Linear Algebra In Machine Learning

Chunpai Wang

1 Norm

A function $f : \mathbb{R}^n \to \mathbb{R}$ is a norm if

1. $f(x) \geq 0, \forall x \in \mathbb{R}^n, f(x) = 0 \Leftrightarrow x = 0$ (Non-negativity)
2. $f(\alpha x) = |\alpha| f(x), \forall x \in \mathbb{R}$ (Homogeneity)
3. $f(x + y) \leq f(x) + f(y)$ (Triangle Inequality)

1. Prove that $\|x\|_P = \sqrt{x^\top P x}$ where $P \succ 0$ define a norm.

Proof. We need to show the non-negativity, homogeneous, and triangle inequality. Since P is positive definite, $x^\top P x > 0$, thus non-negativity holds.

$$\|\alpha x\|_P = \sqrt{\alpha x^\top P \alpha x} = \alpha \sqrt{x^\top P x} = \alpha \|x\|_P$$

, thus homogeneity holds.

$$\begin{aligned} \|x + y\|_P &= \sqrt{(x + y)^\top P (x + y)} = \sqrt{(x + y)^\top P^{1/2} P^{1/2} (x + y)} \\ &= \|(x + y) P^{1/2}\| = \|x P^{1/2} + y P^{1/2}\|, \end{aligned} \tag{1}$$

since $\|x P^{1/2} + y P^{1/2}\|$ is a vector norm, we have

$$\|x P^{1/2} + y P^{1/2}\| \leq \|x P^{1/2}\| + \|y P^{1/2}\| = \sqrt{x^\top P x} + \sqrt{y^\top P y},$$

we get triangle inequality holds. □

2. Prove that Ellipsoid is a convex set, where Ellipsoid is defined as

$$S = \left\{ x | \sqrt{(x - x_c)^\top P (x - x_c)} \leq r \right\} \tag{2}$$

where $x_c \in \mathbb{R}^n, r \in \mathbb{R}_+, P \succ 0$.

Proof. Assume $x, y \in S$, that is

$$\begin{aligned} \sqrt{(x - x_c)^\top P (x - x_c)} \leq r \\ \sqrt{(y - x_c)^\top P (y - x_c)} \leq r \end{aligned} \tag{3}$$

which can be also represented as

$$\begin{aligned} \|(x - x_c)\|_P \leq r \\ \|(y - x_c)\|_P \leq r \end{aligned} \tag{4}$$

we will show that $\lambda x + (1 - \lambda) y \in S$ with $\lambda \in [0, 1]$, that is

$$\|(\lambda x + (1 - \lambda) y - x_c)\|_P \quad \leq r \tag{5}$$

$$
\begin{aligned}
\|(\lambda x + (1-\lambda)y - x_c)\|_P &= \|(\lambda x + (1-\lambda)y - \lambda x_c - (1-\lambda)x_c\|_P \\
&= \|(\lambda(x - x_c) + (1-\lambda)(y - x_c)\|_P \\
&\le \|(\lambda(x - x_c)\|_P + \|(1-\lambda)(y - x_c)\|_P \\
&= \lambda\|(x - x_c)\|_P + (1-\lambda)\|(y - x_c)\|_P \\
&\le \lambda r + (1-\lambda)r \\
&= r
\end{aligned} \tag{6}
$$

Therefore, Ellipsoid is convex. □

2 "Entrywise" Matrix Norms

These norms treat an $m \times n$ matrix as a vector of size mn, and use one of the familiar vector norms. For example,

$$\|A\|_p = \|vec(A)\|_p = (\sum_{i=1}^{m}\sum_{j=1}^{n}|a_{ij}|^2)^{1/2}$$

Let $X \in \mathbb{R}^{m\times n}$. Then

- $\|X\|_F = \sqrt{Tr(X^\top X)} = (\sum_{i=1}^{m}\sum_{j=1}^{n}X_{ij}^2)^{1/2}$
- $\|X\|_{max} = \max_{i,j}|X_{ij}|$

2.1 $L_{2,1}$ Norm

Let $(a_1, ..., a_n)$ be the columns of matrix A. The $L_{2,1}$ norm is the sum of the Euclidean norms of the columns of the matrix:

$$\|A\|_{2,1} = \sum_{j=1}^{n}\|a_j\|_2 = \sum_{j=1}^{n}(\sum_{i=1}^{m}|a_{ij}|^2)^{1/2}$$

The $L_{2,1}$ norm as an error function is more robust since the error for each data point (a column) is not squared. It is used in robust data analysis and sparse coding.

3 Operator Norm or Induced Matrix Norm

Let $\|\cdot\|_a, \|\cdot\|_b$ be norms on space $\mathbb{R}^m$ and $\mathbb{R}^n$, respectively. We define the operator norm on $A \in \mathbb{R}^{m\times n}$ as

$$\|A\|_{a,b} = \max\|Ax\|_a \quad s.t. \|x\|_b \le 1$$

or we can define it as

$$\|A\|_{a,b} = \sup_{\|x\|_b \ne 0}\frac{\|Ax\|_a}{\|x\|_b}$$

We need to verify operator norm is indeed a norm in terms of nonnegativity, homogeneity, and triangle inequality. Operator norm also called induced matrix norm. Note that,

- $\|A\|_a = \|A\|_{a,a}$ the same vector norm in the two same space
- $\|A\|_2 = \|A\|_{2,2} = \sqrt{\lambda_{max}(A^\top A)}$

Proof.

$$\begin{aligned}
\|A\|_2 &= \max \|Ax\|_2 \quad \text{s.t.} \quad \|x\|_2 \leq 1 && (7)\\
&= \max \sqrt{(Ax)^\top Ax} \quad \text{s.t.} \quad \|x\|_2 \leq 1 && (8)\\
&= \max_{\|x\|_2 \leq 1} \sqrt{x A^\top A x} && (9)\\
&\text{Since we have } \lambda_{\min}(A) x^\top x \leq x^\top A x \leq \lambda_{\max}(A) x^\top x && (10)\\
&\leq \max \sqrt{\lambda_{\max}(A^\top A) x^\top x} \quad \text{s.t.} \quad \|x\|_2 \leq 1 && (11)\\
&\leq \sqrt{\lambda_{\max}(A^\top A)} && (12)
\end{aligned}$$

$A^\top A$ is symmetric matrix, thus always positive semidefinite. We can see that when $\|x\|_2 = 1$, it achieves the maximum, which is $\sqrt{\lambda_{\max}(A^\top A)}$. ☐

- $\|A\|_1 = \|A\|_{1,1} = \max_j \sum_i |A_{ij}|$ (maximum column sum)

Proof.

$$\begin{aligned}
\|A\|_{1,1} &= \max_{\|x\|_1 \leq 1} \|Ax\|_1 && (13)\\
&= \max_{\|x\|_1 \leq 1} \sum_{i=1}^{m} |\sum_{j=1}^{n} A_{ij} x_j| && (14)\\
&\leq \max_{\|x\|_1 \leq 1} \sum_{i=1}^{m} \sum_{j=1}^{n} |A_{ij} x_j| && (15)\\
&= \max_{\|x\|_1 \leq 1} \sum_{j=1}^{n} \left[|x_j| \sum_{i=1}^{m} |A_{ij}| \right] && (16)\\
&\leq (\max \sum_{i=1}^{m} |A_{ij}|)(\sum_{j=1} |x_j|) \quad \text{s.t.} \quad \|x\|_1 \leq 1 && (17)\\
&\leq \max_j \sum_{i=1}^{m} |A_{ij}| && (18)
\end{aligned}$$

☐

- $\|A\|_\infty = \max_i \sum_j |A_{ij}|$ (maximum row sum)

Proof.

$$\begin{aligned}
\|A\|_\infty &= \max_{\|x\|_\infty \leq 1} \|Ax\|_\infty && (19)\\
&= \max_{\|x\|_\infty \leq 1} (\max |A_{ij} x_j|) && (20)\\
&\leq (\max_j |x_j|)(\max_i \sum_{j=1}^{n} |A_{ij}|) && (21)\\
&= \|x\|_\infty (\max_i \sum_{j=1}^{n} |A_{ij}|) \quad \text{s.t.} \quad \|x\|_\infty \leq 1 && (22)\\
&\leq \max_i \sum_{j=1}^{n} |A_{ij}| && (23)
\end{aligned}$$

The maximum is achieved at $x_j = sgn(A_{ij})$ ☐

3.1 Submultiplicative Property

$$\|AB\| \leq \|A\|\|B\| \qquad (submultiplicative)$$

Proof.

$$\|AB\| = \max_{x\neq 0} \frac{\|ABx\|}{\|x\|} = \max_{Bx\neq 0} \frac{\|ABx\|}{\|Bx\|}\frac{\|Bx\|}{\|x\|} \leq \max_{y\neq 0} \frac{\|Ay\|}{\|y\|} \max_{x\neq 0} \frac{\|Bx\|}{\|x\|} = \|A\|\|B\|$$

Or we can prove it in this way. First, we will show the property

$$\|Ax\| \leq \|A\|\|x\|.$$

Note here, $\|A\|$ is operator norm, and $\|x\|$ is vector norm. Proof by contradiction, we assume $\|Ax\| > \|A\|\|x\|$, and since $\|x\|$ is scalar and by homogeneity, we have

$$\frac{\|Ax\|}{\|x\|} > \|A\|$$

which contradict the definition of $\|A\| = \max_{x\neq 0} \frac{\|Ax\|}{\|x\|}$. Now, by definition of operator norm we have

$$\begin{aligned}
\|AB\| &= \max \|ABx\| \quad s.t. \quad \|x\| \leq 1 \\
&\leq \max \|A\|\|Bx\| \quad s.t. \quad \|x\| \leq 1 \quad (\|Ax\| \leq \|A\|\|x\|) \\
&= \|A\| \max \|Bx\| \quad s.t. \quad \|x\| \leq 1 \\
&= \|A\|\|B\|
\end{aligned}$$

□

4 Dual Norm

Let $\|x\|$ be any norm, and its dual norm is defined as

$$\|x\|_* = \max_y x^\top y \quad s.t. \quad \|y\| \leq 1$$

or

$$\|x\|_* = \max_{\|y\|\leq 1} x^\top y$$

- $\|x\|_{\infty,*} = \|x\|_1$
- $\|x\|_{2,*} = \|x\|_2$
- $\|x\|_{1,*} = \|x\|_\infty$

4.1 Property

If $\|x\|$ is a norm and $\|x\|_*$ is its dual norm, then $\|z^\top x\| \leq \|z\|\|x\|_*$

5 Schatten Norm

Let $A \in \mathbb{R}^{m\times n}$

$$\|A\|_p = (\sum_{i=1}^{min\{m,n\}} \sigma_i^p(A))^{1/p}$$

We can see that it shares the notation with operator norm and l-p norm, but they are different. We can interpret the schatten norm as the l-p norm on the vector of singular values of matrix.

- $p = 0 : \|A\|_p = rank(A)$

- $p = 1 : \|A\|_p = \sum_i^r \sigma_i(A) = \|A\|_*$ where r is the rank(A), and the $\|A\|_*$ is called **nuclear norm**.
- $p = 2 : \|A\|_p = \sqrt{\sum_{i=1}^r \sigma_i^2} = \sqrt{tr(A^\top A)} = \|A\|_F$ Note that, Frobenius norm is a element-wise matrix norm.
- $p = \infty : \|A\|_p = \sigma_1(A) = \sqrt{\lambda_{max}(A^\top A)} = \|A\|_{2,2} = \|A\|_2$, the $\|A\|_2$ is usually called **spectral norm**.

5.1 Theorem 1

The nuclear norm of a matrix is the dual norm of the its spectral norm.

Proof. Let $A, E \in \mathbb{R}^{m\times n}$, the dual norm of the spectral norm of matrix A is defined as

$$\|A\|_{2,*} = \max_{\|E\|_2 \le 1} \langle E, A\rangle$$

First, we will show that the dual norm of spectral norm is less or equal than the nuclear norm, that is

$$\max_{\|E\|_2 \le 1} \langle E, A\rangle \le \|A\|_*$$

$$\begin{aligned}
\max_{\|E\|_2 \le 1} \langle E, A\rangle &= \max_{\sigma_1(E) \le 1} \langle E, A\rangle \\
&= \max_{\sigma_1(E) \le 1} \langle E, U\Sigma V^\top\rangle \\
&= \max_{\sigma_1(E) \le 1} tr(E^\top U\Sigma V^\top) \\
&= \max_{\sigma_1(E) \le 1} tr(V^\top E^\top U\Sigma) \\
&= \max_{\sigma_1(E) \le 1} \sum_{i=1}^r \sigma_i(V^\top E^\top U\Sigma) \\
&= \max_{\sigma_1(E) \le 1} \sum_{i=1}^r \sigma_i(\langle U^\top EV, \Sigma\rangle) \\
&= \max_{\sigma_1(E) \le 1} \sum_{i=1}^r \sigma_i(A) \cdot \sigma_i(U^\top EV) \\
&= \max_{\sigma_1(E) \le 1} \sum_{i=1}^r \sigma_i(A) \cdot (U^\top EV)_{ii} \\
&\le \max_{\sigma_1(E) \le 1} \sum_{i=1}^r \sigma_i(A) \cdot (U^\top EV)_{11} \\
&= \max_{\sigma_1(E) \le 1} \sigma_1(E) \cdot \sum_{i=1}^r \sigma_i(A) \\
&= 1\ cdot \sum_{i=1}^r \sigma_i(A) \\
&= \|A\|_*
\end{aligned}$$

Now, we will show that the dual norm of spectral is greater than or equal to the nuclear norm, that is

$$\max_{\|E\|_2 \le 1} \langle E, A\rangle \ge \|A\|_*$$

$$
\begin{aligned}
\max_{\|E\|_2 \le 1} \langle E, A \rangle &= \max_{\|E\|_2 \le 1} tr(E^\top A) \\
&= \max_{\|E\|_2 \le 1} tr(E^\top U \Sigma V^\top) \\
&\ge tr(\underbrace{V I^\top U^\top}_{E^\top} \underbrace{U^\top U \Sigma}_{A}) \quad \text{because we restrict } E \\
&= tr(V^\top V I^\top U^\top U \Sigma) \quad \text{because} \;\; tr(AB) = tr(BA) \\
&= tr(\Sigma) \\
&= \sum_{i}^{\min\{m,n\}} \sigma_i \\
&= \|A\|_*
\end{aligned}
$$

Thus, we proved that

$$
\|A\|_{2,*} = \max_{\|E\|_2 \le 1} \langle E, A \rangle = \|A\|_*
$$

□

5.2 Theorem 2

The spectral norm of a matrix is the dual norm of the its nuclear norm.

References

[1] Appendix A.5.5 and C.4.3. Boyd, Stephen, and Lieven Vandenberghe. Convex optimization. Cambridge university press, 2004.

[2]