple meaning to assumption (10a). Consider a singular value decomposition,

$$A \;=\; U \begin{bmatrix} \Sigma \\ 0 \end{bmatrix} V^t\,,$$

    where U and V are orthogonal (square) matrices, and let

$$\delta A \;=\; U \begin{bmatrix} E_1 \\ E_2 \end{bmatrix} V^t\,.$$

    The perturbation is acute if and only if $A_1 = \Sigma + E_1$ is nonsingular [51, p. 139, thm. 3.3]. With these assumptions and notation, and with some condensation and correction,[2] Stewart's error bound can be stated as [51, p. 157, thm. 5.2],

$$\frac{\|\delta x\|_2}{\|x_0\|_2} \;\le\; \|A_1^{-1}\|_2\,\|E_1\|_2 + \|A_1^{-1}\|_2^2 \left(\frac{\|E_2\|_2\,\|r_0\|_2}{\|x_0\|_2} + \|E_2\|_2^2 \right)\,. \tag{49}$$

11. Golub and Van Loan (1983) derived a bound under van der Sluis's assumptions. They assumed,

    (a) A has full column rank,
    (b) $\|\delta A\|_2 \le \varepsilon \|A\|_2$ and $\|\delta b\|_2 \le \varepsilon \|b\|_2$,
    (c) $\varepsilon\,\kappa_2 < 1$,

    from which they showed [22, p. 141, eqn. 6.1-10] [23, p. 228, eqn. 5.3.8],

$$\frac{\|\delta x\|_2}{\|x_0\|_2} \;\le\; \varepsilon \left(\frac{2\,\kappa_2}{\cos(\theta)} + \tan(\theta)\,\kappa_2^2 \right) + \mathcal{O}(\varepsilon^2)\,, \tag{50}$$

    where θ is the angle between b and the column space of A. This bound is recommended by LAPACK [2, p. 50].

[2] The bound as originally stated contains a typographical error: the second occurrence of E_{12} should be E_{21} in [48, p. 654, eqn. 5.4] [51, p. 157, eqn. 5.3].

12. Arioli, Duff, and de Rijk (1989) assumed that each perturbation is small compared to the corresponding entry of A or b,

    (a) A has full column rank,

    (b) $|\delta A| \le \varepsilon\,|A|$ and $|\delta b| \le \varepsilon\,|b|$,

    where, in (12b), the notation $|\cdot|$ applied to a matrix or vector is the like object whose entries are the other's magnitudes, and the inequalities apply entry-by-entry. From these they derived [3, p. 673, eqn. 3.13],

$$|\delta x| \;\le\; \varepsilon\,|A^\dagger|\,(|A|\,|x| + |b|) + \varepsilon\,|\,(A^tA)^{-1}|\,|A^t|\,|r|\,. \tag{51}$$

13. Björck (1989) assumed that each perturbation is small compared to the corresponding entry of a reference matrix and vector, E and f,

    (a) A has full column rank,

    (b) $|\delta A| \le \varepsilon E$ and $|\delta b| \le \varepsilon f$,

    (c) $\varepsilon\,\rho\left(\begin{bmatrix} |A^\dagger|^t\,E^t & |I-\mathcal{P}|\,E \\ |(A^tA)^{-1}|E^t & |A^\dagger|\,E \end{bmatrix}\right) < 1$,

    where the notation $|\cdot|$ has the meaning of assumption (12b), and $\rho(\cdots)$ is the spectral radius of the matrix inside the parentheses. From these he derived [8] [9, p. 240, eqn. 2.5],

$$|\delta x| \;\le\; \varepsilon\,|A^\dagger|\,(f + E\,|x_0|) + \varepsilon\,|\,(A^tA)^{-1}|\,E^t\,|r_0| + \mathcal{O}(\varepsilon^2)\,. \tag{52}$$

14. Higham (1990) assumed [32, p. 394],

    (a) A and $A + \delta A$ have full column rank,

    (b) $|\delta A| \le \varepsilon E$ and $|\delta b| \le \varepsilon f$,

    from which he derived [31, p. 203, eqn. 3.5],

$$|\delta x| \;\le\; \varepsilon\,|A^\dagger|\,(f + E\,|x|) + \varepsilon\,|\,(A^tA)^{-1}|\,E^t\,|r|\,. \tag{53}$$

    The bound can be made relative by applying any absolute norm and dividing, for example [32, p. 394],

$$\frac{\|\delta x\|}{\|x_0\|} \;\le\; \frac{\varepsilon\,\big\||A^\dagger|\,(f + E\,|x|)\big\| + \varepsilon\,\big\|\,|\,(A^tA)^{-1}|\,E^t\,|r|\big\|}{\|x_0\|}\,. \tag{54}$$

Higham [31, p. 202] credits Arioli, Duff, and de Rijk [3] and Björck [8] [9] for independently originating the componentwise bounds, 12 through 14. The difference among them is that Arioli, Duff, and de Rijk's equation (51) uses the approximate r and x, while Björck's equation (52) has the exact r_0 and x_0 as well as E, f, and an $\mathcal{O}(\varepsilon^2)$ term. Note that r_0 and x_0 seem to entail the $\mathcal{O}(\varepsilon^2)$

discrepancy. Higham's equation (53) combines the r-x and E-f versions. The distinction between equations (51) and (52) sometimes is overlooked as when Stewart and Sun [51, p. 163] attribute Higham's version to Björck.

Componentwise bounds have several advantages. They appear to be sharper than normwise bounds [31, p. 203], and with $E = |A|$ and $f = |b|$ they are less sensitive to row and column scalings of the matrix [31, p. 203] [32, p. 394]. Of course, the componentwise bounds incorporate the matrix's nonzero structure [51, p. 158]. Componentwise bounds for linear least squares problems were motivated by Oettli and Prager [42], whose structured backward errors for linear equations were applied to the augmented equation (40), and by Bunch [12], who suggested structured perturbation analyses.

15. Wei (1990) assumed,

    (a) $\kappa_2 \, \varepsilon \leq 1 - 1/\sqrt{2}$,

    where $\varepsilon = \|\delta A\|_2 / \|A\|_2$. For a given solution x of the problem perturbed by δA and δb, Wei showed the unperturbed problem has a solution $x_0 = x - \delta x$ with [64, p. 180, eqn. 2.7],

$$\begin{aligned} \frac{\|\delta x\|_2}{\|x_0\|_2} \;\leq\; & \frac{\kappa_2}{1 - 2\,\kappa_2\,\varepsilon} \left(2\,\kappa_2\,\varepsilon \, \frac{\|r_0\|_2}{\|A\|_2\,\|x_0\|_2} + 2\,\varepsilon + \frac{\|\delta b\|_2}{\|A\|_2\,\|x_0\|_2} \right) \\ & + 2\,\kappa_2\,\varepsilon + \frac{2\,\kappa_2\,\varepsilon}{1 - \sqrt{2}\,\kappa_2\,\varepsilon} \, \frac{\|x\|_2}{\|x_0\|_2} \, . \end{aligned} \tag{55}$$

Wei's bound combines generality and simplicity. All the bounds listed here are derived from perturbation analyses of either the least squares problem or the pseudoinverse. The former tend to give bounds only for full rank problems; the latter tend to give bounds in terms of decompositions of the perturbation matrix. In contrast, equation (55) applies even to rank deficient problems yet it depends on the perturbation matrix only through its norm.

16. Higham (1996) reworked the proof of Wedin's bound, number 6. He used van der Sluis's assumptions to obtain a slightly different result that he graciously attributed to Wedin.[3] Higham assumed,

    (a) A and $A + \delta A$ have full column rank,
    (b) $\|\delta A\|_2 \leq \varepsilon \|A\|_2$ and $\|\delta b\|_2 \leq \varepsilon \|b\|_2$,
    (c) $\varepsilon \, \kappa_2 < 1$,

    from which he obtained [32, p. 392],

$$\frac{\|\delta x\|_2}{\|x_0\|_2} \;\leq\; \frac{\varepsilon\,\kappa_2}{1 - \varepsilon\,\kappa_2} \left\{ 2 + (1 + \kappa_2) \, \frac{\|r_0\|_2}{\|A\|_2\,\|x_0\|_2} \right\} . \tag{56}$$

    Note the hypotheses about $A + \delta A$ is implied by the others.

[3] This and the prominence of [32] as a desk reference have led to some confusion. At least one textbook refers to the bound as Wedin's even though it cannot be found in his work.

It is revealing to examine the literature's error bounds in equation (9)'s context of small matrix perturbations. Table 4 shows the leading terms of all the bounds that can be expanded in power series of $\|\delta A\|_2/\|A\|_2$. How similar they become! Portions of seven bounds colored blue are the same. Of these, the bounds of Björck, Stoer, Wedin, and Lawson and Hanson are identical (for small perturbations) and the simplest.[4] Golub and Wilkinson's bound when it is applied to appropriately scaled least squares problems also belongs to this group, but Table 4 does not include equation (39) because it does not actually appear in [24].

From Theorem 2.1, the common expression found in Table 4 is an upper bound for $\chi_2^{(\text{LS, rel})}(A)$. Whether

$$\frac{\|r_0\|_2\,\sigma_{\max}}{\|x_0\|_2\,\sigma_{\min}^2} + \frac{\sigma_{\max}}{\sigma_{\min}} \;=\; \left(\frac{\|r_0\|_2}{\|x_0\|_2\,\sigma_{\min}} + 1\right)\kappa_2 \tag{57}$$

is attained as the relative spectral condition number has not been determined in the literature. Two approaches have been taken.

- Algebraic reasoning can be used to derive an expression for δx in terms of δA. For error bound 5, Stoer showed in the notation of this paper that [52, p. 176] [53, p. 211],

$$\delta x \;=\; (A^tA)^{-1}\,\delta A^t\,r_0 - (A^tA)^{-1}\,A^t\,\delta A\,x_0 + o(\delta A^2)\,. \tag{58}$$

  Taking norms and dividing by $\|x_0\|_2$ after some manipulation produces an error bound in which the first two terms, above, become the respective terms in equation (57)'s coefficient for $\|\delta A\|_2/\|A\|_2$. Yet it is unclear how to choose δA so that δx attains the bound. A similar expression for δx but in terms of pseudoinverses occurs in Wedin's analysis leading to error bound 6 [62, p. 224].

- Van der Sluis, error bound 9, used geometric reasoning to derive upper and lower bounds on the attainable solution error. In the notation of this paper, these bounds are [59, p. 251, eqns. 5.8–9],

$$\frac{\|\delta x\|_2}{\|x\|_2} \quad \begin{cases} \;\leq\; \varepsilon\left(\dfrac{\|r_0\|_2\,\sigma_{\max}}{\|x_0\|_2\,\sigma_{\min}^2} + \dfrac{\sigma_{\max}}{\sigma_{\min}} + \dfrac{\|b\|_2}{\|x_0\|_2\,\sigma_{\min}}\right) + \mathcal{O}(\varepsilon^2) \\ \;\geq\; \varepsilon\left(\dfrac{\|r_0\|_2\,\sigma_{\max}}{\|x_0\|_2\,\sigma_{\min}^2} + \dfrac{\|b\|_2}{\|x_0\|_2\,\sigma_{\min}}\right) + \mathcal{O}(\varepsilon^2) \end{cases},$$

  where ε is the maximum of the normwise relative perturbations to A and b. The lower bound unfortunately is too weak to determine the condition number exactly [59, p. 250, rem. 5.2].

[4] The bounds of van der Sluis and Higham-Wedin have extra terms that account for b's perturbations if they are present. Wei's bound has an unusual extra term that can be removed by multiplying the whole bound by a small factor that is mentioned but not given in [64].

Table 4: *Leading terms in expansions with respect to* $\|\delta A\|_2/\|A\|_2$ *for several error bounds listed in Section 5.1. The few bounds that have separate notation for perturbations to* b *are evaluated with* $\delta b = 0$. *Blue indicates the consensus evidenced by seven bounds.*

	source of bound	eqn.	leading term with respect to $\|\delta A\|_2/\|A\|_2$
2.	Björck, 1967 [7]	(41)	$\frac{\|r_0\|_2\,\sigma_{\max}}{\|x_0\|_2\,\sigma_{\min}^2} + \frac{\sigma_{\max}}{\sigma_{\min}}$
3.	Hanson and Lawson, 1969 [29]	(42)	$\frac{\|r_0\|_2\,\sigma_{\max}^2}{\|Ax_0\|_2\,\sigma_{\min}^2} + \frac{\sigma_{\max}}{\sigma_{\min}}$
4.	Pereyra, 1969 [44]	(43)	$\frac{\|b\|_2\,\sigma_{\max}}{\|Ax_0\|_2\,\sigma_{\min}} + \frac{\sigma_{\max}^2}{\sigma_{\min}^2} + \frac{\sigma_{\max}}{\sigma_{\min}}$
5.	Stoer, 1972 [52]	(44)	$\frac{\|r_0\|_2\,\sigma_{\max}}{\|x_0\|_2\,\sigma_{\min}^2} + \frac{\sigma_{\max}}{\sigma_{\min}}$
6.	Wedin, 1973 [62]	(45)	$\frac{\|r_0\|_2\,\sigma_{\max}}{\|x_0\|_2\,\sigma_{\min}^2} + \frac{\sigma_{\max}}{\sigma_{\min}}$
8.	Lawson and Hanson, 1974 [7]	(47)	$\frac{\|r_0\|_2\,\sigma_{\max}}{\|x_0\|_2\,\sigma_{\min}^2} + \frac{\sigma_{\max}}{\sigma_{\min}}$
9.	van der Sluis, 1975 [59]	(48)	$\frac{\|r_0\|_2\,\sigma_{\max}}{\|x_0\|_2\,\sigma_{\min}^2} + \frac{\sigma_{\max}}{\sigma_{\min}} + \frac{\|b\|_2}{\|x_0\|_2\,\sigma_{\min}}$
11.	Golub and Van Loan, 1983 [22]	(50)	$\frac{\|r_0\|_2\,\sigma_{\max}^2}{\|Ax_0\|_2\,\sigma_{\min}^2} + \frac{2\,\|b\|_2\,\sigma_{\max}}{\|Ax_0\|_2\,\sigma_{\min}}$
15.	Wei, 1990 [64]	(55)	$\frac{2\,\|r_0\|_2\,\sigma_{\max}}{\|x_0\|_2\,\sigma_{\min}^2} + \frac{4\,\sigma_{\max}}{\sigma_{\min}} + \frac{2\,\|x\|_2\,\sigma_{\max}}{\|x_0\|_2\,\sigma_{\min}}$
16.	Higham-Wedin, 1996 [32]	(56)	$\frac{\|r_0\|_2\,\sigma_{\max}}{\|x_0\|_2\,\sigma_{\min}^2} + \frac{2\,\sigma_{\max}}{\sigma_{\min}} + \frac{\|r_0\|_2}{\|x_0\|_2\,\sigma_{\min}}$

5.2 Condition Numbers

This section derives an expression for the condition number of linear least squares problems, and hence gives the best possible error bound for small perturbations. The method of analysis is to apply Theorem 3.5. To that end, three matters must be addressed.

First is to check the theorem's five hypotheses. Three were checked in Section 4 when proving Theorem 4.2. The remaining two are these.

4. For the function in equation (31), it is easy to evaluate

$$\mathcal{J}_2 F^{(\text{LS})}(v(A), x_0) = -A^t A\,,$$

   so the fourth hypothesis is that A has full column rank.

5. Since by the fourth hypothesis A has full column rank, the same is true of every matrix that is sufficiently nearby. The least squares problems involving these matrices have unique solutions.

In summary, Theorem 3.5's five hypotheses are equivalent to A having full column rank and equation (32)'s matrix being nonsingular, which with the condition on A is equivalent to $b \neq 0$.

The second matter is to make Theorem 3.5's choice of norms. Choosing 2-norms for both the data and solution vectors means that matrices, considered as data vectors, are measured by the Frobenius norm.

Third, applying Theorem 3.5 requires $\mu_F^{(\text{LS})}$ whose equations (20) and (22) are difficult to evaluate, so it is more convenient to apply Corollary 3.6 which only requires $\mu_F^{(\text{LS},\,0)}$. Corollary 4.3 expresses $\mu_F^{(\text{LS},\,0)}$ in terms of a singular value decomposition for A.

With this preparation it is possible to evaluate,

$$\begin{aligned}
\chi_F^{(\text{LS, abs})}(A) &= \limsup_{x \to x_0} \frac{\|x - x_0\|_2}{\mu_F^{(\text{LS},\,0)}(x)} \\
&= \limsup_{x \to x_0} \frac{\|x - x_0\|_2}{\| \left(\|r_0\|_2^2\, I + \|x_0\|_2^2\, \Sigma^2\right)^{-1/2} \Sigma^2\, V^t (x_0 - x)\|_2} \\
&= \limsup_{\delta x \to 0} \frac{\|\delta x\|_2}{\| \left(\|r_0\|_2^2\, I + \|x_0\|_2^2\, \Sigma^2\right)^{-1/2} \Sigma^2\, V^t \delta x\|_2} \\
&= \left\| (\|r_0\|_2^2\, I + \|x_0\|_2^2\, \Sigma^2)^{1/2} \Sigma^{-2} \right\|_2 \\
&= \frac{\left(\|r_0\|_2^2 + \|x_0\|_2^2\, \sigma_{\min}^2\right)^{1/2}}{\sigma_{\min}^2}\,, \qquad (59)
\end{aligned}$$

where $\sigma_{\min}$ is the smallest nonzero singular value of A.

It is also possible to show that equation (59) lies within a small factor of the spectral condition number. Waldén, Karlson, and Sun's equation (21) implies

$$\frac{\|x - x_0\|_2}{\mu_F^{(\mathrm{LS})}(x)} \;\leq\; \frac{\|x - x_0\|_2}{\mu_2^{(\mathrm{LS})}(x)} \;\leq\; \sqrt{2}\,\frac{\|x - x_0\|_2}{\mu_F^{(\mathrm{LS})}(x)}\,.$$

Passing to $\limsup_{x \to x_0}$ replaces the ratios on either end by equation (59). Additionally multiplying by $\|A\|_2/\|x_0\|_2$ converts the quantity in the middle to the relative spectral condition number. (Note that forming the relative condition number introduces the requirement $x_0 \neq 0$.) This leaves,

$$\mathcal{C} \;\leq\; \chi_2^{(\mathrm{LS,\ rel})}(A) \;\leq\; \sqrt{2}\,\mathcal{C}\,, \tag{60}$$

where from equation (59) after some simplification,

$$\mathcal{C} \;=\; \chi_F^{(\mathrm{LS,\ abs})}(A)\,\frac{\|A\|_2}{\|x_0\|_2} \;=\; \left(\frac{\|r_0\|_2^2}{\|x_0\|_2^2\,\sigma_{\min}^2} + 1\right)^{1/2} \kappa_2\,.$$

It is easy to see that the literature's consensus upper bound in equation (57) lies between equation (60)'s limits. Thus the literature's error bounds combined with this paper's results determine what is the likely spectral condition number within a factor of $\sqrt{2}$.

Theorem 5.1 (Condition Numbers for LS). *Suppose Problem 4.1 (LS) has A of full column rank, $b \neq 0$, true solution x_0, and true least squares residual $r_0 = b - Ax_0$. This problem's Frobenius and spectral norm absolute condition numbers are,*

$$\chi_F^{(\mathrm{LS,\ abs})}(A) \;=\; \frac{\left(\|r_0\|_2^2 + \|x_0\|_2^2\,\sigma_{\min}^2\right)^{1/2}}{\sigma_{\min}^2} \;\leq\; \chi_2^{(\mathrm{LS,\ abs})}(A) \;\leq\; \frac{\|r_0\|_2}{\sigma_{\min}^2} + \frac{\|x_0\|_2}{\sigma_{\min}}\,.$$

If additionally $x_0 \neq 0$, then the Frobenius and spectral norm relative condition numbers are,

$$\begin{aligned} \chi_F^{(\mathrm{LS,\ rel})}(A) \quad &= \left(\frac{\|r_0\|_2^2}{\|x_0\|_2^2\,\sigma_{\min}^2} + 1\right)^{1/2} \frac{\|A\|_F}{\sigma_{\min}} \\ \chi_2^{(\mathrm{LS,\ rel})}(A) \quad &\begin{cases} \leq \left(\dfrac{\|r_0\|_2}{\|x_0\|_2\,\sigma_{\min}} + 1\right)\kappa_2 \\ \geq \left(\dfrac{\|r_0\|_2^2}{\|x_0\|_2^2\,\sigma_{\min}^2} + 1\right)^{1/2}\kappa_2 \end{cases} \end{aligned} \tag{61}$$

where $\kappa_2 = \sigma_{\max}/\sigma_{\min}$ is the matrix condition number, and $\sigma_{\max}$ and $\sigma_{\min}$ are the largest and smallest singular values of A. The ratios of the upper to the lower bounds for both spectral condition numbers are at most $\sqrt{2}$.

Corollary 5.2 (Optimal Error Bounds for LS) *Continuing Theorem 5.1, if $x_0 + \delta x$ solves the perturbed problem $\min_u \|b - (A + \delta A)u\|_2$, then*

$$\frac{\|\delta x\|_2}{\|x_0\|_2} \;\leq\; \chi_{2 \text{ or } F}^{(\text{LS, rel})}(A)\, \frac{\|\delta A\|}{\|A\|} \;+\; \mathcal{O}\left(\frac{\|\delta A\|^2}{\|A\|^2}\right),$$

where the matrix norms to be used with $\chi_2^{(\text{LS, rel})}$ and $\chi_F^{(\text{LS, rel})}$ are the spectral and the Frobenius norms, respectively.

Conjecture 5.3 (Spectral Condition Numbers for LS) *The spectral condition numbers are the upper bounds in Theorem 5.1.*

5.3 Dependence on κ_2^2

Section 5.1's first error bound led Golub and Wilkinson to suggest that κ_2^2 is "relevant to some extent" [24, p. 144] to the least squares problem. The prominence of this finding assured that it has been reexamined several times. Since least squares error bounds traditionally have been formulated to exhibit κ_2^2, the discussions have been phrased in terms ameliorating its effect. The most detailed analysis, by van der Sluis, concluded that least squares problems are sensitive to κ_2^2 very rarely.

1. Some authors write Table 4's consensus coefficient of $\|\delta A\|_2/\|A\|_2$ in the normwise relative error bounds as

$$\left(\frac{\|r_0\|_2}{\|x_0\|_2\,\sigma_{\min}} + 1\right)\kappa_2 \;=\; \frac{\|r_0\|_2}{\|A\|_2\,\|x_0\|_2}\,\kappa_2^2 + \kappa_2\,. \tag{62}$$

   They summarize this by saying the bound is sensitive to κ_2^2 unless $\|r_0\|_2$ is small compared to $\|A\|_2\,\|x_0\|_2$, in which case the coefficient is more like κ_2. Björck [7, p. 17] appears to have originated this explanation though it often appears in textbooks without attribution [23, p. 230] [32, p. 393].

2. Van der Sluis wrote the leading coefficient as [59, p. 251, eqn. 5.8],

$$\frac{\|r_0\|_2}{\|x_0\|_2\,\sigma_{\min}}\,\kappa_2 \;=\; \frac{\frac{\|Ax_0\|_2}{\|x_0\|_2}}{\sigma_{\min}}\,\kappa_2\,\tan(\theta)\,, \tag{63}$$

   where θ is the angle between b and $\text{col}(A)$. If $\|Ax_0\|_2/\|x_0\|_2 \approx \|A\|_2$, then equation (63) gains a second factor of κ_2. On the other hand, if x_0 has a comparatively small projections into A's right singular vectors corresponding to the largest singular values, then κ_2^2 "plays no role" [59, p. 251].

   Interestingly, van der Sluis observed that κ_2^2 is more likely to be irrelevant especially when κ_2 is large. This is because b's weight in the space corresponding to a singular value σ_i is magnified in x_0 by σ_i's reciprocal. Indeed, $v_i^t x_0 = u_i^t b/\sigma_i$ where u_i and v_i are left and right singular vectors corresponding to σ_i. The scaling by $1/\sigma_i$ strongly favors the smaller singular values over the larger when $\sigma_{\min} \ll \sigma_{\max}$.

3. Stewart's error bound in the full rank case of equation (49) contains a term that can be rearranged as follows [48, p. 655, eqn. 5.11] [51, p. 158],

$$\|(\Sigma + E_1)^{-1}\|_2^2 \, \frac{\|E_2\|_2 \, \|r_0\|_2}{\|x_0\|_2} \;=\; \frac{\tan(\theta)}{\frac{\|A\|_2 \, \|x_0\|_2}{\|Ax_0\|_2}} \, \frac{\|A\|_2^2}{\sigma_{\min}^2(\Sigma + E_1)} \, \frac{\|E_2\|_2}{\|A\|_2} \,,$$

where E_1 and E_2 are submatrices of the orthogonally transformed δA shown with error bound 10. The same rearrangement can be applied to the leading coefficient of the consensus error bound,

$$\frac{\|r_0\|_2}{\|x_0\|_2 \, \sigma_{\min}} \, \kappa_2 \;=\; \frac{\tan(\theta)}{\frac{\|A\|_2 \, \|x_0\|_2}{\|Ax_0\|_2}} \, \kappa_2^2 \,. \tag{64}$$

The denominator mitigates κ_2^2 because it varies from 1 to κ_2. There is no effect when $\|A\|_2 \, \|x_0\|_2 \approx \|Ax_0\|_2$, equivalently when x_0 has a comparatively large projections into A's right singular vectors corresponding to the largest singular values. This agrees with the finding of van der Sluis whom Stewart credits for suggesting the analysis [48, p. 657] [51, p. 163].

With the benefit of Theorem 5.1, it is possible to examine the dependence on κ_2^2 in terms of the condition number itself. This analysis follows the approach of van der Sluis in equation (63). The term $\|x_0\|_2 \, \sigma_{\min}$ in the condition number is analyzed for situations where it scales with κ_2 and thus contributes an additional factor of κ_2 to the condition number.

Equation (61)'s sharp bounds on the condition number depend on:

- $\sigma_{i=1,\ldots,n}$, the singular values of A,
- x_0, or equivalently, the coefficients $b_{i=1,\ldots,n}$ of $\mathcal{P}b$ in the basis of A's left singular vectors,
- $r_0 = (I - \mathcal{P})b$, the portion of b orthogonal to $\mathrm{col}(A)$.

With this notation,

$$\begin{aligned} \chi_2^{(\text{LS, rel})}(A) \;&\approx\; \left(\frac{\|r_0\|_2}{\|x_0\|_2 \, \sigma_{\min}} + 1 \right) \kappa_2 \\ &=\; \left(\frac{\|r_0\|_2}{\sqrt{\sum_{i=1}^{n} (b_i \sigma_{\min}/\sigma_i)^2}} + 1 \right) \kappa_2 \,. \end{aligned} \tag{65}$$

The first factor in equation (65) depends on $\sigma_{\max}/\sigma_{\min}$ only if the coefficients b_i favor the term where $\sigma_i = \sigma_{\max}$. Randomly distributed coefficients have a high probability that they would overwhelm this term since it is the smallest: $1 \geq (\sigma_{\min}/\sigma_i) \geq (\sigma_{\min}/\sigma_{\max})$.[5] Thus the denominator's sum is proportional to

[5] This is van der Sluis's observation that the chance of κ_2^2 being relevant declines as κ_2 increases.

$\sigma_{\min}/\sigma_{\max}$ if and only if b's projection into the column space of A lies predominantly in the space of the largest singular value, or of singular values clustered there. This is equivalent to $\|Ax_0\|_2 \approx \|A\|_2 \|x_0\|_2$, in which case equation (65) reduces to

$$\left(\frac{\|r_0\|_2}{\sqrt{\sum_{i=1}^{n} (b_i \sigma_{\min}/\sigma_i)^2}} + 1 \right) \kappa_2 \;\approx\; \left(\frac{\|r_0\|_2}{\|\mathcal{P}b\|_2/\kappa_2} + 1 \right) \kappa_2 \tag{66}$$

$$= \left(\tan(\theta)\, \kappa_2 + 1 \right) \kappa_2 \,. \tag{67}$$

The whole expression is furthermore proportional to κ_2^2 only when $\tan(\theta)$ is reasonably larger than $1/\kappa_2$. This is summarized in the following theorem.

Theorem 5.4 (Tangent Theorem) *Suppose Problem 4.1 (LS) has A of full column rank and exact solution $x_0 \neq 0$. The relative spectral condition number is proportional to the squared matrix condition number if and only if:*

1. *$\|Ax_0\|_2 \approx \|A\|_2 \|x_0\|_2$, equivalently, x_0 lies in spaces corresponding to the singular values of A clustered at $\sigma_{\max}$,*
2. *and $\tan(\theta)$ is at least moderately larger than κ_2^{-1},*

where θ is the angle between b and $\mathrm{col}(A)$. If these conditions are satisfied then the constant of proportionality is roughly $\tan(\theta)$ so that $\chi_2^{(\mathrm{LS})}(A) \approx \tan(\theta)\, \kappa_2^2$. Moreover, among least squares problems with the same θ and singular vectors for A, the probability that a randomly chosen problem's condition number will be sensitive to κ_2^2 declines as κ_2 increases.

Theorem 5.4 is essentially van der Sluis's although it is not formally stated in his paper. It reiterates his finding that it is "not realistic" [59, p. 251] to expect the errors of a linear least squares problem to be sensitive to κ_2^2.

This theorem reveals a defect in the textbook description (item 1 in Section 5.3) of when the least squares condition number may be sensitive to κ_2^2. Following the analysis of equation (62), the condition number always can be written,

$$\chi_2^{(\mathrm{LS,\ rel})}(A) \;\approx\; \frac{\|r_0\|_2}{\|A\|_2 \|x_0\|_2}\, \kappa_2^2 + \kappa_2 \,,$$

so κ_2^2 always appears to be present. This is a false dichotomy because both κ_2^2 and its coefficient in this equation vary with the singular values of A. Thus κ_2^2's impact is undetermined even when this coefficient is large. See Example 6.1 in Section 6.2. In contrast, Theorem 5.4 identifies the situations where κ_2^2 appears with a coefficient independent of κ_2,

$$\chi_2^{(\mathrm{LS,\ rel})}(A) \;\approx\; \tan(\theta)\, \kappa_2^2 \,.$$

In only these cases does the condition number vary unambiguously with the square of the matrix condition number.

5.4 Examination of the Problem's Conditioning

Theorem 5.1's equation (61) estimates the spectral condition number sufficiently well to draw definitive conclusions about the conditioning of the problem.

Theorem 5.5 (Well Conditioned LS Problems) *Suppose Problem 4.1 (LS) has A of full column rank and exact solution $x_0 \neq 0$. The problem is well conditioned with respect to perturbations of the matrix if and only if:*

1. *$\|r_0\|_2$ is at most moderately larger than $\|x_0\|_2 \, \sigma_{\min}$, and*
2. *A is well conditioned,*

where $r_0 = b - Ax_0$ is the exact least squares residual, and $\sigma_{\min}$ is the smallest singular value of A.

Proof. From Theorem 5.1, $\chi_2^{(\text{LS, rel})}(A)$ is with a factor of $\sqrt{2}$ from

$$\left(\frac{\|r_0\|_2}{\|x_0\|_2 \, \sigma_{\min}} + 1 \right) \kappa_2 \, ,$$

where κ_2 is the condition number of A. ∎

Stoer [52, p. 177, Zusammenfassung] [53, p. 213, summary] provides the only previous discussion of Theorem 5.5's criteria. Although technically it is impossible to deduce sufficiency from upper bounds alone, Stoer concluded on the basis of error bounds that the opposite criteria would imply the problem is ill conditioned.

A puzzle that is not addressed in the literature is how to reconcile van der Sluis's conclusion that κ_2^2 rarely affects the least squares problem with the view that least squares problems are often difficult to solve accurately. Evidently there is a more commonplace source of ill conditioning than κ_2^2.

Further understanding of ill conditioned problems can be obtained by simplifying equation (61) using an inequality associated with the matrix lower bound [26]: $\|x_0\|_2 \, \sigma_{\min} \leq \|Ax_0\|_2$ when A has full column rank. This leads to a lower bound,

$$\begin{aligned}
\chi_2^{(\text{LS, rel})}(A) \;&\geq\; \sqrt{\frac{\|r_0\|_2^2}{\|x_0\|_2^2 \, \sigma_{\min}^2} + 1} \;\; \kappa_2 \\
&\geq\; \sqrt{\frac{\|r_0\|_2^2}{\|Ax_0\|_2^2} + 1} \;\; \kappa_2 \\
&=\; \frac{\|b\|_2}{\|Ax_0\|_2} \;\; \kappa_2 \\
&=\; \sec(\theta) \, \kappa_2 \, ,
\end{aligned}$$

where θ is the angle between b and the column space of A. An upper bound is similar,

$$\begin{aligned}
\chi_2^{(\text{LS, rel})}(A) &\le \sqrt{2}\,\sqrt{\frac{\|r_0\|_2^2}{\|x_0\|_2^2\,\sigma_{\min}^2}+1}\;\kappa_2 \\
&= \sqrt{2}\,\sqrt{\frac{\|r_0\|_2^2+\|x_0\|_2^2\,\sigma_{\min}^2}{\|x_0\|_2^2\,\sigma_{\min}^2}}\;\kappa_2 \\
&\le \sqrt{2}\,\sqrt{\frac{\|r_0\|_2^2+\|Ax_0\|_2^2}{\|x_0\|_2^2\,\sigma_{\min}^2}}\;\kappa_2 \\
&= \sqrt{2}\,\frac{\|b\|_2}{\|x_0\|_2\,\sigma_{\min}}\;\kappa_2\,.
\end{aligned}$$

Notice that the ratio of the upper bound to the lower bound is at most $\sqrt{2}\,\kappa_2$.

$$\left(\sqrt{2}\,\frac{\|b\|_2}{\|x_0\|_2\,\sigma_{\min}}\,\kappa_2\right)\left(\frac{\|b\|_2}{\|Ax_0\|_2}\,\kappa_2\right)^{-1} = \sqrt{2}\,\frac{\|Ax_0\|_2}{\|x_0\|_2\,\sigma_{\min}} \le \sqrt{2}\,\frac{\sigma_{\max}}{\sigma_{\min}}$$

All this justifies the following theorem.

Theorem 5.6 (Secant Theorem) *Suppose Problem 4.1 (LS) has A of full column rank and exact solution $x_0 \neq 0$. The relative spectral condition number has the following lower and upper bounds,*

$$\sec(\theta)\,\kappa_2 \;\le\; \chi_2^{(\text{LS, rel})}(A) \;\le\; \sec(\theta)\,\kappa_2\,\sqrt{2}\,\frac{\|Ax_0\|_2}{\|x_0\|_2\,\sigma_{\min}}\,, \tag{68}$$

where κ_2 is the spectral condition number of A, and θ is the angle between b and the column space of A. Therefore sufficient conditions for the problem to be ill conditioned are:

1. *b is nearly orthogonal to the column space of A, or*
2. *A is ill-conditioned.*

Theorem 5.6's criteria are sufficient but not necessary because equation (68)'s upper bound is weaker than equation (61)'s.

The theorem identifies a source of ill conditioning that is underappreciated in the literature. Usually r_0 being large (in comparison to $\mathcal{P}b$ or some surrogate for it such as $\|A\|_2\,\|x_0\|_2$) is interpreted to mean that the least squares problem is sensitive to κ_2^2 [59, p. 242, top]. In such cases any ill conditioning is attributed to A. Theorem 5.6 shows to the contrary that θ is a separate source of ill conditioning independent of the distribution of A's singular values.

6 Applications

6.1 Failure of Simple Iterative Refinement

This section examines a famous numerical experiment in terms of this paper's sensitivity analysis. Golub and Wilkinson's [24] example that simple iterative improvement fails for linear least squares problems is one of the most important in numerical analysis, as measured by the research it inspired, see Table 1. This failure is often mentioned in the literature, but because Björck [6] developed a provably effective improvement algorithm shortly thereafter, the simple algorithm is seldom discussed in detail.

Businger and Golub in [13] had noted that the error, $e = x_0 - x$, in a least squares solution, x, also satisfies a least squares problem (and note, with the same true residual, r_0),

$$\|b - Ax_0\|_2 = \|r - Ae\|_2 \quad \text{where} \quad r = b - Ax\,.$$

In principle, e can be determined by solving this problem. There results a computed correction $\bar{e}$, so that x can be iteratively improved by Figure 3's algorithm. This would produce a sequence of corrected solutions x_1, x_2, ...

A similar process had long been used for linear equations [32, p. 232]. There, the sequence converges to a solution that is accurate to nearly the machine's precision and at a rate dependent on κ_2. This is provided $\kappa_2\,\mathbf{u} < 1$ where $\mathbf{u}$ is the roundoff unit of the floating point arithmetic, and the residual calculation uses higher precision arithmetic so that the computed residual $\bar{r} \approx r$ is very accurate [22, p. 75].

When Businger and Golub tested Figure 3's algorithm with the then-new least squares solution method based on Householder transformations, they found that some problems were not solved to full working accuracy [21]. Golub and Wilkinson [24] reported one such case with A and b having integer entries, A of full rank, and $A^t b = 0$. The unique solution of the least squares problem is therefore $x_0 = 0$. The least squares solution method produced a computed $x \neq 0$ and the high precision computed residual $\bar{r}$. One step of Figure 3's algorithm was performed by computing a correction, $\bar{e}$. Since $x_0 = 0$, the true correction

---

1. Form an orthogonal factorization of A.
2. Use the factorization to calculate a solution, x, to $\min_u \|b - Au\|_2$.
3. Use higher precision arithmetic to calculate $\bar{r}$, the residual $b - Ax$.
4. Use the factorization to calculate a solution, $\bar{e}$, to $\min_u \|\bar{r} - Au\|_2$.
5. Replace x by the calculated sum $x + \bar{e}$ and repeat from step 3, as needed.

Figure 3: *Simple iterative improvement of a least squares solution.*

is $e = -x$, but the computed $\bar{e}$'s entries were roughly five orders of magnitude smaller. Thus the iterative improvement step was performed with relative error

$$\frac{\| e - \bar{e} \|_2}{\|e\|_2} \approx 1 - 10^{-5} \approx 1 .$$

This and subsequent iterations made no improvement in the computed solution.

An analysis of this test of Figure 3's algorithm must distinguish between three different corrections to the approximate solution: e, $\tilde{e}$, and $\bar{e}$.

- The "true correction," for which $x + e = x_0$, is the solution of

  $$\min_u \|r - Au\|_2 ,$$

  where $r = b - Ax$.

- The "intended correction" is $\tilde{e}$, the true solution of

  $$\min_u \|\bar{r} - Au\|_2 , \tag{69}$$

  where $\bar{r}$ is the vector computed for r in Figure 3's step 3.

- The "computed correction" is $\bar{e}$, Figure 3 step 4's computed solution for equation (69).

With this notation, Table 5 lists many of the relevant quantities in Golub and Wilkinson's test. These data either are given by them [24, p. 147] or are derived from their data. For example, the computed $\bar{r}$ is available in scientific notation of 11 decimal digits, while the solution of equation (69) can be derived as $\tilde{e} = (A^t A)^{-1} A^t \bar{r}$. This and the other derived quantities in Table 5 are formed by exact, rational arithmetic [66].

Two different explanations for the breakdown of simple iterative refinement focus on the two different parts of the error, $e - \bar{e} = (e - \tilde{e}) + (\tilde{e} - \bar{e})$.

1. This explanation is due to Stewart [50, pp. 320–321]. Regarding $e - \tilde{e}$, the computed approximate residual is,

   $$\bar{r} = r + \delta r = (r_0 + Ae) + \delta r ,$$

   where δr is the error of evaluating r. In this notation equation (69) is

   $$\min_u \left\| (r_0 + Ae + \delta r) - Au \right\|_2 ,$$

   which makes $\tilde{e}$ the solution of $Au = Ae + \mathcal{P}\delta r$ where $\mathcal{P}$ is the orthogonal projection into $\mathrm{col}(A)$. If Ae is subordinate to the rounding error term, $\mathcal{P}\delta r$, then $\tilde{e}$ can't be accurate, so any computed $\bar{e}$ is inaccurate no matter how well equation (69) is solved.

   It is plausible that $\bar{r}$'s rounding errors may cause trouble, but in this

Table 5: *Values of some quantities in the numerical example of Golub and Wilkinson. The matrix A is the first five columns of the inverse Hilbert matrix of order* 6. *The vectors $b = r_0$, $e = -x$, $\bar{e}$, $\bar{r}$ are given in [24, p. 147] as b_1, $-x^{(1)}$, $\delta^{(1)}$, and $r^{(1)}$, respectively. Other quantities are derived from these.*

quantity	value
$\|A\|_F$	8.89×10^6
$\|A\|_2$	8.89×10^6
$\sigma_{\min}$	1.89
κ_2	4.70×10^6
$\|e\|_2 \equiv \|x\|_2$	1.44×10^{-3}
$\|\tilde{e}\|_2$	1.44×10^{-3}
$\|\bar{e}\|_2$	1.17×10^{-8}
$\|e - \tilde{e}\|_2$	6.27×10^{-9}
$\|\tilde{e} - \bar{e}\|_2$	1.44×10^{-3}
$\|Ae\|_2$	2.77×10^{-3}
$\|A\tilde{e}\|_2$	2.77×10^{-3}
$\|r_0\|_2 \equiv \|b\|_2$	8.52×10^3
$\|r\|_2$	8.52×10^3
$\|\bar{r}\|_2$	8.52×10^3
$(\delta r = \bar{r} - r) \quad \|\delta r\|_2$	2.86×10^{-6}
$\|\mathcal{P}\delta r\|_2$	2.86×10^{-6}
$\sec(\theta) = \|\bar{r}\|_2 / \|A\tilde{e}\|_2$	3.08×10^6
$\theta = \angle(\bar{r}, \mathrm{col}(A))$	$\pi/2 - 3.25 \times 10^{-7}$
$\mathbf{u} = 2^{-39}$	1.82×10^{-12}

specific case it is not why the algorithm failed. From Table 5,

$$\frac{\|\mathcal{P}\delta r\|_2}{\|Ae\|_2} \;=\; 1.03 \times 10^{-3}\,,$$

so rounding error does not overwhelm Ae. As a result, equation (69)'s solution $\widetilde{e}$ is a good approximation to the true correction e,

$$\frac{\|e - \widetilde{e}\|_2}{\|e\|_2} \;=\; 4.37 \times 10^{-6}\,.$$

2. Regarding $\widetilde{e} - \bar{e}$, this is the difference between the true and computed solutions of equation (69), in whose study Golub and Wilkinson [24] derived Section 5.1's first error bound. Since $\bar{r} \approx r$, equation (69) has nearly the same least squares residual, r_0, as the original problem. Thus the potentially largest term in equation (37)'s error bound for $\bar{e}$ appears to be $\|r_0\|_2\,\kappa_2^2$, which Golub and Wilkinson noted would vanish only if the original equations are consistent [24, p. 144, middle].

   Indeed, from Table 5 it is true that

$$\frac{\|\widetilde{e} - \bar{e}\|_2}{\|\widetilde{e}\|_2} \;=\; 0.99\,,$$

   so the inability to solve equation (69) explains the test's failure. However, the equation violates Golub and Wilkinson's hypothesis (1b),

$$\left.\begin{array}{rcl} \|A\|_2 &=& 8.89 \times 10^6 \\ \|\bar{r}\|_2 &=& 8.52 \times 10^3 \end{array}\right\} \;\gg\; 1\,,$$

   so their error bound, equation (37), does not actually explain the inability to solve for $\widetilde{e}$ accurately.

It is therefore new to examine simple iterative improvement, equation (69), in terms of a sensitivity analysis of the least squares problem. From Theorem 5.1 and Table 5, the condition number of the solution $\widetilde{e}$ is

$$\chi_2^{(\mathrm{LS,rel})}(\widetilde{e}) \;\approx\; \left(\frac{\|\bar{r}\|_2}{\|\widetilde{e}\|_2\,\sigma_{\min}} + 1\right)\kappa_2 \;=\; 1.47 \times 10^{13}\,.$$

This exceeds the reciprocal of the roundoff unit on the machine used for Golub and Wilkinson's calculations. Thus, even if equation (69) is posed with the smallest representable backward error, the condition number is large enough to account for no accuracy in the computed solution $\bar{e}$. In fact, due to the large error in the computed solution $\bar{e}$, the exact size of its optimal backward error is 1.07×10^7 as determined by Higham's equation (22).

More generally, the sensitivity analysis of this paper supports an explanation for the failure of simple iterative improvement that complements Stewart's case.

If δr, the rounding error of computing $\bar{r}$, is subordinate to Ae, then $\bar{r}$ has been computed sufficiently well to enable equation (69) to determine a correction, in particular, $\bar{r} \approx r$. The least squares residual has a condition number that is no worse than κ_2 [23, p. 230], so for reasonable A and x it is likely that r is fairly accurate, $r \approx r_0$. Altogether $\bar{r} \approx r \approx r_0$, so $\bar{r}$ is nearly orthogonal to $\mathrm{col}(A)$. The secant Theorem 5.6 now says the least squares problem for the correction tends to be unboundedly ill conditioned no matter how well conditioned the original problem may be.

6.2 Ill-Conditioned Without κ_2^2

This section presents a simple example that illustrates many of this paper's results about $\chi_2^{(\mathrm{LS,\ rel})}$ and especially about its relationship to κ_2^2. The example is a modification of Golub and Van Loan's example [22, p. 141] [23, p. 223] which was repeated by Higham [32, p. 393]. In their original version, the first entry of b rather than the second is nonzero.

Example 6.1 (Ill Conditioned Without κ_2^2) *Let*

$$A = \begin{bmatrix} 1 & \\ & \alpha \\ & \end{bmatrix}, \quad \delta A = \begin{bmatrix} & \\ & \epsilon \\ & \end{bmatrix}, \quad b = \begin{bmatrix} \beta \\ 1 \\ \end{bmatrix},$$

where $0 < \epsilon \ll \alpha, \beta < 1$. *In this example,*

$$x_0 = \begin{bmatrix} \\ \frac{\beta}{\alpha} \\ \end{bmatrix}, \quad r_0 = \begin{bmatrix} \\ 1 \\ \end{bmatrix}, \quad x = \begin{bmatrix} \\ \frac{\alpha\beta+\epsilon}{\alpha^2+\epsilon^2} \\ \end{bmatrix} = x_0 + \begin{bmatrix} \\ \frac{\epsilon(\alpha-\beta\epsilon)}{\beta(\alpha^2+\epsilon^2)} \\ \end{bmatrix}.$$

Here are the conclusions to draw from this example.

- The relative spectral condition number's two principal terms can be independently manipulated to make the condition number large,

$$\chi_2^{(\mathrm{LS,\ rel})} \approx \left(\frac{\|r_0\|_2}{\|x_0\|_2 \, \sigma_{\min}} + 1 \right) \kappa_2 = \left(\frac{1}{\beta} + 1 \right) \frac{1}{\alpha}.$$

- Corollary 5.2's optimal error bound for small perturbations is sharp. The actual normwise relative error in this example is,

$$\frac{\|\delta x\|_2}{\|x_0\|_2} = \frac{\epsilon(\alpha - \beta\epsilon)}{\beta(\alpha^2 + \epsilon^2)} \approx \frac{\epsilon}{\beta \, \alpha}.$$

- The tangent Theorem 5.4 correctly predicts that the condition number does not depend on the squared matrix condition number. The theorem's first criterion fails because the quantities

$$\|Ax_0\|_2 = \beta \quad \text{and} \quad \|A\|_2 \, \|x_0\|_2 = \frac{\beta}{\alpha}$$

are not equal.

- The linear least squares problem can be ill conditioned (choose β small) even when its condition number does not depend on the matrix condition number squared.

- The textbook explanation, that the condition number depends on κ_2^2 unless the residual is small compared to the matrix and solution (see item 1 in Section 5.3), is wrong. In this example, equation (62)'s coefficient

$$\frac{\|r_0\|_2}{\|A\|_2 \, \|x_0\|_2} \;=\; \frac{\alpha}{\beta}$$

  can be made arbitrarily large, yet the problem's condition number does not depend on κ_2^2.

6.3 Error Bounds that Overestimate the Error

This section examines the consequences of varying from the literature's consensus in Table 4. Although most of the error bounds in the literature appear to depend on the square of the matrix condition number, upon simplification the dependence vanishes from many.

The bounds that always contain κ_2^2 are those not equivalent to Corollary 5.2's optimal error bound, or in the Stewart's terminology, that omit any mitigation of κ_2^2. These effectively replace $\|A\|_2 \, \|x_0\|_2$ in equation (62) by its lower bound $\|Ax_0\|_2$, or they remove the denominator from equation (64), both of which replace the condition number by a weak upper bound,

$$\tan(\theta) \, \kappa_2^2 + \kappa_2 \, .$$

This unnecessarily overestimates the error when the matrix is at least moderately ill-conditioned but the condition number does not actually depend on κ_2^2 (which are the majority of cases according to van der Sluis [59, p. 251] and Theorem 5.4).

To illustrate this effect, Table 6 specifies two instances of Example 6.1 for which almost all the bounds in Section 5.1's survey have been evaluated. Table 7 shows the ratios of the bounds to the actual error. This overestimating behavior of the bounds that do not mitigate κ_2^2 is apparent even when the matrix is only slightly ill conditioned.

7 Conclusion

7.1 Narrative

It is interesting to note that many basic discoveries in numerical linear algebra were made as a result of linear least squares problems. Solving the normal equations was the original use for Gaussian elimination, by Gauss. See historical references in Table 1. It was also the motivation for Cholesky's [5] version of the

Table 6: *Two instances of Example 6.1 used in Table 7 to compare error bounds.*

	α	β	ϵ	κ_2	$\chi_2^{(\text{LS, rel})}$
case (a)	0.5	0.002	0.00001	2	≈ 1002
case (b)	0.01	0.002	0.00001	100	≈ 50100

Table 7: *Ratio of error bounds to actual error for most of the bounds listed in Section 5.1 applied to two instances of Example 6.1 described in Table 6. Case (a) has a very well conditioned matrix ($\kappa_2 = 2$), while case (b) has a less well conditioned matrix ($\kappa_2 = 100$). The componentwise bound, number 14, is evaluated using the sparsity matrix of $A + \delta A$ as the reference matrix. Blue indicates those bounds whose leading terms in Table 4 resemble Corollary 5.2's optimal first-order error bound.*

			ratio of bound to actual error	
	source of bound	eqn.	case (a)	case (b)
2.	Björck, 1967 [7]	(41)	1.002	1.004
3.	Hanson and Lawson, 1969 [29]	(42)	2.002	100.202
4.	Pereyra, 1969 [44]	(43)	2.006	111.460
5.	Stoer, 1972 [52]	(44)	1.002	1.002
6.	Wedin, 1973 [62]	(45)	1.002	1.003
7.	Abdelmalek, 1974 [1]	(46)	1.000	1.001
8.	Lawson and Hanson, 1974 [36]	(47)	1.002	1.003
9.	van der Sluis, 1975 [59]	(48)	1.502	1.012
10.	Stewart, 1977 [48]	(49)	1.000	1.000
11.	Golub and Van Loan, 1983 [22]	(50)	4.000	102.000
14.	Higham, 1990 [31]	(54)	1.002	1.002
15.	Wei, 1990 [64]	(55)	2.012	2.018
16.	Higham-Wedin, 1996 [32]	(56)	1.504	1.015
	LAPACK, 1992 [2]		4.000	102.000

elimination algorithm, and for interpreting the algorithm as matrix factoring, by Banachiewicz [4] and Dwyer [17]. An unrelated third source for triangular factoring and likely the person who elevated it to folklore was von Neumann.

> "We may therefore interpret the elimination method as the combination of two tricks . . . ." [41, p. 1053]

Matrix factoring became a paradigm for numerical analysis roughly twenty years after these pioneers. It was then realized that numerical problems might be avoided by using orthogonal factoring. For example, Householder [33, p. 341] noted that his transformations could calculate the upper triangular Cholesky factor of A^tA. He did not actually explain how to solve linear least squares problems without forming the normal equations. Instead it was Golub [21] [50, p. 324] who described essentially the algorithm that is still used today.

This awakened interest in the sensitivity analysis of least squares [51, p. 151]. Golub and Wilkinson derived an error bound whose leading term suggested "although the use of orthogonal transformations avoids some of the ill effects inherent in the normal equations, the value of κ_2^2 is still relevant to some extent" [24, p. 144]. Van der Sluis commented that "the serious of this prognosis might be doubted since only an upper bound is given, were it not that numerical experiments seem to confirm it" [59, p. 242]. See the discussion of this experiment in Section 6.1.

Golub and Wilkinson's finding was "something of a shock" [59, p. 241]. Whatever its cause, the negative result meant that orthogonal factorization would not be a panacea for numerical troubles. Some measure of the upset can be taken from the quantity of papers that followed. Many of the error bounds surveyed in Section 5.1 have their origin in this period.

7.2 Summary

This paper has made a thorough review of the literature ensuing from Golub and Wilkinson's result of thirty-five years ago (Table 1). This has been used to guide the application of new methods from real analysis and optimization theory in drawing definitive conclusions about linear least squares problems. All the backward (Tables 2, 3) and forward (Tables 4, 7) error bounds in the literature have been compared and contrasted in a systematic fashion in light of the optimal bounds made possible by this paper's new methods of analysis.

Specifically, some results (Theorem 3.3) from the sensitivity analysis of optimization problems were used to derive asymptotic formulas for the size of optimal backward errors for linear least squares problems. A simple asymptotic expression for the size in the Frobenius norm (Theorem 4.2) showed how the exact formula of Waldén, Karlson, and Sun [61] behaves for small perturbations. This expression and bounds in the literature suggested a computable asymptotic formula (Theorem 4.4) that may provide a means to estimate the optimal backward error in practice.

It was further shown that the size of optimal backward errors, or asymptotic expressions for it, can be used to evaluate condition numbers (Theorem 3.5 and Corollary 3.6). A simple expression was found for the Frobenius norm condition number of full rank problems, and a sharp estimate was derived for the spectral norm condition number of the same problems (Theorem 5.1). This and a consensus of the error bounds in the literature suggested a conjecture for the exact spectral condition number (Conjecture 5.3). The sharp bounds on the condition number were used to prove some criteria, which have appeared in the literature, for the condition number to depend on the square of the matrix condition number (tangent Theorem 5.4), and for the solution to be well conditioned with respect to perturbations of the matrix (Theorem 5.5). A mechanism for ill conditioning that has not been emphasized in the literature was discovered (secant Theorem 5.6).

These theorems and the understanding they provide were then illustrated by examples. They explained the historic numerical experiment of Golub and Wilkinson (Section 6.1). They motivated a simple example that demonstrated fallacies in textbook explanations of ill conditioning (Section 6.2). Finally, they revealed a situation in which some error bounds in the literature unnecessarily overestimate the error (Section 6.3).

7.3 Open Questions

The present work suggests several open questions. The first deals with calculating the size of optimal backward errors for linear least squares.

1. How can Theorem 4.4's calculable estimate $\widetilde{\mu}_F^{(\mathrm{LS})}(x)$ be efficiently evaluated, and how good is it in estimating the optimal backward errors of actual computations?

   A satisfactory answer to both questions would resolve a quarter-century old challenge posed by Stewart and Wilkinson [49, p. 6–7] and reiterated by Higham [31, p. 201].

The next three questions address the conditioning of least squares solutions.

2. Either by example or further analysis, prove or disprove Conjecture 5.3.

   A very interesting answer would be to identify a δA that maximizes Stoer's equation (58) for δx.

3. Does equation (3) also give the condition number of rank deficient linear least squares problems?

4. If the answer to question 3 is no, then rank deficient problems may become ill conditioned in other ways than by increasing the size of equation (3)'s two factors. What are the other ways, if any?

Some questions concern the related matter of the least squares residual.

5. Are the several bounds for the error in the least squares residual the same, or do they have significant differences of the kind that Tables 4 and 7 reveal for the bounds on the error in the solution?

   Error bounds for the residual can be found in [7, p. 16, eqn. 7.7] [22, p. 141, eqn. 6.1-11] [23, p. 228, eqn. 5.3.9] [32, pp. 392, 394, eqns. 19.2, 19.10] [51, p. 160] [62, thm. 5.1].

6. What is the optimal condition number of the linear least squares residual with respect to perturbations of the matrix? Is the residual better conditioned than the solution as is often suggested?

The next set of questions suggest applications of this paper's methods to other kinds of least squares problems.

7. What is the size of the optimal backward errors of linearly constrained linear least squares problems (LLS)? What is the condition number?

   Cox and Higham [14] [15] found approximations for both quantities. It seems likely that the optimal values can be found by the methods of this paper. Use Theorem 3.3 to asymptotically estimate the size of the optimal backward errors, and then use Corollary 3.6 to evaluate the condition number.

8. Theorem 3.3 and Corollary 3.4 can be used to asymptotically estimate the size of the optimal backward error in the coefficients of least squares problems with Toeplitz coefficient matrices. That is, a coefficient should receive the same perturbation wherever it appears in the matrix. (This is a form of structured backward error.) Does the asymptotic size of the optimal backward error have a simple formula in terms of the Toeplitz coefficients?

   This generalizes the problem for linear equations which was considered by the Highams [30] and by Varah [60].

9. Continuing the previous question, either by an explicit formula or numerically, examine the stability of the many specialized algorithms for solving Toeplitz linear least squares problems that are cited in [27, p. 364, top].

   Gu [28] [27, p. 365, lines 9–11] applied this a posteriori approach to stability analysis using Higham's equation (22) for the optimal size of *unstructured* backward errors. For structured backward errors do the conclusions about which algorithms are stable differ from his?

Lastly are two questions of basic importance to numerical analysis.

10. Can Demmel's conjecture about condition numbers can be reconciled with the linear least squares problem?

    For many problems in numerical linear algebra, Demmel [16] has shown

that the condition number is related to the distance to the nearest ill-posed problem. Specifically, he suggests that the reciprocal of the condition number is a measure of the relative distance to the nearest ill-posed problem.

For example, if A is nonsingular, then Section 3.3's example shows that the relative spectral condition number of solving $Ax = b$ is

$$\frac{1}{\chi_2^{(\text{LE, rel})}} \;=\; \frac{\sigma_{\min}}{\sigma_{\max}} \;=\; \frac{\sigma_{\min}}{\|A\|_2}\,.$$

The numerator, $\sigma_{\min}$, is known to be the distance (as measured by the spectral matrix norm) to the nonsingular matrix nearest A, while the reciprocal's denominator makes this distance relative to the size of A.

Demmel's conjecture is not obviously true for the linear least squares problems. Theorem 5.1's sharp bounds for $\chi_2^{(\text{LS, rel})}$ give,

$$\frac{1}{\chi_2^{(\text{LS, rel})}} \;\approx\; \frac{\sigma_{\min}}{\|A\|_2} \left(\frac{\|r_0\|_2}{\|x_0\|_2\, \sigma_{\min}} + 1 \right)^{-1}\,.$$

This quantity seems unrelated to the distance to the nearest ill posed problem. It is known that $\sigma_{\min}$ is the spectral distance to the nearest rank deficient matrix, and $\|A\|_2$ again makes the distance relative to the size of A. However, the final term can make the supposed distance arbitrarily small by varying r_0.

11. Equations (11) and (12) show that, for an approximate solution $x \approx x_0$ of a numerical problem $F(y_0, x_0) = 0$, the size of optimal backward errors asymptotically equals a norm of the numerical problem's residual, $F(y_0, x)$. If the backward errors are measured by the 2-norm, then this norm of the residual is defined by

$$\left\| [\mathcal{J}_1 F(y_0, x_0)]^\dagger \cdot \right\|_2\,.$$

The norm is in fact unique [25, thm. 6.3]. It is appropriate to call it the von Neumann norm because the backward errors' dependence on the residual was originally noted by von Neumann and Goldstine [41, p. 1093], and more recently by Sun [54, p. 358].

For the linear least squares problem, whose residual function is $F(A, x) = A^t(b - Ax) = A^t r$, equation (33) shows that the von Neumann norm is

$$\|\cdot\|_F^{(\text{LS vN})} \;=\; \left\| (\|r_0\|_2^2\, I + \|x_0\|_2^2\, A^t A)^{-1/2} \cdot \right\|_2\,.$$

That is, the Frobenius norm size of optimal backward errors for the approximate solution x asymptotically equals $\|A^t r\|_2^{(\text{LS vN})}$. What are the von Neumann norms of other numerical problems?

Nomenclature

A	coefficient matrix in linear equations or linear least squares problems
$\mathcal{B}_c(r)$	the open ball with center c and radius r in whatever space is indicated
b	inhomogeneous vector in linear equations or linear least squares problems
b_i	coefficient corresponding to σ_i of $\mathcal{P}b$ expanded in a basis of A's left singular vectors
$\mathrm{col}(A)$	the column space of A
$\mathcal{D}F(y_0)$	the Fréchet derivative of f evaluated at y
$\mathcal{J}_1 F(y_0, x_0)$	the Jacobian matrix of derivatives with respect to F's first block of variables evaluated at (y_0, x_0)
$\mathcal{J}_2 F(y_0, x_0)$	like $\mathcal{J}_1 F(y_0, x_0)$ but for the second block of variables
$F(y, x)$	the residual function of data, y, and solutions, x, that defines a numerical problem
$F^{(\mathrm{LE})}$	$= Ax - b$, the residual function (of data A and solution x) for linear equations $Au = b$
$F^{(\mathrm{LS})}$	$= A^t(b - Ax)$, the residual function (of data A and solution x) for the linear least squares problem $\min_u \|b - Au\|$, see equation (31)
κ_2	$= \sigma_{\max}/\sigma_{\min}$, the spectral condition number of A
$\lambda_{\min}(M)$	smallest (most negative) eigenvalue of symmetric matrix M
$\mu(x)$	size of the optimal backward error of a numerical problem for the approximate solution x, equation (4)
$\mu^{(0)}(x)$	a function that asymptotically equals $\mu(x)$ at x_0 and that is of the form $\|F(y_0, x)\|^{(\mathrm{vN})}$ where the von Neumann norm $\|\cdot\|^{(\mathrm{vN})}$ is independent of x, see Theorem 3.3, equation (11), and open question 11

$\mu_{2 \text{ or } F}^{(\text{LE or LS})}(x)$ $= \mu(x)$ for the coefficient matrix of:

- (LE) linear equations, see equation (19),
- (LS) least squares, equations (20) and (22)

and if present the 2 or F signifying that the matrix perturbations are measured in the spectral or Frobenius norms, respectively

$\widetilde{\mu}_F^{(\text{LS})}(x)$ a function that asymptotically equals $\mu_F^{(\text{LS})}(x)$ at x_0 and that is computable because it does not depend on x_0, see Theorem 4.4

$\mu_F^{(\text{LS, } 0)}(x)$ $= \mu^{(0)}(x)$ for the coefficient matrix of least squares problems with the matrix perturbations measured in the Frobenius norm, see equations (33) and (34) in Theorem 4.2 and Corollary 4.3

$o(\cdots)$ Landau's little o notation for a quantity that converges to 0 more quickly than the expression inside the parentheses

$\mathcal{P}$ orthogonal projection into $\text{col}(A)$

σ_i a nonzero singular value of A

$\sigma_{\max}(M)$ largest singular value of matrix M, but if no matrix is specified then of matrix A

$\sigma_{\min}(M)$ smallest nonzero singular value of matrix M, but if no matrix is specified then of matrix A

r $= b - Ax$, approximate residual of linear equations or linear least squares problem

r_0 $= b - Ax_0$, true residual of linear least squares problem

$\rho(M)$ spectral radius (largest magnitude of any eigenvalue) of matrix M

θ angle between b (or the inhomogeneous vector in a least squares problem) and the column space of A

$\mathbf{u}$ arithmetic precision, machine epsilon, roundoff unit, unit roundoff

$v(A)$ vector that lists the matrix's entries column-by-column, see Figure 1

x	approximate solution of a numerical problem, especially linear equations or linear least squares
x_0	true solution of a numerical problem, especially linear equations or linear least squares
$\chi^{(\text{abs or rel})}(y_0)$	absolute or relative condition number of a numerical problem with respect to perturbations of the data y_0, equations (5) and (6) respectively
$\chi^{(\text{LE, abs or rel})}(A)$	$= \|A^{-1}\| \, \|x_0\|$ or $\|A^{-1}\| \, \|A\|$, respectively, the absolute or relative condition number of linear equations $Au = b$ with respect to perturbations of A
$\chi_{2 \text{ or } F}^{(\text{LS, abs or rel})}(A)$	absolute or relative condition number of the linear least squares problem for perturbations to the solution measured by the 2-norm, and perturbations to A measured by the spectral or Frobenius norms, see Theorem 5.1
y_0	data of a numerical problem, $F(y, x) = 0$, for which x_0 is a true solution
$\bar{}$	overbar indicating a computed value approximating the quantity underneath
*	adjoint, used rather than ^{t} in equation (11)'s subexpression $\|\mathcal{J}_1 F(y_0, x_0)^* f\|$ to emphasize that the norm is for the space of functionals dual to the matrix's column space
^{t}	matrix transpose
†	matrix pseudoinverse
$\simeq$	relationship of asymptotic equality among functions defined in the neighborhood of a point, x_0, see Definition 3.1
$\dot{\leq}$	used by Stoer to indicate that his error bound neglects terms that are second order in δA and δb, see equation (44)

References

[1] N. N. Abdelmalek. On the solution of the linear least squares problems and pseudo-inverses. *Computing*, 13(3–4):215–228, 1974.

[2] E. Anderson, Z. Bai, C. Bischof, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney, S. Ostrouchov, and D. Sorensen. *LAPACK Users' Guide*. Society for Industrial and Applied Mathematics, Philadelphia, 1992.

[3] M. Arioli, I. S. Duff, and P. P. M. de Rijk. On the augmented system approach to sparse least-squares problems. *Numerische Mathematik*, 55(6):667–684, 1989.

[4] T. Banachiewicz. Études d'analyse pratique. Cracow Observatory Reprint 22, University of Cracow, Cracow, 1938. Cited in [18, p. 103].

[5] Bénoit. Note sur une méthode ... (Procédé du Commandant Cholesky). *Bulletin géodésique*, 24(2):5–77, April, May, June 1924. The author is identified only as Commandant Bénoit.

[6] A. Björck. Iterative improvement of linear least squares solutions I. *BIT*, 7(4):257–278, 1967.

[7] A. Björck. Solving linear least squares problems by Gram-Schmidt orthogonalization. *BIT*, 7(1):1–21, 1967.

[8] A. Björck. Componentwise backward errors and condition estimates for linear least squares problems. Technical Report LiTH-MATH-R-1989-13, Department of Mathematics, Linköping University, Linköping, 1989. This document is cited in [51] and as a manuscript dated a year earlier in [31].

[9] A. Björck. Component-wise perturbation analysis and error bounds for linear least squares solutions. *BIT*, 31(2):238–244, 1991. This evidently is the published form of [8].

[10] A. Björck. *Numerical Methods for Least Squares Problems*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, 1996.

[11] J. F. Bonnans and A. Shapiro. *Perturbation Analysis of Optimization Problems*. Springer Series in Optimization Research. Springer Verlag, New York, 2000.

[12] J. R. Bunch. The weak and strong stability of algorithms in numerical linear algebra. *Linear Algebra and Its Applications*, 88/89:49–66, April 1987.

[13] P. Businger and G. H. Golub. Linear least squares solutions by Householder transformations. *Numerische Mathematik*, 7:269–276, 1965.

[14] A. J. Cox and N. J. Higham. Accuracy and stability of the null space method for solving the equality constrained least squares problem. *BIT*, 39(1):34–50, March 1999.

[15] A. J. Cox and N. J. Higham. Backward error bounds for constrained least squares problems. *BIT*, 39(2):210–227, June 1999.

[16] J. W. Demmel. The condition number and the distance to the nearest ill-posed problem. *Numerische Mathematik*, 51(3):251–289, July 1987.

[17] P. S. Dwyer. A matrix presentation of least squares and correlation theory with matrix justification of improved methods of solution. *The Annals of Mathematical Statistics*, 5:82–89, 1944.

[18] P. S. Dwyer. *Linear Computations*. John Wiley & Sons, New York, 1951.

[19] L. Eldén. Perturbation theory for the least squares prolem with linear equality constraints. *SIAM Journal on Numerical Analysis*, 17(3):338–350, 1980.

[20] C. F. Gauss. *Theory of the Combination of Observations Least Subject to Errors, Part One, Part Two, Supplement*, volume 11 of *Classics in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, 1995. Tanslated to English by G. W. Stewart.

[21] G. H. Golub. Numerical methods for solving linear least squares problems. *Numerische Mathematik*, 7:206–216, 1965.

[22] G. H. Golub and C. F. Van Loan. *Matrix Computations*. The Johns Hopkins University Press, Baltimore, first edition, 1983.

[23] G. H. Golub and C. F. Van Loan. *Matrix Computations*. The Johns Hopkins University Press, Baltimore, second edition, 1989.

[24] G. H. Golub and J. H. Wilkinson. Note on the iterative refinement of least squares solutions. *Numerische Mathematik*, 9(2):139–148, December 1966.

[25] J. F. Grcar. Differential equivalence classes for metric projections and optimal backward errors. Technical Report LBNL-51940, Lawrence Berkeley National Laboratory, 2002. Submitted for publication.

[26] J. F. Grcar. A matrix lower bound. Technical Report LBNL-50635, Lawrence Berkeley National Laboratory, 2002. Submitted for publication.

[27] M. Gu. Backward perturbation bounds for linear least squares problems. *SIAM Journal on Matrix Analysis and Applications*, 20(2):363–372, 1999.

[28] M. Gu. New fast algorithms for structured linear least squares problems. *SIAM Journal on Matrix Analysis and Applications*, 20(1):244–269, 1999.

[29] R. J. Hanson and C. L. Lawson. Extensions and applications of the Householder algorithm for solving linear least squares problems. *Mathematics of Computation*, 23(108):787–812, October 1969.

[30] D. J. Higham and N. J. Higham. Backward error and condition of structured linear systems. *SIAM Journal on Matrix Analysis and Applications*, 13(1):162–175, January 1992.

[31] N. J. Higham. Computing error bounds for regression problems. In P. J. Brown and W. A. Fuller, editors, *Statistical Analysis of Measurement Error Models and Applications*, volume 112 of *Contemporary Mathematics*, pages 195–208. American Mathematical Society, Providence, 1990.

[32] N. J. Higham. *Accuracy and Stability of Numerical Algorithms*. Society for Industrial and Applied Mathematics, Philadelphia, first edition, 1996.

[33] A. S. Householder. Unitary triangularization of a nonsymmetric matrix. *Journal of the Association for Computing Machinery*, 5:339–342, 1958.

[34] D. Kahaner, C. Moler, and S. Nash. *Numerical Methods and Software*. Prentice Hall, Englewood Cliffs, 1989.

[35] R. Karlson and B. Waldén. Estimation of optimal backward perturbation bounds for the linear least squares problem. *BIT*, 37(4):862–869, December 1997.

[36] C. L. Lawson and R. J. Hanson. *Solving Least Squares Problems*. Series in Automatic Computation. Prentice-Hall, Englewood Cliffs, 1974.

[37] P. Lötstedt. Perturbation bounds for the linear least squares problem subject to linear inequality constraints. *BIT*, 23:500–519, 1983.

[38] P. Lötstedt. Solving the minimal least squares problem subject to bounds on the variables. *BIT*, 24:206–224, 1984.

[39] A. N. Malyshev. Optimal backward perturbation bounds for the LSS problem. *BIT*, 41(3):430–432, 2001.

[40] A. N. Malyshev and M. Sadkane. Computation of optimal backward perturbation bounds for large sparse linear least squares problems. *BIT*, 41(4):739–747, December 2002.

[41] J. von Neumann and H. H. Goldstine. Numerical inverting of matrices of high order. *Bulletin of the American Mathematical Society*, 53(11):1021–1099, November 1947. Reprinted in [57, v. 5, pp. 479–557].

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

[43] C. C. Paige. Computer solution and perturbation analysis of generalized linear least squares problems. *Mathematics of Computation*, 33(145):171–183, January 1979.

[44] V. Pereyra. Stability of general systems of linear equations. *aequationes mathematicae*, 2(2–3):194–206, 1969.

[45] J. R. Rice. A theory of condition. *SIAM Journal on Numerical Analysis*, 3(2):287–310, 1966.

[46] J. L. Rigal and J. Gaches. On on the compatibility of a given solution with the data of a linear system. *Journal of the Association of Computing Machinery*, 14(3):543–548, 1967.

[47] R. D. Skeel. Scaling for numerical stability in Gaussian elimination. *Journal of the Association for Computing Machinery*, 26(3):494–526, July 1979.

[48] G. W. Stewart. On the perturbation of pseudo-inverses, projections and linear least squares problems. *SIAM Review*, 19(4):634–662, October 1977.

[49] G. W. Stewart. Research development and LINPACK. In J. R. Rice, editor, *Mathematical Software III*, pages 1–14. Academic Press, New York, 1977.

[50] G. W. Stewart. *Matrix Algorithms 1: Basic Decompositions*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, 1998.

[51] G. W. Stewart and J.-G. Sun. *Matrix Perturbation Theory*. Academic Press, San Diego, 1990.

[52] J. Stoer. *Einführung in die Numerische Mathematik I*. Heidelberger Taschenbücher. Springer Verlag, Berlin, first edition, 1972.

[53] J. Stoer and R. Bulirsch. *Introduction to Numerical Analysis*. Springer Verlag, New York, second edition, 1980. The original text is [52].

[54] J.-G. Sun. Backward perturbation analysis of certain characteristic subspaces. *Numerische Mathematik*, 65(3):357–382, July 1993.

[55] J.-G. Sun. Optimal backward perturbation bounds for the linear least-squares problem with multiple right-hand sides. *IMA Journal of Numerical Analysis*, 16(1):1–11, January 1996.

[56] J.-G. Sun. On optimal backward perturbation bounds for the linear least-squares problem. *BIT*, 37(1):179–188, March 1997.

[57] A. H. Taub, editor. *John von Neumann Collected Works*. Macmillan, New York, 1963.

[58] A. M. Turing. Rounding-off errors in matrix processes. *The Quarterly Journal of Mechanics and Applied Mathematics*, 1(3):287–308, September 1948.

[59] A. van der Sluis. Stability of the solutions of linear least squares problems. *Numerische Mathematik*, 23:241–254, 1975.

[60] J. M. Varah. Backward error estimates for Toeplitz systems. *SIAM Journal on Matrix Analysis and Applications*, 15(2):408–417, April 1994.

[61] B. Waldén, R. Karlson, and J.-G. Sun. Optimal backward perturbation bounds for the linear least squares problem. *Numerical Linear Algebra With Applications*, 2(3):271–286, 1995.

[62] P.-A. Wedin. Perturbation theory for pseudo-inverses. *BIT*, 13(2):217–232, 1973.

[63] P.-A. Wedin. Perturbation theory and condition numbers for generalized and constrained linear least squares problems. Report UMINF 125.85, University of Umea, Umea, 1985. Report S-901-87, Institute of Information Processing, University of Umea, Umea, 1987. These versions of the document are cited in [32] and [51], respectively.

[64] M. Wei. Perturbation of the least squares problem. *Linear Algebra and Its Applications*, 141:177–182, November 1990.

[65] J. H. Wilkinson. *Rounding Errors in Algebraic Processes*. Prentice Hall, Englewood Cliffs, New Jersey, 1963.

[66] S. Wolfram. *The Mathematica Book*. Wolfram Media / Cambridge University Press, Champaign and Cambridge, third edition, 1996.