so K is a square root of M^*M. There is only one positive semidefinite square root of any positive semidefinite matrix, so the $K = Q\Sigma Q^*$ we specified is the unique choice.

If M, and hence K, is invertible the matrix $U = MK^{-1}$ is also unique.

Likewise, $K' = \sqrt{M^*M}$ is always unique.

When M is normal we have $K = K'$ and so $M = UK = KU$.

Any complex matrix M can be written as $\boldsymbol{M = UK = K'U}$ where U is **unitary** and K and K' are **positive semidefinite.**

$K = \sqrt{MM^*}$ and $K' = \sqrt{M^*M}$ so the positive semidefiniteness condition makes them unique.

These are called **Polar Decompositions of M.**

Note that if M is normal then $K = K'$ and $M = UK = KU$.

If M (and hence K) is invertible U is unique.

67. Operator Norm and Error in Matrix Equations

Operator Norm.

A **norm** on any real vector space V is a function

$$\|\cdot\| \colon V \to [0, \infty)$$

with the following properties.

(1) $\|\boldsymbol{v}\| = 0$ if and only if $\boldsymbol{v} = 0$.

(2) $\|c\boldsymbol{v}\| = |c|\,\|\boldsymbol{v}\|$ for all real c and any vector $\boldsymbol{v}$.

(3) $\|\boldsymbol{v} + \boldsymbol{w}\| \leq \|\boldsymbol{v}\| + \|\boldsymbol{w}\|$ for any pair of vectors $\boldsymbol{v}$ and $\boldsymbol{w}$.

We defined norm earlier for inner product spaces, and created the norm using the inner product. But norms can arise elsewhere too.

67.1. ***Exercise***. *Prove that for any norm*

$$|\,\|\boldsymbol{v}\| - \|\boldsymbol{w}\|\,| \leq \|\boldsymbol{v} - \boldsymbol{w}\|.$$

If $M \in \mathbb{M}_{m\times n}(\mathbb{R})$, a real $m \times n$ matrix, we know that M has a singular value decomposition. Specifically, there is an orthogonal $m \times m$ matrix P and orthogonal $n \times n$ matrix Q and diagonal Σ so that

$$M = P\Sigma Q.$$

The diagonal entries of Σ, the numbers $\sigma_1, \ldots, \sigma_n$, are the singular values. They form non-negative and decreasing sequence.

We define $\|\boldsymbol{M}\|_{\boldsymbol{op}} = \sigma_1$, the first (and largest) of these singular values. This is called the **operator norm of M**.

67.2. ***Exercise***. *Suppose $c_1 > c_2 > \cdots > c_k > 0$. Show that c_1 is the maximum of all sums of the form $c_i b^i$ where the b^i are non-negative and add to 1.*

We can use the exercise to create an alternative definition of $\|M\|_{op}$, which does not mention the SVD decomposition. The claim is that

$$\|\boldsymbol{M}\|_{\boldsymbol{op}} = \textbf{ maximum value of } \{\, \|M\mathbf{u}\| \mid \mathbf{u} \textbf{ is a unit vector} \,\}.$$

Since P and Q in the equation $M = P\Sigma Q$ are orthogonal, they don't change the length of vectors. So

$$\|M\boldsymbol{v}\| = \|P\Sigma Q\boldsymbol{v}\| = \|\Sigma Q\boldsymbol{v}\| = \|\Sigma\boldsymbol{w}\|$$

for vector $\boldsymbol{w} = Q^{-1}\boldsymbol{v} = Q^T\boldsymbol{v}$, which has the same length as $\boldsymbol{v}$.

$$\|M\boldsymbol{v}\|^2 = \|\Sigma\boldsymbol{w}\|^2 = (\Sigma\boldsymbol{w}) \cdot (\Sigma\boldsymbol{w}) = \sum_{i=1}^{n} (\sigma_i)^2 (w^i)^2.$$

Since $\sum_{i=1}^{n} (w^i)^2 = 1$, if we are looking for the maximum value which can be obtained it must be that the only nonzero coefficients among the w^i correspond to the singular value σ_1, in which case the sum itself is σ_1, as claimed.

We still have not verified that $\|\cdot\|_{op}$ is actually a norm.

It is obvious that it satisfies items (1) and (2) of the requirements for a norm. It remains to verify (3), the triangle inequality.

Suppose K and M are $m \times n$ matrices and $\boldsymbol{v}$ is a unit vector for which $\|K + M\|_{op} = \|(K + M)\boldsymbol{v}\|$. Then we have

$$\|K + M\|_{op} = \|(K + M)\boldsymbol{v}\| \leq \|K\boldsymbol{v}\| + \|M\boldsymbol{v}\| \leq \|K\|_{op} + \|M\|_{op}$$

and the triangle inequality holds for operator norm.

Operator norm has a feature not possessed by many norms. It is called **sub-multiplicative** by virtue of the fact that

$$\|\boldsymbol{KM}\|_{\boldsymbol{op}} \leq \|\boldsymbol{K}\|_{\boldsymbol{op}} \|\boldsymbol{M}\|_{\boldsymbol{op}}$$

whenever K and M are compatible matrices for multiplication in this order. Since most vector spaces fail to possess a multiplication between members, there is no possibility of this inequality for most norms.

Assuming $\|KM\|_{op} \neq 0$ select $\boldsymbol{v}$ so that $\|KM\|_{op} = \|KM\boldsymbol{v}\|$. Then

$$\|KM\|_{op} = \|KM\boldsymbol{v}\| = \|M\boldsymbol{v}\| \left\| K\left(\frac{M\boldsymbol{v}}{\|M\boldsymbol{v}\|}\right)\right\| \leq \|K\|_{op}\|M\|_{op}.$$

This implies that $\mathbf{1} \leq \|\boldsymbol{A}\|_{op}\|\boldsymbol{A}^{-1}\|_{op}$ for invertible $\boldsymbol{A}$.

Suppose $m \times n$ matrix B has entries $b_{i,j}$ and rows R_i. Suppose every entry of B is smaller than ε in magnitude and $\boldsymbol{w}$ is a unit vector. Let $\boldsymbol{y} = B\boldsymbol{w}$. Then

$$\begin{aligned}\|\boldsymbol{y}\|^2 = \boldsymbol{y}\cdot\boldsymbol{y} = y^i y^i = \sum_{i=1}^{m}(R_i\boldsymbol{w})^2 = \sum_{i=1}^{m}(R_i^T\cdot\boldsymbol{w})^2 \leq \sum_{i=1}^{m}(\|R_i^T\|\,\|\boldsymbol{w}\|)^2 \\ = \sum_{i=1}^{m}(\|R_i^T\|)^2 = \sum_{i=1}^{m}\sum_{j=1}^{n} b_{i,j}^2 < mn\varepsilon^2.\end{aligned}$$

This implies that $\|B\|_{op}$ cannot exceed $\varepsilon\sqrt{mn}$. Together with the triangle inequality we have

$$\|A\|_{op} - \varepsilon\sqrt{mn} \;\leq\; \|A+B\|_{op} \;\leq\; \|A\|_{op} + \varepsilon\sqrt{mn}.$$

In other words, two matrices whose entries are all close have operator norms that are close too.

On the other hand, suppose a single entry of matrix $B = C - A$ exceeds δ. Suppose that entry is $b_{i,j}$. Then $B\mathbf{e}_j$, which is the jth column of B, has magnitude exceeding δ. Since $\mathbf{e}_j$ is a unit vector that means $\|B\|_{op} > \delta$.

So two matrices $\boldsymbol{A}$ and $\boldsymbol{C}$ are "close" in the sense that $\|\boldsymbol{C}-\boldsymbol{A}\|_{op}$ is small if and only if they are "close" in the sense that every entry in $\boldsymbol{A}$ is close to the corresponding entry in $\boldsymbol{C}$.

Condition Number and Error.

If A is an invertible matrix with operator norm we define the **condition number** of A to be

$$\boldsymbol{K_A} = \|A\|_{op}\|A^{-1}\|_{op}$$

.

Any of the standard mathematical software packages will produce this condition number.

67.3. ***Exercise***. *Suppose A is invertible with non-increasing positive singular values $\sigma_1, \sigma_2, \ldots, \sigma_n$. Show that $K_A = \frac{\sigma_1}{\sigma_n}$.*

Condition numbers cannot be smaller than 1, but a very large condition number indicates that the matrix has features that interfere with the accuracy of a computation in situations where finite-precision arithmetic (usually to 16 places) rather than "perfect arithmetic" is being performed, and suggests that any calculation involving the matrix should be used with caution. Such matrices are called **ill-conditioned.**

We suppose given a matrix equation $A\,\boldsymbol{x} = \boldsymbol{b}$ for invertible matrix A. The solution is given by $\boldsymbol{x} = A^{-1}\boldsymbol{b}$ in terms of the target $\boldsymbol{b}$.

It may be that entries of coefficient matrix A are known exactly but there is variation $\delta\boldsymbol{b}$ in knowledge of the target vector. We don't know $\delta\boldsymbol{b}$ exactly, but we do have an idea of the range of possible variation: the maximum possible magnitude of the entries of $\delta\boldsymbol{b}$.

We suppose $\boldsymbol{y} = \boldsymbol{x} + \delta\boldsymbol{x}$ where $A\,\boldsymbol{y} = \boldsymbol{c} = \boldsymbol{b} + \delta\boldsymbol{b}$.

So $\boldsymbol{y}$ is the solution for erroneous target $\boldsymbol{c}$, and so probably differs from the solution $\boldsymbol{x}$ we really would like to know for the true target $\boldsymbol{b}$.

But by how much could it differ?

We calculate that

$$A\,\boldsymbol{y} = A\,(\boldsymbol{x} + \delta\boldsymbol{x}) = \boldsymbol{c} = \boldsymbol{b} + \delta\boldsymbol{b} \quad\Longrightarrow\quad A\,\delta\boldsymbol{x} = \delta\boldsymbol{b} \quad\Longrightarrow\quad \delta\boldsymbol{x} = A^{-1}\delta\boldsymbol{b}.$$

So $\|\delta\boldsymbol{x}\| \le \|A^{-1}\|_{op}\,\|\delta\boldsymbol{b}\|$. And also $A\,\boldsymbol{y} = \boldsymbol{c}$ so $\|A\|_{op}\|\boldsymbol{y}\| \ge \|\boldsymbol{c}\|$.

Then we have

$$\frac{\|\delta\boldsymbol{x}\|}{\|\boldsymbol{y}\|} \le \|A\|_{op}\|A^{-1}\|_{op}\frac{\|\delta\boldsymbol{b}\|}{\|\boldsymbol{c}\|} = K_A\frac{\|\delta\boldsymbol{b}\|}{\|\boldsymbol{c}\|}$$

which is an expression for the worst-case relative "error" of calculated solution $\boldsymbol{y}$ in terms of the relative uncertainty of the "target" vector $\boldsymbol{c}$ and the condition number of exactly known matrix A.

We now suppose that $\boldsymbol{b}$ is known exactly but matrix A has variation δA. We calculate our solution using slightly erroneous but invertible matrix $C = A + \delta A$.

We don't actually know δA but we do understand our ability to do the measurements that produced the entries of C, so we have an estimate of the maximum size of the entries of δA, and we assume δA to be small enough so that $C + \delta B$ is invertible whenever $\|\delta B\|_{op}$ does not exceed the maximum possible value of $\|\delta A\|_{op}$. We have matrix C in hand and know it is invertible, and with this supposition the (unknown but exact) matrix A is too.

Suppose, again, that $\boldsymbol{x}$ is the solution we really want, $A\,\boldsymbol{x} = \boldsymbol{b}$, and $\boldsymbol{y} = \boldsymbol{x} + \delta\boldsymbol{x}$ is the solution we *have*, with

$$C\,\boldsymbol{y} = (A + \delta A)\boldsymbol{y} = (A + \delta A)(\boldsymbol{x} + \delta\boldsymbol{x}) = \boldsymbol{b}.$$

Then $(A+\delta A)\boldsymbol{y} = \boldsymbol{b}$

$$\implies A\,\boldsymbol{x} + A\,\delta\boldsymbol{x} + \delta A\,\boldsymbol{y} = \boldsymbol{b} \quad \implies \quad A\,\delta\boldsymbol{x} = -\delta A\,\boldsymbol{y}$$

$$\implies \quad \delta\boldsymbol{x} = A^{-1}\delta A\,\boldsymbol{y} \quad \implies \quad \|\delta\boldsymbol{x}\| \le \|A^{-1}\|_{op}\|\delta A\|_{op}\|\boldsymbol{y}\|.$$

This now produces

$$\frac{\|\delta\boldsymbol{x}\|}{\|\boldsymbol{y}\|} \le \|A\|_{op}\|A^{-1}\|_{op}\frac{\|\delta A\|_{op}}{\|A\|_{op}} = K_A\frac{\|\delta A\|_{op}}{\|A\|_{op}}$$

expressing the solution error relative to the disturbed solution in terms of the condition number of matrix A and the relative matrix error.

The problem here is that in this situation we don't *have* A, so we can't calculate K_A. We do have $C = A+\delta A$, and so $\|A\| = \|C-\delta A\| \ge \|C\|-\|\delta A|$.

$$\frac{\|\delta\boldsymbol{x}\|}{\|\boldsymbol{y}\|} \le \|A^{-1}\|_{op}\|\delta A\|_{op} = \|C\|_{op}\|C^{-1}\|_{op}\frac{\|A\|_{op}\|A^{-1}\|_{op}}{\|C\|_{op}\|C^{-1}\|_{op}}\frac{\|\delta A\|_{op}}{\|A\|_{op}}$$

$$= K_C\frac{K_A}{K_C}\frac{\|\delta A\|_{op}}{\|A\|_{op}} \le \left(\frac{K_A}{K_C}\right)K_C\frac{\|\delta A\|_{op}}{\|C\|_{op}-\|\delta A\|_{op}}$$

The final expression gives the the relative solution error in terms of quantities actually in our possession, except for the factor $\frac{K_A}{K_C}$.

If δA is small $K_C = \frac{\sigma_1}{\sigma_n}$ might be close to K_A, where the σ_i are the singular values for C. However if σ_n is close to zero even a small δA could move the smallest singular value for C relatively closer to zero and blow up the factor $\frac{K_A}{K_C}$.

So if the least singular value σ_n for C is large compared to δA the estimate

$$\frac{\|\delta\boldsymbol{x}\|}{\|\boldsymbol{y}\|} \le K_C\left(\frac{\|\delta A\|_{op}}{\|C\|_{op}-\|\delta A\|_{op}}\right),$$

is likely to be useful. Otherwise you should be very cautious about using the approximate solution $\boldsymbol{y}$.

I should emphasize here that *this does not mean the solution is necessarily bad or useless.* This is worst-case scenario. It is a warning flag.

In the exercise below we investigate a typical situation where $\frac{K_A}{K_C}$ is near one so we *can* use this uncertainty estimate.

67.4. ***Exercise***. *Suppose* $\|C\|_{op} = 1.5$ *and* $K_C = 1000$ *and the uncertainty in the entries of* C *produces maximum possible* $\|\delta A\|_{op} = 10^{-4}$. *The least singular value of* C *is large compared to* 10^{-4}.

You solve $C\boldsymbol{y} = \boldsymbol{b}$ *and find* $\|\boldsymbol{y}\| = 2.3$.

The top entry y^1 *of solution vector* $\boldsymbol{y}$ *is* 1.3423192. *How many of those decimal places should you trust?*