Matrix Computation I: Norms, Operators, and Numerical Size

N-20260311

§1.1 Why size matters

Matrix computation starts from a simple question: how large can an error become after a linear operation acts on it? A vector norm measures the size of data; a matrix norm measures the size of an operator. In machine learning language, these are the quantities behind amplification, Lipschitz constants, gradient growth, and numerical stability.

A matrix is an operator

$$x \longmapsto Ax.$$

To control it we need both

$$\text{vector size} \qquad \text{and} \qquad \text{operator size}.$$

§1.2 Metrics and completeness

A metric on a set X is a function $d : X \times X \to [0, \infty)$ satisfying positivity, symmetry, and the triangle inequality. It defines convergence. A sequence is Cauchy if

$$\forall \varepsilon > 0,\ \exists N, \quad m, n > N \Rightarrow d(x_m, x_n) < \varepsilon.$$

A metric space is complete if every Cauchy sequence converges in it. Completeness means that limiting procedures stay inside the space.

The discrete metric is a useful warning: not every metric expresses the geometry we want. For numerical work, metrics induced by norms are the central examples.

§1.3 Normed vector spaces

A norm satisfies

$$\|x\| = 0 \Leftrightarrow x = 0, \qquad \|\alpha x\| = |\alpha|\|x\|, \qquad \|x + y\| \le \|x\| + \|y\|.$$

It induces the metric

$$d(x, y) = \|x - y\|.$$

A complete normed vector space is a Banach space.

In finite dimension, all norms are equivalent: for two norms $\|\cdot\|_a, \|\cdot\|_b$, there exist constants $c, C > 0$ such that

$$c\|x\|_a \le \|x\|_b \le C\|x\|_a.$$

Thus finite-dimensional convergence does not depend on the chosen norm, although numerical constants do.

§1.4 Standard vector norms

For $x = (x_1, \dots, x_n)^T$,

$$\|x\|_1 = \sum_i |x_i|, \qquad \|x\|_2 = \left(\sum_i |x_i|^2 \right)^{1/2}, \qquad \|x\|_\infty = \max_i |x_i|.$$

More generally,

$$\|x\|_p = \left(\sum_i |x_i|^p \right)^{1/p}, \qquad 1 \le p < \infty.$$

As $p \to \infty$,

$$\|x\|_p \to \|x\|_\infty.$$

Holder's inequality gives

$$\sum_i |x_i y_i| \le \|x\|_p \|y\|_q, \qquad \frac{1}{p} + \frac{1}{q} = 1,$$

and Minkowski's inequality gives the triangle inequality for p-norms.

§1.5 Matrix norms

A matrix norm satisfies the norm axioms and, for computation, should also be submultiplicative:

$$\|AB\| \le \|A\| \|B\|.$$

This is what allows repeated operations to be bounded:

$$\|A^k\| \le \|A\|^k.$$

Entrywise examples include

$$m_1(A) = \sum_{i,j} |a_{ij}|, \qquad \|A\|_F = \left(\sum_{i,j} |a_{ij}|^2 \right)^{1/2}.$$

The Frobenius norm also satisfies

$$\|A\|_F^2 = \operatorname{tr}(A^H A).$$

§1.6 Unitary invariance

A matrix U is unitary if $U^H U = I$. It preserves Euclidean length:

$$\|Ux\|_2 = \|x\|_2.$$

The Frobenius norm is unitarily invariant:

$$\|UAV\|_F = \|A\|_F$$

for unitary U, V. This follows from cyclic invariance of trace:

$$\|UAV\|_F^2 = \operatorname{tr}(V^H A^H A V) = \operatorname{tr}(A^H A).$$

Unitary transformations are changes of orthonormal coordinates, not changes of size.

§1.7 Induced norms

Given vector norms on domain and codomain, the induced matrix norm is

$$\|A\| = \max_{x \neq 0} \frac{\|Ax\|}{\|x\|} = \max_{\|x\|=1} \|Ax\|.$$

It is the maximal amplification factor of A. The standard formulas are

$$\|A\|_1 = \max_j \sum_i |a_{ij}|, \qquad \|A\|_\infty = \max_i \sum_j |a_{ij}|,$$

and

$$\|A\|_2 = \sigma_{\max}(A) = \sqrt{\lambda_{\max}(A^H A)}.$$

The 2-norm is the spectral norm.

§1.8 Normal matrices

A matrix is Hermitian if $A^H = A$, skew-Hermitian if $A^H = -A$, and normal if

$$A^H A = A A^H.$$

Normal matrices are unitarily diagonalizable. If

$$A = U \Lambda U^H,$$

then

$$\|A\|_2 = \max_i |\lambda_i|.$$

For nonnormal matrices, eigenvalues may seriously underestimate transient amplification.

§1.9 Compatibility

A vector norm and a matrix norm are compatible if

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

Every induced matrix norm is compatible with the vector norm that induces it. Compatibility is the bridge from operator size to error propagation:

$$\|A(x + \delta x) - Ax\| \leq \|A\| \|\delta x\|.$$

§1.10 Worked constrained maxima

Let

$$A = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 2 \end{bmatrix}.$$

Then A is Hermitian with eigenvalues $2, 2, 0$. Hence

$$\|A\|_2 = 2.$$

The column sums are all 2, so $\|A\|_1 = 2$. Therefore

$$\max_{\|x\|_1=4} \|Ax\|_1 = 8.$$

If U is unitary and $\|Ux\|_2 = 3$, then $\|x\|_2 = 3$, hence

$$\max_{\|Ux\|_2=3} \|Ax\|_2 = 6.$$

The method is: translate the constraint into a norm ball and apply the induced norm.

§1.11 ML-facing interpretation

A linear layer has Lipschitz constant at most $\|W\|_2$ in Euclidean norm. A stack obeys

$$\|W_L \cdots W_1\|_2 \le \prod_{\ell=1}^{L} \|W_\ell\|_2.$$

This inequality is crude but important: moderate layerwise amplification can accumulate over depth. Spectral normalization, orthogonal initialization, residual scaling, and weight decay are all connected to the same numerical question: how large can a perturbation become?

§1.12 What each norm is measuring

object	what it measures	typical use
metric	distance between two points	convergence and continuity
vector norm	size of one vector	error magnitude
induced matrix norm	$\max_{x\neq 0} \|Ax\|/\|x\|$	worst-case amplification
Frobenius norm	$\sqrt{\sum_{ij} \|a_{ij}\|^2}$	entrywise energy
spectral norm	$\sigma_{\max}(A)$	Euclidean Lipschitz bound

The worked constrained-maximum examples explain the difference between Frobenius and spectral size. Frobenius norm counts all entries at once, while the spectral norm asks for the single input direction that is stretched the most. In the linear-layer interpretation, $\|W\|_2$ is therefore the relevant bound for adversarial or perturbation amplification, whereas $\|W\|_F$ is a global weight-size penalty. The inequality

$$\|W_L \cdots W_1\|_2 \le \prod_{\ell=1}^{L} \|W_\ell\|_2$$

should be read as a warning about composition: even if each layer is controlled separately, the product controls how errors can accumulate through the whole map.