1 Solving Equations: Robustness to Noise

This lecture introduces basic principles of *robustness* with respect to *additive noise* for solving systems of linear equations in several variables. The main ideas here are similar to those used in designing optimal linear filters and controllers (H2 and H-Infinity optimization, Kalman filtering, etc.), but the presentation is greatly simplified since it does not involve a dynamical systems aspect. The lecture can also serve as a brief tutorial on linear algebra basics.

1.1 Exact Solutions of Linear Equations

This section deals with *exact* solutions of systems of linear equations, defining linearity and formulating its most basic properties.

1.1.1 Linearity and Linear Equations

The problems to be studied in this section can be written symbolically in the form

$$LF = M,$$

(1.1)

where $L, M$ are two given *linear transformations*, and $F$ is a linear transformation to be found. Equation (1.1) can also be represented by the block diagram on Figure 1.1.

![Figure 1.1: Equation Solving: a Block Diagram](image)

There are two useful ways of interpreting equation (1.1) (and, similarly, block diagram of Figure 1.1). In terms of *matrix algebra*, $L$ and $M$ are viewed as given matrices of dimensions $k$-by-$m$ and $k$-by-$d$ respectively, and $F$ is the $m$-by-$d$ matrix to be found. The (most common) situation is when $L, M$ are real matrices, and $F$ is required to be real as well: the setup to be considered here. The case when $L, M$ and $F$ are arbitrary complex matrices is very similar.

Since (1.1) means that $LFv = Mv$ for every $d$-by-1 column real matrix $v$ (i.e. a *vector* from $\mathbb{R}^d$), it is reasonable to view $v$ as a vector input data variable, $u = Fv \in \mathbb{R}^m$ as the
decision made by "solver" $F$, and $e = LFv - Mv \in \mathbb{R}^k$ as the error. For exact equation solving, $e$ is required to be zero for all $v \in \mathbb{R}^d$.

An alternative (more powerful but also more abstract) coordinate-free interpretation of (1.1) views $L, M$, and $F$ as general linear functions $L : U \mapsto E$, $M : V \mapsto E$, $F : V \mapsto U$ on real vector spaces $V, U, E$ (and $LF$ denotes the composition $L \circ F$ of $L$ and $F$). Briefly speaking, a real vector space is a set on which operations of addition and scaling (multiplication by real numbers) are defined, satisfying the usual axioms of commutativity, distributivity, multiplication by zero and the unit, etc. If $X$ and $Y$ are real vector spaces, a function $f : X \mapsto Y$ is called linear when it satisfies the superposition law

$$f(c_1 x_1 + c_2 x_2) = c_1 f(x_1) + c_2 f(x_2) \quad \forall c_1, c_2 \in \mathbb{R}, x_1, x_2 \in X.$$  

Formally, the second interpretation of (1.1) is more general than the first one, because the set $\mathbb{R}^n$ is naturally a real vector space for every $n$, and, for a fixed $m$-by-$n$ real matrix $M$, the function $f : \mathbb{R}^n \mapsto \mathbb{R}^m$ defined by $f(v) = Mv$ is linear. On the other hand, most applications deal with finite dimensional vector spaces $X$, for which every element $x \in X$ can be uniquely represented in the form

$$x = c_1 x_1 + c_2 x_2 + \cdots + c_n x_n, \quad c = \begin{bmatrix} c_1 \\ \vdots \\ c_n \end{bmatrix} \in \mathbb{R}^n, \quad (1.2)$$

where $\{x_1, \ldots, x_n\}$ is a fixed set of elements of $X$, called a basis in $X$. When a one-to-one representation (1.2) is possible, the real vector space $X$ is said to have finite dimension $n$, and there is little practical difference between $X$ and $\mathbb{R}^n$. Moreover, when $X$ and $Y$ are two real vector spaces with bases $B_X = \{x_1, \ldots, x_n\} \subset X$ and $B_Y = \{y_1, \ldots, y_m\} \subset Y$ then every linear function $f : X \mapsto Y$ is uniquely represented by its matrix with respect to $B_X$ and $B_Y$: an $m$-by-$n$ matrix $M$ such that $h = Mc$ whenever

$$h = \begin{bmatrix} h_1 \\ \vdots \\ h_m \end{bmatrix} \in \mathbb{R}^m, \quad c = \begin{bmatrix} c_1 \\ \vdots \\ c_n \end{bmatrix} \in \mathbb{R}^n$$

are such that $y = f(x)$ for

$$y = h_1 y_1 + \cdots + h_m y_m, \quad x = c_1 x_1 + c_2 x_2 + \cdots + c_n x_n.$$  

Accordingly, in such situation there is no essential difference between linear functions and matrix-vector multiplication.

The coordinate-free interpretation is very helpful in understanding the basic principles of linearity, and also in proving theoretical statements in linear algebra. The matrix algebra approach provides a direct link to computer-aided calculations, as they are typically defined and performed in a matrix format.
Example 1.1 Let $k$ be a positive integer. Consider the algebraic equation

$$(1 + s)^k p(s) + (1 - s)^k p(-s) = q(s^2), \quad (1.3)$$

where $p$ and $q$ are polynomials of degree less than $k$ with real coefficients. The task of solving (1.3) for $p$ while $q$ is given can be easily recognized as a special case of equation (1.1) (a coordinate-free interpretation), where $L: U \mapsto E$ and $M: V \mapsto E$ are defined as follows: $U = V$ is the real vector space of all real polynomials of degree less than $k$; $E$ is the real vector space of all even real polynomials of degree less than $2k$; function $L$ maps polynomial $u(s)$ to polynomial

$$(Lu)(s) = (1 + s)^k u(s) + (1 - s)^k u(-s);$$

and function $M$ maps polynomial $v(s)$ to polynomial

$$(Mv)(s) = v(s^2).$$

To get a matrix algebra representation of the problem, one can use bases

$$B_U = B_V = \{1, s, \ldots, s^{k-1}\}, \quad B_E = \{1, s^2, s^4, \ldots, s^{2k-2}\}$$

(though this is not necessarily a good choice from a robustness prospective, to be discussed later). For example, for $k = 2$, linear function $M$ maps 1 and $s$ to 1 and $s^2$ respectively. Similarly, linear function $L$ maps 1 and $s$ to

$$(1 + s)^2 + (1 - s)^2 = 2 + 2s^2$$

and

$$(1 + s)^2 s - (1 - s)^2 s = 4s^2.$$

Hence the matrices of $M$ and $L$ in the selected bases are

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 2 & 0 \\ 2 & 4 \end{bmatrix}$$

respectively. Now (1.3) can be solved numerically in MATLAB using something like

$$M = \begin{bmatrix} 2 & 0 \\ 2 & 4 \end{bmatrix}, \quad L = \text{eye}(2), \quad F = M \backslash L,$$

which yields a matrix representation of $F$:

$$F = \begin{bmatrix} 0.5 & 0 \\ -0.25 & 0.25 \end{bmatrix}.$$

1.1.2 Basic Theorems on Solving Linear Equations

The following theorems on solving linear equations will be used throughout the class. Recall that the kernel $\ker f$ of a linear function $f: X \mapsto Y$ is the set of all $x \in X$ such that $f(x) = 0$.

**Theorem 1.1** Let $V, U, E$ be real vector spaces. Let $L: U \mapsto E$ and $M: V \mapsto E$ be linear functions. Then the following conditions are equivalent:
(a) there exists a linear function $F : V \mapsto U$ such that $M = L \circ F$;

(b) $f \circ M = 0$ for every linear function $f : E \mapsto \mathbb{R}$ such that $f \circ L = 0$.

Since $M = L \circ F$ implies $f \circ M = f \circ L \circ F$, implication $(a) \iff (b)$ is easy. The really useful part of Theorem 1.1 is the implication $(b) \iff (a)$.

**Theorem 1.2** Let $V, U, E$ be real vector spaces. Let $L : U \mapsto E$ and $M : V \mapsto E$ be linear functions. Assume that $L$ and $M$ satisfy condition (a) (or, equivalently, (b)) of Theorem 1.1. Then the following conditions are equivalent:

(a) there exists unique function $F : V \mapsto U$ such that $M = L \circ F$;

(b) $\ker L = \{0\}$.

Theorem 1.2 can be viewed as a *uniqueness* condition for the setup of Theorem 1.1. The following statement can be viewed as an easier-to-use special case of Theorem 1.2.

**Theorem 1.3** Let $V, U, E$ be real vector spaces of finite dimensions $d$, $m$, and $k$ respectively. Let $L : U \mapsto E$ be a linear function. Then the following conditions are equivalent:

(a) a linear function $F : V \mapsto U$ such that $M = L \circ F$ exists and is unique for every linear function $M : V \mapsto E$;

(b) $m = k$ and $\ker L = \{0\}$.

Theorem 1.3, essentially, means that a linear system of equations in which the numbers of scalar equations (the dimension of $E$) and scalar variables (the dimension of $U$) are equal has a solution if and only if the kernel of associated linear transformation consists of one element.

For finite dimensional vector spaces, Theorems 1.1-1.3 can be proven using the fact that, with an appropriate selection of bases in $U$ and $E$, the matrix of $L$ will have zeroes everywhere except on the main diagonal. Such “diagonalization” can be achieved using the method of Gaussian elimination, which, with some modifications, is also used for finding approximate solutions of linear equations.

**Example 1.2** Theorem 1.2 can be used to prove existence and uniqueness of solutions of the polynomial equations from Example 1.1. Indeed, the dimensions of $U$ and $E$ are equal. On the other hand, the kernel of $L$ consists of polynomials $u = u(s)$ of degree less than $k$ satisfying equation

$$(1 + s)^k u(s) = -(1 - s)^k u(-s).$$

According to it, $u$ is divisible by $(1 - s)^k$. Because of the degree constraint, this implies $u \equiv 0$. 

4
1.2 Approximation Quality and Robustness to Additive Noise

Consider again the block diagram on Figure 1.1. In many applications, one has to deal with the situation when the number of equations exceeds the number of variables, and hence $e$ cannot be made zero for some values of $v$. Accordingly, the task of designing $F$ to “minimize” $e$ has to be addressed. This section discusses quadratic forms as a tool for quantifying both absolute and relative quality in minimizing $e$ as a function of $v$.

1.2.1 Quadratic Forms and Quadratic Norms

Quadratic forms are used throughout robust control to measure distances, set up design objectives, quantify approximation errors, define Lyapunov functions, etc. The matrix algebra view of a quadratic form is as follows. Let $^t$ (the “prime” sign) denote transposition of a matrix (Hermitian conjugation for complex matrices). A quadratic form on $\mathbb{R}^n$ is a function $\sigma : \mathbb{R}^n \mapsto \mathbb{R}$ defined by the expression $\sigma(x) = x^tRx$, where $R$ is a given $n$-by-$n$ matrix satisfying the symmetry constraint $R = R^t$. The definition generalizes to the coordinate-free framework in the following way. Let $V$ be a real vector space. A quadratic form on $V$ is a function $\sigma : V \mapsto \mathbb{R}$ defined according to $\sigma(x) = b(x,x)$, where $b : V \times V \mapsto \mathbb{R}$ is a symmetric bilinear function, i.e. such that $b(v_1,v_2) = b(v_2,v_1)$ and

$$b(c_1v_1 + c_2v_2,v) = c_1b(v_1,v) + c_2b(v_2,v)$$

for all $v,v_1,v_2 \in V$, $c_1,c_2 \in \mathbb{R}$.

It is easy to see that for a given $n$-by-$n$ matrix $R$ the function $b : \mathbb{R}^n \times \mathbb{R}^n \mapsto \mathbb{R}$ defined by $b(v_1,v_2) = v_1^tRv_2$ is symmetric bilinear. Conversely, if $V$ is a finite dimensional real vector space with basis $B_V = \{v_1,\ldots,v_n\}$ then every symmetric bilinear function $b : V \times V \mapsto \mathbb{R}$ is completely determined by the numbers $R_{ij} = b(v_i,v_j)$. Moreover, if $R$ is the (symmetric) $n$-by-$n$ matrix defined by the entries $R_{ij}$, then for every

$$v = c_1v_1 + \cdots + c_nv_n \in V, \quad c = \begin{bmatrix} c_1 \\ \vdots \\ c_n \end{bmatrix} \in \mathbb{R}^n$$

the equality $b(v,v) = c^tRc$ takes place ($R = R^t$ is called the matrix of $\sigma$ in this case). It is clear that, as in the case of real vector spaces and linear functions, the matrix algebra and coordinate-free approaches to quadratic forms are equivalent for most practical purposes.

A quadratic form $\sigma : V \mapsto V$ is called positive definite when $\sigma(v) > 0$ for all $v \neq 0$, and positive semi-definite when $\sigma(v) \geq 0$ for all $v$. Accordingly, when $V = \mathbb{R}^n$ and $\sigma(v) = v^tRv$ (i.e. $R$ is the matrix of $\sigma$), $R$ is called positive (semi-)definite whenever $\sigma$ is. We will use notation $\sigma \succ 0$ and $R \succ 0$ for positive definiteness, and $\sigma \succeq 0$, $R \succeq 0$ for positive semi-definiteness.

When a quadratic form $\sigma : V \mapsto V$ is positive semi-definite, the expression

$$|v|_\sigma \overset{\text{def}}{=} \sigma(v)^{1/2}$$

(1.4)
satisfies the triangle inequality
\[ |v_1 + v_2|_\sigma \leq |v_1|_\sigma + |v_2|_\sigma. \]

Accordingly, when \( \sigma \) is positive definite, \( |v|_\sigma \) can be used to quantify length of \( v \). Similarly, when \( R = R' \) is a positive semi-definite symmetric \( n \)-by-\( n \) matrix, \( |v|_R \) for \( v \in \mathbb{R}^n \) will denote the square root of \( v'Rv \). In the special case when \( R = I_n \) is an identity matrix (1’s on the diagonal, 0’s elsewhere), \( |v| \), the so-called Euclidean norm of \( v \), will stand for \( |v|_R \). As a rule, we will use \(|e|\) to quantify the amount of error associated with a particular vector \( e \) in the block diagram on Figure 1.1.

### 1.2.2 Induced Gain

While \(|e|\) can be used an an error measure for a specific \( v \), in general \( e \) in Figure 1.1 is a function of \( v \), while \( v \) is expected to be “arbitrary” within certain limits. Accordingly, to quantify quality of a particular equation solver \( F \), one would like to introduce a measure of relative sensitivity of \( e \) with respect to \( v \). In general, one would like to have such measure defined for general transformations \( \Delta : \mathbb{R}^d \mapsto \mathbb{R}^k \) (where \( \Delta = LF - M \) in Figure 1.1), not only the linear ones, or even for maps \( \Delta : V \mapsto E \), where \( V, E \) are two real vector spaces with fixed positive definite quadratic forms.

The two alternative approaches to defining relative error measures are relying upon, respectively, the worst case and average scenarios. A typical example of the worst case definition is that of operator norm (another term for it is gain) induced by the Euclidean length.

**Definition 1.1** Let \( \Delta : \mathbb{R}^d \mapsto \mathbb{R}^k \) be a function. The (Euclidean length induced) gain \( \|\Delta\| \) of \( \Delta \) is the minimal upper bound of the ratio of lengths of \( \Delta(v) \) and \( v \):
\[
\|\Delta\| \overset{\text{def}}{=} \sup_{v \in \mathbb{R}^d, v \neq 0} \frac{|\Delta v|}{|v|} = \inf \left\{ \gamma > 0 : \gamma^2 |v|^2 - |\Delta(v)|^2 \geq 0 \ \forall \ v \in \mathbb{R}^d \right\}.
\]

For a \( k \)-by-\( d \) real matrix \( \Delta \), \( \|\Delta\| \) is the gain of the function \( v \mapsto \Delta v \).

According to the definition, the gain is defined for all functions \( \Delta \), and is either a non-negative number, or plus infinity.

**Example 1.3** The function \( \Delta : \mathbb{R} \mapsto \mathbb{R} \) defined by \( \Delta(v) = \sin(3v) \) has Euclidean length induced gain of 3. Indeed, since \( |\sin(x)| \leq |x| \) for all \( x \), the gain is not larger than 3. On the other hand, \( \sin(3v)/v \) converges to 3 as \( v \neq 0 \) approaches zero, and hence the gain is as large as 3.

**Example 1.4** The function \( \Delta : \mathbb{R} \mapsto \mathbb{R} \) defined by \( \Delta(v) = \cos(v) \) has Euclidean length induced gain of \(+\infty\). Indeed, \( \cos(v)/v \) converges to \( \infty \) as \( v > 0 \) approaches zero.
Example 1.5 The function $\Delta : \mathbb{R}^2 \rightarrow \mathbb{R}$ defined by $\Delta([v_1; v_2]) = v_1v_2$ has Euclidean length induced gain of $+\infty$. Indeed, for $v_1 = v_2 = h \rightarrow \infty$ the ratio $|D(v)|/|v| = |h|/\sqrt{2}$ approaches $\infty$.

Computing the gain for some multivariable functions could be a challenge. However, there is a nice formula covering the case of linear transformations.

Theorem 1.4 The Euclidean length induced gain of a real $k$-by-$d$ matrix $\Delta$ equals the square root of the largest eigenvalue of matrix $\Delta'\Delta$.

Since square roots of the eigenvalues of $\Delta'\Delta$ are called singular values (or singular numbers) of $\Delta$, the Euclidean induced gain of a matrix $\Delta$ is frequently referred to as its largest singular number, and denoted as $\sigma_{\text{max}}(\Delta)$.

Example 1.6 The linear functions

$$
\Delta_+ \left( \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} \right) = \begin{bmatrix} v_1 + v_2 \\ v_1 + v_2 \end{bmatrix} \quad \text{and} \quad \Delta_- \left( \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} \right) = \begin{bmatrix} v_1 + v_2 \\ v_1 - v_2 \end{bmatrix}
$$

have matrices

$$
\Delta_+ = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \quad \text{and} \quad \Delta_- = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}
$$

respectively. As can be checked using the norm.m or eig.m functions of MATLAB, their gains are 2 and $\sqrt{2}$ respectively.

1.2.3 Relative Sensitivity Based on Case Averaging

Using averaging to define relative sensitivity of a function $\Delta : \mathbb{R}^d \rightarrow \mathbb{R}^k$ means, essentially, computing the expected value of a length measure of $e = \Delta(v)$ while assuming a particular distribution of $v$. Indeed, every random variable $\xi$ can be safely interpreted as an integrable (strictly speaking, measurable) function of a hidden “case” parameter $\theta \in [0, 1]$, in which case the expected value $E[\xi]$ of $\xi$ is the integral of $\xi$ over $[0, 1]$.

A commonly used relative sensitivity measure based on averaging is defined for a measureable function $\Delta : \mathbb{R}^d \rightarrow \mathbb{R}^k$ as square root of $E[|\Delta(v)|^2]$, where $v$ is a normalized Gaussian random vector ($E[v] = 0$, $E[vv'] = I_d$). For a linear transformation $v \mapsto \Delta(v) = \Delta v$, the expected value can be calculated explicitly according to

$$
E[|\Delta v|^2] = E[v'\Delta'\Delta v] = E[\text{tr}\{vv'\Delta'\Delta\}] = \text{tr}\{E[vv']\Delta'\Delta\} = \text{tr}\{\Delta'\Delta\},
$$

i.e. equals the square $\|\Delta\|_F^2$ of the Frobenius norm $\|\Delta\|_F$ of $\Delta$ ($\|\Delta\|_F$ is defined as square root of the sum of squares of all elements of $\Delta$).
Example 1.7 Consider the situation when a constant scalar parameter $h$ is measured $n$ times, each measurement $y_i = h + w_i$ a result of adding noise $w_i$ to the true value of $h$. Assuming that the noise to signal ratio $a_i$ in the $i$-th measurement is known, it is natural to model $h$ and $y_i$ according to $y = Lv$ and $h = Mv$, where

$$y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \quad L = \begin{bmatrix} a_1 & 0 & 1 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & a_n \end{bmatrix}, \quad M = \begin{bmatrix} 0 & \cdots & 0 & 1 \end{bmatrix},$$

and $v$ is a normalized $(n+1)$-dimensional Gaussian random variable with zero mean. Assuming that a weighted sum

$$u = F_1 y_1 + \cdots + F_n y_n = F y, \quad F = \begin{bmatrix} F_1 & \cdots & F_n \end{bmatrix}$$

is used to produce an estimate $u$ of $h$, the estimation error $e = h - u$ depends on $v$ according to $e = \Delta v$, where $\Delta = M - FL$. The expected value of $|e|^2$ can be calculated as the square of the Frobenius norm of $\Delta$:

$$E[|e|^2] = \|\Delta\|_F^2 = a_1^2 F_1^2 + \cdots + a_n^2 F_n^2 + (F_1 + \cdots + F_n - 1)^2.$$

It serves as a measure of relative sensitivity of $e$ with respect to $v$ in an average sense, and provides a reasonable assessment of quality of estimator $F$ subject to the setup assumptions.