LINEAR ALGEBRA, NUMERICS

MINSEOK SONG

1. Determinant

Begin by visualizing an $n \times n$ matrix as a linear transformation. In mathematics, it's often enlightening to view objects not just for what they are, but for the roles they play—in this case, as functions. Through this lens, we can perceive the determinant as a function: it ingests n column vectors from $\mathbb{R}^n$ present in the matrix and produces a real number. But what does this number represent? At its essence, the determinant can be seen as an indicator of oriented volume.

(1) **Sign Inversion**: Interchanging rows (or columns) of the matrix inverts the sign of the determinant.
(2) **Linearity**: The determinant is linear in relation to each column and row.
(3) **Identity Matrix**: The determinant of the identity matrix is 1.

With this understanding, we recognize the inherent logic in the definition of the determinant. Moreover, when extended to continuous domains, this understanding paves the way to the concept of the wedge product—an essential tool for generalizing integration, which is fundamentally about calculating the "volume" of more complex structures, often referred to as manifolds.

We define

$$f \wedge g = \frac{1}{k!l!} A(f \otimes g)$$

where

$$Af = \sum_{\sigma \in S_k} (\text{sgn } \sigma) \sigma f$$

and

$$f \otimes g(v_1, \ldots, v_{k+l}) = f(v_1, \ldots, v_k) g(v_{k+1}, \ldots, v_{k+l}).$$

It follows that

$$(\alpha^1 \wedge \cdots \wedge \alpha^k)(v_1, \ldots, v_k) = det[\alpha^i(v_j)]$$

The formulation of $f \wedge g$ is meticulously designed to encapsulate the inherent attributes of the determinant. Specifically:

(1) The anticommutative nature is reflected in the property 1.
(2) The linearity is mirrored in property 2.
(3) The normalization constant $\frac{1}{k!l!}$ embodies property 3.

Remark 1.
- The above characterizations intuitively and rigorously (by simply checking that $\det(A)\det(B)$ satisfies three characterizations) demonstrate why $\det(AB) = \det(A) \times \det(B)$.
- From this, we can see that the determinant of orthonormal matrix is -1 or 1, and in turn that the determinant is a multiplication of all singular values by SVD.
- These singular values represent the extent of stretching in each of the n directions.
- The logarithm function translates multiplication into addition and possesses inherent concavity. As a result, for a positive definite matrix A, $\log\det(A)$ is concave. A more rigorous justification can be derived by verifying $g''(t) \leqslant 0$ for the function $g(t) = f(Z + tV)$ where $Z, V \in S^n$.

2. Pseudo Inverse, rank-r approximation

Definition 1. *Moore-Penrose pseudo-inverse for a matrix $A \in \mathbb{C}^{m\times n}$ is defined as a matrix $X \in \mathbb{C}^{n\times m}$ satisfying*

- $(AX)^* = AX$
- $(XA)^* = XA$
- $XAX = X$
- $AXA = A$

Remark 2.
- Uniqueness can be seen by computing SVD and inverse each singular value, which is the most obvious thing to do.
- Geometrically, this is a least squares problem (L^2 norm): there exists a unique vector x such that Ax is closest to b.
- We can use different pseudo-inverse by using different norm, say L^∞ norm.
- Related concept is rank-r approximation in Young-Eckart theorem. The key of the proof is by exploring the space spanned by the right singular vectors $v_1, \dots, v_{r+1}$ (and connecting it to kernel of B) and deriving contradiction from that (for example by dimensionality argument).
- Eckart-Young holds for any unitarily invariant norm. Intuitively, unitarily invariant norm is the one that "respects" the decomposition of SVD: proof can be found here: `https://cklixx.people.wm.edu/uinorm-note.pdf`

3. Condition number

Definition 2. *The absolute condition number of a problem f is*
$\lim_{\epsilon\to 0^+} \sup_{\|\delta x\|\leqslant\epsilon} \frac{\frac{\|\delta f(x)\|}{\|f(x)\|}}{\frac{\|\delta x\|}{\|x\|}}$

- Consider the invertible matrix A. Say $f(x) := A^{-1}x$. The condition number will be $\|A^{-1}\|\frac{\|x\|}{\|A^{-1}x\|}$. To eliminate the dependency, we consider the "worst case" so that we get $\|A\|\|A^{-1}\|$.
- important to note that condition number may depend on x. Unless f is continous, there is no reason to believe that the condition number is a constant.

$*$ We may illustrate some relationship with relative error.

- Assume that A is a nonsingular matrix, and let $0 \neq b \in \mathbb{R}^n$. Assume that $\|Ax\| \leqslant \|A\|\|x\|$ for all $A \in \mathbb{R}^{n\times n}$ and all $x \in \mathbb{R}^n$.
- Suppose $(A + \Delta A)(\mathbf{x} + \Delta\mathbf{x}) = \mathbf{b}$ and $\hat{\mathbf{x}} = \mathbf{x} + \Delta\mathbf{x}$. We have that
$$\frac{\|\Delta\mathbf{x}\|}{\|\hat{\mathbf{x}}\|} \leqslant \kappa(A)\frac{\|\Delta A\|}{\|A\|}. \tag{1}$$
- Suppose $(A + \Delta A)(\mathbf{x} + \Delta\mathbf{x}) = \mathbf{b}$ and $\hat{\mathbf{x}} = \mathbf{x} + \Delta\mathbf{x}$. Further, suppose that $\frac{\|\Delta A\|}{\|A\|} < \frac{1}{\kappa(A)}$. We have
$$\frac{\|\Delta\mathbf{x}\|}{\|\mathbf{x}\|} \leqslant \frac{\kappa(A)\frac{\|\Delta A\|}{\|A\|}}{1 - \kappa(A)\frac{\|\Delta A\|}{\|A\|}}. \tag{2}$$
- Suppose $(A + \Delta A)\hat{\mathbf{x}} = \mathbf{b} + \Delta\mathbf{b}$ where $\hat{\mathbf{b}} = \mathbf{b} + \Delta\mathbf{b} \neq \mathbf{0}$ and $\hat{\mathbf{x}} = \mathbf{x} + \Delta\mathbf{x} \neq \mathbf{0}$. We have
$$\frac{\|\Delta\mathbf{x}\|}{\|\hat{\mathbf{x}}\|} \leqslant \kappa(A)\left(\frac{\|\Delta A\|}{\|A\|} + \frac{\|\Delta\mathbf{b}\|}{\|\hat{\mathbf{b}}\|} + \frac{\|\Delta A\|}{\|A\|}\frac{\|\Delta\mathbf{b}\|}{\|\hat{\mathbf{b}}\|}\right). \tag{3}$$

- Suppose $(A+\Delta A)\hat{\mathbf{x}} = \mathbf{b}+\Delta\mathbf{b}$ where $\hat{\mathbf{b}} = \mathbf{b}+\Delta\mathbf{b} \neq \mathbf{0}$ and $\hat{\mathbf{x}} = \mathbf{x}+\Delta\mathbf{x} \neq \mathbf{0}$. Further suppose that $\frac{\|\Delta A\|}{\|A\|} < \frac{1}{\kappa(A)}$. We then have that

$$\frac{\|\Delta\mathbf{x}\|}{\|\mathbf{x}\|} \leqslant \frac{\kappa(A)\left(\frac{\|\Delta A\|}{\|A\|}+\frac{\|\Delta\mathbf{b}\|}{\|\mathbf{b}\|}\right)}{1-\kappa(A)\frac{\|\Delta A\|}{\|A\|}}. \tag{4}$$

- As expected, condition number and the relative error of the matrix control the bound of relative error of x. The smaller, the better.
- The condition involving $\kappa(A)$ is an upper bound of condition number.

4. Floating Point Arithmetic

The representation of floating-point numbers in many modern systems follows the IEEE 754 standard. This standard provides a consistent methodology to represent and manipulate real numbers on digital computing systems.

Definition 3. *Given a floating point representation, x, of the form:*

$$x : \pm a_1 a_2 \dots a_{11} b_1 b_2 \dots b_{52}$$

We can interpret x based on its bit patterns:

- *For **normalized** values, when $a_1, \dots, a_{11}$ are not all 0's nor all 1's:*

$$x = \pm(1.b_1 b_2 \dots b_{52})_2 \times 2^{(a_1 a_2 \dots a_{11})_2 - 1023}$$

- *For **subnormal** values when $a_1, \dots, a_{11}$ are all 0's:*

$$x = \pm(0.b_1 b_2 \dots b_{52})_2 \times 2^{(a_1 a_2 \dots a_{11})_2 - 1022}$$

(1) The usage of subnormal is to prevent underflow.
(2) This is an excellent example of engineering design to represent real numbers given a fixed number of bits.
(3) The subnormal value is always less than or equalt o 2^{-1022}, and the normalized value if greater than or equal to that, justifying such a choice.

Definition 4. *The machine epsilon is the gap between 1 and the next largest floating point number. It is $2^{-52} \approx 10^{-16}$ for the double format.*

$*$ *It follows that $|round(x) - x| \leqslant \frac{2^{-52}}{2} \cdot 2^e \leqslant \frac{2^{-52}}{2}|x|$.*

Definition 5. *Largely, we may use two kinds of error; absolute error, $\epsilon_{abs} = \|x - \hat{x}\|$ or $\epsilon_{rel} = \frac{\|x-\hat{x}\|}{\|x\|}$.*

- One might ask, why would we use x instead of $\hat{x}$ in the denominator. This is purely due to interpretability (just more natural).

On fl notation. How do we round? Well, we do

$$\text{round}(x) = \begin{cases} x_- & \text{if } d_- < d_+ \text{ or } (d_- = d_+ \text{ and } \text{lsb}(x_-) = 0) \\ x_+ & \text{if } d_+ < d_- \text{ or } (d_+ = d_- \text{ and } \text{lsb}(x_+) = 0) \end{cases}$$

- We can check that when $x \in [N_{min}, N_{max}]$, we have $\epsilon_{rel} \leqslant \epsilon_{machine}$.
- Trefethen and Bau generalize this in the case $x = \pm(m/\beta^t)\beta^e$ where $\beta^{t-1} \leqslant m \leqslant \beta^t - 1$, in which case we have the machine epsilon $\frac{1}{2}\beta^{1-t}$ (this is an idealized case, since we do not have any restriction on the size of bits).

5. Matrix decomposition

5.1. LU decomposition.