Recap: Last Time

We looked at a backward stability proof in gory detail. — The technique is quite straight-forward, albeit somewhat tedious.

— We replace the floating point operators $\oplus$, $\ominus$, $\otimes$, and $\oslash$ with exact mathematical operations $+$ relative error terms, i.e. $x \oplus y \leadsto (x + y)(1 + \epsilon)$, where $|\epsilon| \leq \epsilon_{\text{mach}}$.

— Then we interpret the error as perturbations on the appropriate part of the problem formulation (so that that computed solution is the exact solution to a nearby problem).

As we used the backward substitution algorithm for the detailed backward stability proof; we now turn to the least squares problems for a detailed discussion on conditioning...

...and we recall that Accuracy(conditioning, stability), so these are all important pieces in the larger "numerics jigsaw puzzle."
Once again, we return to the least squares problem.

\[ \mathbf{b} \]
\[ \mathbf{r} = \mathbf{b} - \mathbf{A}\mathbf{x} \]
\[ \mathbf{y} = \mathbf{A}\mathbf{x} = \mathbf{P}\mathbf{b} \]

We measure everything in the two-norm, and let \( \| \cdot \| = \| \cdot \|_2 \); formally we are trying to solve

Given \( \mathbf{A} \in \mathbb{C}^{m \times n} \) of full rank, \( m \geq n \), \( \tilde{\mathbf{b}} \in \mathbb{C}^m \), find \( \tilde{\mathbf{x}} \in \mathbb{C}^n \) such that \( \| \tilde{\mathbf{b}} - \mathbf{A}\tilde{\mathbf{x}} \| \) is minimized.

The conditioning of these problems depend on a combination of

1. The conditioning of square systems of equations
2. The geometry of orthogonal projections.

The topic is subtle, and has nontrivial implications for the **stability** (and ultimately, the **accuracy**) of least squares algorithms.

From our previous discussion of least squares problem we know

\[
\tilde{\mathbf{x}} = \mathbf{A}^\dagger \tilde{\mathbf{b}}, \quad \text{where} \quad \mathbf{A}^\dagger = (\mathbf{A}^* \mathbf{A})^{-1} \mathbf{A}^* \\
\mathbf{A}\tilde{\mathbf{x}} = \tilde{\mathbf{y}}, \quad \text{where} \quad \tilde{\mathbf{y}} = \mathbf{P}\tilde{\mathbf{b}}, \quad \mathbf{P} = \mathbf{A}\mathbf{A}^\dagger
\]

\( \mathbf{P} \) is the orthogonal projector onto \( \text{range}(\mathbf{A}) \), and \( \mathbf{A}^\dagger \in \mathbb{C}^{m \times m} \) is the pseudo-inverse of \( \mathbf{A} \).

Conditioning is the measure of sensitivity of solutions to perturbations in the data.

Our data are

\[
\mathbf{A} \in \mathbb{C}^{m \times n}, \quad \text{and} \quad \tilde{\mathbf{b}} \in \mathbb{C}^m,
\]

and the solution is either the vector \( \tilde{\mathbf{x}} \in \mathbb{C}^n \), or the vector \( \tilde{\mathbf{y}} = \mathbf{P}\tilde{\mathbf{b}} \) (depending on our point of view / application).

We end up with four combinations of input/output-perturbations:

<table>
<thead>
<tr>
<th>Input, Output →</th>
<th>( \tilde{\mathbf{y}} )</th>
<th>( \tilde{\mathbf{x}} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \tilde{\mathbf{b}} )</td>
<td>( \kappa(\tilde{\mathbf{b}} \rightarrow \tilde{\mathbf{y}}) )</td>
<td>( \kappa(\tilde{\mathbf{b}} \rightarrow \tilde{\mathbf{x}}) )</td>
</tr>
<tr>
<td>( \mathbf{A} )</td>
<td>( \kappa(\mathbf{A} \rightarrow \tilde{\mathbf{y}}) )</td>
<td>( \kappa(\mathbf{A} \rightarrow \tilde{\mathbf{x}}) )</td>
</tr>
</tbody>
</table>

We are going to express all the condition-numbers using three dimensionless parameters — \( \kappa(A) \), \( \theta \), and \( \eta \).

\( \kappa(A) \) is our old friend the condition number of the matrix \( \mathbf{A} \)

\[
\kappa(A) = \frac{\sigma_1}{\sigma_n}.
\]

\( \theta \) is the angle between \( \tilde{\mathbf{b}} \) and \( \tilde{\mathbf{y}} = \mathbf{A}\tilde{\mathbf{x}} = \mathbf{P}\tilde{\mathbf{b}} \),
Three Dimensionless Parameters

η is a measure of how much \( \|\vec{y}\| \) falls short of its maximum value, given \( \|A\| \) and \( \|\vec{x}\| \):

\[
\eta = \frac{\|A\| \|\vec{x}\|}{\|\vec{y}\|} = \frac{\|A\| \|\vec{x}\|}{\|A\vec{x}\|}.
\]

These parameters lie in the ranges

\[
\kappa(A) \in [1, \infty), \quad \theta \in \left[0, \frac{\pi}{2}\right], \quad \eta \in [1, \kappa(A)),
\]

and

\[
\cos(\theta) = \frac{\|\vec{y}\|}{\|\vec{b}\|}, \quad \theta = \cos^{-1}\left(\frac{\|\vec{y}\|}{\|\vec{b}\|}\right).
\]

A Note on the Theorem

In the special case \( m = n \), the least squares problem reduces to a square non-singular problem, with \( \theta = 0 \), and the table looks like

<table>
<thead>
<tr>
<th>↓ Input, Output →</th>
<th>( \vec{y} )</th>
<th>( \vec{x} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \vec{b} )</td>
<td>( \frac{1}{\cos \theta} )</td>
<td>( \frac{\kappa(A)}{\eta \cos \theta} )</td>
</tr>
<tr>
<td>( A )</td>
<td>( 0 )</td>
<td>( \frac{\kappa(A)}{\cos \theta} )</td>
</tr>
</tbody>
</table>

Since \( A \) is square + full rank, \( \vec{y} \) is already in the range, so no projection is needed; hence the condition number is 0.

Least Squares Problems... Conditioning Theorem

Theorem

Let \( \vec{b} \in \mathbb{C}^m \) and \( A \in \mathbb{C}^{m \times n} \) of full rank be given. The least squares problem, \( \min_{\vec{x} \in \mathbb{C}^n} \| \vec{b} - A\vec{x} \| \) has the following 2-norm relative condition numbers describing the sensitivities of \( \vec{y} \) and \( \vec{x} \) to perturbations in \( \vec{b} \) and \( A \):

\[
\begin{array}{c|c|c}
\downarrow \text{Input, Output} & \vec{y} & \vec{x} \\
\hline
\vec{b} & 1 & \frac{\kappa(A)}{\eta \cos \theta} \\
A & 0 & \frac{\kappa(A)}{\cos \theta} \\
\end{array}
\]

The results in the first row are exact, being attained for certain perturbations \( \delta \vec{b} \), and the results in the second row are upper bounds.

Simplifying the Proof Using the SVD

We have shown (a long, long time ago) that every matrix has a singular value decomposition.

Let \( U\Sigma V^* = A \) be the SVD of \( A \). We can use \( U \) and \( V \) to get a convenient basis in which we prove the theorem. Since 2-norm perturbations are not changed by a unitary change of basis, the perturbation behavior of \( A \) is the same as that of \( \Sigma \).

Without loss of generality we can assume that

\[
A = \begin{bmatrix}
\Sigma & 0 \\
0 & 0
\end{bmatrix} = \begin{bmatrix} A_1 & \sigma_n \\
0 & 0
\end{bmatrix}.
\]
Moving Along...

Now with
\[
\bar{b} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}, \quad \bar{b}_1 \in \mathbb{C}^n, \; \bar{b}_2 \in \mathbb{C}^{m-n}
\]
the projection of \( \bar{b} \) onto \( \text{range}(A) \) is trivial

\[
\bar{y} = P\bar{b} = \begin{bmatrix} \bar{b}_1 \\ 0 \end{bmatrix}
\]

Now, \( A\bar{x} = \bar{y} \) has the unique solution \( \bar{x} = A_1^{-1}\bar{b}_1 \).

We note that the orthogonal projector, and the pseudo-inverse of \( A \) take the forms

\[
P = \begin{bmatrix} I_{n \times n} & 0 \\ 0 & 0 \end{bmatrix}, \quad A^\dagger = \begin{bmatrix} A_1^{-1} & 0 \end{bmatrix}
\]

---

Proof — Part#1: Sensitivity of \( \bar{y} \) to Perturbations in \( \bar{b} \)

\( \bar{y} = P\bar{b} \) is a linear differentiable map; and the Jacobian is \( P \) itself, with \( \|P\| = 1 \).

For a differentiable map \( x : \rightarrow f(\bar{x}) \) the condition number is

\[
\kappa = \frac{\|J(\bar{x})\|}{\|f(\bar{x})\|/\|\bar{x}\|}.
\]

Here we have,

\[
\kappa(\bar{b} \rightarrow \bar{y}) = \frac{\|P\|}{\|\bar{y}\|/\|\bar{b}\|} = \frac{1}{\cos \theta} \quad \square.
\]

---

Proof — Part#2: Sensitivity of \( \bar{x} \) to Perturbations in \( \bar{b} \)

\( \bar{x} = A^\dagger \bar{b} \) is also linear, with Jacobian \( J = A^\dagger \), so

\[
\kappa(\bar{b} \rightarrow \bar{x}) = \frac{\|A^\dagger\|}{\|\bar{x}\|/\|\bar{b}\|} = \|A^\dagger\| \|\bar{b}\|/\|\bar{y}\| \|\bar{x}\| = \|A^\dagger\| \frac{1}{\cos \theta} \frac{\|A\|}{\eta}
\]

Finally, we recognize \( \kappa(A) = \sigma_1 \cdot \frac{1}{\sigma_n} = \|A\|/\|A^\dagger\| \) (in this case), and we have

\[
\kappa(\bar{b} \rightarrow \bar{x}) = \frac{\kappa(A)}{\eta \cos \theta} \quad \square
\]

That concludes the easy parts of the proof...

---

Perturbations in \( A \)

Perturbations in \( A \) affect the least squares problem in two ways

- The mapping of \( \mathbb{C}^n \) onto \( \text{range}(A) \) is distorted.
- \( \text{range}(A) \) is also altered.

The changes in \( \text{range}(A) \) introduce a “tilt” of the space; and the question is what is the maximal tilt \( \delta \alpha \) induced by a perturbation \( \delta A \)?

The image of the unit sphere in \( \mathbb{R}^n, S^{n-1} \) is \( A S^{n-1} \), a hyper-ellipse that “lies flat” in \( \text{range}(A) \).

We grab a point \( \bar{p} = A\bar{v} \) on the hyper-ellipse (hence \( \|\bar{v}\| = 1 \), since \( \bar{v} \in S^{n-1} \); we introduce a perturbation \( \delta \bar{p} \perp \text{range}(A) \).

We can express this as a rank-1 matrix perturbation \( \delta A = (\delta \bar{p})\bar{v}^* \leftrightarrow (\delta A)\bar{v} = \delta \bar{p}, \) and \( \|\delta A\| = \|\delta \bar{p}\| \).
Now, clearly, if we want to maximize the tilt, we should grab the hyper-ellipse as close to the origin as possible.

Hence, we let \( \tilde{p} = \sigma_n \tilde{u}_n \) (the minor semi-axis in \( A^{n-1} \)).

Now, since we have \( A \) in a convenient diagonal form, \( \tilde{p} \) is the last column of \( A \), \( \tilde{v}^* = (0, 0, \ldots, 0, 1) \), and \( \delta A \) a perturbation below the diagonal in this (last) column.

\[
\tilde{p} = \begin{bmatrix} 0 \\ \vdots \\ \sigma_n \\ 0 \\ 0 \\ \vdots \\ \delta p_{n+1} \\ \vdots \\ \delta p_m \\ 0 \end{bmatrix}, \quad \delta \tilde{p} = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ \delta A_{n+1,n} \\ \vdots \\ \delta A_{m,n} \\ \vdots \\ \vdots \\ \delta A_{n,n} \end{bmatrix}
\]

The tilting angle induced by this perturbation is

\[
\tan(\delta \alpha) = \frac{\|\delta \tilde{p}\|}{\sigma_n}.
\]

We have

\[
\tan(\delta \alpha) = \frac{\|\delta \tilde{p}\|}{\sigma_n}.
\]

Further,

\[
\|\delta \tilde{p}\| = \|\delta A\|, \quad \delta \alpha \leq \tan(\delta \alpha),
\]

Hence,

\[
\delta \alpha \leq \frac{\|\delta A\|}{\sigma_n} = \frac{\|\delta A\|}{\|A\|} \kappa(A).
\]

We are now ready to proceed with the proof...

Since \( \tilde{y} \) is the orthogonal projection of \( \tilde{b} \) onto \( \text{range}(A) \), it is determined by \( \tilde{b} \) and \( \text{range}(A) \) alone. Therefore we can study changes on \( \tilde{y} \) induced by tiltings \( \delta \alpha \) of \( \text{range}(A) \).

No matter how we tilt \( \text{range}(A) \), \( \tilde{y} \in \text{range}(A) \) must be orthogonal to \( \tilde{y} - \tilde{b} \). — As \( \text{range}(A) \) varies, the point \( \tilde{y} \) moves along a sphere of radius \( \|\tilde{b}\|/2 \) centered at the point \( \tilde{b}/2 \).
Proof — Part#3: Sensitivity of \( \bar{y} \) to Perturbations in \( A \)

Tilting range(\( A \)) in the plane \( \bar{b} - \bar{b} - \bar{y} \) by an angle \( \delta \alpha \) changes that angle \( 2\theta \) at the central point \( \bar{b}/2 \) by \( 2\delta \alpha \).

The corresponding change \( \delta \bar{y} \), is the base of an isosceles triangle with central angle \( 2\delta \alpha \), and edge length \( ||\bar{b}||/2 \). Hence,

\[
||\delta \bar{y}|| = ||\bar{b}|| \sin(\delta \alpha)
\]

Tilting range(\( A \)) in any other plane results in a similar geometry in a different plane and perturbations smaller by a factor as small as \( \sin \theta \).

For arbitrary perturbations we have

\[
||\delta \bar{y}|| \leq ||\bar{b}|| \sin(\delta \bar{a}) \leq ||\bar{b}|| \delta \alpha
\]

Combining with previous results give us \( \kappa(A \rightarrow \bar{y}) \)

\[
\frac{||\delta \bar{y}||}{||\bar{b}||} \kappa(A) = \frac{||\bar{y}||}{||b||} \frac{||\delta A||}{||A||} \kappa(A) \iff \frac{||\delta \bar{y}||}{||\bar{y}||} / \frac{||\delta A||}{||A||} \leq \frac{\kappa(A)}{\cos \theta} \quad \Box
\]

Proof — Part#4: Sensitivity of \( \bar{x} \) to Perturbations in \( A \)

We now analyze the most interesting relationship of the theorem; the sensitivity of the least squares solution to perturbations in \( A \).

We write perturbations in two parts

\[
\delta A = \begin{bmatrix} \delta A_1 \\ \delta A_2 \end{bmatrix}, \quad \delta A_1 \in \mathbb{C}^{n \times n}, \quad \delta A_2 \in \mathbb{C}^{(m-n) \times n}
\]

First, we look at the effects of \( \delta A_1 \): these perturbations change the mapping of \( A \) in its range, but does not change range(\( A \)) itself, and hence not \( \bar{y} \). We get

\[
(A_1 + \delta A_1)\bar{x} = \bar{b}
\]

The condition number for this operation is simply (as before)

\[
\kappa(A_1 \rightarrow \bar{x}) = \frac{||\delta \bar{x}||}{||\bar{x}||} / \frac{||\delta A_1||}{||A_1||} \leq \kappa(A_1) = \kappa(A)
\]

Next, we consider the effects of \( \delta A_2 \). This perturbation tilts range(\( A \)) without changing the mapping of \( A \) within this space.

The point vector \( \vec{b}_1 \), and the point \( \bar{y} = [\vec{b}_1 \vec{b}]^{*} \) are perturbed, but \( A_1 \) is not. This corresponds to perturbing \( \bar{b}_1 \) in \( \bar{x} = A_1^{-1}\bar{b}_1 \), for which the condition number takes the form

\[
\kappa = \frac{||\delta \bar{x}||}{||\bar{x}||} \frac{||\delta \bar{b}_1||}{||\bar{b}_1||} \leq \frac{\kappa(A_1)}{\eta(A_1, \bar{x})} = \frac{\kappa(A)}{\eta}
\]

since...

\[
\frac{||\delta \bar{x}||}{||\bar{x}||} \frac{||\delta \bar{b}_1||}{||\bar{b}_1||} \leq \frac{||J(\bar{x})||}{||\bar{x}||} = \frac{||A_1^{-1}|| ||\bar{b}_1||}{||A_1|| ||\bar{x}||} = \frac{\sigma_1}{\sigma_n} \frac{||A_1 \bar{x}||}{||A_1|| ||\bar{x}||}
\]

In order to close this argument out, we must relate \( \delta \bar{b}_1 \) to \( \delta A_2 \... \)

The vector \( \vec{b}_1 \) is \( \bar{y} \) expressed in the coordinates of range(\( A \)). Therefore, the only changes in \( \bar{y} \) that are realized as changes in \( \vec{b}_1 \) are those that are parallel to range(\( A \)).
Proof — Part#4: Sensitivity of $\bar{x}$ to Perturbations in $A$

If $\text{range}(A)$ is tilted by $\delta \alpha$ in the $\overline{0}-\overline{b}-\overline{y}$ plane, the resulting perturbation $\delta \overline{y}$ is not parallel to $\text{range}(A)$, but at an angle $\frac{\pi}{2} - \theta$, therefore

$$\| \delta \overline{b}_1 \| = \| \delta \overline{y} \| \sin \theta \leq \| \overline{b} \| \delta \alpha \sin \theta.$$

If $\text{range}(A)$ is tilted in a direction orthogonal to the $\overline{0}-\overline{b}-\overline{y}$ plane, $\delta \overline{y}$ is parallel to $\text{range}(A)$, and we get $\| \delta \overline{y} \| \leq \| \overline{b} \| \delta \alpha \sin \theta$, and since $\| \delta \overline{b}_1 \| \leq \| \delta \overline{y} \|$, we have

$$\| \delta \overline{b}_1 \| \leq \| \overline{b} \| \delta \alpha \sin \theta, \quad \text{same argument as for } \kappa(A \to \overline{y}).$$

We now have all the pieces to the puzzle... all we need is a bit of glue!

One Final Comment

Clearly, finding the least squares solution $\bar{x}$ is a tough problem:

— The condition number contains the square of the condition number of the matrix $A$.

— Even for moderately ill-conditioned matrices, the least squares problem quickly becomes very ill-conditioned.

Next time we connect the conditioning results derived here with the stability (or lack thereof) of some numerical algorithms applied to the least squares problem.

Think about: Problem TB-18.2.