# A Tutorial Overview of Vector and Matrix Norms

Prepared for the Numerical Analysis Seminar at U.C. Berkeley, 2010-2013 ,
and for lectures 8 & 15 Oct. 2010 at the Univ. of Texas @ Arlington

by W. Kahan, Prof. Emeritus
Mathematics Dept., & Computer Science Dept.
University of California @ Berkeley

This has been posted at
<www.eecs.berkeley.edu/~wkahan/MathH110/NormOvrv.pdf>

# Abstract

Intended for new graduate students whose experience as undergraduates may have prepared them inadequately to apply norms to numerical error-analyses and to proofs of convergence, this tutorial surveys norms for finite-dimensional real spaces in a way that may ease a transition to the infinite-dimensional spaces of Functional Analysis. Among the topics covered are some more useful than is implied by their absence from most curricula. The notation is mostly standard but interpreted in ways not always taught to undergraduates, so attendees may prepare for the tutorial by reading just a few of my lecture notes for Math. H110 posted at

<eecs.berkeley.edu/~wkahan/MathH110/2dspaces.pdf> and <.../pts.pdf> in that order, and afterwards <.../geo.pdf> and <.../geos.pdf> skimmed lightly.

This tutorial omits proofs; almost all can be found in <.../NORMlite.pdf>, <.../GIlite.pdf>, and a few other places cited.

This tutorial's pages have been posted at <.../NormOvrv.pdf> .

# Contents



… continued …

… Contents continued …

# What are Norms for?

They provide vector spaces and their linear operators with measures of
size, length and distance
only a little more general than what we already use routinely in everyday life.

"A little more general" $\Rightarrow$ more widely applicable than our most familiar notions
but still often conforming to our intuitions about them.

**Examples …**

**1•** A "Backward" Error-Analysis of a computed approximation $\mathbf{Y}$ to $f(\mathbf{X})$ :
$\mathbf{Y} + \Delta\mathbf{Y} = f(\mathbf{X} + \Delta\mathbf{X})$ ; are $\Delta\mathbf{Y}$ and $\Delta\mathbf{X}$ negligible?
Depends upon what we can infer about $\|\Delta\mathbf{Y}\|$ and $\|\Delta\mathbf{X}\|$ . vector norms

**2•** Convergence Analysis of an Iteration towards a Fixed-Point $\mathbf{z}$ :
$\mathbf{x}_{n+1} := f(\mathbf{x}_n) = f^{[n+1]}(\mathbf{x}_0) \rightarrow ? \rightarrow \mathbf{z} = f(\mathbf{z})$ ?
Depends upon what we can infer about derivative … $\|f\,`(\mathbf{z})\|$ . matrix norm

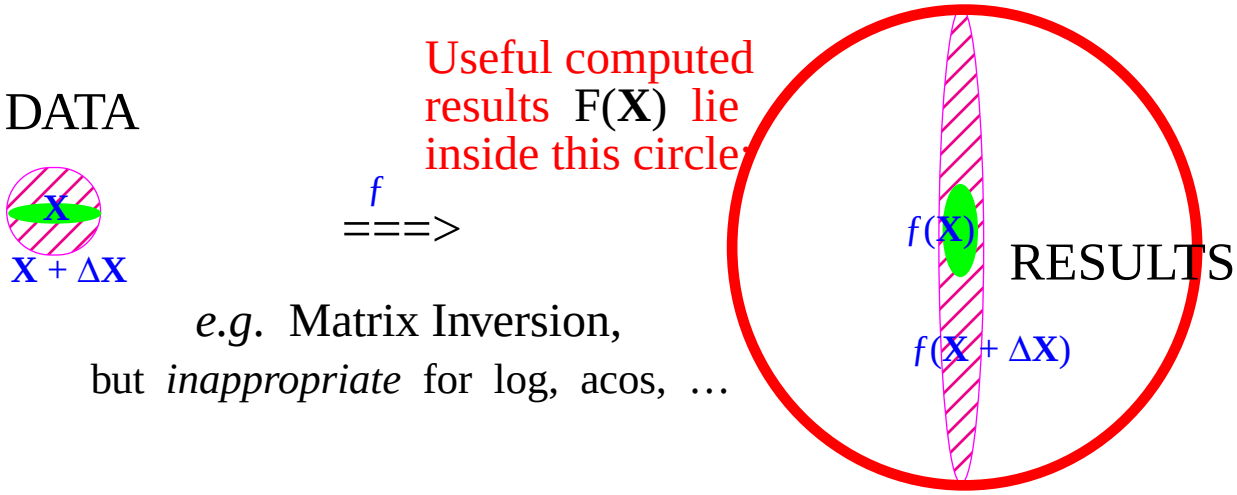
• • •

## **Example 1:** A "Backward" Error-Analysis

A program F(**X**) is intended to compute vector-valued $f(\mathbf{x})$ for vector inputs **x** .
Actually, computed **Y** := F(**X**) only approximates **y** := $f(\mathbf{x})$ . *HOW WELL* **?**

We deem program F to be "Backward Stable" numerically just ***If*** we have proved that $\|F(\mathbf{X}) - f(\mathbf{X})\|$ is at worst slightly bigger than $\|f(\mathbf{X} + \Delta\mathbf{X}) - f(\mathbf{X})\|$ can be for some unknown roundoff-induced perturbation $\Delta\mathbf{X}$ whose $\|\Delta\mathbf{X}\|$ is at worst slightly bigger than negligible compared with $\|\mathbf{X}\|$ for all inputs **X** in a region big enough to be useful.

Our chosen norm may exaggerate uncertainties in actual input data and computed results whose correlations the norm disregards.




*e.g.* Matrix Inversion,
but *inappropriate* for log, acos, …

If F is "backward stable" but computed F(**X**) is very wrong, do we blame the victim $f$ for "ill condition" at **X** ?

Error-Analyses tend to excessive pessimism partly because they allow for unlikely conspiracies among rounding errors, and partly because the chosen norms are often not the most suitable.

# **Example 2:** Convergence Analysis of an Iteration

Given a smooth map $f(\mathbf{x})$ of a vector-space to itself, and a starting vector $\mathbf{x}_0$, let

$$\mathbf{x}_{n+1} := f(\mathbf{x}_n) = f(f(\mathbf{x}_{n-1})) = \ldots = f^{[n+1]}(\mathbf{x}_0) \quad \text{for } n = 0, 1, 2, 3, \ldots \text{ in turn.}$$

Does $\mathbf{x}_n \to \mathbf{z}$ fast enough from every $\mathbf{x}_0$ near enough to a Fixed-Point $\mathbf{z} = f(\mathbf{z})$ ?

Yes if and only if a $\mathbf{z}$ exists and every $|\text{eigenvalue of } f`(\mathbf{z})|$ is sufficiently less than $1$.

But we don't know $\mathbf{z}$ yet, much less the eigenvalues of the derivative $f`(\mathbf{z})$. Jacobian

Instead we explore conditions upon $f$ easier to test. For instance, maybe $f$ is a …

**Contractive Map**: $\|f(\mathbf{y}) - f(\mathbf{x})\|/\|\mathbf{y} - \mathbf{x}\| < \lambda < 1$ whenever distinct $\mathbf{x}$ and $\mathbf{y}$ lie in some *sufficiently large* region $\mathbb{X}$ .

Then *either* $\|\mathbf{x}_n - \mathbf{z}\| \le \lambda^n \cdot \|\mathbf{x}_0 - \mathbf{z}\| \to 0$ so $\mathbf{x}_n \to \mathbf{z} = f(\mathbf{z})$ uniquely in $\mathbb{X}$ ,

*or* ultimately $\mathbf{x}_n$ escapes from $\mathbb{X}$ , which is too small to hold a fixed-point.

And test the Contractive hypothesis: Is $\|\mathbf{x}_{n+1} - \mathbf{x}_n\|/\|\mathbf{x}_n - \mathbf{x}_{n-1}\| < 1$ ? Until roundoff …

THE CATCH: All this makes sense only for an appropriately chosen norm $\|\ldots\|$ .

That is the trouble with norms: There are so many of them, apt choice may be hard.

# Part I

# Vector Norms

# The trouble with norms is that there are so many of them.

To be a vector norm, $\|\ldots\|$ need satisfy only three requirements …

- Positivity: $\infty > \|\mathbf{x}\| > 0$ for every vector $\mathbf{x}$ except $\|\mathbf{o}\| = 0$ .
- Homogeneity: $\|\lambda\cdot\mathbf{x}\| = |\lambda|\cdot\|\mathbf{x}\|$ for every scalar $\lambda$ . Let's keep $\lambda$ real.
- Triangle Inequality: $\|\mathbf{x} + \mathbf{y}\| \le \|\mathbf{x}\| + \|\mathbf{y}\|$ for all $\mathbf{x}$ and $\mathbf{y}$ . Equality need not imply parallelism.

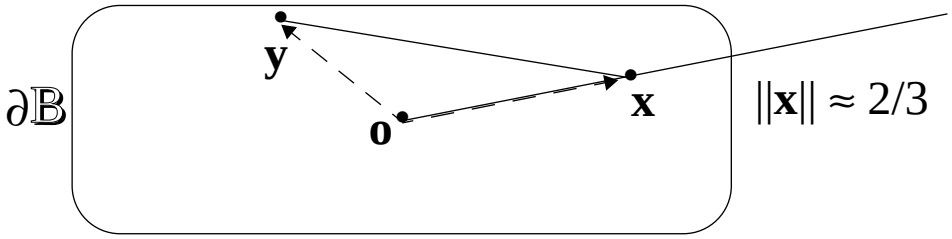
If $\|\mathbf{x}\|$ is a norm, so is $|||\mathbf{x}||| := \|\mathbf{L}^{-1}\cdot\mathbf{x}\|$ for any fixed invertible linear operator $\mathbf{L}$ .

If $\|\mathbf{x}\|$ and $|||\mathbf{x}|||$ are norms, so are $\max\{\|\mathbf{x}\|, |||\mathbf{x}|||\}$ , $\sqrt{(\|\mathbf{x}\|^2 + |||\mathbf{x}|||^2)}$ , $\|\mathbf{x}\| + |||\mathbf{x}|||$ , … .

The **Unit Ball** of a norm $\|\mathbf{x}\|$ is the region $\mathbb{B} := \{ \mathbf{x} : \|\mathbf{x}\| \le 1 \}$ . This $\mathbb{B}$ turns out to be closed, bounded, centrally symmetric ( $\mathbb{B} = -\mathbb{B}$ ) and convex with $\mathbf{o}$ strictly inside. "Convex" means that, if $\mathbf{x}$ and $\mathbf{y}$ lie in $\mathbb{B}$ , so does $\zeta\cdot\mathbf{x} + (1-\zeta)\cdot\mathbf{y}$ for $0 \le \zeta \le 1$ .

Line segment joining "points" $\mathbf{x}$ and $\mathbf{y}$

Conversely, ***any*** region $\mathbb{B}$ closed, bounded, centrally symmetric and convex with $\mathbf{o}$ strictly inside is the Unit Ball of the norm $\|\mathbf{x}\| := \inf ( |\xi| \text{ for which } \mathbf{x}/\xi \text{ lies in } \mathbb{B} )$ .



# The trouble with norms is that there are so many of them.

## How is an appropriate one to be chosen?

Scalar value $\|\mathbf{x}\|$ has to be computable at a tolerable cost. *Computable* from what? Computable from scalar *components* of a representation x of (abstract?) vector $\mathbf{x}$.

An applied mathematician's first challenge is to choose suitable *parameters*
(variables, coordinates, basis vectors, …)
to represent his problem's entities (on paper, on a blackboard, in his computer, …).

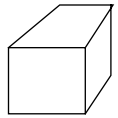
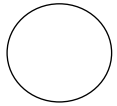
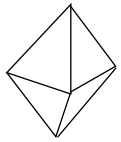
Let $\mathbf{B} := [\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3, \ldots, \mathbf{b}_n]$ be a ***Basis*** for the (abstract?) space of vectors $\mathbf{x}$, so each vector $\mathbf{x} = \mathbf{B}\cdot x = \sum_j \mathbf{b}_j \cdot \xi_j$ is represented by a column-vector $x = [\xi_1; \xi_2; \xi_3; \ldots; \xi_n]$ in MATLAB's notation. Thus, basis $\mathbf{B}$ is an invertible linear map from a space of columns x to the (abstract?) space of vectors $\mathbf{x} = \mathbf{B}\cdot x$; each $\mathbf{x}$ has its own column $x = \mathbf{B}^{-1}\cdot\mathbf{x}$.

Then $\|\mathbf{x}\|$ will be computed as $\|\mathbf{x}\| := \|x\|$ from the components $\xi_j$ of column x.

If a basis $\mathbf{B}$ accepted first is unsatisfactory, a new basis $\overline{\mathbf{B}} := \mathbf{B}\cdot C^{-1}$ can be chosen; here C is an invertible matrix. And the new representative of $\mathbf{x} = \mathbf{B}\cdot x = \overline{\mathbf{B}}\cdot\overline{x}$ is then column $\overline{x} = C\cdot x$, whence the new formula to compute $\|\mathbf{x}\| = \overline{\|}\,\overline{x}\,\overline{\|}$ becomes $\overline{\|}\,\overline{x}\,\overline{\|} = \|C^{-1}\cdot\overline{x}\|$.

Don't Memorize! Re-derive!

Having chosen (or accepted) a basis in which vector **x** is represented by its column $x = [\xi_1; \xi_2; \xi_3; \ldots; \xi_n]$ , we can choose a formula by which to compute $\|\mathbf{x}\| := \|x\|$ from the components $\xi_j$ . Here are three familiar formulas and their Unit Balls for $n = 3$ :

- $\|x\|_\infty := \max |\xi_j|$ . Ball is a solid Cube.  Postal box
- $\|x\|_2 := \sqrt{(\sum |\xi_j|^2)}$ . Ball is a solid Sphere.  Euclidean length!
- $\|x\|_1 := \sum |\xi_j|$ . Ball is a solid Octahedron  Taxicab norm

These norms are unchanged by permutation of the elements of x ; these norms treat each element the same as any other. What if some elements need closer scrutiny than others?

Let W be an invertible diagonal matrix; $|||x||| := \|W^{-1}\cdot x\|$ is a norm too, maybe better.

Let C be an invertible matrix of a change of basis; $|||x||| := \|C^{-1}\cdot x\|$ is a norm too.

A change of variables often induces a change of norm. Or is it *vice-versa* ?

A change of variables often induces a change of norm. Or is it *vice-versa* ?

# How should a norm be chosen for error-analysis?

***Ideally***, an appropriate norm would be what I call ***Equitable***, *i.e.*, so chosen that …

All errors of the same norm have roughly equal (in)significance.

This ideal is often not achievable before a computation's results have been inspected, and not always achievable afterwards, but always worth attempting.

**Example:** The control of electric power distribution networks involves voltages of widely diverse magnitudes, ranging from those on cross-country transmission lines to those in radio receivers of the microwave transmissions needed to control power generation.

| Locations | | Voltage Ranges | Better Units |
|---|---|---|---|
| 1 | Cross-Country Transmission Towers | 250,000 - 1,000,000 | Megavolts |
| 2 | Substation to Transformers on Telephone Poles | 2,000 - 10,000 | Kilovolts |
| 3 | In-house wall sockets to appliances | 110 - 240 | Volts |
| 4 | Computer power supplies to transistors | 2 - 12 | Volts |
| 5 | Transistor circuits' on-off variations | 0.001 - 0.01 | Millivolts |
| 6 | Inputs to radio receivers' antennas | 0.000,001 - 0.000,01 | Microvolts |

Changes by 0.001 **Volt** are negligible in locations 1 - 4 , devastating in the last two.
Changes by 0.001 **Unit** are negligible in all locations.

A change of variables often induces a change of norm. Or is it *vice-versa* ?

# How should a norm be chosen for convergence analysis?

Ideally, the norm should be so chosen that a *Stationary Iteration* …

$x_{n+1} := f(x_n) \to z = f(z)$ converges monotonically in norm.

This ideal is often not achievable, but always worth attempting.

**Local Convergence:** If $1 > \lambda > |$each eigenvalue of $f`(z)|$, norms $\|\ldots\|$ exist such that

$\|x_{n+1} - z\| \le \lambda \cdot \|x_n - z\| \le \lambda^{n+1} \cdot \|x_0 - z\| \to 0$ for every $x_0$ close enough to $z$ .

In principle such a norm can be constructed by changing to a new basis obtained from the eigendecomposition of $f`(z)$ , but this merely raises hopes for something more practical.

**Global Convergence:** Rather than by shrinkage of a norm, convergence must be proved (if true) by shrinkage of a *Lyapunov Function* whose Level-lines/surfaces form a nested family shrinking onto $z$ , and whose shapes need not be centrally symmetric and convex.

**Global Convergence** of a *Non-Stationary Iteration* $x_{n+1} := f_n(x_n) \to z = f_m(z)$ for all $m$ can rarely be proved using norms; rare exceptions are mostly matrix computations like *Conjugate Gradients*. Other global convergence proofs need some other monotonicity.

E.g.: §6 of <eecs.berkeley.edu/~wkahan/Math128/GnSymEig.pdf> uses a monotonic determinant, not a norm at all.

And now for something entirely different:

# Why are vector spaces like potato chips?

# Why are vector spaces like potato chips?

# **Because you cannot have just one.**

Each space of vectors $\mathbf{x}$ comes with its *Dual-Space* of *Linear Functionals* $\mathbf{w}^T$ :

- *Scalar Product* $\mathbf{w}^T\cdot\mathbf{x}$ is a scalar, real for real spaces. Complex: $\mathbf{w}^*\cdot\mathbf{x}$ or $\mathbf{w}^H\cdot\mathbf{x}$
- $\mathbf{w}^T$ acts linearly upon vectors $\mathbf{x}$ and $\mathbf{y}$ : $\mathbf{w}^T\cdot(\lambda\cdot\mathbf{x} + \mu\cdot\mathbf{y}) = \lambda\cdot\mathbf{w}^T\cdot\mathbf{x} + \mu\cdot\mathbf{w}^T\cdot\mathbf{y}$ .
- $\mathbf{x}$ acts linearly upon $\mathbf{v}^T$ and $\mathbf{w}^T$ : $(\lambda\cdot\mathbf{v}^T + \mu\cdot\mathbf{w}^T)\cdot\mathbf{x} = \lambda\cdot\mathbf{v}^T\cdot\mathbf{x} + \mu\cdot\mathbf{w}^T\cdot\mathbf{x}$ .

So the linear functionals $\mathbf{w}^T$ form a vector space *Dual* or *Conjugate* to the space of vectors $\mathbf{x}$ . Each space is dual to the other, and they have the same finite dimension.

But among infinite dimensional spaces, many are properly contained within their dual's dual.

Since $\mathbf{x}$ need not be a column, $\mathbf{w}^T$ need not be a row, much less the "Transpose" of a vector $\mathbf{w}$ . Except for *Euclidean* and other *Inner-Product* spaces, there is no necessary relation between $\mathbf{w}^T$ and $\mathbf{w}$ , just as Miss Carla and Master Carlo need not be related.

Compare *Contravariant* and *Covariant Tensors*.

Distinctions between dual spaces are obscured by other notations for $\mathbf{w}^T\cdot\mathbf{x}$ like $\mathbf{w}'\cdot\mathbf{x}$, $\mathbf{w}\bullet\mathbf{x}$, $\langle\mathbf{x}, \mathbf{w}\rangle$, $\langle\mathbf{x}|\mathbf{w}\rangle$, $(\mathbf{x}, \mathbf{w})$, … devised originally for *Euclidean* and other *Inner-Product* spaces, each one *Isomorphic* to its dual. Many mathematicians expect context to disambiguate those notations.

*E.g.*: The space dual to $\mathbf{x}$'s contains the derivative $\mu`(\mathbf{x})$ of any scalar function $\mu(\mathbf{x})$ because scalar $d\mu(\mathbf{x}) = \mu`(\mathbf{x})\cdot d\mathbf{x}$ .

# The Dual-Space's Norm

The natural dual norm for functional $\mathbf{w}^T$ is $\|\mathbf{w}^T\| := \max |\mathbf{w}^T\cdot\mathbf{x}|/\|\mathbf{x}\|$ over $\mathbf{x} \neq \mathbf{o}$ .

From that follows $\|\mathbf{x}\| = \max |\mathbf{w}^T\cdot\mathbf{x}|/\|\mathbf{w}^T\|$ over $\mathbf{w}^T \neq \mathbf{o}^T$ . Non-trivial proof in general

**Examples** of dual norms for rows and columns: … in MATLAB's notation …

Say column $x = [\xi_1; \xi_2; \xi_3; \ldots; \xi_n]$ , and row $w^T = [\omega_1, \omega_2, \omega_3, \ldots, \omega_n]$ .

- Dual to $\|x\|_\infty := \max |\xi_j|$ is $\|w^T\|_\infty = \sum |\omega_j|$
- Dual to $\|x\|_2 := \sqrt{(\sum |\xi_j|^2)}$ is $\|w^T\|_2 = \sqrt{(\sum |\omega_j|^2)}$
- Dual to $\|x\|_1 := \sum |\xi_j|$ is $\|w^T\|_1 = \max |\omega_j|$

Hölder's Inequality: $|\mathbf{w}^T\cdot\mathbf{x}| \leq \|\mathbf{w}^T\|\cdot\|\mathbf{x}\|$ for all $\mathbf{w}^T$ and $\mathbf{x}$ . *Cf.* Cauchy's for $\|\ldots\|_2$

# Vectors and Functionals Dual with respect to the Norm

$\mathbf{z}$ and $\mathbf{u}^T$ are called "Dual w.r.t. the norm" when $\mathbf{u}^T\cdot\mathbf{z} = \|\mathbf{u}^T\|\cdot\|\mathbf{z}\|$ and $\|\mathbf{u}^T\| = \|\mathbf{z}\| \neq 0$.

This duality relation is also called "Polarity". Its $\mathbf{z}$ and $\mathbf{u}^T$ determine each other nonlinearly and perhaps nonuniquely in general, but linearly and uniquely only in Euclidean and in Inner-Product spaces to be described imminently. Examples …

- Grad $\mu(\mathbf{x})$ is a vector in $\mathbf{x}$'s space dual to derivative $\mu`(\mathbf{x})$ in the dual-space.
  From Hölder's Inequality, $\mu(\mathbf{x})$ changes fastest in the direction of Grad $\mu(\mathbf{x})$, and $\|\text{Grad } \mu(\mathbf{x})\| = \|\mu`(\mathbf{x})\|$ .

- Row(s) $w^T = [\omega_1, \omega_2, \omega_3, \ldots, \omega_n]$ **dual** to column $x = [\xi_1; \xi_2; \xi_3; \ldots; \xi_n]$ w.r.t. $\|x\|$ :

» For $\|x\|_\infty$ : If $|\xi_j| < \|x\|_\infty$ then $\omega_j := 0$
else let $\text{sign}(\omega_j) := \text{sign}(\xi_j)$ or $0$, and $\sum |\omega_j| := \|x\|_\infty$ .
This dual $w^T$ is unique just when only one $|\xi_j| = \|x\|_\infty$ .

» For $\|x\|_2$ : $w^T := x^T$ (its transpose **!**) uniquely. (Complex conjugate transpose for complex spaces)

» For $\|x\|_1$ : If $\xi_j = 0$ then choose any $\omega_j$ with $|\omega_j| \leq \|x\|_1$
else $\omega_j := \|x\|_1\cdot\text{sign}(\xi_j)$ .
This dual $w^T$ is unique just when every $\xi_j \neq 0$ .

## Changing a vector space's Basis changes the Dual-space's Basis:

Column $x$ that represents vector $\mathbf{x} = \mathbf{B}\cdot x$ in the basis $\mathbf{B} := [\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3, \ldots, \mathbf{b}_n]$ has its counterpart in row $w^T$ that represents functional $\mathbf{w}^T = w^T\cdot\mathbf{B}^{-1}$ in the same basis. This notation preserves the scalar product $\mathbf{w}^T\cdot\mathbf{x} = w^T\cdot x$ as a matrix product in a natural way.

Just as the "columns" $\mathbf{b}_j$ of $\mathbf{B}$ are basis vectors for the space of vectors $\mathbf{x}$, the "rows" of $\mathbf{B}^{-1}$ are basis vectors $\mathbf{e}_j^T$ (functionals) for the dual-space of linear functionals $\mathbf{w}^T$. "Evaluation" functional $\mathbf{e}_j^T$ extracts $\mathbf{x}$'s component $\xi_j = \mathbf{e}_j^T\cdot\mathbf{x}$ of column $x = \mathbf{B}^{-1}\cdot\mathbf{x}$.

Changing basis to $\overline{\mathbf{B}} := \mathbf{B}\cdot C^{-1}$ changes column $x$ representing vector $\mathbf{x} = \mathbf{B}\cdot x = \overline{\mathbf{B}}\cdot\overline{x}$ to $\overline{x} = C\cdot x$, and changes row $w^T$ representing $\mathbf{w}^T = w^T\cdot\mathbf{B}^{-1} = \overline{w^T}\cdot\overline{\mathbf{B}}^{-1}$ to $\overline{w^T} = w^T\cdot C^{-1}$.

- When can " $\ldots^T$ " be construed as an operator instead of merely a suffix?

  When does row $\overline{w^T} = w^T\cdot C^{-1}$ match the transpose $(\overline{w})^T = w^T\cdot C^T$ of column $\overline{w} = C\cdot w$ ?

  They match only if $C^T = C^{-1}$ is an *Orthogonal* matrix. That is what happens when $\mathbf{B}$ and $\overline{\mathbf{B}}$ are *Orthonormal* bases for a Euclidean space, which is the prototypical instance of an *Inner-Product Space* ...

# Real Inner-Product Spaces

A real Inner-Product Space possesses a symmetric scalar product $\mathbf{x}\bullet\mathbf{y} = \mathbf{y}\bullet\mathbf{x}$ linear in each of $\mathbf{x}$ and $\mathbf{y}$ separately, and with $\mathbf{x}\bullet\mathbf{x} > 0$ except $\mathbf{o}\bullet\mathbf{o} = 0$. Therefore the space is normed: $\|\mathbf{x}\| := \sqrt{\mathbf{x}\bullet\mathbf{x}}$. But its Unit Ball is generally an *Ellipsoid*. ( Complex $\mathbf{x}\bullet\mathbf{y} = \overline{\mathbf{y}\bullet\mathbf{x}}$ .)

An inner-product space has a natural linear map between it and its dual space. Vector $\mathbf{x} = \mathbf{B}\cdot x$ represented by column $x$ maps to linear functional $\mathbf{x}\bullet = x^T\cdot M\cdot\mathbf{B}^{-1}$ represented by row $x^T\cdot M$ for some *symmetric positive definite* matrix $M = M^T$, the space's *Metric*. $M$ comes from the basis vectors in $\mathbf{B} := [\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3, \ldots, \mathbf{b}_n]$ thus: $M_{ij} = \mathbf{b}_i\bullet\mathbf{b}_j$. Similarly $\mathbf{w}\bullet = w^T\mathbf{B}^{-1}$ represented by row $w^T$ maps to $\mathbf{w} = \mathbf{B}\cdot M^{-1}\cdot w$. Now representative columns $x = \mathbf{B}^{-1}\cdot\mathbf{x}$ and $y = \mathbf{B}^{-1}\cdot\mathbf{y}$ figure in the formula to obtain $\mathbf{y}\bullet\mathbf{x} = y^T\cdot M\cdot x$; and $\|\mathbf{x}\| = \sqrt{(x^T\cdot M\cdot x)}$ but $\|\mathbf{w}\bullet\| = \sqrt{(w^T\cdot M^{-1}\cdot w)}$. Now $\mathbf{x}$ and $\mathbf{x}\bullet$ are duals w.r.t. $\|\ldots\|$.

A Euclidean space's metric $M$ is the Identity matrix $I$, which simplifies everything.

Changing basis $\mathbf{B}$ to $\overline{\mathbf{B}} := \mathbf{B}\cdot C^{-1}$ changes metric $M$ to $\overline{M} = C^{-1T}\cdot M\cdot C^{-1}$. Therefore

Every real inner-product space is a Euclidean space disguised perhaps by a non-orthonormal basis $\mathbf{B}$ .

$\mathbf{B}$ changes to an orthonormal basis $\overline{\mathbf{B}} := \mathbf{B}\cdot C^{-1}$ when $C^T\cdot C = M$. ... Gram-Schmidt, Cholesky

# What distinguishes an Inner-Product or Euclidean space from all other *Normed* vector spaces?

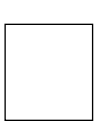
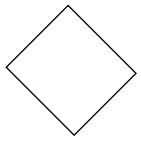
The ***Parallelogram Law***: $\|\mathbf{x} + \mathbf{y}\|^2 + \|\mathbf{x} - \mathbf{y}\|^2 \equiv 2\|\mathbf{x}\|^2 + 2\|\mathbf{y}\|^2$ . … Jordan-von Neumann Th'm

And then $\mathbf{x}\bullet\mathbf{y} \equiv ( \|\mathbf{x} + \mathbf{y}\|^2 - \|\mathbf{x} - \mathbf{y}\|^2 )/4$ .

For a proof see <www.eecs.berkeley.edu/~wkahan/MathH110/QF.pdf>

- Every Inner-Product space is ***Isomorphic*** to its Dual-space.

But every *other* normed n-space is *NOT Isomorphic* to its dual space except possibly if dimension $n \le 2$ .

 $\|\ldots\|_\infty \leftrightarrow \|\ldots\|_1$ 

Generally, flat spots on the surface of a normed space's unit ball match up with vertices of the dual space's unit ball, and *vice-versa*.

# The Derivative of a Vector Norm

( Say dimension $n > 1$ .)

It is determined uniquely only where the norm's unit ball is smooth (no vertex nor edge), as is an inner-product space's ellipsoidal unit ball, everywhere but at $\mathbf{o}$ .

Then $d\|\mathbf{z}\| = \mathbf{u}^T \cdot d\mathbf{z} / \|\mathbf{z}\|$ in which $\mathbf{u}^T$ is the linear functional dual to $\mathbf{z}$ w.r.t. $\|\ldots\|$ .

More generally, $d\|\mathbf{z}\|$ is an extremum of $\mathbf{u}^T \cdot d\mathbf{z} / \|\mathbf{z}\|$ over all the linear functionals $\mathbf{u}^T$ dual to $\mathbf{z}$ w.r.t. $\|\ldots\|$ .

The foregoing formula helps to provide a short proof of the most useful special case of …

**Auerbach's Theorem:** Any n-space's unit ball $\mathbb{B}$ can be circumscribed by at least one *Parallelepiped* $\mathbb{P}$ whose $2n$ faces touch $\mathbb{B}$ at their midpoints.

One such $\mathbb{P}$ is a circumscribing parallelepiped of minimum volume. And if a new basis $\mathbf{P}$ is chosen from just the vectors that join $\mathbf{o}$ to the midpoints of $\mathbb{P}$'s faces, then the column $x := \mathbf{P}^{-1} \cdot \mathbf{x}$ that represents n-dimensional vector $\mathbf{x}$ in this new basis $\mathbf{P}$ has

$$\|x\|_1 / n \le \|x\|_\infty \le \|\mathbf{x}\| \le \|x\|_1 \le n \cdot \|x\|_\infty \quad \text{for every } \mathbf{x} .$$

- Therefore any n-dimensional norm $\|\ldots\|$ can be approximated within a factor $n^{\pm 1/2}$ by either $\|\ldots\|_\infty$ or $\|\ldots\|_1$ , each of which costs at most $n$ operations to compute after the change of basis (whose one-time cost may be of order $n^3$).

An Ellipsoid is the level surface of a Positive Definite Quadratic Form; see pp. 62-63.

## Fritz John's Ellipsoid Theorem: (1948)

Any n-dimensional normed space's unit ball $\mathbb{B}$ can be circumscribed by an ellipsoid $\mathbb{E}$ tightly enough that $\sqrt{n}\cdot\mathbb{B} \supseteq \mathbb{E} \supseteq \mathbb{B}$ .

One such $\mathbb{E}$ is a circumscribing ellipsoid of minimum volume. And if a new basis $\mathbf{E}$ is chosen from the principal axes of $\mathbb{E}$ then the column $x := \mathbf{E}^{-1}\cdot\mathbf{x}$ that represents n-dimensional vector $\mathbf{x}$ in this new basis $\mathbf{E}$ has $\|\mathbf{x}\|/\sqrt{n} \le \|x\|_2 \le \|\mathbf{x}\|$ for every $\mathbf{x}$ .

- Therefore ***any*** n-dimensional norm $\|\ldots\|$ can be approximated within a factor $n^{\pm 1/4}$ by $\|\ldots\|_2$ at a cost of order n after the change of basis.

Often a change of basis involving nothing more onerous than choices of suitable units (recall the voltages' example on p. 12) allows one of the cheap norms $\|\ldots\|_\infty$, $\|\ldots\|_2$ or $\|\ldots\|_1$ to provide adequate error-estimates.

Fritz John's Ellipsoid Theorem has vastly many other useful implications some of which are mentioned in my posted lecture notes <…/MathH110/NORMlite.pdf>

However, if dimension n is too big or infinite, different norms may defy comparison.

# Part II

# Matrix Norms

All the linear operators **L** mapping one given normed space of vectors **x** to another constitute a third vector space and therefore can be subjected to any of a vast horde of norms. We prefer that each such **Matrix Norm** satisfy these *four* requirements:

- Positivity: $\infty > \|\mathbf{L}\| > 0$ for every **L** except $\|\mathbf{O}\| = 0$ .
- Homogeneity: $\|\lambda\cdot\mathbf{L}\| = |\lambda|\cdot\|\mathbf{L}\|$ for every scalar $\lambda$ . Let's keep $\lambda$ real.
- Triangle Inequality: $\|\mathbf{L} + \mathbf{K}\| \le \|\mathbf{L}\| + \|\mathbf{K}\|$ for all **L** and **K** .
- Compatibility: $\|\mathbf{L}\cdot\mathbf{x}\| \le \|\mathbf{L}\|\cdot\|\mathbf{x}\|$ for all **L** and **x** . Subordination

## Overloaded Notation

Note that the last requirement involves three norms all written "||…||" . This notation is heavily overloaded, obliging readers to disambiguate norms by close attention to the linguistic type and context of each norm's argument, just as we learned to distinguish $\|\mathbf{w}^T\|$ from $\|\mathbf{x}\|$ . A *Compatible* Matrix norm is compatible with one or two vector norms and consequently with their dual norms: Compatibility implies $\|\mathbf{w}^T\cdot\mathbf{L}\| \le \|\mathbf{w}^T\|\cdot\|\mathbf{L}\|$ too.

This *Compatibility* requirement is trifling because it can always be met by scaling up the norm offered for linear maps **L** : If $\mu := ( \max \|\mathbf{L}\cdot\mathbf{x}\|$ over $\|\mathbf{L}\| = \|\mathbf{x}\| = 1 ) > 1$ then replace $\|\mathbf{L}\|$ by a scaled-up norm $|||\mathbf{L}||| := \|\mathbf{L}\|\cdot\mu$ to make it compatible.

(Then if identity **I** maps a space to itself but with a different norm, $\|\mathbf{I}\|$ can have any positive value**!** )

# The trouble with norms is that there are so many of them.

## How is an appropriate one to be chosen? … to be computed?

Scalar value $\|\mathbf{L}\|$ has to be computable at a tolerable cost. *Computable* from what? Computable from scalar *components* of a representation L of (abstract?) operator **L** .

Let **B** be a basis for vectors **x** in the domain of **L** . and **E** a basis for vectors **y** in the target-space containing the range of **L** . Then matrix $L := \mathbf{E}^{-1}\cdot\mathbf{L}\cdot\mathbf{B}$ represents **L** , and $\|\mathbf{L}\|$ must be computed from its elements $L_{ij}$ . These change when bases are changed.

More generally, most of a second college course in linear algebra concerns the question

"What can be inferred about a linear operator **L** from its matrix L given only the natures of **L**'s domain- and target-spaces but not their bases?"

If only the spaces' dimensions are known, only *Rank*(**L**) := Rank(L) can be inferred.
If **L** maps an otherwise undistinguished space to itself, only **L**'s *Jordan Normal Form*.
If **L** maps one Euclidean space to another, only the *Singular Values* of **L** .
If **L** is a *Symmetric* map between a space and its dual space, only **L**'s *Inertia*.

see Sylvester's ***Inertia*** Theorem on p. 63

If too little is known about the domain- and target-spaces of linear operator **L** , not much about it can be inferred (numerically) from arbitrary matrix representations L of **L** .

Here are three Matrix norms for m-by-n matrices L that take account only of its elements $L_{ij}$ , not of the operator **L** that L represents, so these norms' theory is shallow:

- $\|L\|_\Sigma := \sum_i \sum_j |L_{ij}|$ extends from $\|x\|_1$
- $\|L\|_F := \sqrt{( \sum_i \sum_j |L_{ij}|^2 )}$ … the *Frobenius* norm, the earliest matrix norm.
- $\|L\|_\mu := n \cdot \max_{i,j} |L_{ij}|$ extends from $\|x\|_\infty$

• Compatibilities: $\|L\|_\Sigma$ & $\|x\|_1$ , $\|L\|_F$ & $\|x\|_2$ , $\|L\|_\mu$ & $\|x\|_\infty$ , among others.

Each of these norms possesses separately a property called *Multiplicative Dominance* :

$\|K \cdot L\| \le \|K\| \cdot \|L\|$ whenever matrices K and L can be multiplied.

This property implies *Compatibility* with inherited n-by-1 vector norms. Consequently Multiplicative Dominance is so useful that we shall require it of all Matrix norms.

And we shall get it from *Operator Norms* defined thus:

- $\|\mathbf{L}\| := ( \max \|\mathbf{L} \cdot \mathbf{x}\|$ over $\|\mathbf{x}\| = 1 ) = ( \max \|\mathbf{w}^T \cdot \mathbf{L}\|$ over $\|\mathbf{w}^T\| = 1 )$
  $= ( \max |\mathbf{w}^T \cdot \mathbf{L} \cdot \mathbf{x}|$ over $\|\mathbf{w}^T\| = \|\mathbf{x}\| = 1 )$ with dual $\|\ldots^T\|$ and $\|\ldots\|$

***Danger*** : Here five distinguishable norms are all written the same way: "‖…‖" .

Operator Norms are also called "*Sup* Norms" and "*LUB*-norms" (Least Upper Bound).

# Operator Norms

- $\|\mathbf{L}\| := (\max \|\mathbf{L}\cdot\mathbf{x}\| \text{ over } \|\mathbf{x}\| = 1) = (\max \|\mathbf{w}^T\cdot\mathbf{L}\| \text{ over } \|\mathbf{w}^T\| = 1)$
  $= (\max |\mathbf{w}^T\cdot\mathbf{L}\cdot\mathbf{x}| \text{ over } \|\mathbf{w}^T\| = \|\mathbf{x}\| = 1)$ with $\|\ldots^T\|$ dual to $\|L\cdot\ldots\|$

Later we shall see how to test whether a given Matrix norm $\|L\|$ is an Operator Norm too.

For now observe that Operator Norms possess Multiplicative Dominance in this way:
$\|\mathbf{K}\cdot\mathbf{L}\| \le \|\mathbf{K}\|\cdot\|\mathbf{L}\|$ if Domain($\mathbf{K}$) $\supseteq$ Range($\mathbf{L}$) and both have the same vector norm.

And for linear maps $\mathbf{L}$ of a space to itself each Operator Norm satisfies $\|\mathbf{L}\| \ge |\lambda|$ for every eigenvalue $\lambda$ of $\mathbf{L}$; and $\|\mathbf{I}\| = 1$.

A few Operator Norms can be computed at modest cost from the matrix L representing $\mathbf{L}$ in suitably chosen bases for its domain- and target-spaces. Because there are many uses for different norms, they must be distinguished by subscripts that clutter the notation:

Let $\|\mathbf{L}\|_{\alpha\beta} := (\max \|\mathbf{L}\cdot\mathbf{x}\|_\alpha/\|\mathbf{x}\|_\beta \text{ over } \mathbf{x} \ne \mathbf{o}) = (\max \|\mathbf{w}^T\cdot\mathbf{L}\|_\beta/\|\mathbf{w}^T\|_\alpha \text{ over } \mathbf{w}^T \ne \mathbf{o}^T)$
except we abbreviate $\|\mathbf{L}\|_{\alpha\alpha}$ to $\|\mathbf{L}\|_\alpha$. Most subscripts $\alpha$ and $\beta$ will be 1, 2 or $\infty$.

Actually we compute $\|L\|_{\alpha\beta}$ from elements $L_{ij}$ of $\mathbf{L}$'s matrix $L := \mathbf{E}^{-1}\cdot\mathbf{L}\cdot\mathbf{B}$, just as we got $\|x\|_\beta$ from elements $\xi_i$ of $x := \mathbf{B}^{-1}\cdot\mathbf{x}$, and $\|w^T\|_\alpha$ from $\omega_j$ of $w^T = \mathbf{w}^T\cdot\mathbf{E}$.

# Formulas for Five Familiar Operator Norms of Matrices

Here L is an m-by-n matrix of elements $L_{ij}$ :

- $\|L\|_\infty = \max_i \sum_j |L_{ij}| = \|L^T\|_1$ , the biggest Row-Sum of Magnitudes
- $\|L\|_1 = \max_j \sum_i |L_{ij}| = \|L^T\|_\infty$ , the biggest Column-Sum of Magnitudes
- $\|L\|_{\infty 1} = \max_i \max_j |L_{ij}| = \|L^T\|_{\infty 1} = \|L\|_\mu/n = \|L^T\|_\mu/m$ , the biggest Magnitude
- $\|L\|_{\infty 2} = \max_i \sqrt{( \sum_j |L_{ij}|^2 )}$ , the biggest Euclidean Row-Length
- $\|L\|_2$ = ( the biggest *Singular Value* of L ) = $\|L^T\|_2 = \sqrt{}$( biggest Eigenvalue of $L^T \cdot L$ )

Also $\|L\|_2$ = the biggest Eigenvalue of $\begin{bmatrix} O & L^T \\ L & O \end{bmatrix}$ from the *Singular Value Decomposition.*

And $\|L\|_2 \le \sqrt{(\|L\|_1 \cdot \|L\|_\infty)}$ . Do you see why?

$\|L\|_{12}$ , $\|L\|_{21}$ , $\|L\|_{1\infty}$ and $\|L\|_{2\infty}$ are little used because their computations cost too much.

If dimensions are not too big, then, just as any vector norm $\|\ldots\|$ can be approximated by $\|\ldots\|_1$, $\|\ldots\|_2$ or $\|\ldots\|_\infty$ perhaps after a change of basis, so can any Operator Norm …

# Maximized Ratios of Norms

Set $\mu_{\alpha\beta} := \max_{\mathbf{x} \neq \mathbf{o}} \|\mathbf{x}\|_\alpha / \|\mathbf{x}\|_\beta = \max_{\mathbf{w}^T \neq \mathbf{o}^T} \|\mathbf{w}^T\|_\beta / \|\mathbf{w}^T\|_\alpha$ when $\alpha \neq \beta$ .

All such ratios $\mu_{\alpha\beta}$ are finite for finite dimensions $n$ , but may grow with $n$ . For instance, $\mu_{\infty 1} = \mu_{\infty 2} = \mu_{21} = 1$ ; $\mu_{12} = \mu_{2\infty} = \sqrt{n}$ ; $\mu_{1\infty} = n$ . Generally $\mu_{\alpha\beta} \cdot \mu_{\beta\alpha} > 1$ .

There are analogous ratio-maxima for m-by-n Matrix norms; for instance
$\mu_{F\Sigma} := \max_{L \neq O} \|L\|_F / \|L\|_\Sigma = 1$ ; $\mu_{\Sigma F} = \sqrt{m \cdot n}$ ; $\mu_{\mu F} = \mu_{\mu\Sigma} = n$ ; $\mu_{F\mu} = \sqrt{m/n}$ ; $\mu_{\Sigma\mu} = m$ .

Operator Norms inherit their ratio-maxima from their associated vector norms:

Recalling that $\|\mathbf{L}\|_{\alpha\beta} := \max_{\mathbf{x} \neq \mathbf{o}} \|\mathbf{L} \cdot \mathbf{x}\|_\alpha / \|\mathbf{x}\|_\beta$ , set $\mu_{\alpha\beta\gamma\delta} := \max_{\mathbf{L} \neq \mathbf{O}} \|\mathbf{L}\|_{\alpha\beta} / \|\mathbf{L}\|_{\gamma\delta}$ .
Then it so happens that $\mu_{\alpha\beta\gamma\delta} = \mu_{\alpha\gamma} \cdot \mu_{\delta\beta}$ .

Caution: If $\mathbf{L}$ is not square, then ratios $\mu_{\ldots}$ must take different dimensions into account.

These ratio-maxima $\mu_{\ldots}$ reveal how well or badly one norm can approximate another.

The trouble with norms is that there are so many of them.
How is an appropriate one to be chosen?

# How should a Matrix norm be chosen for error-analysis?

***Ideally***, an appropriate norm would be what I call ***Equitable***, *i.e.*, so chosen that …

All errors of the same norm have roughly equal (in)significance.

This ideal is often not achievable before a computation's results have been inspected, and not always achievable afterwards, but always worth attempting.

For example, given two diagonal matrices $V_t$ and $V_d$ with wildly diverse magnitudes on their diagonals, but such that the nonzero elements of $V_t \cdot K \cdot V_d$ span a modest range of magnitudes (such a K is called a "Graded" matrix), then variations $\Delta K$ in K may be gauged more equitably by norm $|||\Delta K||| := ||V_t \cdot \Delta K \cdot V_d||$ than by a familiar $||\Delta K||$ .

If $||\ldots||$ is an Operator Norm, so is $|||\ldots|||$ , but derived from vector norms after basis changes tantamount to scaling the vectors: $|||x||| := ||V_d^{-1} \cdot x||$ for x in the **d**omain-space; $|||y||| := ||V_t \cdot y||$ for y in the **t**arget-space of $\Delta K$ whose $|||\Delta K||| = \max_{x \neq o} |||\Delta K \cdot x|||/|||x|||$ .

Another example: Compute a norm $[\![\Delta K]\!] := ||\,\overline{\Delta K}\,||$ from the elements $\overline{\Delta K}_{ij} := \Delta K_{ij}/\Omega_{ij}$ scaled by a given array of positive *Weights* $1/\Omega_{ij}$ . When $||\ldots|| = ||\ldots||_{\infty 1}$ , the biggest magnitude, $[\![\ldots]\!]$ is called an "Elementwise" norm, especially if every $\Omega_{ij} := |K_{ij}| > 0$ .

# When is a preassigned Matrix norm also an Operator Norm ?

Operator Norms have properties that some mere Matrix norms lack. For instance, the norm of a rank-1 matrix or operator is $||\mathbf{x}\cdot\mathbf{w}^T|| = ||\mathbf{x}||\cdot||\mathbf{w}^T||$ for every Operator Norm but not for every compatible Matrix norm.

To simplify the exploration that follows, suppose a given Matrix norm $[\![\ldots]\!]$ for square n-by-n matrices possesses Positivity, Homogeneity and the Triangle Inequality but, for lack of defined vector norms, perhaps not Compatibility.

Instead we assume that $[\![\ldots]\!]$ possesses Multiplicative Dominance: $[\![K\cdot L]\!] \le [\![K]\!]\cdot[\![L]\!]$, which can always be realized by scaling $[\![\ldots]\!]$ if necessary. Then choose a fixed n-row $r^T \ne o^T$ and define a (perhaps new) vector norm to be $||x|| := [\![x\cdot r^T]\!]$, the same for both the domain- and target-spaces of the n-by-n matrices for which $[\![\ldots]\!]$ was given. Now this vector norm induces an Operator Norm $||L|| := \max_{x \ne o} ||L\cdot x||/||x||$. We find that this $||L|| \le [\![L]\!]$ for all n-by-n matrices L . Thus, there is a sense in which …

Operator Norms are minimal among Multiplicatively Dominant norms.

And $||L|| = [\![L]\!]$ for all L only when the given $[\![\ldots]\!]$ was already an Operator Norm.

*Eg.* $[\![L]\!] = ||L||_F \Rightarrow ||L|| = ||L||_2$ ; $[\![L]\!] = ||L||_\mu \Rightarrow ||L|| = ||L||_\infty$ ; $[\![L]\!] = ||L||_\Sigma \Rightarrow ||L|| = ||L||_1$ .

# Orthogonally Invariant Matrix Norms

These are for m-by-n matrices L that represent operators **L** that map one Euclidean space to another. These norms satisfy $\|L\| = \|Q\cdot L\cdot P\|$ for all Orthogonal matrices $Q^T = Q^{-1}$ and $P^T = P^{-1}$. (Analogous Unitarily Invariant norms are for complex unitary spaces, and especially for Hilbert spaces whose dimensions are infinite.)

These norms $\|L\|$ depend solely upon the *Singular Values* of L, which are the biggest $\min\{m, n\}$ eigenvalues of $\begin{bmatrix} O & L^T \\ L & O \end{bmatrix}$ (these come in $\pm$ pairs). In fact, every such norm is $\|L\| = \|v\|$ in which column vector v is a column of the singular values of L and $\|\ldots\|$ can be any vector norm.

Almost all useful orthogonally invariant matrix norms are *Symmetrical Cross-Norms* :

- "Symmetrical" means the order of the singular values in v does not matter to $\|v\|$.
- "Cross-Norms" means that the norm of a rank-1 matrix is $\|x\cdot w^T\| = \|x\|_2\cdot\|w^T\|_2$.

Among these the most common by far are the Multiplicatively Dominant …

- $\|L\|_2 = \|v\|_\infty$ =the biggest singular value of L, the natural Operator Norm.
- $\|L\|_F = \|v\|_2 = \sqrt{(\text{Trace}(L^T\cdot L))} = \sqrt{(\sum(\text{squared singular values}))}$, the Frobenius norm.

$$\|C\cdot L\|_F \le \|C\|_2\cdot\|L\|_F \le \|C\|_F\cdot\|L\|_F$$

- $\|L\|_N = \|v\|_1 = \sum$ (singular values), the "Nuclear" norm, used for some optimizations.

# Dual Norms for Dual Matrix Spaces

The vector space of all real m-by-n matrices L has its Dual-space consisting of n-by-m real matrices $K^T$ with the scalar product $\sum_i \sum_j K_{ij} \cdot L_{ij} = \text{Trace}(K^T \cdot L) = \text{Trace}(L \cdot K^T)$ .

Given a Matrix norm $\|L\|$ for m-by-n matrices L , the natural norm for their dual-space of n-by-m matrices $K^T$ is $\|K^T\| := \max_{L \neq O} \text{Trace}(K^T \cdot L) / \|L\|$ .

This places a severe strain upon our subscripting notation. For example …

Dual to the norm $\|L\|_\infty = \max_i \sum_j |L_{ij}|$ is $\|K^T\|_{(\infty)} := \sum_i \max_j |K_{ij}| \geq \|K^T\|_\infty = \|K\|_1$ .

The strain is less severe for our orthogonally invariant norms; for example …

Dual to the norm $\|L\|_F = \sqrt{(\text{Trace}(L^T \cdot L))}$ is $\|K^T\|_F = \|K\|_F$ .

Dual to $\|L\|_2 = \max(\text{singular value}(L))$ is $\|K^T\|_N = \sum(\text{singular values}(K^T)) = \|K\|_N$.

This last duality explains partially why the Nuclear norm figures in some optimizations that can be reformulated advantageously in a dual-space. In certain optimizations, the K that minimizes $\|K\|_N$ subject to appropriate constraints also minimizes Rank(K) .

Do not confuse Dual Norms for Dual Matrix Spaces with …

## Norms for Linear Maps £ from a Real Vector Space to its Dual:

Scalar $\mathbf{£}\cdot\mathbf{x}\cdot\mathbf{y}$ is linear in $\mathbf{x}$ and $\mathbf{y}$ separately. … (conjugate-linear in y if space is complex)

$\mathbf{£}\cdot\mathbf{x}\cdot\mathbf{y}$ is called a ***Bilinear Form*** , also written $£(\mathbf{x}, \mathbf{y})$ .

Matrix L representing $\mathbf{£}$ depends upon Basis $\mathbf{B} = [\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_n]$ : $L_{ij} = \mathbf{£}\cdot\mathbf{b}_j\cdot\mathbf{b}_i$ so

$\mathbf{£}\cdot\mathbf{x}\cdot\mathbf{y} = y^T\cdot L\cdot x$ for columns $x := \mathbf{B}^{-1}\cdot\mathbf{x}$ and $y := \mathbf{B}^{-1}\cdot\mathbf{y}$ . … ( … = y'·L·x if complex)

Change of Basis from $\mathbf{B}$ to $\overline{\mathbf{B}} := \mathbf{B}\cdot C^{-1}$ changes L to $\overline{L} := (C^{-1})^T\cdot L\cdot C^{-1}$ ; then …

$\overline{L}$ and L are called ***Congruent***. See also pp. 62-3 for more about Congruence.

Operator Norms: $\|\mathbf{£}\| := \max |\mathbf{£}\cdot\mathbf{x}\cdot\mathbf{y}|$ over $\|\mathbf{x}\| = \|\mathbf{y}\| = 1$ .

$\|L\|_{\underline{\alpha}\alpha} := \max |y^T\cdot L\cdot x|$ over $\|x\|_\alpha = \|y\|_\alpha = 1$ . $\quad \|L\|_{\underline{\alpha}\alpha} = \|L^T\|_{\underline{\alpha}\alpha}$

Examples:

$\|L\|_{\underline{1}1} = \|L\|_{\infty 1} = \max_i \max_j |L_{ij}|$ .

$\|L\|_{\underline{2}2} = \|L\|_2 =$ biggest singular value of L .

$\|L\|_{\underline{\infty}\infty} = \|L\|_{1\infty} \le \sum_i \sum_j |L_{ij}|$ and $\|L\|_{\underline{\infty}\infty} \ge \max\{ \|L\|_1, \|L\|_\infty \}$ .

$\mathbf{£}$ is called *Symmetric* when $\mathbf{£}\cdot\mathbf{x}\cdot\mathbf{y} \equiv \mathbf{£}\cdot\mathbf{y}\cdot\mathbf{x} = y^T\cdot L\cdot x \equiv x^T\cdot L\cdot y$ , whereupon $L^T = L$ .

See also p. 62.

# Part III:
# Matrix Norms and Matrix Inverses

## Condensed Review of Parts I & II :

**1.** Try to avoid choosing a norm ||…|| badly;
take care to choose appropriate Bases, Coordinates, Variables.

Ideally, perturbations with the same norm would be about equally (in)significant.

The choice of $\|...\|_1$ , $\|...\|_2$ or $\|...\|_\infty$ matters most when dimensions are big.

**2.** Among Matrix norms, the Operator Norms $\|\mathbf{L}\| := \max_{\mathbf{x}\neq\mathbf{0}} \|\mathbf{L}\cdot\mathbf{x}\|/\|\mathbf{x}\|$ have the most useful properties ... multiplicative ..., minimality ... .

# Matrix Norms and Matrix Inverses

Sensitivity of Inverses to (Infinitesimal) Perturbations:

Changing $\mathbf{L}$ to $\mathbf{L} + \delta\mathbf{L}$ changes $\mathbf{L}^{-1}$ to $\mathbf{L}^{-1} + \delta(\mathbf{L}^{-1})$ ;

$\delta(\mathbf{L}^{-1}) = -\mathbf{L}^{-1}\cdot\delta\mathbf{L}\cdot\mathbf{L}^{-1}$ so $\|\delta(\mathbf{L}^{-1})\| \le \|\delta\mathbf{L}\|\cdot\|\mathbf{L}^{-1}\|^2$, and equality is achievable.

***Condition Number*** (for Inversion): $\kappa(\mathbf{L}) := \|\mathbf{L}^{-1}\|\cdot\|\mathbf{L}\|$ , an amplification factor;

$\|\delta(\mathbf{L}^{-1})\|/\|\mathbf{L}^{-1}\| \le \kappa(\mathbf{L})\cdot\|\delta\mathbf{L}\|/\|\mathbf{L}\|$ and equality is achievable.

Also $\|\delta(\mathbf{L}^{-1})\|/\|\mathbf{L}^{-1}\| \ge (1/\kappa(\mathbf{L}))\cdot\|\delta\mathbf{L}\|/\|\mathbf{L}\|$ , because $\kappa(\mathbf{L}^{-1}) = \kappa(\mathbf{L})$ .

Perhaps $\kappa(\mathbf{L})$ would be better regarded as a *Distortion* factor.

Sensitivity to Perturbations in Data $\{\mathbf{L}, \mathbf{c}\}$ of Solutions $\mathbf{x}$ of equation $\mathbf{L}\cdot\mathbf{x} = \mathbf{c}$ :

$\|\delta\mathbf{x}\|/\|\mathbf{x}\| \le \kappa(\mathbf{L})\cdot\left(\|\delta\mathbf{L}\|/\|\mathbf{L}\| + \|\delta\mathbf{c}\|/\|\mathbf{c}\|\right)$ , and equality is achievable.

If & when it succeeds,

Backward Error-Analysis maps rounding errors to induced perturbations in data.

$\Rightarrow$ As matrix inversion's rounding errors propagate they get amplified by $\kappa(\mathbf{L})$ .

***Ill-Condition*** means a *HUGE* amplification factor $\kappa(\mathbf{L})$ . *cf.* p. 6.

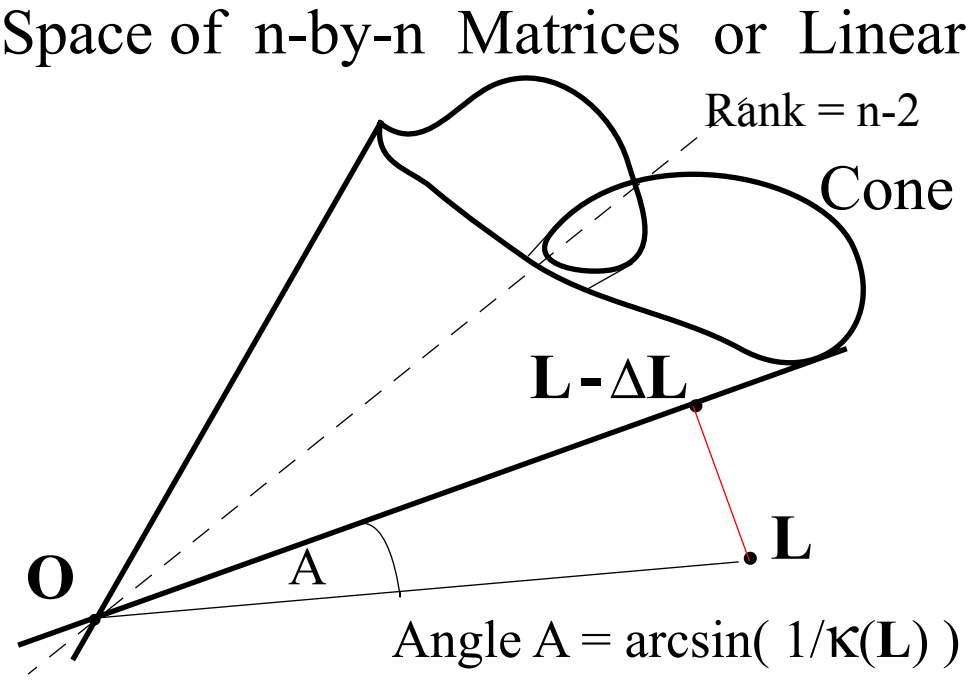
# Operator Norms of Inverses

**Theorem:** For any Operator Norm $\|...\|$, … known to Banach ?

$$\|\mathbf{L}^{-1}\| = 1/\big(\min \|\Delta\mathbf{L}\| \text{ over } \mathrm{Rank}(\mathbf{L}-\Delta\mathbf{L}) < \mathrm{Rank}(\mathbf{L})\big) \ .$$

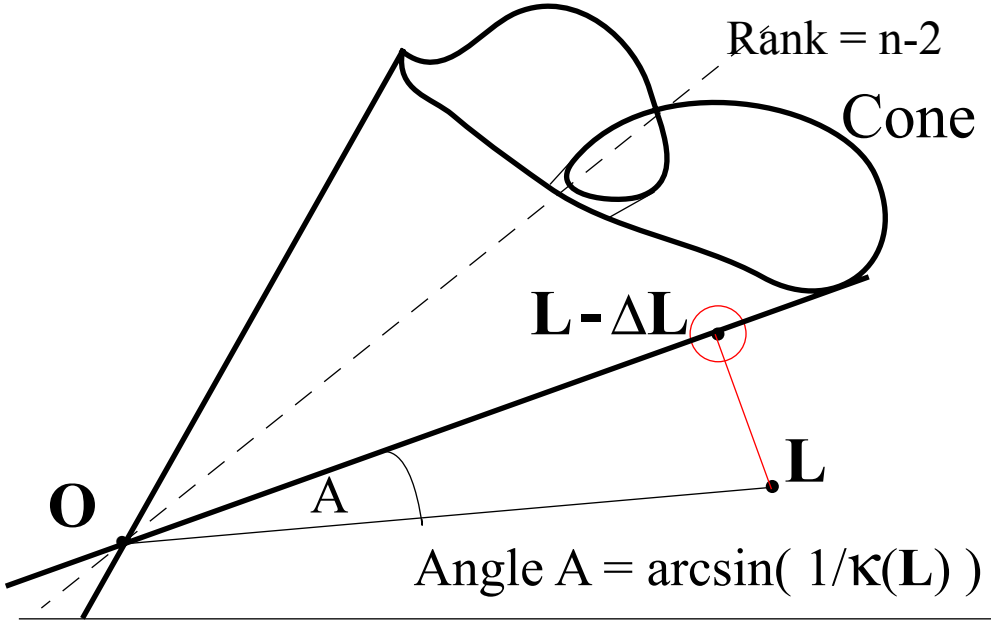
Condition Number (for Inversion): $\kappa(\mathbf{L}) := \|\mathbf{L}^{-1}\|\cdot\|\mathbf{L}\|$ amplifies perturbations' effects

---




An Ill-Conditioned Matrix **L** is extremely close to Singular Matrices

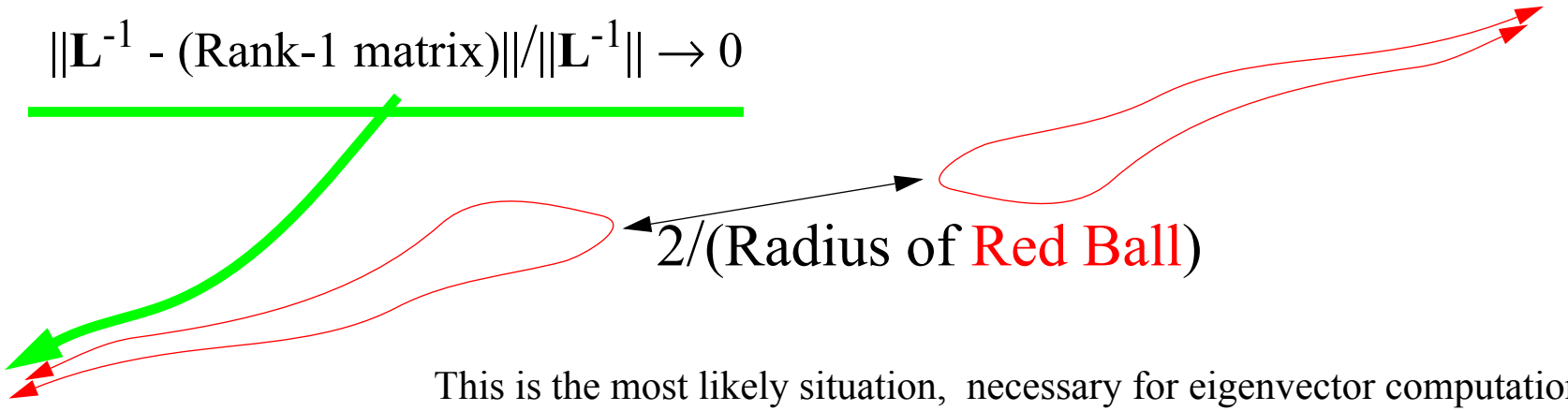
# Space of n-by-n Matrices or Linear Operators




How does $\mathbf{L}^{-1}$ behave as $\mathbf{L}$ approaches a Singular matrix ?

Where are the inverses of the little Red Ball ?

$\|\mathbf{L}^{-1} - (\text{Rank-1 matrix})\|/\|\mathbf{L}^{-1}\| \to 0$




This is the most likely situation, necessary for eigenvector computation.

**Equilibration** invokes ***Diagonal Scaling*** to help compute the solution x of an equation " $C \cdot x = b$ " more nearly as accurately as the data deserve, though limited by the available arithmetic's precision: Choose apt diagonals $V_t$ & $V_d$ ; replace C by $\underline{C} := V_t \cdot C \cdot V_d$ , b by $\underline{b} := V_t \cdot b$ ; solve $\underline{C} \cdot \underline{x} = \underline{b}$ ; get $x := V_d \cdot \underline{x}$ .

How are "apt diagonals" $V_t$ & $V_d$ to be chosen?

- To avoid introducing extraneous roundoff, restrict diagonals' elements to powers of the arithmetic's radix (2 for Binary, 10 for Decimal).
- If the uncertainties in the elements of the data {C, b} are known, diagonals should ideally be chosen to make some common norm, say $\|\ldots\|_\infty$ , *Equitable* for scaled data: *i.e.*, every $\Delta\underline{C}$ with the same $\|\Delta\underline{C}\|_\infty$ is very roughly equally (in)consequential or (in)significant or … . This ideal may be unattainable, especially before solution x has been estimated.
- If the uncertainties in the elements of the data {C, b} are unknown, diagonals should ideally be chosen to roughly minimize some common condition number, say $\kappa_p(\underline{C}) = \|\underline{C}\|_p \cdot \|\underline{C}^{-1}\|_p$ for p in $\{1, 2, \infty\}$ , of scaled data. This ideal usually costs too much; *e.g.* see p. 59. There is an exception:

**Equilibration** continues … minimize condition number … one exception:

A. van der Sluis' Theorem (1969):

Suppose H is an N-by-N *Positive-definite* symmetric or Hermitian matrix. See p. 63

Then $\kappa_2((\mathrm{Diag}(H))^{-1/2}\cdot H\cdot(\mathrm{Diag}(H))^{-1/2}) \le N\cdot\min_{\text{diagonal } V} \kappa_2(V\cdot H\cdot V)$ .

- Other equilibration schemes abound, all somewhat mysterious. For instance:
  Compute diagonals $V_t$ & $V_d$ to turn every row- and column-sum of magnitudes of $\underline{C} := V_t\cdot C\cdot V_d$ into 1 , making $|\underline{C}|$ *Doubly Stochastic*. This computation is iterative; usually it converges fast but at times appallingly slowly, especially when equilibration is most needed.

Gaussian Elimination is affected by equilibration only through its effect upon the order of pivot selection. This effect may be thwarted if column exchanges are disallowed, and then computed results can be undeservedly grossly inaccurate, remediable only by ***Iterative Refinement***. See striking examples posted on my <…/Math128/FailMode.pdf>

**Preconditioning** resembles equilibration's attempt to reduce $\kappa(V_t\cdot C\cdot V_d)$ a lot but allows non-diagonal choices of $V_t$ & $V_d$ , and rarely computes $V_t\cdot C\cdot V_d$ explicitly. Preconditioning is an essential step in the fast solution of discretized continuum problems by iterative methods like ***Conjugate Gradients***.

# Iterative Refinement

## attenuates ill effects of Ill-Condition or BIG dimensions or ... ?

Given F and c, let G stand for operations performed to solve "F·z = c" for z ;
*e.g.*, triangular factorization of F , *etc.*, or Conjugate Gradient iteration, ... .

Computation yields instead $x := G \cdot c \approx z$ ... roundoff accumulates, iteration stops, ...

Let ***Residual*** $r := c - F \cdot x$ , computed ***as accurately as is affordable***. Then reuse (if saved) operations in G to get $y := x + G \cdot r \approx z$ more closely than $x \approx z$ .

If it works, why does it work ? $y - z = ( I - G \cdot F ) \cdot (x - z)$ + more roundoff,
and if $G \approx F^{-1}$ roughly then we expect $\|I - G \cdot F\| << 1$ , so $\|y - z\| << \|x - z\|$ .

It might not work if ...

F is too Ill-Conditioned, within too few rounding errors of singular, or
G is too inaccurate, or *cf.* <www.eecs.berkeley.edu/~wkahan/Math128/FailMode.pdf>
Residual r is too inaccurate, drowned perhaps in its own rounding errors
for lack of extra-precise accumulation of **r** .

# Diagonal Dominance

Some matrices are obviously invertible.

Suppose $\mathbf{L}$ maps a space linearly to itself, and $\|\mathbf{L}\| < 1$. Then $(\mathbf{I} - \mathbf{L})^{-1}$ exists because Liouville's series $(\mathbf{I} - \mathbf{L})^{-1} = \mathbf{I} + \mathbf{L} + \mathbf{L}^2 + \mathbf{L}^3 + \mathbf{L}^4 + \ldots$ converges.

Square matrix B is said to have Rows Dominated by its Diagonal when *every* $|b_{ii}| > \sum_{j \neq i} |b_{ij}|$ ; then $B^{-1}$ exists. ( $L := I - \mathrm{Diag}(B)^{-1} \cdot B$ has $\|L\|_\infty < 1$ )

Square matrix B is said to have Columns Dominated by its Diagonal when *every* $|b_{jj}| > \sum_{i \neq j} |b_{ij}|$ ; then $B^{-1}$ exists. ( $L := I - B \cdot \mathrm{Diag}(B)^{-1}$ has $\|L\|_1 < 1$ )

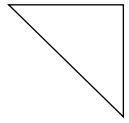
Often B is invertible though " = " replaces almost all " > " signs, but not always.

*Gaussian Elimination* generally needs *Pivotal Exchanges* to ensure numerical stability, but they turn out to be unnecessary for matrices dominated by their diagonals and also for *symmetric Positive definite* matrices. … see p. 63.

# Schur Complements

$\mathbf{S} := Z - E \cdot C^{-1} \cdot D$ is the *Schur Complement* of C in $B := \begin{bmatrix} C & D \\ E & Z \end{bmatrix}$.

Without pivotal exchanges, Gaussian Elimination reduces B through successive Schur Complements of B's leading principal submatrices, like C,

in the course of Triangular (LU) Factorization: $B \longrightarrow \begin{bmatrix} U_C & U_D \\ O & \mathbf{S} \end{bmatrix} \longrightarrow U =$ 

- All Schur Complements in Diagonally Dominant matrices B are also Diagonally Dominant, and do not grow much in norm.

- All Schur Complements in Symmetric Positive Definite matrices B are also Symmetric Positive Definite, and do not grow at all in norm.

What matters most is "… do not grow …" lest elements of Z be corrupted.

The choice of norm implied in "grow" can affect "corrupted" drastically;
*cf.* pp. 3-6 of <…/Math128/FailMode.pdf>

When does equation " $\mathbf{F}\cdot\mathbf{x} = \mathbf{y}$ " have at least one solution $\mathbf{x}$ ?

And if it exists, when is solution $\mathbf{x}$ unique ?

# Ivar Fredholm's Alternatives:

Valid also in many an infinite-dimensional space; **no determinants!**

- At least one solution $\mathbf{x}$ exists if & only if
$$\mathbf{w}'\cdot\mathbf{y} = 0 \text{ whenever } \mathbf{w}'\cdot\mathbf{F} = \mathbf{o}'$$
as $\mathbf{w}'$ runs through the vector space dual to $\mathbf{F}$'s target space.

- If it exists, solution $\mathbf{x}$ is unique if & only if $\mathbf{F}\cdot\mathbf{z} \neq \mathbf{o}$ whenever $\mathbf{z} \neq \mathbf{o}$
as $\mathbf{z}$ runs through $\mathbf{F}$'s domain.

**Proof**: The canonical form of $\mathbf{F}$ under *Equivalence* is $\begin{bmatrix} \mathrm{I} & \mathrm{O} \\ \mathrm{O} & \mathrm{O} \end{bmatrix} = \mathbf{R}^{-1}\cdot\mathbf{F}\cdot\mathbf{C}$ in which dimension(I) = rank($\mathbf{F}$) . (Some O's may be empty.) Now switch given equation " $\mathbf{F}\cdot\mathbf{x} = \mathbf{y}$ " to " $\mathbf{R}^{-1}\cdot\mathbf{F}\cdot\mathbf{C}\cdot(\mathbf{C}^{-1}\cdot\mathbf{x}) = (\mathbf{R}^{-1}\cdot\mathbf{y})$ " in canonical form; *etc*.

# Generalized Inverses and Pseudo-Inverses

Suppose matrix F is non-invertible because it is not square and/or Rank(F) is less than both dimensions. At least one ***Generalized Inverse*** G always exists such that if equation "$F \cdot x = y$" has solution(s) x then $x := G \cdot y$ is a solution.

The Generalized Inverses of F are the solutions G of "$F \cdot G \cdot F = F$".

**Example:** The ***Least-Squares*** problem "Choose x to minimize $\|F \cdot x - g\|_2$" always has at least one solution x and, if more than one, then the one that also minimizes $\|x\|_2$ is $x := F^\dagger \cdot g$ in which $F^\dagger$ is F's *Moore-Penrose* ***Pseudo-Inverse*** , a Generalized Inverse of F .

"$F \cdot F^\dagger \cdot F = F$ , $F^\dagger \cdot F \cdot F^\dagger = F^\dagger$ , $(F^\dagger \cdot F)^T = F^\dagger \cdot F$ , $(F \cdot F^\dagger)^T = F \cdot F^\dagger$ " characterize $F^\dagger$ but $F^\dagger$ is best computed from the ***Singular Value Decomposition*** of F .

F and $F^\dagger$ are *Reciprocal*, each a Generalized Inverse of the other; but non-reciprocal Generalized Inverses of F may exist too. ...

# How Big Must Generalized Inverses Be ?

Most are arbitrarily big; when some $Z \neq O$ satisfies either $F \cdot Z = O$ or $Z \cdot F = O$, then any Generalized Inverse $G$ of $F$ yields infinitely many others: $G \pm \lambda \cdot Z$ .

## Generalized Inverses of **some** matrices F are **all *HUGE*** :

**Theorem:** Every Generalized Inverse $G$ of $F$ has *cf.* picture on p. 37

$$\|G\| \geq 1/\big(\min \|\Delta F\| \text{ over } \mathrm{Rank}(F - \Delta F) < \mathrm{Rank}(F)\big) \ .$$

Equality can occur for Pseudo-Inverses gauged by the $l_2$ Operator Norm $\|...\|_2$ :

$$\|F^\dagger\|_2 = 1/\big(\min \|\Delta F\|_2 \text{ over } \mathrm{Rank}(F - \Delta F) < \mathrm{Rank}(F)\big)$$
$$= 1/(\text{the least nonzero singular value of } F) \ .$$

For Operator Norms $\|...\|$ generally, the best that can be said appears to be ...

**Theorem:** $F$ has at least one Generalized Inverse $G$ with

$$\|G\| \leq \sqrt{\big(\mathrm{Rank}(F)\big)}/\big(\min \|\Delta F\| \text{ over } \mathrm{Rank}(F - \Delta F) < \mathrm{Rank}(F)\big) \ .$$

I hope this theorem has a short proof; mine is much too long.

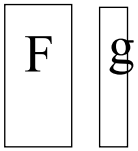
# Use of Generalized Inverses can be Dangerous !

*E.g.*: Pseudo-Inverse $F^\dagger = (F^T \cdot F)^{-1} \cdot F^T$ or $F^T \cdot (F \cdot F^T)^{-1}$, whichever inverse exists, unless neither inverse exists, in which case use ... . SVD ?

Limit Formula: $F^\dagger = \lim_{\alpha \to 0+} (F^T \cdot F + \alpha \cdot I)^{-1} \cdot F^T$ ?

**Over-determined** x :

Choose x with minimum $\|x\|_2$ to minimize $\|F \cdot x - g\|_2$ .




Solution: $x = F^\dagger \cdot g$ unless columns of F are too nearly linearly dependent.

Remedies for HUGE and HYPERSENSITIVE $F^\dagger$ :

QR factorization, Doubled precision, Better basis for x orthogonal polynomials

*cf.* pp. 15-16 of <…/HilbMats.pdf>

**A Bad Idea:** *Tychonoff Regularization*: $x := (F^T \cdot F + \alpha \cdot I)^{-1} \cdot F^T \cdot g$ for a small $\alpha$ .
If good values for *Regularization Parameter* $\alpha$ exist they depend upon NOISE.

**Better Idea:** Choose basis for x so that $\|\Delta x\|_2$ is ***appropriate***, then compute SVD(F) to diagonalize " $F \cdot x \approx g$ " and delete elements below noise levels.

Limit Formula: $F^{\dagger} = \lim_{\alpha \to 0+} F^{T} \cdot (F \cdot F^{T} + \alpha \cdot I)^{-1}$ ?

**Under-determined** x :

Choose x with minimum $\|x\|_2$ to satisfy $F \cdot x \approx g$ .


Solution: $x = F^{\dagger} \cdot g$ unless rows of F are too nearly linearly dependent.

Remedies for HUGE and HYPERSENSITIVE $F^{\dagger}$ :

Doubled precision? Discard redundant rows ? Change basis in Range(F) ?

**A Bad Idea:** *Tychonoff Regularization*: $x := F^{T} \cdot (F \cdot F^{T} + \alpha \cdot I)^{-1} \cdot g$ for a small $\alpha$ .
If good values for *Regularization Parameter* $\alpha$ exist they depend upon NOISE.

**Better Idea:** Choose basis for g so that $\|\Delta F\|_2$ is *Equitable*; then compute
SVD(F) to diagonalize " $F \cdot x \approx g$ " and delete elements below noise levels.
Alternatively, seek x with the fewest "very nonzero" components.

High noise levels can leave no solution x determined unambiguously. Then …

Sometimes NO RESULT is better than a BAD RESULT
when NO GOOD RESULT exists.

# A Generality

Often a computational task amounts to solving "$F(z) = o$" for $z$ given $F$.
Errors and uncertainties due to roundoff etc. make $F(z + \delta z) = \delta F$ instead.

Consequent uncertainty or error in $z$ is $\delta z \approx F`(z)^{-1} \cdot \delta F$.

Uncertainty in data that specifies $F$ is amplified in $z$ by as much as $\|F`(z)^{-1}\|$.

HUGE amplification $\|F`(z)^{-1}\|$ can occur *only* if DATA puts $F$ CLOSE to a SINGULARITY :

- A Pole (infinite value) of $F`^{-1}$.
- Confluence (collapse of dimension; multiplicity of mapping $F$, of zeros $z$).
- Exponentially-growing functions of variables unobviously near $\infty$.

Common! See also my web page's <…/WrongR.pdf>

Changes of variables/parameters/bases, perhaps nonlinear, can alter closeness of $F$ to a singularity, or change the norm that measures closeness, sometimes so drastically as to change amplification of error by many orders of magnitude.

Good choices are worth the thought devoted to them.

# Part IV:

# Matrix Norms and Eigenvalues

# Matrix Norms exceed Eigenvalues

*Eigenvalue* $\lambda$ of $\mathbf{L}$ has $\mathbf{L}\cdot\mathbf{v} = \lambda\cdot\mathbf{v}$ for some *Eigenvector* $\mathbf{v} \neq \mathbf{o}$; therefore

$$|\lambda| \leq \|\mathbf{L}\| \text{ for every compatible } \|...\|,$$
including every Operator Norm.

*Eigenvalue* $\lambda$ of $\mathbf{L}$ also has $\mathbf{w}'\cdot\mathbf{L} = \lambda\cdot\mathbf{w}'$ for some *Left Eigenvector* $\mathbf{w}' \neq \mathbf{o}'$ in the space of linear functionals $\mathbf{w}'$ *Dual* to Domain($\mathbf{L}$) = Target-Space($\mathbf{L}$). The spaces, like $\lambda$, may be complex regardless of whether $\mathbf{L}$ is real.

Eigenvalues $\lambda$ are the zeros of $\mathbf{L}$'s *Characteristic Polynomial* $\det(\lambda\cdot\mathrm{I} - \mathbf{L})$.
Computing the Characteristic Polynomial explicitly is usually a numerically bad way to determine eigenvalues.

Eigenvalues $\lambda$ are Continuous Functions of $\mathbf{L}$;
but Perhaps Not Differentiable,
unless $\lambda$ is a *Simple* eigenvalue where …
$$\det(\lambda\cdot\mathrm{I} - \mathbf{L}) = 0 \neq d\det(\lambda\cdot\mathrm{I} - \mathbf{L})/d\lambda .$$

# A Computed Eigenvector's Error is an *Angle*:

Let $\mathbf{x}$ be a computed approximation to a desired eigenvector $\mathbf{v}$ ;
then the error is not $\|\mathbf{x}-\mathbf{v}\|$ but the difference between two *Subspaces* ,
one spanned by scalar multiples of $\mathbf{x}$ , the other ... of $\mathbf{v}$ .

The relevant error is the (unsigned) *Angle* $\angle(\mathbf{x}, \mathbf{v})$ between the two subspaces,
usually in complex spaces where $x'$ is the complex conjugate transpose of $x$ .

How to compute $\angle(\mathbf{x}, \mathbf{v}) := \arccos\big( |x'\cdot v|/(\|x\|_2\cdot\|v\|_2) \big)$ :
*NOT FROM THIS FORMULA !* Why not?

First choose a basis for which $\|\Delta\mathbf{x}\|_2$ is an *Equitable* measure of perturbations.

Then replace $x$ by $x/\|x\|_2$ and $v$ by $v/\|v\|_2$ , so now $\sqrt{x'\cdot x} = \|x\|_2 = \|v\|_2 = 1$ .

Compute $\angle(\mathbf{x}, \mathbf{v}) := 2\cdot\arcsin\big( \|x\cdot(x'\cdot v/|x'\cdot v|) - v\|_2/2 \big)$ . Presubstitute 1 for $(0/|0|)$ .

A generalization works for ***Angles*** between higher-dimensional ***Subspaces***; see p. 74.
|Accurate| for *all* angles. For ***Signed*** Angle between two vectors see p. 15 of my <.../Cross.pdf>

# (Unobviously) Clustered Eigenvalues of a Matrix

not Real Symmetric nor Hermitian nor Orthogonal nor Unitary nor "Normal"
can be hypersensitive to perturbations

**Example:** $\mathbf{L} := \begin{bmatrix} 0 & \mathbf{10} & 0 & 0 & 0 & 0 \\ 0 & 0 & \mathbf{10} & 0 & 0 & 0 \\ 0 & 0 & 0 & \mathbf{10} & 0 & 0 \\ 0 & 0 & 0 & 0 & \mathbf{10} & 0 \\ 0 & 0 & 0 & 0 & 0 & \mathbf{10} \\ \xi & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$ ; $\det(\lambda \cdot \mathrm{I} - \mathbf{L}) = \lambda^6 - 10^5 \cdot \xi$ .

6 Eigenvalues $\lambda$ jump from 0 at $\xi = 0$ to $|\lambda| = 1$ at $\xi = 1/10^5$. Clustered?

Eigenvectors of a Matrix can be hypersensitive to perturbations, especially if it is close to a matrix with Repeated Eigenvalues.

**Example:** 6 eigenvectors of **L** at tiny $\xi \neq 0$ collapse to one at $\xi = 0$.

Whenever a matrix's ***Jordan Normal Form*** is not merely diagonal, it is a Discontinuous Function of the matrix's elements, and thus very hard to compute.

# Gershgorin's Circles enclose Eigenvalues

Given a square matrix $L$, its eigenvalues $\lambda$ can be located within the union of a family of circular disks in the complex plane: ... Row version. A Column version exists too.

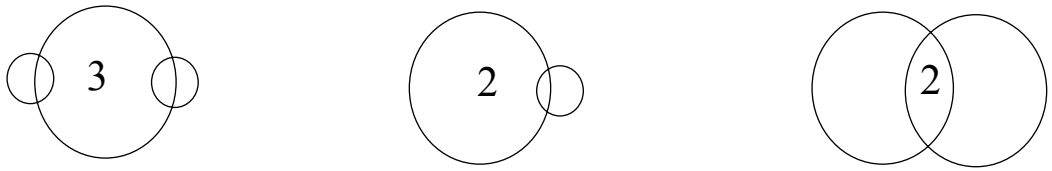
Disk $D_i$ is centered at $L_{ii}$ with radius $\Sigma_{j\neq i} |L_{ij}| = \Sigma \left| \text{ } i^{th} \text{ row's off-diagonals } \right|$.

## Gershgorin's Circles Theorem :

Every eigenvalue $\lambda$ of $L$ lies in the union of all the disks $D_i$.

Each disjoint component of that union contains
as many eigenvalues $\lambda$ as it contains disks.

Numbers of Eigenvalues:




**Proof**: If no disk $D_i$ contains $\eta$ it cannot be an eigenvalue since $\eta I - L$ is *Diagonally Dominant* (p. 42). To count eigenvalues use continuity as disks expand from their centers.

This theorem is useful mainly for matrices with *Prominent* if not *Dominant* diagonals.

Sometimes replacing $L$ by $V^{-1} \cdot L \cdot V$ for a suitable matrix $V$, perhaps diagonal, perhaps consisting of approximate eigenvectors, can reduce the sizes of at least some of the disks.

Diagonal Prominence (rather than Dominance) also admits cheaper estimates of

## Extreme Singular Values:

(Biggest Singular Value of L) = $\max_{\mathbf{x}\neq\mathbf{0}} ||L\cdot\mathbf{x}||_2/||\mathbf{x}||_2 = ||L||_2 = ||L'||_2$ .

(Least Singular Value of square L) = $\min_{\mathbf{x}\neq\mathbf{0}} ||L\cdot\mathbf{x}||_2/||\mathbf{x}||_2 = 1/||L^{-1}||_2$ .

These *Eigenvalues* cost work $O(\min\{m, n\}^3)$ to compute closely for m-by-n matrices L.

### Cheaper estimates of $||L||_2$ :

(Biggest Euclidean row-length) = $||L||_{\infty 2} \le ||L||_2 \le \sqrt{m}\cdot||L||_{\infty 2}$ , and

$$\sqrt{(\,||L||_1\cdot||L||_\infty\,)}/\sqrt[4]{m\cdot n} \le ||L||_2 \le \sqrt{(\,||L||_1\cdot||L||_\infty\,)}\ .$$

The blue "≤" inequalities come closer as the diagonal of L becomes more prominent.
(The 2nd-last "≤" is tight "=" when L is a *Hadamard Matrix*; cf. MATLAB's `hadamard.m` .)

### Cheaper underestimates of the least ($n^{th}$) singular value of an n-by-n matrix L :

If possible, first permute its rows and columns to make its diagonal prominent. Then scale its rows or columns by factors of magnitude 1 to make the diagonal positive. Now

($n^{th}$ Singular Value of L) ≥ $\min_k \{\, L_{kk} - \sum_{j\neq k} |L_{jk} + \overline{L}_{kj}|/2 \,\}$ ; hope it's positive.

*Cf.* Chas. R. Johnson's "A Gersgorin-type Lower Bound for the Smallest Singular Value" pp.1-7 *Lin. Alg. & Appl.* **112** (1989)

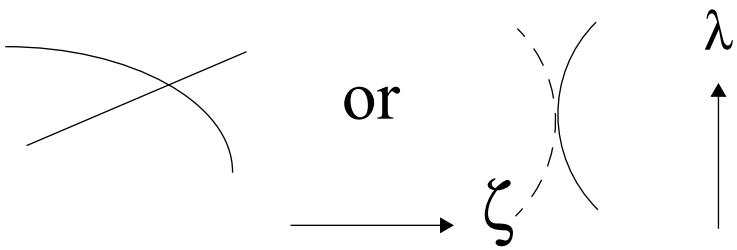
Appending rows and/or columns to n-by-n L cannot decrease its $n^{th}$ singular value.

# Sensitivity of Eigenvalues to Tiny Perturbations

Say $\lambda$ is an eigenvalue of $\mathbf{L}$ with eigenvectors $\mathbf{x}$ and $\mathbf{w}'$, perhaps all complex: $\mathbf{L}\cdot\mathbf{x} = \lambda\cdot\mathbf{x}$ and $\mathbf{w}'\cdot\mathbf{L} = \mathbf{w}'\cdot\lambda$. Then

$$d\lambda/d\zeta = \mathbf{w}'\cdot(d\mathbf{L}/d\zeta)\cdot\mathbf{x}/\mathbf{w}'\cdot\mathbf{x}$$

*UNLESS* $\lambda$ is a multiple eigenvalue, in which case $d\lambda/d\zeta$ may be multi-valued or infinite:




When $\lambda$ is a *simple* eigenvalue, $\underline{\kappa} := \|\mathbf{w}'\|\cdot\|\mathbf{x}\|/|\mathbf{w}'\cdot\mathbf{x}|$ is its

***Condition Number.***

$\underline{\kappa}$ can be arbitrarily big unless $\mathbf{L}$ *stays* special, say "Normal"
(*i.e.* $\mathbf{L}'\cdot\mathbf{L} = \mathbf{L}\cdot\mathbf{L}'$ … Real Symmetric, Hermitian, Unitary, Orthogonal … $\underline{\kappa} = 1$)

"Stays"? If $\mathbf{L}$ is Hermitian but not $d\mathbf{L}$, $\underline{\kappa}$ can be $\approx 1+2\cdot\log(\text{dimension})/\pi$.

# Sensitivity of Eigenvalues to Perturbations, cont'd

Suppose matrix $L$ is diagonalizable by similarity (as almost all square matrices are) and let $\Lambda = X^{-1}\cdot L\cdot X$ be the diagonal matrix of eigenvalues of $L$; ***a*** matrix of eigencolumns is $X$.

How much can $\Lambda$ ***NOT*** change if $L$ changes to $L + \Delta L$ ?

Now (probably non-diagonal) $\Lambda + \Delta\Lambda := X^{-1}\cdot(L + \Delta L)\cdot X$, so

$$\|\Delta\Lambda\|_1 = \|X^{-1}\cdot\Delta L\cdot X\|_1 \le \|X^{-1}\|_1\cdot\|\Delta L\|_1\cdot\|X\|_1 = \kappa_1(X)\cdot\|\Delta L\|_1 .$$

Here $\kappa_1(X)$ is a (perhaps excessive) condition number for inversion of $X$. Now Gershgorin's Circles Theorem applied to $\Lambda + \Delta\Lambda$ implies …

**Bauer-Fike Theorem:**

No eigenvalue of $L + \Delta L$ can differ from an eigenvalue of $L$ by more than $\min\{\,\kappa_1(X)\cdot\|\Delta L\|_1\,,\ \kappa_\infty(X)\cdot\|\Delta L\|_\infty\,\}$ .

# The Perron-Frobenius Theory of Nonnegative Matrices

Let $P$ be any square matrix with elements all positive; we shall write "$P > O$".

Then $P$ has a positive eigenvector, its "Perron Vector" $p > o$, and a positive eigenvalue $\rho > 0$ with $P \cdot p = \rho \cdot p$; and $P$'s "Perron Root" $\rho$ is a simple eigenvalue strictly bigger than the magnitude of every other eigenvalue of $P$.

**Proof:** The simplex $\mathbb{S} := \{s: s \geq o \ \& \ ||s||_1 = 1\}$ is closed, convex and bounded, and mapped into itself continuously by function $\pi(s) := P \cdot s / ||P \cdot s||_1$. ***Brauer's Fixed-Point Theorem*** provides a fixed-point $p = \pi(p)$ strictly inside $\mathbb{S}$, and $\rho = ||P \cdot p||_1$. Let $V := \mathrm{Diag}(p)$ and $B := V^{-1} \cdot P \cdot V$; then $B$ and $P$ have the same eigenvalues, and $\rho = ||B||_\infty$ is the Perron Root of $B$ with Perron eigenvector $b = [1; 1; 1; \ldots; 1]$. Jacobi's formula: $d\,\mathrm{Det}(\lambda I - B)/d\lambda = \mathrm{Trace}(\mathrm{Adj}(\lambda I - B)) > 0$ @ $\lambda = \rho$ because principal submatrices of $\rho I - B$ are *diagonally dominant*, so $\rho$ is a simple eigenvalue. Except $\rho$, every other eigenvalue $ß$ of $B$ has $|ß| < ||B||_\infty = \rho$, strictly $<$ when the eigenvector is considered.

Let $C$ be any square matrix with $|C| \leq P$ elementwise. Then no eigenvalue of $C$ can exceed in magnitude the Perron Root $\rho$ of $P$.

If, instead of $P > O$, we have merely $P \geq O$, then Perron Root $\rho$ need not be simple; it and/or elements of eigenvector $p$ may vanish, and other eigenvalues of $P$ may match $\rho$ in magnitude.

Application of Perron-Frobenius theory to

# Optimal Diagonal Equilibration

(See the bottom of p. 39.)

Suppose we seek diagonal matrices $\Lambda$ and $V$ to minimize the condition number

$$\kappa_\infty(\Lambda^{-1}\cdot C\cdot V^{-1}) := \|(\Lambda^{-1}\cdot C\cdot V^{-1})^{-1}\|_\infty\cdot\|(\Lambda^{-1}\cdot C\cdot V^{-1})\|_\infty \ .$$

**Theorem**:

$$\min\big(\ \kappa_\infty(\Lambda^{-1}\cdot C\cdot V^{-1}) \text{ over all diagonal } \Lambda \text{ and } V\ \big) \ = \ \text{Perron Root of } |C^{-1}|\cdot|C|$$

and the minimum is achieved (at some computational cost) by setting

$$\Lambda := \text{Diag( Perron Vector of } |C|\cdot|C^{-1}| \text{ )} \quad \text{and}$$

$$V := \text{Diag( Perron Vector of } |C^{-1}|\cdot|C| \text{ ) .}$$

… due to F.L. Bauer [1963] “Optimally Scaled Matrices”, pp. 73-87 of ***Numerische Mathematik* 5**.

From the late 1950s Fritz Bauer and his students in Mainz and Munich were major contributors to the applications of norms to numerical analysis. Bauer also contributed to the design of *Algol 60*, an early programming language more humane than most. Later he became “Mr. Computing” to the Bavarian government. Recently he authored a fascinating text on *Cryptology* (Springer). As of this writing he is still active at age 90, having barely survived 1942 - 1945 in the *Wehrmacht*.

# Part V:

# Matrix Norms and Real Symmetric Matrices' Eigenvalues

The best source about eigensystems of real symmetric and Hermitian matrices is B.N. Parlett's book ***The Symmetric Eigenvalue Problem*** [1998] 426 pp., SIAM, Philadelphia

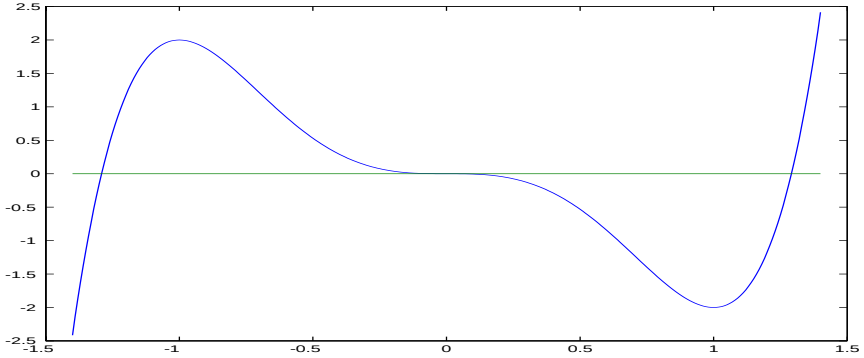
Real Symmetric Matrices' Eigenvalues are all …

# Stationary Values of Quotients of Real Quadratic Forms

- The ***Stationary Values*** (also called ***Critical Values***) of a function $f(\mathbf{x})$ are the values it takes where its derivative $f`(\mathbf{x})$ vanishes.
- The ***Stationary Points*** (also called ***Critical Points***) of a function $f(\mathbf{x})$ are the arguments $\mathbf{x}$ at which its derivative $f`(\mathbf{x})$ vanishes.

Instances are maxima and minima, but these are far from the only instances.

*e.g.*, $f(\xi) := 3\xi^5 - 5\xi^3$ takes all real values on the real $\xi$-axis,
takes a locally maximum value $f(-1) = +2$ ,
takes a locally minimum value $f(+1) = -2$ ,
and takes another stationary value $f(0) = 0$ neither maximum nor minimum.



Real Symmetric Matrices' Eigenvalues are all …

# Stationary Values of Quotients of Real Quadratic Forms

- A ***Real Quadratic Form*** $\Phi(\mathbf{x}) := x'\cdot H\cdot x$ , where column $x$ represents $\mathbf{x}$ in some basis, and matrix $H = H'$ represents a symmetric bilinear operator: $\mathbf{H}\cdot\mathbf{x}\cdot\mathbf{y} = \mathbf{H}\cdot\mathbf{y}\cdot\mathbf{x} = y'\cdot H\cdot x$ . For a real vector space, $y'$ and $H'$ are transposes, and $H$ is the matrix of a linear map $\mathbf{H}$ from $\mathbf{x}$'s vector space to its dual.
  Abstractly, " $\Phi(\mathbf{x}+\mathbf{y}) + \Phi(\mathbf{x}-\mathbf{y}) \equiv 2\Phi(\mathbf{x}) + 2\Phi(\mathbf{y})$ " characterizes quadratic forms $\Phi$ . *cf.* p. 20.
  Example: 2nd derivative in ***Taylor Series*** for scalar $\mu(\mathbf{z}+\mathbf{x}) = \mu(\mathbf{z}) + \mu`(\mathbf{z})\cdot\mathbf{x} + \mu''(z)\cdot\mathbf{x}\cdot\mathbf{x}/2 + \ldots$

  (Complex spaces, for which $y'$ and $H'$ are complex conjugate transposes, and $H = H'$ is ***Hermitian*** instead of real symmetric, and $\mathbf{H}\cdot\mathbf{x}\cdot\mathbf{y}$ is complex conjugate to $\mathbf{H}\cdot\mathbf{y}\cdot\mathbf{x}$ , will not be treated here. Besides, the treatment of complex spaces would differ only slightly from real.)

Typically $\Phi(\mathbf{x})$ is some kind of *Energy*,— Kinetic Energy if $\mathbf{x}$ stands for velocities or momenta, Elastic Energy if for infinitesimal displacements from equilibrium, *etc*.

How does a change of basis from $\mathbf{B}$ to $\overline{\mathbf{B}} := \mathbf{B}\cdot C^{-1}$ affect $H$ ? *cf.* pp. 18-19

$x$ changes to $\overline{x} := C\cdot x$ but $w'$ to $\overline{w}' := w'\cdot C^{-1}$ , and $H$ to $\overline{H} := C'^{-1}\cdot H\cdot C^{-1}$ , whence $\Phi(\mathbf{x}) = x'\cdot H\cdot x = \overline{x}'\cdot\overline{H}\cdot\overline{x}$ . ( $\Phi(\mathbf{x})$ need not be defined upon the vector space dual to $\mathbf{x}$'s .)

The relation between $H$ and $\overline{H} := C'^{-1}\cdot H\cdot C^{-1}$ is called a ***Congruence***.

# Diagonalization of a Real Quadratic Form by Congruence

Let real quadratic form $\Phi(\mathbf{x}) := x'\cdot H\cdot x$ in some chosen basis, no matter which.

Lagrange showed that Infinitely Many Congruences $C'^{-1}\cdot H\cdot C^{-1} = V$ have a diagonal $V$.

Let $n_+$ count positive diagonal elements of $V$, $n_\mathbf{0}$ … zero …, $n_-$ … negative … .

Sylvester's ***Inertia*** Theorem: Every diagonal $V$ congruent to $H$ has the same ***Inertia***$(H) := \{n_+, n_\mathbf{0}, n_-\}$ .

Some authors use the word "Signature" instead of "Inertia", a word chosen by Sylvester.

Geometrically, $n_+ = \max\{$ dimension of every *Subspace* on which $\Phi(\mathbf{x}) > 0$ for $\mathbf{x} \neq \mathbf{o}\}$ ;

$n_- = \max\{$ … $\Phi(\mathbf{x}) < 0$ for $\mathbf{x} \neq \mathbf{o}\}$ ; $n_\mathbf{0} = \text{dimension}(\mathbf{x}) - n_+ - n_-$ .

Inertia distinguishes shapes of Ellipsoids *vs*. Hyperboloids of one or two sheets.

## Nomenclature for $\Phi$ and Symmetric $H$

| Names | | $n_+$ positives | $n_\mathbf{0}$ zeros | $n_-$ negatives |
|---|---|---|---|---|
| Nonnegative Definite | Positive Definite | ≥ 1 | 0 | 0 |
| | Positive Semidefinite | ≥ 1 | ≥ 1 | 0 |
| | Indefinite | ≥ 1 | | ≥ 1 |
| | Degenerate or Singular | | ≥ 1 | |
| Nonpositive Definite | Negative Semidefinite | 0 | ≥ 1 | ≥ 1 |
| | Negative Definite | 0 | 0 | ≥ 1 |

# Stationary Values of Quotients of Real Quadratic Forms

Given two real quadratic forms $\Phi(\mathbf{x}) := x'\cdot H\cdot x$ and $\Psi(\mathbf{x}) := x'\cdot M\cdot x$, we seek all

Stationary Values $\rho$ of the ***Rayleigh Quotient*** $\rho(\mathbf{x}) := \Phi(\mathbf{x})/\Psi(\mathbf{x})$

as $\mathbf{x}$ runs through all nonzero vectors in a real space.

Each stationary value of $\rho(\mathbf{x})$ and its stationary point $\mathbf{x}$ turn out to satisfy

$$H\cdot x = \rho(\mathbf{x})\cdot M\cdot x \quad \text{and} \quad M\cdot x \neq o\ .$$

Hence this $\rho$ and $x$ are solutions of a ***Symmetric Generalized Eigenproblem***.

If $M^{-1}$ exists, each such $\rho$ is an eigenvalue of $B := H\cdot M^{-1}$. However, not all eigenvalues ß of B need be stationary values of $\rho(x)$, not even if all are real.

*e.g.*, $H := \begin{bmatrix} 2 & 0 \\ 0 & 0 \end{bmatrix}$, $M := \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$, $x := \begin{bmatrix} \xi \\ \eta \end{bmatrix}$, $\rho(\mathbf{x}) = \xi^2/(\xi\eta)$ ; $B = H\cdot M^{-1} = \begin{bmatrix} 0 & 2 \\ 0 & 0 \end{bmatrix}$, ß = 0 . No stationary $\rho$ .

*e.g.*, $H := \begin{bmatrix} 2 & 0 \\ 0 & -2 \end{bmatrix}$, $M := \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$, $x := \begin{bmatrix} \xi \\ \eta \end{bmatrix}$, $\rho(\mathbf{x}) = \xi/\eta - \eta/\xi$ ; $B = H\cdot M^{-1} = \begin{bmatrix} 0 & 2 \\ -2 & 0 \end{bmatrix}$, ß = $\pm 2\mathbf{\imath}$ . No stationary $\rho$ .

*Every* real square matrix $B = H\cdot M^{-1}$ for *some* symmetric $H = H'$ and $M = M'$.

Therefore, without further restrictions, Symmetric Generalized Eigenproblems and unrestricted nonsymmetric eigenproblems fall prey to the same pathologies; *cf.* pp. 51-56.

Consequently we shall impose further restrictions upon $\Phi$ and $\Psi$ in what follows. …

# Simultaneous Diagonalization by Congruence

A powerful motive to find all the stationary points $\mathbf{x}$ of the Rayleigh Quotient $\rho(\mathbf{x}) := x'\cdot H\cdot x/x'\cdot M\cdot x$ is that, ***IF*** they are linearly independent and numerous enough to constitute the columns of an invertible matrix $C^{-1}$, they provide a new coordinate system (basis) that transforms $H$ and $M$ into diagonal matrices $\overline{H} := C'^{-1}\cdot H\cdot C^{-1}$ and $\overline{M} := C'^{-1}\cdot M\cdot C^{-1}$ simultaneously by the same congruence.

The eigenvalues (stationary values) $\rho$ are the diagonal elementwise quotients $\overline{H}./\overline{M}$.
They are often identified with squares of resonant frequencies of vibration.

The columns of $C^{-1}$ are often identified with "Natural Modes" of vibration.

## What conditions suffice for simultaneous diagonalization by congruence?

- John Milnor's Criterion: If $x = o$ whenever $x'\cdot H\cdot x = x'\cdot M\cdot x = 0$, and if the dimension of vectors $x$ is $n \neq 2$, then some linear combination of real symmetric $H$ and $M$ is positive definite. If $n = 2$ ? *cf.* 2d. *e.g.* on p. 64.
  (The case $n = 2$ is unexceptional when $x$ is complex and $H$ and $M$ Hermitian.)

- If some linear combination of real symmetric $H$ and $M$ is positive definite, $H$ and $M$ can be diagonalized simultaneously by a congruence $C^{-1}$ …

Suppose some linear combination, say $\eta\cdot H + \mu\cdot M$, of real symmetric H and M is positive definite, so some congruence diagonalizes H and M simultaneously.

How then may such a congruence $C^{-1}$ be found?

Choose a second linear combination, say $\alpha\cdot H - \beta\cdot M$, independent of the first so $\alpha\cdot\mu + \beta\cdot\eta \neq 0$. A good choice would have $\|\alpha\cdot H - \beta\cdot M\| << \| \,|\alpha\cdot H| + |\beta\cdot M|\, \|$ by virtue of substantial cancellation, but this may not be feasible. Thus a given

Symmetric Generalized Eigenproblem " $H\cdot x = \rho\cdot M\cdot x$ " is converted into a ***Definite*** Symmetric Generalized Eigenproblem " $\overline{H}\cdot y = \overline{\rho}\cdot\overline{M}\cdot y$ " in which $\overline{H} := \alpha\cdot H - \beta\cdot M = \overline{H}'$ and $\overline{M} := \eta\cdot H + \mu\cdot M = \overline{M}'$ is positive definite.

Their eigenvalues are related: $\overline{\rho} = (\alpha\cdot\rho - \beta)/(\eta\cdot\rho + \mu)$; $\rho = (\beta + \mu\cdot\overline{\rho})/(\alpha - \eta\cdot\overline{\rho})$.

A way to compute them and a desired congruence: Cholesky factorize $\overline{M} = U'\cdot U$ and compute an eigendecomposition of $W := U'^{-1}\cdot\overline{H}\cdot U^{-1} = W' = Q\cdot\Omega\cdot Q'$ with an orthogonal $Q' = Q^{-1}$ and diagonal $\Omega$ of eigenvalues $\overline{\rho}$. Now $C^{-1} := U^{-1}\cdot Q$.

Thus does MATLAB's `eig`. Since such eigenproblems can be pathological, their error-analysis isn't yet tidy enough for a succinct and memorable overview. See instead Ren-Cang Li's §15.4 in L. Hogben's *Handbook* … cited under *Further Reading*, and my …/Math128/GnSymEig.pdf> . No comparable numerical scheme is known to find an *Indefinite* Symmetric Generalized Eigenproblem's congruence when it exists. What follows concerns problems that are easier and better understood. …

# The (ordinary) **Real Symmetric Eigenproblem**

… is the special case $M = I$ of the Definite Symmetric Generalized Eigenproblem:

Given a real symmetric $H = H'$, we seek its eigenvalues $\theta$ and eigenvectors $q$ satisfying $H\cdot q = \theta\cdot q$. All of them constitute a diagonal $\Theta$ and an orthogonal $Q = Q'^{-1}$ satisfying $Q'\cdot H\cdot Q = \Theta$. The eigenvalues can be ordered in two ways:

• Ascending: $\theta_1 \le \theta_2 \le \theta_3 \le \ldots \le \theta_n$. • Descending: $\theta_1 \ge \theta_2 \ge \theta_3 \ge \ldots \ge \theta_n$.

These two orderings are relevant to the identification of eigenvalues as stationary values of the ***Rayleigh Quotient*** $\rho(\mathbf{x}) := x'\cdot H\cdot x/x'\cdot x$ via the …

## Courant-Fischer Minimax Principle:

Ascending order, $\theta_k = \mathrm{Min}_{\text{subspaces } \mathbb{S} \text{ of dimension } k}\ \mathrm{Max}_{\text{nonzero } \mathbf{x} \text{ in } \mathbb{S}}\ \rho(\mathbf{x})$.

Descending order, $\theta_k = \mathrm{Max}_{\text{subspaces } \mathbb{S} \text{ of dimension } k}\ \mathrm{Min}_{\text{nonzero } \mathbf{x} \text{ in } \mathbb{S}}\ \rho(\mathbf{x})$.

Let a perturbation $\Delta H = \Delta H'$ that changes eigenvalues of $H + \Delta H$ to $\Theta + \Delta\Theta$ have Inertia$(\Delta H) = \{\ \pi := n_+\ ,\ \zeta := n_{\mathbf{0}}\ ,\ \ \nu := n_-\ \}$. (See p. 63.) Then ordered …

Ascending, $\theta_k + \Delta\theta_k \le \theta_{k+\pi}$ for $1 \le k \le n-\pi$; $\theta_m + \Delta\theta_m \ge \theta_{m-\nu}$ for $\nu < m \le n$.

Descending, $\theta_i + \Delta\theta_i \ge \theta_{i+\nu}$ for $1 \le i \le n-\nu$; $\theta_j + \Delta\theta_j \le \theta_{j-\pi}$ for $\pi < j \le n$.

… useful mainly when $\pi$ and/or $\nu$ is small like 0, 1 or 2.

# Absolute and Relative Perturbations

Let perturbation $\Delta H = \Delta H'$ change eigenvalues of $H + \Delta H$ to $\Theta + \Delta\Theta$ with the same ordering as $\Theta$. Then $\|\Delta\Theta\| \le \|\Delta H\|$ for every Orthogonally Invariant norm $\|\ldots\|$. (See p. 32.) The norms usually chosen are $\|\ldots\|_2$ and $\|\ldots\|_F$.

The foregoing inequality $\|\Delta\Theta\| \le \|\Delta H\|$ is satisfactory when $H$ is the matrix of a symmetric linear operator from a Euclidean space to itself. More generally if $H$ is the matrix of a symmetric linear operator from a normed vector space to its dual, no orthogonally invariant norm need be *Equitable*; different perturbations with the same norm $\|\Delta H\|$ may differ utterly in significance and effect.

Graded matrices; *cf.* p. 30. Better coordinates; *cf.* p. 12. Other norms; *cf.* Li & Mathias [1999]

## Ostrowski's Refinement of Sylvester's Inertia Theorem: (See p. 63)

If the eigenvalues $\Theta$ of $H$ change to eigenvalues $\overline{\Theta}$ of $\overline{H} := C'^{-1}\cdot H\cdot C^{-1}$ with the same ordering, $\mathrm{Inertia}(\overline{\Theta}) = \mathrm{Inertia}(\Theta)$ and every $\overline{\theta}_j \ne 0$ in $\overline{\Theta}$ has

$$1/\|C'\cdot C\|_2 \le \overline{\theta}_j/\theta_j \le \|(C'\cdot C)^{-1}\|_2 .$$

Typically this is applied with $C$ close to $I$, so $\overline{H}$ is *Relatively* close to $H$.

# Partial Eigensystems and Spectral Gaps

The ***Spectrum*** of n-by-n real symmetric $H = H'$ is the *Multiset* of its n eigenvalues $\theta_j$ (some of which may be repeated) which we assume to be ordered *Descending*, say. Let $\mathbb{E}(H)$ denote this so ordered spectrum $\{ \theta_1 \geq \theta_2 \geq \theta_3 \geq \ldots \geq \theta_n \}$ of $H$.

Especially when dimension n is big, occasions arise to compute only those eigenvalues of H in some interval separated from the rest of $\mathbb{E}(H)$ by a *Gap* or two. To this end, suppose n-by-m matrix F has $m < n$ *linearly independent* columns approximating (perhaps poorly) m eigenvectors of H, and suppose m-by-m real symmetric matrix M (not necessarily diagonal) has a spectrum $\mathbb{E}(M) = \{\mu_1 \geq \mu_2 \geq \mu_3 \geq \ldots \geq \mu_m\}$ thought to approximate part of $\mathbb{E}(H)$. Let ***Residual*** $R := H \cdot F - F \cdot M$ and let $\beta := \|F^\dagger\|_2 \cdot \|R\|_2$.

Computable $\boxed{R} := \boxed{H}\,\boxed{F} - \boxed{F}\,\boxed{M}$ For $\|F^\dagger\|_2$ see p. 44.

**Theorem**: Among n eigenvalues $\theta_j$ of H there are m each of which lies in a different (though perhaps overlapping) interval $\mu_i - \beta \leq \theta \leq \mu_i + \beta$ for $i = 1, 2, 3, \ldots, m$.

Eigenvectors of H are orthogonal. If their estimates in F are too far from orthonormal, $\|F^\dagger\|_2$ may be excessively big. A remedy for this is *Reorthogonalization*: …

## Reorthogonalization:

One way replaces F by Q from the QR factorization $F = Q \cdot U$ with upper-triangular U and $Q' \cdot Q = I$. Other ways, one sometimes faster, one closer to F, are explored in <www.eecs.berkeley.edu/~wkahan/Math128/NearestQ.pdf> .

After a (closely) orthonormal n-by-m matrix Q replaces F, the new residual becomes $R := H \cdot Q - Q \cdot M$, and $\beta := \|R\|_2$. Then, as before, the m eigenvalues $\mu_i$ of $\mathbb{E}(M)$ estimate m of the n eigenvalues $\theta_j$ in $\mathbb{E}(H)$ thus:

> Among n eigenvalues $\theta_j$ in $\mathbb{E}(H)$ there are m each of which lies in a different (possibly overlapping) interval wherein $|\mu_i - \theta| \le \beta$ for $i = 1, 2, 3, \ldots, m$.

Now R can have its norm $\beta := \|R\|_2$ minimized by the choice $M := Q' \cdot H \cdot Q$. Then those m eigenvalues $\theta_j$ fall into much narrower intervals when $\beta$ is much tinier than *Spectral Gaps* between the rest of $\mathbb{E}(H)$ and those m or their estimates $\mu_i$. To describe these Spectral Gaps we perform a (notional, not necessarily computed) change of coordinates:

Let n-by-n orthogonal $[Q, \overline{Q}]$ be obtained from *any* n–m orthonormal columns $\overline{Q}$ orthogonal to Q, so n-by-n $[Q, \overline{Q}]' \cdot [Q, \overline{Q}] = I$. This $[Q, \overline{Q}]$ provides a new orthonormal coordinate system in which the linear operator formerly represented by H is now represented by $[Q, \overline{Q}]' \cdot H \cdot [Q, \overline{Q}]$. Of course, spectrum $\mathbb{E}([Q, \overline{Q}]' \cdot H \cdot [Q, \overline{Q}]) = \mathbb{E}(H)$.

To define ***Spectral Gaps***, let …

$[Q, \overline{Q}]'\cdot H\cdot[Q,\overline{Q}] =: \begin{bmatrix} M & B \\ B' & W \end{bmatrix}$ and set $Y := \begin{bmatrix} M & O \\ O' & W \end{bmatrix}$. Now $\begin{bmatrix} O \\ B' \end{bmatrix}$ represents $R := H\cdot Q - Q\cdot M$ in the new coordinates; $\beta := \|R\|_2 = \|B\|_2$. Usually only $M = Q'\cdot H\cdot Q$ and $R$ are computed.

Let names for the relevant ordered spectra be …

$\mathbb{E}(H) = \mathbb{E}([Q, \overline{Q}]'\cdot H\cdot[Q, \overline{Q}]) = \{ \theta_1 \ge \theta_2 \ge \theta_3 \ge \ldots \ge \theta_n \}$ which we wish to estimate.

$\mathbb{E}(M) = \{ \mu_1 \ge \mu_2 \ge \mu_3 \ge \ldots \ge \mu_m \}$ which are our $m$ computed estimates.

$\mathbb{E}(W) = \{ \omega_1 \ge \omega_2 \ge \omega_3 \ge \ldots \ge \omega_{n-m} \}$ for which rough estimates will be needed.

$\mathbb{E}(Y) = \{ \eta_1 \ge \eta_2 \ge \eta_3 \ge \ldots \ge \eta_n \} = \mathbb{E}(M) \cup \mathbb{E}(W)$ as multi-sets. $|\theta_i - \eta_i| \le \beta$.

For $i = 1, 2, 3, \ldots, n$ define the ***Gaps*** $\gamma_i$ between spectra $\mathbb{E}(M)$ and $\mathbb{E}(W)$ thus:

If $\eta_i \in \mathbb{E}(M)$ then $\gamma_i := \min_j |\eta_i - \omega_j|$ else if $\eta_i \in \mathbb{E}(W)$ then $\gamma_i := \min_j |\eta_i - \mu_j|$.

Let Gap $\gamma := \min_i \gamma_i$. Usually $\mathbb{E}(M)$ and $\mathbb{E}(W)$ are disjoint, and then $\gamma > 0$.

Then Chi-Kwong Li & Ren-Cang Li [2005] proved VERY GENERALLY that *every*

$$|\theta_i - \eta_i| \le \beta^2/\left(\gamma_i/2 + \sqrt{(\beta^2 + \gamma_i^2/4)}\right) \le \beta^2/\left(\gamma/2 + \sqrt{(\beta^2 + \gamma^2/4)}\right) \le \min\{\beta, \beta^2/\gamma\}.$$

When $\beta << \gamma$ these inequalities estimate that part of $\mathbb{E}(H)$ approximated by $\mathbb{E}(M)$ far more tightly than $\beta$ because ***somehow*** the rest of $\mathbb{E}(H)$ is known to be farther away.

***SOMEHOW ?***

Error-bounds on three previous pages bound *Absolute* errors $\theta_i - \eta_i$ . What about

## ***Relative Errors*** $\log(\theta_i/\eta_i)$ ?

Again, n-by-m matrix Q has orthonormal columns, $M := Q'\cdot H\cdot Q$, $R := H\cdot Q - Q\cdot M$.
Now, provided $\|R\cdot M^{-1}\|_2 < \psi < 1$, the m eigenvalues $\mu_i$ in $\mathbb{E}(M)$ estimate m of the n eigenvalues $\theta_j$ in $\mathbb{E}(H)$ thus: *cf.* p. 70

> Among n eigenvalues $\theta_j$ in $\mathbb{E}(H)$ there are m each of which lies in a different (possibly overlapping) interval wherein $|\log(\theta/\mu_i)| < \psi$ for $i = 1, 2, 3, \ldots, m$.

These bounds are advantageous only when $\|R\cdot M^{-1}\|$ is a lot smaller than $\|R\|\cdot\|M^{-1}\|$ .

**Deflation** replaces $H = H' = \begin{bmatrix} M & B \\ B' & W \end{bmatrix}$ by $Y = \begin{bmatrix} M & O \\ O' & W \end{bmatrix}$ when $\|B\|_2$ is deemed small enough.

Its motivation is to repace a big eigenproblem by two smaller ones, maybe not both much smaller.
Replacing $\mathbb{E}(H)$ by $\mathbb{E}(Y)$ incurs *Absolute* errors bounded by $\|B\|_2$, or by $\|B\|_2^2/\gamma$ if the gap $\gamma$ is known to be big enough. *Relative* errors are bounded by $\psi < 1$ whenever

***both*** $\|M^{-1}\cdot B\|_2 < \psi$ ***and*** $\|B\cdot W^{-1}\|_2 < \psi$ .

Only in special situations can these Relative error-bounds be worth what they cost to compute.

# Relative Error Bounds for Deflated Singular Value Problems

Let triangular $S := \begin{bmatrix} D & E \\ O' & F \end{bmatrix}$ and $Z := \begin{bmatrix} D & O \\ O' & F \end{bmatrix}$ have singular value multisets respectively

$$\mathbb{S}(S) = \{ \sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_n \} \text{ and } \mathbb{S}(Z) = \{ \zeta_1 \geq \zeta_2 \geq \ldots \geq \zeta_n \} = \mathbb{S}(D) \cup \mathbb{S}(F)$$

$$\text{wherein } \mathbb{S}(D) = \{ \delta_1 \geq \delta_2 \geq \ldots \geq \delta_m \} \text{ and } \mathbb{S}(F) = \{ \phi_1 \geq \phi_2 \geq \ldots \geq \phi_{n-m} \} .$$

Z comes from S via Deflation. Every Absolute Error $|\sigma_j - \zeta_j| \leq \|E\|$ .

## What about Relative Errors $\log(\sigma_j/\zeta_j)$ ?

If ***either*** $\|D^{-1}\cdot E\| < 2\psi < 1$ ***or*** $\|E\cdot F^{-1}\| < 2\psi < 1$ then every $|\log(\sigma_j/\zeta_j)| < \psi$ .

It persists with $0/0 := 1$ even if some $\zeta_j = 0$ so long as either $\|D^{-1}\cdot E\|$ or $\|E\cdot F^{-1}\|$ exists.

**Example:**

Let n-by-n $S := \text{bidiag}\begin{bmatrix} s & s & \ldots & s & s & e \\ 1 & 1 & \ldots & \ldots & 1 & 1 & f \end{bmatrix} = \begin{bmatrix} D & \mathbf{e} \\ \mathbf{o}' & f \end{bmatrix}$ in which the pair $\begin{bmatrix} s \\ 1 \end{bmatrix}$ is missing from only the first and last columns, and $s > f >> 1 > e > 0$ .

$$\sigma_1 \approx s + 1 >> \sigma_n \approx (s^2 - 1)/\sqrt{(s^{2n} - n\cdot s^2 + n - 1)} \quad \text{for } s > 3 \text{ and } n > 3 .$$

Deleting $\mathbf{e}$ causes relative error $< \psi$ if $e < 2\psi\cdot f$ though this $e$ can exceed $\sigma_n$ hugely.

For relative error-bounds improved substantially by a knowledge of gaps like $\gamma$, see Kahan [2012"].

# Spectral Gaps, Invariant Subspaces, and Angles

Although the n eigenvectors of n-by-n H = H' can be chosen to be orthonormal, those belonging to eigenvalues repeated or too tightly clustered are partially indeterminate, at least numerically. Instead the ***Invariant Subspace*** spanned by the cluster's eigenvectors is determined accurately whenever the cluster is separated from the rest of the spectrum by a sufficiently wide gap. The ***Angles*** between a subspace and an approximation to it are described best in Davis & Kahan [1969]; the best way to compute them is on p. 7 of <www.eecs.berkeley.edu/~wkahan/Math128/NearestQ.pdf> .

Let the columns of n-by-m Q be an orthonormal ($Q'\cdot Q = I$) basis for an approximation to the invariant subspace belonging to a tight cluster of m eigenvalues of H . As before,

$$[Q, \overline{Q}]'\cdot H\cdot[Q, \overline{Q}] = \begin{bmatrix} M & B \\ B' & W \end{bmatrix}; \quad Y := \begin{bmatrix} M & O \\ O' & W \end{bmatrix}; \quad R := H\cdot Q - Q\cdot M = \overline{Q}\cdot B' \text{ so } \beta := \|R\|_2 = \|B\|_2 .$$

$\mathbb{E}(H) = \mathbb{E}([Q, \overline{Q}]'\cdot H\cdot[Q, \overline{Q}]) = \{ \theta_1 \ge \theta_2 \ge \theta_3 \ge \ldots \ge \theta_n \}$ . Recall $M := Q'\cdot H\cdot Q$.

$\mathbb{E}(M) = \{ \mu_1 \ge \mu_2 \ge \mu_3 \ge \ldots \ge \mu_m \}$ approximates a tight cluster in $\mathbb{E}(H)$ .

$\mathbb{E}(W) = \{ \omega_1 \ge \omega_2 \ge \omega_3 \ge \ldots \ge \omega_{n-m} \}$ approximates the rest of $\mathbb{E}(H)$ .

Gap $\gamma := \min_i \gamma_i = \min_{i,j} |\mu_i - \omega_j|$ . This may be of little use unless $\gamma >> 2\beta$ .

Let $\hat{A}$ be the biggest angle between Range(Q) and the cluster's invariant subspace;

then $\hat{A} \le \arcsin(2\beta/\gamma)/2$ if $\gamma > 2\beta$ .

When $\mathbb{E}(W)$ is all on just one side of $\mathbb{E}(M)$ , then $\hat{A} \le \arctan(2\beta/\gamma)/2 \le \beta/\gamma$ .

# Miscellany:

• Elaborations of Perturbation Theory: Ren-Cang Li's §15 in Hogben [2007]

• Compute Eigenvalues to High Relative Accuracy: Z. Drmač's §46 in Hogben
… of Graded Real Symmetric Matrices, Positive Definite Matrices (only ?)
… of Specially Structured Matrices: see his citations of P. Koev *et al.*
Another such example: pp. 12-13 of <…/HilbMats.pdf>

• Computational Methods and Perturbation Theories of Singular Values of L :
… same as for eigensystem of $\begin{bmatrix} O & L^T \\ L & O \end{bmatrix}$ amended by some simplifications.
see R. Mathias' §17 in Hogben [2007]

**Regions in the Complex Plane** associated with **L** acting on a complex space:

• ε-*Pseudo-Spectrum* of **L** : $\{ \zeta$ for which $\|(\zeta\cdot\mathbf{I} - \mathbf{L})^{-1}\| > 1/\varepsilon \}$ . Usually $\|\ldots\|_2$
… includes spectrum of **L** ; see M. Embree's §16 in Hogben [2007]

• *Field of Values* of, or *Numerical Range* of **L** : $\{ \mathbf{w}'\cdot\mathbf{L}\cdot\mathbf{x}/\mathbf{w}'\cdot\mathbf{x} \}$ as $\mathbf{w}'$ and $\mathbf{x}$ run through all nonzero pairs *Dual* (p. 17) with respect to the space's norm.
… includes spectrum of **L** . Convex set for $\|\ldots\|_2$ ; see C.K. Li's §18 in Hogben

# Citations and Further Reading

A huge encyclopedic survey of facts citing the literature for their proofs is ***Handbook of Linear Algebra*** ed. by Leslie Hogben [2007] 1504 pp., Chapman & Hall/CRC.

For a treatment of finite-dimensional linear algebra that prepares the diligent reader for infinite dimensions, try ***Finite-Dimensional Linear Analysis, a Systematic Presentation in Problem Form*** by I.M. Glazman & Ju.I. Ljubič, translated and edited by G.P. Barker & G. Kuerti [1974] MIT Press. This huge text's exposition consists of about 1500 problems with hints but none accompanied by a full solution. You must do it yourself.

***Normed Linear Spaces*** 3rd ed. by M.M. Day [1973], Springer-Verlag, is a short (211 pages) brutally compressed overview of the situation in infinite-dimensional spaces. The last chapter is a nine-page reader's guide to the literature up to 1972.

For more about unitarially invariant Cross-Norms and Symmetric Gauge Functions applicable to linear operators upon Hilbert spaces, see ch. V of ***Norm Ideals of Completely Continuous Operators*** by R. Schatten [1960], Springer-Verlag.

B.N. Parlett's book ***The Symmetric Eigenvalue Problem*** [1998] 426 pp., SIAM, Philadelphia, is the best source about the properties and computations of eigensystems of real symmetric and Hermitian matrices.

Citations & Further Reading continues …

… Citations & Further Reading continued …

N.J. Higham [2002] ***Accuracy & Stability of Numerical Algorithms*** 2nd ed., ~700 pp,. SIAM, Philadelphia, is the best text treating error-analysis of roundoff.

Often better than bounds upon norms of errors are elementwise bounds mentioned herein at the bottom of p. 30. For more about them see Higham's "A survey of componentwise perturbation theory in numerical linear algebra" pp. 49-77 in …

W. Gautschi (ed.) ***Mathematics of Computation 1943-1993,*** *A half century of computational mathematics*; Proc. of the *Math. of Comp*. 50th Anniv. Symposium, 9 - 13 Aug. 1993 in Vancouver B.C., American Math Soc.

Chi-Kwong Li & Roy Mathias [1999] "The Lidskii-Mirsky-Wielandt Theorem — additive and multiplicative versions" pp. 377-413 of ***Numerische Mathematik*** **81**, is a superb survey with elegant proofs of matrix norms' relations with Hermitian matrices.

Chandler Davis & W.M. Kahan [1969] "Some New Bounds on Perturbations of Subspaces" pp. 863-868 in ***Bulletin Amer. Math. Soc.*** **75** #4. This describes bounds upon angles between subspaces in a readable way, far more so than the most often cited …

Davis & Kahan [1970] "The Rotation of Eigenvectors by a Perturbation. III" pp. 1-46 in ***SIAM J. Numer. Anal*** **7** #1. Used herein on p. 74.

Citations & Further Reading continues …

… Citations & Further Reading continued …

Chi-Kwong Li & Ren-Cang Li [2005] "A note on eigenvalues of perturbed Hermitian matrices" pp. 183-190 in ***Linear Algebra and its Applications* 395** , cited herein on p. 71.

"Deflations Preserving Relative Accuracy" by W. Kahan [2012"] was posted recently at <www.eecs.berkeley.edu/~wkahan/4June12.pdf> and fully at …/ma221/Deflate.pdf> .

Mentioned at the bottom of p. 56 is the possibly heightened sensitivity of the eigenvalues of an Hermitian matrix to non-Hermitian perturbations. For more about this see …

"Spectra of Operators with Fixed Imaginary Parts" by Andrzej Pokrzywa [1981], pp. 359-364 in ***Proc. Amer. Math. Soc*. 81** #3, and …

"Arbitrary Perturbations of Hermitian Matrices" by Arnold Schönhage [1979], pp. 143-9 in ***Linear Algebra and its Applications* 24**

# Epilogue

I learned what little I know about norms *etc*. over half a century ago, and later turned in a different direction. Consult appropriate Math. Dept. professors for this century's understandings of normed and more general metric spaces.

As an error-analyst, I have chosen mostly applications to error-analyses to illustrate how a norm can be used. Scattered among them are attempts to awaken an awareness of how important, despite its difficulty, is choosing an appropriate thing to gauge with a norm. Choosing an *Equitable* norm (pp. 12, 30, 39, 48, 52, 68) raised that issue. However, …

Many a situation cannot be comprehended in a single number.

These situations abound in Scientific and Engineering computations, and in almost all human endeavors; see a few surprising military examples in …
J.G. Roche & B.D. Watts [1991] "Choosing Analytic Measures" pp. 165-209 in ***J. Strategic Studies*** **14** #2. Disregard their mathematically naive and irrelevant uses of "linear" and "chaos".

Still, when a decision is needed, it often amounts to distilling one number out of many:
Pass a test? Accept a candidate? Choose a purchase? Launch an enterprise? …
Such unavoidable decisions must occasionally be mistaken or, less often, very lucky.