linear algebra

- vectors identify points in space → is one way to think of them. each element gives a coordinate on a different axis.
- for a matrix $A_{i,:}$ identifies all elements of row i. $A_{:,i}$ identifies the i-th column.
- higher dimensional matrices are called tensors.
- Broadcasting (the same idea from the numpy notes) is where we fit a smaller dimensional tensor to a bigger dimensional one in a way that makes sense so as to allow the computation in context. $C = A + b$, where A is a matrix and b is a vector is the same as copying b as many times as the columns of A into a matrix and doing the addition.
- element-wise product, as opposed to matrix product, is called Hadamard product.
- $(AB)^T = B^T A^T$.
- The summary of numerical problems on computers is that each number is represented by a finite number of bits and for some operations the propagated errors are small vs for others where its intolerable, for the application at hand.
- Inverting (look at linear algebra notes) is not very computationally efficient, and also numerically not great. It is more of a theoretical tool when it comes to super large matrices.
- Determining whether $Ax = b$ has a solution amounts to testing whether b is in the column span of A. Column span is also called column space or range.
- Norms:
  - to measure the size of a vector.
  - L^P norm is given by $||x||_p = \sum_i (|x_i|^p)^{\frac{1}{p}}$
  - Notice that for all norms, we take absolute value before raising it to a power.
  - More rigorously, a norm is any function that satisfies:
    - $f(\mathbf{x}) = 0$ *implies* $\mathbf{x} = 0$.
    - $f(\mathbf{x} + \mathbf{y}) \leq f(\mathbf{x}) + f(\mathbf{y})$ (triangle inequality)
    - $f(\alpha\mathbf{x}) = |\alpha| f(\mathbf{x})$
  - L^2 norm is called euclidean norm → is $x^T x$.
  - the L^1 norm has the property that its derivative is constant, so it grows at the same rate everywhere.
  - we sometimes measure the size of a vector by counting the number of nonzero elements. some authors informally refer to this as "L^0" norm.
  - The L^∞ norm refers to the largest element in the vector.
  - Frobenius norm is simply the sum of squared elements of a matrix → a rather obscure measure for size of a matrix.
- Diagonal matrices (with non-zero elements only along diagonals) are of special interest because they are super easy to multiply and invert (think why, its easy). We might derive some general ML algorithm but then restrict some assumptions to be diagonal matrices in order to scale the algorithm.
- in $\mathbb{R}^n$, we can have at most n vectors of non-zero norm mutually orthogonal → refer to linear algebra notes to see why.
- For an orthogonal matrix, $A^{-1} = A^T$.
- eigendecomposition lets us decompose matrices in ways that show us information about their functional properties that are not obvious from their usual representation as an array of elements.

- we can decompose symmetric real matrices A into $A = Vdiag(\lambda)V^{-1}$, where V is a set of linearly independent eigenvectors and λ is a diagonal matrix with eigenvalues → see linear algebra notes for proof of this.
- If A is real and symmetric, we can show that the decomposition can be $A = Q\Lambda Q^T$, where Q is orthogonal, not just linearly independent.
- The eigendecomposition allows us to write A^k as $Vdiag(\lambda)^k V^{-1}$. This lets us discover what happens to a matrix as we keep repeatedly multiplying it with A → which dimensions keep getting stretched and which ones shrink.
- A matrix is singular iff any of the eigenvalues are 0.
- A matrix with all positive eigenvalues is positive definite which means $x^T Ax > 0$ for all x.
- Singular Value decompsition (SVD) is a more general method applicable to any real matrix, not just to symmetric ones. It gives us $A = UDV^T$, where U, V are orthogonal. U is the eigenvectors of AA^T and V is the eigenvectors of $A^T A$.
- The SVD reveals similar properties of a matrix as does the eigendecomposition.
- Moore Penrose Pseudoinverse of a singular A is $A^+ = VD^+U^T$ where D^+ is defined by taking the reciprocal of non zero elements of D. Verify that $AA^+ = I$. Of course, A^+ is not unique, unlike if A were invertible.
- The determinant is the product of all eigenvalues. If determinant is 0, then space is contracted completely along at least one dimension, causing it to lose all volume.
- Trace is sum of diagonal elements. it can be shown that $Tr(ABC) = Tr(BCA) = Tr(CAB)$.
- PCA is like lossy compression but with the least sum of L2 losses with the reconstructed vectors. We use the k largest eigenvalues → this reduces loss the most. → the math is pretty easy to workout here, see cs 229 notes.