LINEAR ALGEBRA, NUMERICS

MINSEOK SONG

Contents

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 $Z, V \in S^n$.

Fact 1. $A^{-1} = \frac{1}{det(A)} adj(A)$, with $adj(A)_{ij} = (-1)^{j+i} det(A_{-j,-i})$

Proof. We can simply check that the above three properties hold. The multiplication involves pairing the rows of A with the columns of $adj(A)$. This is precisely why the adjugate is formed with transposed cofactors. Further, excluding (i, j) index in the definition enables linearity. □

2. Pseudo Inverse, rank

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*

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

- Uniqueness can be seen by computing SVD and inverse each singular value, which is canonical.
- Geometrically, this is a least squares problem (L^2 norm): there exists a unique vector x such that Ax is closest to b.
- There is a generalization using different norm, for example L^∞.

Theorem 1 (Eckhart-Young)**.** *Let $A = \sum_{i=1}^{rank(A)} \sigma_i u_i v_i^*, \sigma_1 \geqslant \cdots \geqslant \sigma_{rank(A)} > 0$. Then* $\arg_X \min_{rank(X)\leqslant r} \|A - X\|_2 = \sum_{i=1}^{r} \sigma_i u_i v_i^*$. *Furthermore, we have*

$$\min_{rank(X)\leqslant r} \|A - X\|_2 = \sigma_{r+1}$$

Proof. The key of the proof is by exploring the space spanned by the right singular vectors $v_1, \ldots, v_{r+1}$ (and connecting it to kernel of B) and deriving contradiction from that (for example by dimensionality argument). □

Remark 2. 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`

Fact 2. The dimension of row vector and column vector is same.

Proof. Decompose $A = BC$ where B (m by k) consists of basis of column vectors and C (k by n) consists of coefficient for each column vector of A. On the other hand, we can view each row of A as linear combination of row of C with B giving the coefficients. This implies that $dim(rowsp(A)) \leqslant dim(colsp(A))$. Apply the same principle to A^T and we get the desired equality. □

Fact 3. $dim(AB) \leqslant \min(dim(A), dim(B))$

Fact 4. For Hermitian A, we have $\max\|\langle Ax, x\rangle\| = \|A\|_2$.

- This comes from the fact that for Hermitian matrix, eigenvalue and singular value coincide, so when the maximum is achieved, x and Ax will have the same direction.

3. Condition number

Definition 2. *The absolute condition number of a problem f is*

$$\kappa(f) = \lim_{\epsilon \to 0^+} \sup_{\|\delta x\| \leqslant \epsilon} \frac{\|\delta f(x)\|/\|f(x)\|}{\|\delta x\|/\|x\|}.$$

- 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 rule of thumb here is fordward error $\lesssim$ condition number $\times$ backward error.
- 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 to 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).

4.1. 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. SVD decomposition.

- The proof is brief and illustrative.

Proof. (1) Apply spectral theorem on AA^T. Then we get $A^TA = V\Lambda V^T = \sum_{i=1}^{n} (\sigma_i)^2 v_i v_i^T$.
(2) It follows that $A^TAv_i = (\sigma_i)^2 v_i$.
(3) Put $u_i = \dfrac{Av_i}{\sigma_i}$.
(4) By construction, we have $U = AV\Sigma^{-1}$.

□

- The choice of u_i comes from the fact that we know it is an eigenvector of AA^T, so it has to be of form Av_i, and σ_i is just a normalizing factor.
- Time complexity is $O(mn^2)$ where $n < m$. Truncated SVD where for getting the first k singular values takes $O(mnk)$.
- In practice, we would prefer randomized SVD.
- Without any duplication of singular value, the corresponding eigenvector is unique upto complex number multiplication... now if we have any duplicates, it's unique upto rotation. The following illustrates this point.

$$u_1\sigma_1 v_1^T + u_2\sigma_1 v_2^T = \begin{pmatrix} u_1 & u_2 \end{pmatrix} \begin{pmatrix} \sigma & \\ & \sigma \end{pmatrix} \begin{pmatrix} v_1^T \\ v_2^T \end{pmatrix} = \sigma \begin{pmatrix} u_1 & u_2 \end{pmatrix} \begin{pmatrix} v_1^T \\ v_2^T \end{pmatrix} = \sigma \begin{pmatrix} u_1 & u_2 \end{pmatrix} QQ^T \begin{pmatrix} v_1^T \\ v_2^T \end{pmatrix}$$

5.2. LU decomposition.

- We decompose by doing Gaussian elimination, resulting in $A = \Pi LU$.
- The Π factor is due to row swapping. We could also do $\Pi AQ = LU$ where Q corresponds to column swapping.
- Indeed these swappings are purely due to numerical consideration (roundoff error). In general we do not want pivoting element to be small.
- When Π is an identity matrix, we get $A = LU$.

5.3. LDU decomposition.

- From LU, we make the diagonal elements of L and U all 1's.
- This makes it easier for us to take inverse of LU.

5.4. Schur decomposition.

- Refer to here.
- The idea is that we can use "block" LDU decomposition.

5.5. Cholesky factorization.

- Cholesky factorization gives $A = LL^T$ for Hermitian, positive-definite matrix.
- This is somewhat analogous to $x = (\sqrt{x})^2$.
- Uniqueness is guaranteed for positive definite matrix. However, there is no uniqueness for positive semidefinite.
- For example, for $\begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}$, we could have set $L = \begin{pmatrix} 0 & 0 \\ \cos\theta & \sin\theta \end{pmatrix}$ or $L = \begin{pmatrix} 0 & 0 \\ \sin\theta & \cos\theta \end{pmatrix}$
- We could've gotten an algorithm to find F with $A = FF^T$, that is (from lec13 note by),

$$f_{kk} == (a_{kk} - \sum_{j=1}^{k-1} f_{kj}^2)^{1/2}, f_{ik} = \frac{a_{ik} - \sum_{j=1}^{k-1} f_{kj} f_{ij}}{f_{kk}}, i = k+1, \ldots, n$$

- The positiveness of the term inside comes from positive definiteness of the original matrix (could be done via induction).

5.6. **QR decomposition.**

- Poor man's SVD
- In the $A = QR$ decomposition, the matrix Q is an orthogonal matrix, and the spanning set of the first k leading columns of K is same as the space spanning by columns of A.
- In case A is $m \times n$ dimensional and A has rank r, we can decompose
  (1) $m \times r$, $r \times n$ (this is a reduced form capturing the essence of A in terms of its rank).
  (2) $m \times n$, $n \times n$ (this is referred to as thin or reduced QR decomposition, derived from the next kind).
  (3) $m \times m$, $m \times n$ (this is the full QR decomposition).
- We also have $A\Pi = Q \begin{pmatrix} R_1 & S \\ 0 & 0 \end{pmatrix}$ and in addition $A = Q \begin{pmatrix} L & 0 \\ 0 & 0 \end{pmatrix} U^*$ where U and Q are orthogonal, by taking $\begin{pmatrix} R_1^* \\ S^* \end{pmatrix} = Z \begin{pmatrix} R_2 \\ 0 \end{pmatrix}$
- The detail is canonical.
- There are several ways to compute: Gram-Schmidt process, Householder transformations, and Givens rotations.

5.7. **Jordan Canonical form.**

- We have
$$T^{-1}AT = J = \begin{pmatrix} J_1 & & & 0 \\ & \ddots & & \vdots \\ & & J_k & \\ 0 & \cdots & & J_n \end{pmatrix}$$
where
$$J_i = \begin{pmatrix} \lambda_i & 1 & & & 0 \\ & \lambda_i & 1 & & \vdots \\ & & \ddots & \ddots & \\ & & & \lambda_i & 1 \\ 0 & \cdots & & & \lambda_i \end{pmatrix} \in \mathbb{C}^{n_i \times n_i}$$
called Jordan block.
- Jordan form is unique upto permutations of the blocks.

5.8. **CUR decomposition.**

- Also called interpolative decomposition.
- The main advantage of this decomposition is interpretability, since it uses the column from actual data matrix.
- First choose k columns and rows, denoted by C and R. Calculate $U = \tilde{U}^+$, and we can approximate as $A = CUR$.
- However, this is not suited for sparse matrix.

6. Tensor decomposition

7. Error analysis, stability

Definition 6. *(1) Backward error is the smallest Δx such that $f(x + \Delta x) = y^*$.*
(2) Forward error is $y^* - y$.
(3) Truncation error is the error incurred by Taylor polynomial.

- *Some intuition of why Taylor polynomial might not converge: the condition for which the error decreases as dimension increases is*

$$|x - x_0| < (n+1)|\frac{f^{(n)}(\xi_{n-1})}{f^{(n+1)}(\xi_n)}$$

but this might well limit to zero. One example given frequently is e^{-1/x^2} *around* 0 *and* 0 *on* 0*, which grows very slowly around* 0 *(slower than any polynomial).*

(4) Roundoff error is the error caused by floating point approximation.

- Backward error is said to be stable if $|\frac{\Delta x}{x}|$ is small for every input x.
- This is not practical; what if y^* is not even in the range of f?
- An algorithm is said to be numerically stable (mixed forward-backward stability) if for any $x \in X$, the computed $\hat{y}$ satisfies

$$\hat{y} + \Delta y = f(x + \Delta x), |\Delta x| \leqslant \delta|x|, |\Delta y| \leqslant \epsilon|y|.$$

for small $\delta, \epsilon > 0$.
- This says that our computed value is almost the right answer for almost the right data.
- In machine learning framework, we often speak of approximation, integration, and optimization error. Approximation error is a distance between true value and an output of hypothesis class. Integration error is a distance between the output of hypothesis class and the output of ERM. Optimization error is a distance between the output of ERM and the hypothesis from of our choice of algorithm.

8. Norms

8.1. **Nuclear norm.**

- $\|A\|_* = \sup_{\sigma_1(Q) \leqslant 1} \langle Q, A \rangle$: this is like saying that the dual norm of spectral norm is nuclear norm.
- Using this, we can prove that the nuclear norm is actually a norm.

9. miscellaneous facts

Fact 5. determinant of A is equal to the product of eigenvalues.

Proof. Note that $det(A - \lambda I) = \Pi_i(\lambda_i - \lambda)$ and plug in $\lambda = 0$. □