1 9th of November 2018 — F. Poloni

1.1 Conditioning

Two lectures ago we introduced the QR factorization to solve least squares problems and we noticed that it has a computational complexity which is much worse thant the normal equations method.

Why did we introduce the QR factorization to solve the least squares problem, then? Altough it's more complex computationally speaking, it's much better than the normal equation for what concerns accuracy. Let's see why through an example:

Example 1.1. *Let $A \in \mathcal{M}(4, 3, \mathbb{R})$ s.t.:*

$$A = \begin{pmatrix} 1 & 1 & 2 \\ 1 & 2 & 3 \\ 3 & 1 & 4 \\ 1 & 2 & 3 + 10^{-8} \end{pmatrix}$$

In this case the normal equations method doesn't even find the order of magnitude correctly.

At this point we may indroduce the problem of **sensitivity**:

Definition 1.1 (Sensitivity)**.** *We call **sensitivity** the measure of how much the output of a problem changes when we perturb its input.*

As an example, let $f(x, y) = x + 2y$. If we perturb the second parameter of f as follows $\widetilde{y} = y + \delta$, we can compute the variation in the output value of the function $\boldsymbol{v}$ as

$$\mathbf{v} = f(x, \widetilde{y}) - f(x, y) = x + 2(y + \delta) - (x + 2y) = 2\delta$$

A good example of this behaviour is the temperature of water coming from the shower: in particular, when we rotate little the knob the water becomes too cold or too hot very fast. This function is plotted in Figure 1.1.

Definition 1.2 (Absolute condition measure)**.** *The **absolute condition number** of a function f is the **maximum** possible output change / input change ratio in the limit for a **small** change of the input.*

$$\kappa_{abs}(f, x) = \lim_{\varepsilon \to 0} \sup_{|\tilde{x} - x| \leq \varepsilon} \frac{|f(\tilde{x}) - f(x)|}{|\tilde{x} - x|}$$

We would like to focus on what this definition means.

Why are we interested in the limit of a very small change?

If we zoom-in a continuous function is gets basically linear (key idea of derivative) and then the ratio between the difference on the outputs and the one of the inputs is approximatively the derivative, as shown in Figure 1.2.

We take a point x and we cosider a ball of radius ε and we compute the change in the output over the change in the input, then we take the maximum.

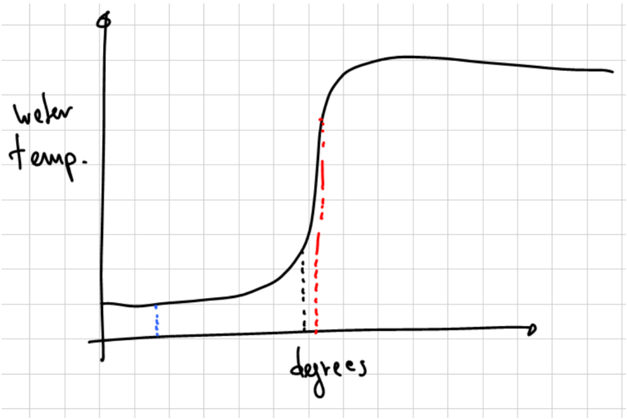

FIGURE 1.1: Geometric idea of temperature of the water in the shower

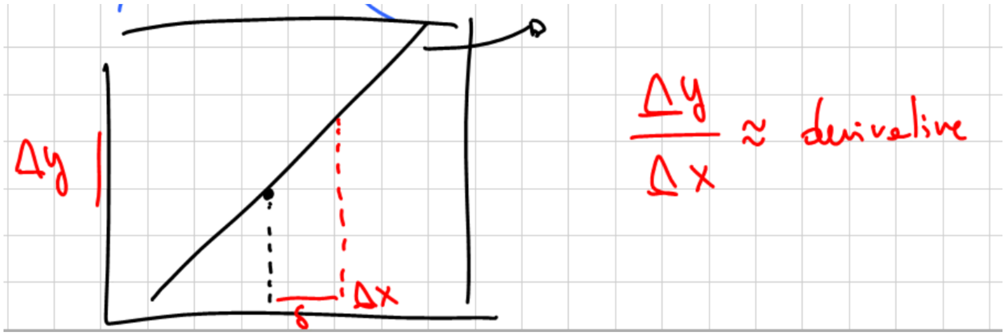

FIGURE 1.2: Geometric idea behind the derivative, as "zoom" of the function in a certain point.

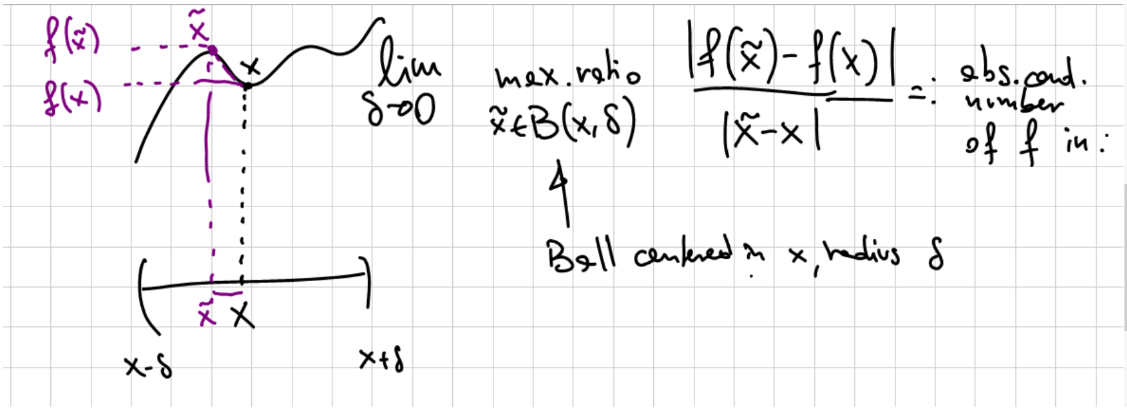

FIGURE 1.3: Geometric idea behind absolute condition number.

Example 1.2. *Let $f(x) = x^2$.*

If we perturb the input x to $x + \delta$, then $f(\tilde{x}) = (x + \delta)^2 = x^2 + 2x\delta + \delta^2$, then we obtain *the ratio*

$$r = \frac{|f(\widetilde{x}) - f(x)|}{|\widetilde{x} - x|} = \frac{|2x\delta + \delta^2|}{|\delta|} = |2x + \delta|$$

If we denote ε the radius of the ball, we obtain the following

$$\max_{\substack{|\delta|<|\varepsilon| \\ \widetilde{x}\in B(x,\varepsilon)}} r = |2x| + |\varepsilon|$$

then

$$\lim_{\varepsilon\to 0} \max_{|\delta|<|\epsilon|} r = \lim_{\varepsilon\to 0} |2x| + |\varepsilon| = |2x|$$

Example 1.3. *It's more interesting to see a multivariate function:*

Let $f(x) = x^T Qx$, for instance for $x \in \mathbb{R}^2$ so that we can plot its graph in $\mathbb{R}^3$.

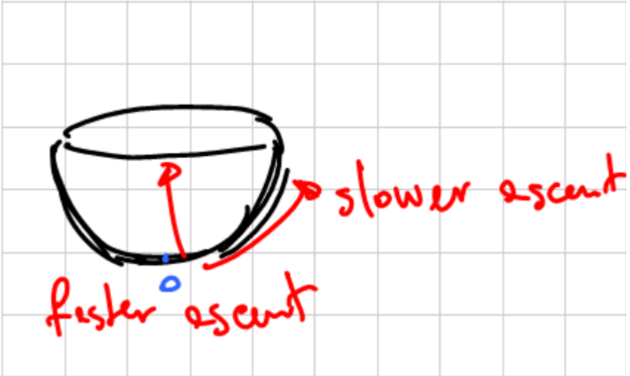

FIGURE 1.4: Paraboloid

We shall take a general example where the cross-section are ellipses, so that there is a direction of faster and slower ascent; this is not just a circular "cup" seen in perspective. Note that these directions of faster ascent and lower ascent correspond to the eigenvectors of the matrix Q.

In this case the absolute condition number is $\lim_{\varepsilon\to 0} \max_{\widetilde{x}\in B(x,\varepsilon)} \frac{\|f(\widetilde{x})-f(x)\|}{\|\widetilde{x}-x\|}$, and one can see that the output/input ratio varies with the direction in which $\widetilde{x}$ is, so we have to take a maximum in the whole ball $B(x, \varepsilon)$.

At this point, an observation is mandatory: **any** absolute measure doesn't take into account the values of the function in other points, so we want to define the following

Definition 1.3 (Relative error)**.** *The **relative error** of an approximation $\widetilde{x}$ to a quantity x is $\frac{\|\widetilde{x}-x\|}{\|x\|}$.*

Here are some examples of good and bad accuracy:

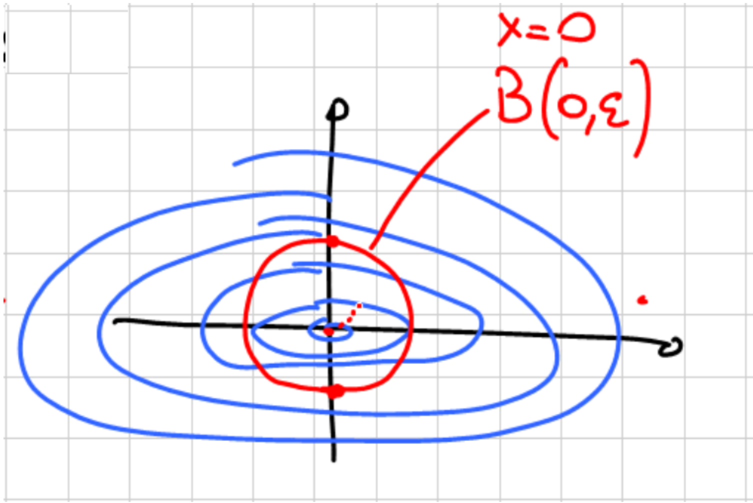

Figure 1.5: Level curves of a quadratic function ("seen from above").

- $\frac{|\tilde{x}-x|}{|x|} \approx 1$: **very bad** accuracy; it's just a number with the same order of magnitude.
- $\frac{|\tilde{x}-x|}{|x|} \approx 10^{-3}$: about 3 correct significant digits.
- $\frac{|\tilde{x}-x|}{|x|} \approx 10^{-16}$: about 16 correct digits; we **can't do better** typically (with `double` precision arithmetic).

Definition 1.4 (Relative condition number)**.** *The* ***relative condition number*** *of a function f is defined as*

$$\kappa_{rel}(f,\mathbf{x}) = \lim_{\delta\to 0} \sup_{\frac{\|\tilde{\mathbf{x}}-\mathbf{x}\|}{\|\mathbf{x}\|}\leq\delta} \frac{\frac{\|f(\tilde{\mathbf{x}})-f(\mathbf{x})\|}{\|f(\mathbf{x})\|}}{\frac{\|\tilde{\mathbf{x}}-\mathbf{x}\|}{\|\mathbf{x}\|}} = \kappa_{abs}(f,\mathbf{x})\frac{\|\mathbf{x}\|}{\|f(\mathbf{x})\|},$$

i.e., we replace the absolute error $\|\tilde{\mathbf{x}}-\mathbf{x}\|$ with the relative error.

1.1.1 Conditioning of linear systems

At this point we would like to compute the condition number of solving a linear system, i.e., the condition number of the function $f(A,b) = A^{-1}b$, perturbing the inputs A and b, one at a time.

Perturbing b We want to compute the limit of the relative error $\frac{\left\|f(A,\widetilde{b})-f(A,b)\right\|}{\|f(A,b)\|}$, so we set $x = A^{-1}b$ and $\tilde{x} = A^{-1}\tilde{b}$, and we estimate *output error* $= \frac{\left\|\widetilde{x}-x\right\|}{\|x\|} =?$

1.
$$\begin{aligned} \|\widetilde{x} - x\| &= \left\|A^{-1}\widetilde{b} - A^{-1}b\right\| \\ &= \left\|A^{-1}(\widetilde{b} - b)\right\| \\ &\le \left\|A^{-1}\right\| \left\|\widetilde{b} - b\right\| \end{aligned} \tag{1.1}$$

2. Since $\|b\| = \|Ax\| \le \|A\| \, \|x\|$ we have $\frac{\left\|\widetilde{x}-x\right\|}{\|x\|} \le \|A^{-1}\| \, \|A\| \, \frac{\left\|\widetilde{b}-b\right\|}{\|b\|}$

In the end, since *input error* $= \frac{\left\|\widetilde{b}-b\right\|}{\|b\|}$ we obtain

$$\kappa_{rel}(f, x) = \lim_{\varepsilon \to 0} \frac{output\ error}{input\ error} \le \lim_{\varepsilon \to 0} \left\|A^{-1}\right\| \|A\| = \left\|A^{-1}\right\| \|A\|$$

We denote $\kappa(A) = \|A^{-1}\| \, \|A\|$ the **condition number of** A;

Perturbing A Given $Ax = b$ we obtain $(A + \Delta_A)(x + \Delta_x) = b$, where $\widetilde{A} = A + \Delta_A$ and $\widetilde{x} = x + \Delta_x$. Then we can expand as follows

$\cancel{Ax} + \Delta_A x + A\Delta_x + \Delta_A \Delta_x = \cancel{b}$

We can stop taking into account $\Delta_A \Delta_x$, since it's a sort of second order term ($\Delta_A \Delta_x = o(\|\Delta_A\| \, \|\Delta_x\|)$), so we get the following

$$\Delta_A x + A\Delta_x = 0$$

$$\Delta_x = -A^{-1}\Delta_A x$$

then $\|\Delta_x\| \le \|A^{-1}\| \, \|\Delta_A\| \, \|x\|$, which implies $\frac{\|\Delta_x\|}{\|x\|} \le \|A^{-1}\| \, \frac{\|\Delta_A\|}{\|A\|}$.

We obtain that *relative output error* $\le \kappa(A) \cdot$ *relative input error*.

We only proved an inequality, but it turns out that it is tight: for every A and b there is a possible choice of the perturbation $\tilde{x}$ that attains equality.

In the end, in both cases, the error in the output is the error in the input (namely b or A) times the condition number.

1.2 Condition number, SVD, and distance to singularity

Fact 1.1. $\kappa(A) = \frac{\sigma_1}{\sigma_n}$, *i.e.,* $\kappa(A)$ *is the ratio between the smallest and the largest singular value.*

So we can say that if a matrix is close to a singular matrix, then its condition number is going to be large.

Proof. Let $A = USV^T$, then $\|A\| = \left\|USV^T\right\| = \|S\| = \sigma_1$, since

$$S = \begin{pmatrix} \sigma_1 & & & & & \\ & \sigma_2 & & & & \\ & & . & & & \\ & & & . & & \\ & & & & . & \\ & & & & & \sigma_n \end{pmatrix} \text{ and } \sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_n.$$

It's also true that $\|A^{-1}\| = \left\|(USV)^{-1}\right\| = \left\|VS^{-1}U^T\right\| = \|S^{-1}\| = \sigma_n$, since

$$S = \begin{pmatrix} \frac{1}{\sigma_1} & & & & & \\ & \frac{1}{\sigma_2} & & & & \\ & & . & & & \\ & & & . & & \\ & & & & . & \\ & & & & & \frac{1}{\sigma_n} \end{pmatrix} \text{ and } \sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_n.$$

In the end $\kappa(A) = \frac{\|A\|}{\|A^{-1}\|} = \frac{\sigma_1}{\sigma_n}$ □

Fact 1.2. *The relative distance between A and the closest singular matrix is $\frac{1}{\kappa(A)}$.*

Do you recall?

Eckart-Young theorem: the closest matrix to A that has rank $\leq n-1$ is:

$$\hat{A} = U \begin{pmatrix} \sigma_1 & & & & & & \\ & \sigma_2 & & & & & \\ & & . & & & & \\ & & & . & & & \\ & & & & . & & \\ & & & & & \sigma_n & \\ & & & & & & 0 \end{pmatrix} V^T$$

Proof.

$$\begin{aligned}
\left\|A-\hat{A}\right\| &= \left\| U\left(\begin{pmatrix} \sigma_1 & & & & \\ & \sigma_2 & & & \\ & & \ddots & & \\ & & & & \sigma_n \end{pmatrix} - \begin{pmatrix} \sigma_1 & & & & \\ & \sigma_2 & & & \\ & & \ddots & & \\ & & & \sigma_{n-1} & \\ & & & & 0 \end{pmatrix} \right) V^T \right\| \\
&= \left\| U \begin{pmatrix} 0 & & & \\ & \ddots & & \\ & & 0 & \\ & & & \sigma_n \end{pmatrix} V^T \right\| \\
&= \sigma_n
\end{aligned} \tag{1.2}$$

Thus, $\left\|A-\hat{A}\right\| = \sigma_n$. We know already that $\|A\| = \sigma_1$, so we just need to take the ratio. □

We analyzed the conditioning of linear systems, but the main problem we want to study in this course is least squares problem.

1.3 Conditioning of least squares problem

We need two quantities to be able to measure the conditioning of least squares problem:

- $\kappa(A)$ Let $A \in \mathcal{M}(m, n, \mathbb{R})$, with $m > n$ (tall, thin A). We *define* $\kappa(A) = \frac{\sigma_1}{\sigma_n}$. Note that we cannot use the other definition $\kappa(A) = \|A\| \, \|A^{-1}\|$, since A^{-1} does not exist for a non-square A. However, one can verify that $\|A\| \, \|A^{+}\| = \frac{\sigma_1}{\sigma_n} = \kappa(A)$, where A^{+} is the pseudoinverse.

  Observation 1.1. *Note that* $\frac{1}{\kappa(A)}$ *is the relative distance to the closest* $\hat{A}$ *without full column rank.*

- θ The second quantity needed is the angle between Ax and b, see Figure 1.6. $\theta = \arccos \frac{\|Ax\|}{\|b\|}$

Now we can express the theorem:

Theorem 1.3 (Trefethen, Bau)**.** *Consider the linear least squares problem* $\min \|Ax - b\|$, *with* $A \in \mathbb{R}^{m \times n}$ *with full column rank. Its relative condition number with respect to the input* b *is*

$$\kappa_{rel, b \to x} \leq \frac{\kappa(A)}{\cos \theta}$$

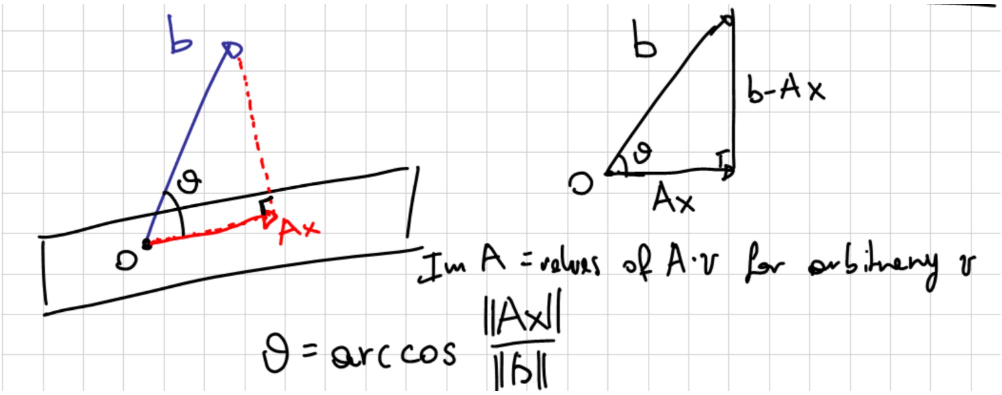
b
ϑ
Ax
O
b
b-Ax
ϑ
O
Ax
Im A = values of A·v for arbitrary v
ϑ = arccos ||Ax|| / ||b||

FIGURE 1.6: The triangle in the figure (the one whose cathets are Ax and $b - Ax$) is a square triangle.

and with respect to A it is

$$\kappa_{rel,A\to x} \leq \kappa(A) + \kappa(A)^2 \tan\theta$$

where θ is the angle such that $\cos\theta = \frac{\|Ax\|}{\|b\|}$.

At this point we have two condition numbers and they both depend on $\kappa(A)$ and θ.

Observation 1.2.

SPECIAL CASE 1: $\theta \approx 90°$ *We can see from the figure that a big change of b induces a small perturbation of Ax. No matter what the conditioning of A is, a small (relative) perturbation in b can change a large (relative) perturbation in x and Ax, see Figure 1.7.*

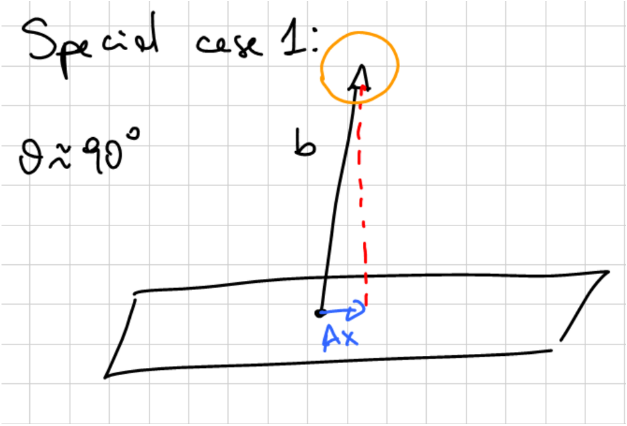

FIGURE 1.7: Special case 1

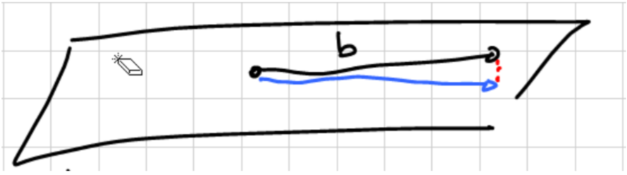

Figure 1.8: Special case 2.

Special case 2: $\theta \approx 0°$ *When b is almost in plane with $Im(x)$. In this case cond $\approx \kappa(A)$, see Figure 1.8.*

General case: θ far from $0°$ and $90°$ *In the more general case, cond $\approx \kappa(A)^2$.*