1) Definitions & Properties

Linear Dependence - If we have pivot vars, we are linearly independent, else we're not.

For a matrix R^{nm} n = rows, m = cols. Rows are across the matrix, cols are down the matrix.

1.1) Determinants: For A, B ∈ $R^{n\times n}$

1) $det(A^T) = det(A)$
2) det(AB) = det(BA) = det(A)det(B). det(kA) = k^ndet(A)

det(A) = 0 ⇔ No Inverse.
Else, $det(A^{-1}) = 1/det(A)$.

Dets only exist for Sq Matrices, we get them by getting REF and multiplying by lead diag.
REF ops have effects on the det:
1) Swapping row multiplies it by -1
2) Adding/subbing rows does nothing
3) If any two rows are equal, or lead diag = 0, then det = 0.
4) Multiplying by scalar also increases det by a scalar.

Trace - sum of diag elems.
Rank of invertible matrices - nxn matrix is invertible ⇔ rank is n ⇔ cols are lin indep, rows too
Singular - A square matrix is non-singular if the columns are linearly indep, ie: if rk(A) = n, or det(A) =/= 0. Else its singular.
Vector Space - a set U =/= ∅ is a vector space if U is closed under addition and scalar multiplication: 1) For all u, v ∈ U, u + v ∈ U
2) For all u ∈ U and c ∈ R, cu ∈ U
Vector Subspace: subset of a Vector Space
Generating set: Our vector subspace X is a generating set of U it could express every vector U in the Vector Space as a linear combo of its vectors.
A **basis** is a minimal generating set. A **simple basis** is one with as many 0s, as possible, get by transp. -> RREF -> transp. and taking the pivot columns span as our simple basis.
Dimension = num basis vectors.
M1) Finding Basis: get the REF and take original vector of each pivot column. Span of this = basis.
M2) Change of Basis Matrix:
We just represent each basis vector in terms of the other basis, and each representation is one of our columns.
Euclidean norm: magnitude of vector: root of sum of squares.
Parallelogram Law: ∀u, v ∈ R^n, $||u+v||^2+||u-v||^2=2||u||^2+2||v||^2$
Angle between vectors
cos(x) = u.v / ||u|| ||v||

2) Orthogonality
Take vectors u,v **Orthogonal**: u.v = 0. This means they are perpendicular.
Orthonormal: orthogonal ∧ u &v's magnitudes are 1.
Orthogonal Matrix: A ∈ $R^{n\times n}$ is invertible with $A^{-1} = A^T$
Orthogonal Subspaces: (aka: U⊥V)
For ∀u ∈ U, ∀v ∈ V, u·v = 0.
Cauchy Schwarz inequality:
u, v ∈ R n , |u · v| ≤ ||u|| ||v||
Triangle inequality:
∀u, v ∈ R n , ||u + v|| ≤ ||u|| + ||v||
u ⊥ v ⇔ $||u + v||^2 = ||u||^2 + ||v||^2$

3) Linear Maps
A map from one vector subspace to another is linear if:
For all u, v ∈ V , f(u + v) = f(u) + f(v)
For all u ∈ V and c ∈ R, f(cu) = cf(u)
A basis change matrix from u to v is a linear map. Example: (map from 2 dimensions to 3)
$f(x)=\begin{pmatrix} y \\ x \\ y+x\end{pmatrix} = \begin{bmatrix}0 & 1\\ 1 & 0\\ 1 & 1\end{bmatrix}\begin{pmatrix}x\\y\end{pmatrix}$
We can compose basis change and linear maps
For an f: $R^n → R^m$ linear map, the **Image Space**: refers to the set of points mapped to in R^m. The **Kernel/Null** space is the set of points in the N space that is 0.
Rank Nullity Theorem: For linear map f:$R^n → R^m$ with Mat A ∈ R^{mn}: dim(im(f)) + dim(ker(f)) = rk(A) + dim(ker(A)) = n

4) Eigenvalues and Eigenvectors
Eigenvalues = solutions to the Char. Polynomial (CP) (A-λI).
Spectrum: set of EigVals of a Mat. Tr(A) = sum of eigenvalues.
det(A) = product of eigenvalues. **Algebraic Multiplicity (AM) of an eigenvalue** = num times solves CP. **Geometric Multiplicity (GM)** = dim of the eigenspace of eigenvalue.
1 <= GM <= AM <= Dim of Matrix.
Eigenspace of an eigenvalue = span of its eigenvectors.
Similar Matrices: Two matrices A and B are similar if there is a matrix P such that A= PBP^{-1}
Cayley Hamilton Theorem: Take CP of A; take A as lambda, we get the 0 matrix. We find inverses this way: $A = \begin{pmatrix}1 & -1\\ 2 & 1\end{pmatrix}$ CP: $x^2-2x+3= 0$.
Then: $-A^2+2A=3I$ ⇔ A*1/3(-A+2I) = I
A^{-1}=1/3(-A+2I) = I. So A^{-1}=1/3(2I-A)
The Rank of a Matrix = Rank of its Transpose.
A^TA and A have the same nullspace, easily provable.
Orthogonal transformations do not change the norm. Orthogonal Matrices have columns which are perpendicular to each other
Transpose - Write the rows as columns. $(AB)^T = B^TA^T$
For a matrix to be diagonalisable, the number of eigenvectors must be equal to its dimension. (Diagonalizable matrices are always square).
An Eigenvector scales a matrix by its associated Eigenvalue.
Matrix A is orthogonal ⇔ columns of A are perpendicular

Computational Techniques

1) Vector Norms:
Computing size on a vector.
Properties: For vector x, y in Rn, scalar k
1) ||x|| > 0. (for non-zero x)
2) ||kx|| = ||k|| ||x||
3) ||x+y|| <= ||x|| + ||y||

L_p norm is $||v||_p = (\sum_{i=1}^n |v_i|^p)^{1/p}$
L_1 = Sum of absolute vector elems
L_2 = Euclidean Norm
L_∞ = Max absolute vector entry.
Any two vector norms $||\cdot||_a$ and $||\cdot||_b$ in $\mathbb{R}^n$ are equivalent. Formally:
$\exists r, s \in \mathbb{R}. r > 0 \wedge s > 0 \wedge \forall x \in \mathbb{R}^n. r||x||_a \leq ||x||_b \leq s||x||_a$
$L_1, L_2, L_\infty : \mathbb{R}^n \to \mathbb{R}$ are **equivalent** as below:
1. $||x||_\infty \leq ||x||_2 \leq ||x||_1$
2. $||x||_2 \leq \sqrt{n}||x||_\infty$
3. $||x||_1 \leq \sqrt{n}||x||_2$

2) Matrix Norms (MN):
1) ||A|| > 0   2) ||kA|| = ||k|| ||A||
3) ||A + B|| ≤ ||A|| + ||B||
Sub-Multiplicative:
In this course we only use sub-multiplicative MNs, where
4) ||AB|| ≤ ||A|| ||B||
Computing the Norms:
L_1 = max absolute column sum (add up abs values of each entry)
L_2 = Largest singular value of A (on a square matrix, this is the biggest eigen. Else A^TA/AA^T)
L_∞ = Maximum absolute row sum
For vector norms $||\cdot||_a$ and $||\cdot||_b$:
Consistent: A matrix norm is consistent with the vector norms if
$\forall A \in \mathbb{R}^{m\times n}, x \in \mathbb{R}^n. ||Ax||_b \leq ||A|| ||x||_a$
Compatible:
A matrix norm is compatible with the vector norms if
$\forall A \in \mathbb{R}^{m\times n}, x \in \mathbb{R}^n. ||Ax||_b = ||A|| ||x||_a$
Subordinate Matrix Norm: For a vector norm $||\cdot|| : \mathbb{R}^n \to \mathbb{R}$, the subordinate matrix norm $||A|| : \mathbb{R}^{m\times n} \to \mathbb{R}$ is:
$\forall A \in \mathbb{R}^{m\times n}. ||A|| = \max\{||Ax|| \mid x \in \mathbb{R}^{m\times n}, ||x|| = 1\}$
A vector norm is compatible with its subordinate matrix norm AKA: for all A^{mn} and x^n, ||Ax|| <= ||A||||x||. For p = 1, 2 and infinity the MN is subordinate to the vector norm. Thus the matrix norm is also compatible with the vector norm.

3) Complex Linear Maps
Complex Num * Conjugate > 0. **Complex Conjugate of a Matrix:** Compute the conjugate of each element (if complex).
If a vector, scalar or matrix = its conjugate, then its real.
Standard Inner Product <u, v>: Of two vectors u, v ∈ C^n = $\bar{u}^Tv$.
The inner product is **conjugate** symmetric: <u, v> = $\overline{<u, v>}$
Standard Norm on u: $\sqrt{\bar{u}^Tu}$
SIP > 0; if it is 0 then u = 0. **Complex Eigenvalues and Eigenvectors:** The eigenvalues of a real matrix can be complex. This means our eigenvectors are complex too.

4) Least Squares Method
Endomorphism: A linear map where the domain and codomain are the same f: $R^n → R^n$
Automorphism: bijective endomorphism.
Automorphism
f: $R^n → R^n$ ⟺ f injective (ker(f)={0}) ⟺ f surjective (im(f) = R^n)
Projection of a Subspace:
Let U ⊂ R^m be an n dimensional subspace generated by an ordered basis $(u_1,\cdots,u_n)$. Let U = $[U_1,...,U_n]$
The **orthogonal projection** Π_U on U is the following endomorphism:
$\Pi_U: R^m → R^m$
$v → \Pi_U(v) = U(U^TU)^{-1}U^Tv$
im(A) ⊥ ker(A^T)
Unique Vector Decomposition: Let A ∈ R^{mn} For all vectors b ∈ R^m, there exists unique b_i ∈ im(A), and unique b_k ∈ ker(A^T) such that: b = b_i + b_k.
Let A ∈ R^{mn} and b ∈ R^m. Suppose Ax = b has no solution for x ∈ R^n, i.e., b ∉ im(A).
LSM finds x ∈ R^n such that
$||Ax - b||_2$, or equivalently $||Ax - b||^2_2$ is minimised.
$||Ax-b||_2$ is minimised when
$||Ax-b_i||_2 = 0$ ⇔ Ax = b_i.

5) Linear Regression
We have a set of points (y_i, a_i), y a real number, and a is a real vector of dimension n.
We want to find the model of best fit with parameters s_0 ∈ R and s ∈ R^n, so the sum of the errors squared is minimized:
$$\sum_{i=1}^m (s_0 + s * a_i - y_i)^2$$
We require $s_0 + s \cdot a_i \approx y_i$
The sum of the squared errors is this:
$\sum_{i=1}^m (s_0 + s*a_i - y_i)^2 = ||Az - y||^2_2$

6) Spectral Decomposition of Symmetric Matrices
Properties of Orthogonal Matrices:
1) *Orthogonal Matrix transformations preserve Euclidean length of vectors:*
∀u ∈ R^n, $||Qu||_2 = ||u||_2$
2) *Orthogonal transformations preserve the magnitude of the angle between vectors:*
∀u, v ∈ R^n, $\widehat{Qu\,Qv} = \widehat{u\,v}$
3) The transformations performed by an orthogonal matrix can be interpreted as a change of basis or a series of rotations and reflections. Orthogonal transformations don't change the norm.
4) det = 1 or -1.
5) All eigenvalues have modulus 1

Properties of Symmetric Matrices:
Defined by $A^T = A$.
1) If A is a real symmetric matrix, then all its Eigenvalues are real.
2) If A is a real symmetric matrix, then for each Eigenvalue the algebraic multiplicity and geometric multiplicity are equal.
3) If A is an n by n real symmetric matrix, eigenvectors for distinct eigenvalues are orthogonal.

6.1) Spectral Theorem
If A is a real, symmetric matrix then it can be diagonalised like so:
$A = QDQ^T = QDQ^{-1}$
where: Q is an **orthogonal** matrix
D is the **diagonal** eigenvalue matrix.

7) Singular Value Decomposition
A is real symmetric matrix.
Positive Definite: A is positive definite iff: $\forall x \in R^n - \{0\}, x^TAx > 0$.
Positive Semi-Definite iff:
$\forall x \in R^n - \{0\}, x^TAx >= 0$.
Theorem: Positive Definiteness in Terms of Eigenvalues:
1) Positive definite ⇔ all eigenvalues are strictly positive
2) Positive semi definite ⇔ eigenvalues are non negative
Properties of A^TA and AA^T
Take arbitrary $A \in R^{mn}$
1) $A^TA \in R^{nn}$ and $AA^T \in R^{mm}$ are both **symmetric** and **positive-semi definite.**
This gives rise to SVD - a more general decomposition than spectral, which takes more work to get into but still exhibits useful properties.

8) SVD Definition:
Take arbitrary $A \in R^{mn}$. SVD of A is any decomposition of the form:
$A = USV^T$
$U \in R^{mm}$ and $V \in R^{nn}$ are orthogonal.
$S \in R^{mn}$ is a diagonal matrix, $S = diag(\sigma_1, \sigma_2, \sigma_3, ..., \sigma_p)$ where p = min(m, n).
$\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_p \geq 0$ (**We write largest first**)
The values $\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_p \geq 0$, are known as the singular values of A
NOTE for symmetric matrices, singular values are simply the absolute values of the eigenvalues.
Useful Properties of SVD:
1) SVD: $A = USV^T$
If $U = [u_1, ..., u_m]$ and $V = [v_1, ..., v_n]$, then
$A = \sigma_1 u_1 v_1^T + \sigma_2 u_2 v_2^T + \cdots + \sigma_r u_r v_r^T$
2) **rk A = num singular positive values in S**
3) For $A \in R^{mn}$ $||A||_2$ = largest singular value of A.
The positive singular values of A = positive square roots of the eigenvalues of AA^T or A^TA
5) The span of the first r columns of U = im(A).
The span of the last m-r columns is ker(A), where r is the rank
6) **The Principal Axes and First Principal Axis of a Collection of Samples:** Assume $A \in R^{mn}$ represents m samples of n dimensional data. The principal axes of A are the columns of V, i.e., v_i for $1 \leq i \leq r$. The first (principal) axis of the collection of m samples is defined as: $w(1) = \operatorname{argmax}_{||w||=1} w^TA^TAw$
7) **The Principal Axes of a Singular Value Decomposition** Assume $A \in R^{mn}$ represents m samples of n dimensional data. Assume we have the following singular value decomposition for A: $A = USV^T$, where $V = [v_1, ... v_n]$ Then, v_1 is the first (principal) axis of A. It corresponds to the column vector of V associated with the largest singular value of A.
8) **The Principal Components (Scores of a Selection of Samples)**
The columns of US, i.e., $\sigma_i u_i$ for $1 \leq i \leq r$, are the principal components or scores of A.

9) Generalised Eigenvectors
Take a square matrix $A \in R^{nn}$.
Generalised Eigenvector: A nonzero vector $V \in C^n$ **of rank m** associated with eigenvalue λ when $(A - \lambda I)^m v = 0$ and $(A - \lambda I)^{m-1} v =/= 0$
Any eigenvector associated with its eigenvalue is a generalised eigenvalue of rank 1.
An eigenvector of A associated with λ and multiplied by the matrix A - λI, results in the zero vector.
If A has a generalized eigenvector of rank 1 associated with λ, then when you multiply it by A - λI, the result will be another eigenvector associated with λ.
If A has a generalized eigenvector of rank 2 associated with λ, then when you multiply it by A - λI, the result will be a linear combination of a rank 1 generalized eigenvector and another eigenvector associated with λ. This pattern continues for higher-rank generalized eigenvectors. Example:
Take $A = \begin{bmatrix}1&1&1\\0&1&0\\0&0&1\end{bmatrix}$ CP: $(1-\lambda)^3$,
$\lambda_1 = 1$ (with AM = 3). We end up with two linearly indep EVecs: $(0,1,-1)^T$, $(1,0,0)^T$. There is one more EVec $v_3 = (0,0,1)^T$. We compute

10) Jordan Normal Form
A matrix is in Jordan Normal Form if it is of this form:
$$\begin{bmatrix} J_{k_1}(\lambda_1) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & J_{k_n}(\lambda_n)\end{bmatrix}$$
Each $J_{ki}(\lambda_i)$ is a Jordan block of size k_i with a diagonal (not always unique) coefficient λ_i:
$$\begin{bmatrix}\lambda_i & \cdots & 0\\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_i\end{bmatrix}$$
Even if $A \in R^{nn}$, a JNF might not be in R^{nn} but rather in C^{nn}.
Blocks in JNF and Multiplicities of Eigenvalues: The algebraic multiplicity of an eigenvalue λ is the sum of the sizes of blocks with λ on the diagonal. The geometric multiplicity of λ is the number of blocks with λ on the diagonal.

11) Cholesky Decomposition
A matrix A^{nn} is **lower triangular** $\forall i < j, A_{ij} = 0$
A matrix A^{nn} is **upper triangular** $\forall i > j, A_{ij} = 0$
These matrices exhibit useful properties. The equation Ax = b can easily be solved on them, by first getting x_1, and then x_2 with direct substitution and so on, on lower triangular matrices. For upper triangular we get x_n first, and go backwards.
Additional Properties of Symmetric Matrices:
Let $A \in R^{nn}$ be a symmetric matrix.
1) If A is **positive definite**, all its diagonal elements are **strictly positive**.
2) If A is **positive semi-definite**, all its diagonal elements are **non-negative**.
3) If A is positive definite then max $(A_{ii}, A_{jj}) > |A_{ij}|$. If A is positive semi definite then max $(A_{ii}, A_{jj}) >= |A_{ij}|$. Thus, the largest coefficient of A is on its diagonal.
4) If A is positive definite then the 1x1, 2x2, ... mxm... matrices in the upper left corner of A are also positive definite. Same holds for semi-definite.
We can quickly notice non-positive-**semi definite** matrices - if we have a symmetric matrix with a negative diagonal element, it cannot be PSD. Also: if we see a matrix element $|A_{ij}| > \max(A_{ii}, A_{jj})$ then it's not PSD (e.g if the 3rd element in the first row = 3, 1st in 1st row = 2, 3rd in 3rd row = 1, then we violate rule 3 as above).
Cholesky Decomposition: any decomposition of a real square matrix A^{nn} of the form $A = LL^T$, where L is a lower triangular matrix.
If A is not positive semi-definite then it doesn't have a Cholesky Decomposition. If A is positive semi definite then it always has a Cholesky Decomposition. Also there exists a version of L with strictly positive diagonal elements.

12) QR Decomposition
QR decomposition is a matrix factorisation method, decomposing our matrix A into an orthogonal matrix Q and an upper triangular matrix R such that A = QR.
Using the Gram-Schmidt Process:
$a_1, ..., a_n$ are the vector columns of A: $[a_1 ... a_n]$.
Assume they're linearly independent.
1) We apply GS process constructing an orthonormal basis $(e_1 ... e_n)$ s.t $span(e_1,...,e_n) = span(a_1,..., a_n)$.
2) Selecting Q and R: Set $Q = [e_1, ..., e_n]$. It is semi orthogonal (e.g $Q^TQ = I$). We then set R like so:
$$\begin{bmatrix} e_1a_1 & e_1a_2 & \cdots & e_1a_n\\ 0 & e_2a_2 & \cdots & e_2a_n \\ \vdots & 0 & \ddots & \vdots \\ 0 & \cdots & 0 & e_na_n\end{bmatrix}$$

13) QR Algorithm
Efficiently finds the eigenvalues of a matrix.
1) We start with our matrix A
2) Use GS method to get the QR decomposition: A = QR.
3) Compute the product RQ, set A = RQ.
4) Repeat steps 2-3, until A converges to a diagonal matrix D, whose diagonal entries are the eigenvalues of A.
Properties of the QR Decomposition:
1) For k ∈ N, A_k is similar to A.
2) For k ∈ N, we have $A_k = Q_k^TAQ_k$ from above. So, A_k and A have the same eigenvalues and v is an eigenvector of A_k if and only if Q_kv is an eigenvector of A
3) The sequence (A_k) converges to an upper triangular matrix under certain conditions. Useful because of the following property
4) (if A is symmetric) The eigenvalues of an upper triangular matrix are simply its diagonal elements. So, the QR decomposition is easily findable, and from this, we can converge it to an upper triangle matrix, from which eigenvalues are easily findable!!
5) 2) $A^n = Q_1 ... Q_nR_1 ... R_n$

13.2) Householder Maps
Suppose we have a hyper-plane P going through the origin with unit normal $u \in R^m$, i.e., $P = \{x \in R^m : u \cdot x = 0\}$. The Householder matrix defined by $H_u = I - 2uu^T$ induces reflection wrt P.
Properties of a Householder Matrix:
1) H_u is **involutory**: $H_u = H_u^{-1}u$.
2) H_u is **orthogonal**: $H^Tu = H^{-1}u$.
3) H_u **preserves the euclidian length of vectors**: $||H_u(x)|| = ||x||$.
4) The **eigenvalues** are only 1 or -1.
5) The **eigenvectors** are vectors perpendicular to the hyperplane P reflects across - e.g any vectors in the hyperplane P.
6) H_u **preserves Euclidian length and angles** between vectors as it is orthogonal. Because of this any rotations and reflections are orthogonal transformations (because they preserve the aforementioned things).
7) The orthogonal projection Q on the hyperplane P is given by: $Q = I - uu^T$ with $Q^2 = Q$ and $Q = Q^T$

14) Application of the QR Decomposition to Symmetric Matrices
If A is symmetric, so are all the A_k. If A is symmetric, the algorithm converges, under certain conditions, to a diagonal matrix, hence the eigenvectors of A are in effect the columns of $Q^\sim_k$ for large enough k.

14.1) Convergence Under Specific Conditions: The LU Decomposition:
A non-singular matrix $A \in R^{nn}$ can be factorised as A = LU
L is a lower triangular matrix
U is an upper triangular matrix
We can do this *if and only if* A can be reduced to its row echelon form *without swapping any two rows.*
Properties of the LU Decomposition:
1) Uniqueness of the LU Deomposition when A is non singular: If A = LU and the diagonal elements of L are all one, the decomposition is unique.
2) Let $A \in R^{nn}$ be a symmetric positive definite matrix with distinct eigenvalues $\lambda_1 > \lambda_2 > ... > \lambda_n > 0$ with eigendecomposition $A = Q\Lambda Q^T$. Suppose $Q^T = LU$ with unit lower triangular L and the diagonal elements of U are positive. Then $A_k \to \Lambda$

15) Fixed Points
1) Convergence of a Sequence of Real Numbers: Let $(a_n)_{n\in N} \in R^N$ be a sequence of real numbers and $l \in R$. The sequence (a_n) is said to converge to its limit l, $\lim_{n\to\infty} a_n = l$ if and only if: $\forall e > 0, \exists N \in N$ such that $\forall n > N, |a_n - l| < e$
The steps we take to show a sequence of real numbers converges:
1) Find the limit l
2) Take e > 0
3) Find N ∈ N such that $|a_n - l| < e$ for n > N, the value of N will usually depend on e, in an inversely proportional relationship
2) Cauchy Sequence: Let $(a_n)_{n\in N} \in R^N$ a sequence of real numbers. Then (a_n) is said to be a Cauchy sequence if and only if: $\forall e > 0, \exists N \in N$ such that $\forall n, m > N, |a_n - a_m| < e$
This gives rise to the Cauchy Test: Let $(a_n)_{n\in N} \in R^N$ a sequence of real numbers. Then (a_n) is convergent if and only if it is a Cauchy sequence
3) Metric Spaces: A metric space is a tuple (S, d) where S is a non-empty set and d is a metric over S. d : S × S → R is such that:
∀x, y ∈ S, d(x, y) ≥ 0 (positivity)
∀x, y ∈ S, d(x, y) = 0 ⟺ x = y (reflexivity)
∀x, y ∈ S, d(x, y) = d(y, x) (symmetry)
∀x, y, z ∈ S, d(x, y) ≤ d(x, z) + d(z, y) (tri. ineq.)
4) Metric in a normed Vector Space: V is a vector space equipped with the norm ||.||.
Let d be the function $d: \begin{cases} V \times V \to \mathbb{R} \\ (x, y) \to ||x - y||\end{cases}$
d is a metric space.
From here we consider the concepts seen in R but generalised to a metric space (S, d).
5) Convergence in a Metric Space:
Convergence in a metric space. Let (S, d) be a metric space and (a_n) a sequence in S. Then a_n is said to converge to a limit l ∈ S iff : ∀e > 0, ∃N ∈ N such that ∀n > N, $d(a_n, l) < e$
6) Cauchy Sequence in a Metric Space: Let (S, d) be a metric space and (a_n) a sequence in S. (a_n) is a Cauchy sequence iff: ∀e > 0, ∃N ∈ N such that ∀n, m > N, $d(a_n, a_m) < e$
7) Cauchy Test on a Metric Space: Let (S, d) be a metric space and (a_n) a sequence in S: if (a_n) is convergent, then it is a Cauchy sequence.
8) Complete Space: Let (S, d) be a metric space. Then it is said to be a complete space if and only if every Cauchy sequence in S converges to a value in S.
9) Completeness of L1, L2 and L∞ Norms: For any k > 0, R^k equipped with any of the three metrics induced by L1, L2 or L∞ is complete. The space (C[a, b], || · ||∞) is complete.
10) Only Cauchy Sequences Converge in Complete Metric Spaces: Let (S, d) be a complete metric space and (a_n) a sequence in S. (a_n) is converges iff it's a Cauchy Seq.
11) Fixed Point: Let S be a non-empty set and f: S → S a function from S to itself. p ∈ S is a fixed point if: f(p) = p
12) Contraction: Let (S, d) be a metric space and f: S → S. f is called a contraction of S (or a contracting map) **if there exists 0 ≤ α < 1** called the **contraction constant** s.t: ∀x, y ∈ S, d(f(x), f(y)) ≤ αd(x, y).
13) The Fixed-Point Theorem: Let (S, d) a complete metric space and f a contraction of S. Then f has a unique fixed point. This leads to the theorem:
14) Uniqueness of Solutions of Differential Equations: Differential Equations have unique solutions: there exists a unique function x: $(t_0 - d, t_0 + d)$ satisfying: dx/dt = f(x(t), (t)) with $x(t_0) = x_0$.
$d: \begin{cases}\frac{dx}{dt} = \dot{x}(t) = f(x, t) \\ x(t_0) = x_0\end{cases}$
Where d = min(a,b/(B+bc)) with B = sup{|f(x, t)| : $|x-x_0|<b$, $|t-t_0|<a$}>0

16) Condition Numbers
Condition numbers are the measure of sensitivity of a system to small fluctuations in input.
1) Calculating Condition Numbers:
$$\kappa(P) = \max_\epsilon \frac{||s(d) - s(d+\epsilon)||}{||\epsilon||}$$
2) Relative Condition Number:
$$\frac{||d||}{||s(d)||}\frac{||s(d) - s(d+\epsilon)||}{||\epsilon||}$$
3) Condition Numbers on Square Matrices: **Let** A be a non-singular matrix. The Condition Number of A, $\kappa(A) = ||A^{-1}|| \, ||A||$ = ***Relative Condition Number***.
4) Condition Numbers of a Linear Equation: Ax = b = Condition Number of Coeff Mat A.
Big norm != ill conditioned, we might just be working with big numbers!!
5.1) The Pseudoinverse of a Matrix:
Ax = b ⇔ $A^TAx = A^Tb$ ⇔ $x = (A^TA)^{-1}A^Tb$.
The matrix $(A^TA)^{-1}A^T$ is known as the **pseudo inverse** of A, and denoted A^+.
5.2) Cond Numbers of Non-Square Matrices: Compute the Pseudoinverse of A. Then to compute the condition number of a non square matrix A: $k(A) = ||A^+|| \, ||A||$
6) Conditioning of a Problem:
To decide if a problem is well or ill conditioned, we can use this Rule of Thumb to guide us: For a condition number κ(A), you lose about **log10(κ(A)) significant figures** in accuracy. If we have an accurate implementation to 12 dp, if the condition number is 1,000,000 then we'll lose about 6 dp in precision.

17) Iterative Solutions to Linear Equations
$A \in R^{nn}$ and $b \in R^n$. We want to solve Ax = b.
Gaussian Elimination $O(n^3)$ is too expensive. We can approx. the solution with iteration.
1) Theorem of Convergence of Sequences: Let $M \in R^{nn}$ and $c \in R^n$. Let ||.|| be a consistent norm on R^{nn}, if ||M|| < 1 then the sequence (x_k) defined by $x_{k+1} = Mx_k + c$ converges for any starting point x_0.
So, to solve Ax+b we solve the equation x = Mx+c. Where $M = -G^{-1}R$, $c = G^{-1}b$.
This is in the form required for our Theorem. A solution to the equation, thanks to fixed point theory would be when $x_{k+1} = x_k$
2) Choosing Efficient Choices of Splitting:
1. We want $G^{-1}Rx$ and $G^{-1}b$ to be easy to compute
2. We want ||M|| is small (for fast convergence to a solution).
We'll assume A has no 0s on it's diagonal (if it does we can do a change of basis to achieve it):
3) The ADLU Split:
We can write A = D + L + U, where D is the diagonal of A, L, U are the lower and upper triangle parts of A respectively.
4) The Jacobi Method:
1) Let R = L+U
2) We want to solve, for some $b \in R^n$
Ax = b
Ax = b ⇔ (D+R)x = b ⇔ $x = -D^{-1}Rx + D^{-1}b$
⇔ x = Mx + c. ($M = -D^{-1}R$, $c = D^{-1}b$)
This splitting is good as x^{k+1} is easily computed as:
D^{-1} = reciprocal of each diag element as diags, R = the combination of the upper and lower triangle matrices, so: $Mx = -D^{-1}Rx =$
$$-D^{-1}\begin{bmatrix}\sum_{j\neq1} a_{1,j}x_j\\ \sum_{j\neq2} a_{2,j}x_j \\ \vdots \\ \sum_{j\neq n} a_{n,j}x_j\end{bmatrix} = \begin{bmatrix} -\frac{1}{a_{1,1}}\sum_{j\neq1}a_{1,j}x_j \\ -\frac{1}{a_{2,2}}\sum_{j\neq2}a_{2,j}x_j\\ \vdots \\ -\frac{1}{a_{n,n}}\sum_{j\neq n}a_{n,j}x_j\end{bmatrix}$$
(This makes sense as $-D^{-1}$ is the negative of the reciprocal of each diagonal element, and we multiply this by L+U, which is a matrix of elements except the diagonal).
Since $x^{(k+1)} = Mx^{(k)} + c$, to solve $Ax = b$ we have
$$x_i^{(k+1)} = \frac{1}{a_{i,i}}\left(b_i - \sum_{j\neq i} a_{i,j}x_j^{(k)}\right)$$
5) Gauss-Seidel Method:
Reuse the ADLU splitting, but do A = (D+L) + U. Then Ax = b ⇔ (D+L)x = b - Ux ⇔ x = Mx + c,
Where $M = -(D+L)^{-1}U$, $c = (D+L)^{-1}b$.
D+L is a lower triangle matrix with a diagonal D, U is an upper triangle matrix.
And then $x^{k+1} = Mx^k + c$ ⇔ $(D+L)x^{k+1} = -Ux^k + b$
$$\forall i \in [1,n] \sum_{j\leq i} a_{i,j}x_j^{(k+1)} = -\sum_{j>i}a_{i,j}x_j^{(k)} + b_i$$
So the i^{th} element of $x^{(n+1)}$ can be computed thanks to A, $x^{(n)}$, b and the k^{th} elements of $x^{(n+1)}$ for k < i. As the update is computed with more recent quantities, the convergence is faster with the GS method.
These methods are only guaranteed to converge as A is strictly row diagonally dominant - AKA: $\forall i, |a_{ii}| > \sum_{j\neq i}|a_{ij}|$
1) Influence of the Condition Number on Convergence: These techniques converge faster for low condition numbers. If big, it might even diverge.
2) **Irreducible:** A matrix is irreducible if it can't take the form: $\begin{bmatrix}A_{11} & A_{12}\\ 0 & A_{22}\end{bmatrix}$ where A_{11} and A_{22} are block matrices.
3) If A is weakly row diagonally dominant and irreducible, Jacobi and G-S still converge.

18) Iterative Techniques for computing Eigenvectors and Eigenvalues
It's $O(n^4)$ to get eigenvectors, (Solve CP; GE for EVecs). Need numerical methods:
The Dominant Eigenvalue of A is the eigenvalue with the largest modulus, a **dominant eigenvector** is an eigenvector of a dominant eigenvalue.
Power Iteration Theorem:
Let A ∈ R^{nn} be a diagonalisable matrix with eigenvalues of distinct modulus. Let λ ∈ R be the dominant eigenvalue.
We consider the sequence (x_k) defined by
$x_{k+1} = Ax_k / ||Ax_k||$ and $x_0 \in R^n \setminus \{0\}$.
Then, as k→infinity, x_k → v, and $||Ax_k||$ → |λ|. This convergence to v is not rigorous - if λ is negative then we oscillate between v and -v.
Power Iteration Limitations:
1. x_0 is chosen at random, it's possible for it to be such that α1 = ... = αp = 0. In this case, the iteration yields second dominant eigenvalue and eigenvector. We need to make sure there's a non zero component in the corresponding eigenspace - usually happens in computers due to floating point error.
2. Eigenvalues might not have distinct modulus, in which case the power iteration will converge to a linear combination of the corresponding eigenvectors.
3. Convergence might be slow if dominant evec isn't too dominant.

18.1) Inverse Power Iteration
Finds the smallest eigenvalue and its eigenvector. We do this by taking the inverse of A, and then performing power method. This yields the eigenvalue 1/λ, where λ = smallest eigenvalue of A.
Inverse Power Iteration Theorem:
Let A ∈ R^{nn} be a diagonalisable non-singular matrix with eigenvalues of distinct modulus. Let λ ∈ R be the eigenvalue with the smallest modulus. We consider the sequence (x_k) defined by x_{k+1} =
$A^{-1}x_k / ||A^{-1}x_k||$ and $x_0 \in R^n \setminus \{0\}$. Then,
$x_k \xrightarrow[k\to\infty]{} v$ and $||A^{-1}x_k|| \xrightarrow[k\to\infty]{} \left|\frac{1}{\lambda}\right|$

18.2) Shifts
Let A ∈ R^{nn} and s ∈ R. The matrix A-sI is the **shifted matrix** and with the following property:
Let A ∈ R^{nn} and s ∈ R. λ ∈ R is an eigenvalue of A if and only if λ-s is an eigenvalue of A-sI with the same eigenvectors. Thanks to the shifted matrix, it is possible to focus on a particular eigenvalue of the matrix A.
Theorem of Shifts: A ∈ R^{nn} is a diagonalisable matrix with eigenvalues of distinct modulus and s ∈ R. Suppose that the shifted matrix A - sI is non-singular (it is non-singular as long as s is not an eigenvalue of A), then by performing inverse power iterations, we can find the eigenvalue of A that it the closest to s.
Thus we can use the Inverse Power Method and Shifts to discover all eigenvalues.

18.3) Rayleigh Quotient
Let A ∈ R^{nn} and x ∈ $R^n \setminus \{0\}$. Rayleigh quotient R(A, x) is given by: x^TAx / x^Tx, where we apply it on each x in our sequence of power iteration.
When using an iterative technique to find the EVal/EVec of a matrix the Rayleigh quotient monitors the convergence to the eigenvalue as it gives access to an approximation of the eigenvalue (rather than the modulus).

18.4) Deflation
We can use the power method to get the max eigenvalue. But it's a problem to get the next eigenvalue, as we have to perform a shift - but we don't actually know what to shift by.
Instead we can **deflate our matrix into an $A^{(n-1)(n-1)}$** matrix, removing the dominant eigenvector. Let λ_1, ..., λ_n be the eigenvalues of A, ordered according to their magnitude with λ_1 being the dominant one. Define H as a non-singular matrix such that:
$Hx_1 = \alpha e_1$ where α ∈ R \ {0} and e1=$[1,0,..., 0]^T$ is the first vector in the standard basis. We then have:
$HAH^{-1}e_1 = HA\, x_1/\alpha = H\, \lambda_1/\alpha\, x_1 = \lambda_1 e_1$
The first column of HAH^{-1} is $[\lambda_1, 0, ... 0]^T$ and so
$HAH^{-1} = [\lambda_1\, b^T, 0\ B]$, where
B is an (n-1) by (n-1) matrix with eigenvalues λ_2, ... , λ_n.
They also have the property:
Let λ_2 be the second dominant eigenvector of A with λ_2 =/= λ_1.
An eigenvector x_2 of A corresponding to eigenvalue λ_2 is: $x_2 = H^{-1}[\beta\ z_2]$
With $\beta = b^Tz_2 / \lambda_2 - \lambda_1$ and z_2 is a dominant eigenvector of B

19) Functions of Several Variables
Functions of N Variables are rules which assign to each ordered tuple of real numbers $(x_1, x_2, ... x_n)$ a unique real number denoted by $f(x_1, x_2 ... x_n)$.
Range = set of values f takes, **domain** = set of values taken by params. The graph of f is the set of all points (x, y, z) in $\mathbb{R}^{n+1}$.

19.1) Level Contours
The level curves of a function on two variables are the curves with equation
$f(x, y) = k$, where k is a constant (in the range of f) A level curve $f(x, y) = k$ is the set of all points in the domain of f at which f takes on a given value k.
These look like rings at a height k, around the parts of the graph which assume that value. Like contour maps, the surface is steep where level curves are close.

19.2) Partial Derivatives
For a function of two variables, the partial derivatives are denoted by
$f_x(x, y) = \lim_{h\to0} f(x+h, y) - f(x,y)/h$
$f_y(x, y) = \lim_{h\to0} f(x, y + h) - f(x, y)/h$
(AKA just differentiate with respect to 1 and treat the other as a constant).
Geometrically, the partial derivatives of a 2 valued functions are tangent lines that are constant to the dimension which they "ignore" (as seen by T_1 and T_2 which are tangents to C_1 and C_2 (both curves are in the range) at P.
They represent the rate of change of z, with respect to x or y, when the other is fixed. The higher derivatives are computed as imagined
Clairaut's Theorem: Suppose f is defined on a disk D that contains the point (a, b). If the functions f_{xy} and f_{yx} are both continuous on D then $f_{xy}(a, b) = f_{yx}(a, b)$

19.3) Directional Derivatives
For a function $z = f(x, y)$, the partial derivatives f_x and f_y, represent the rates of change in z in the x - and y - directions, i.e., in the directions of unit vectors **i** and **j**. (calculated in the same way as before).
The rate of change of an arbitrary unit vector can be found with basic trig.
Our tangent line represents the rate of change of z at a specific point, the direction of which depends on the variable we diff with respect to. (e.g, how much do we move in x and y as we progress?)
Theorem: If f is a differentiable function of x and y, then f has a directional derivative in the direction of a unit vector **u** = a, b:
$D_uf(x, y) = f_x(x, y)a + f_y(x, y)b$
If the unit vector **u** makes an angle θ with the positive x-axis, then we can write **u** = <cos θ, sin θ> and the formula becomes $D_uf(x, y) =$
$f_x(x, y) \cos\theta + f_y(x, y) \sin\theta$

20) The Gradient Vector
$D_uf(x,y) = \langle f_x(x,y), f_y(x, y)\rangle \cdot u$
The first vector in the dot product is the gradient vector, denoted as grad f or ∇f.
It is given by this formula:
$\nabla f(x,y) = \langle f_x(x, y), f_y(x, y)\rangle$
$= \partial f/\partial x\,\mathbf{i} + \partial f/\partial y\,\mathbf{j}$
Tangent Plane: The tangent plane to a surface at a given point is a flat plane that locally approximates the surface at that point, and is parallel to the plane spanned by the partial derivatives of the surface at that point. We have these for 3 valued functions.
The gradient vector $\nabla F(x_0, y_0, z_0)$ gives the direction of fastest increase of f.
$\nabla F(x_0, y_0, z_0)$ is orthogonal to the level surface S of f through P.
If we consider a topographical map of a hill and let $f(x, y)$ represent the height above sea level at a point with coordinates (x, y) then **a curve of steepest ascent** can be drawn as shown in the figure by making it perpendicular to all of the contour lines.
Maxima and Minima:
Local Maximum & Local Min; Absolute Max and Absolute Min - remember these terms.

21) Gradient Based Optimization
A quadratic n variable function looks like:
$Q(x_1,..., x_n) = b_{11}x_1^2 + b_{12}x_1x_2 + \cdots + b_{ij}x_ix_j + \cdots + b_{nn}x_n^2$
We can write this more conveniently with matrices: $Q(x) = x^TAx$, where x is a column vector with elements x_i, 1 <= i <= n, and A is a symm. matrix with $a_{ij} = b_{ij} + b_{ji} / 2$
Note $x_ix_j = x_jx_i$. Eg:
$Q(x_1,x_2) = x_1^2 - ax_1x_2 + bx_2x_1 - cx_1x_3 + 5x_2^2$ is expressed as
$x = [x_1, x_2, x_3]^T$ $\quad A = \begin{bmatrix}1 & \frac{a+b}{2} & -\frac{c}{2}\\ \frac{a+b}{2} & 5 & 0\\ -\frac{c}{2} & 0 & 0\end{bmatrix}$
And then $x^TAx = Q(x_1, x_2)$

23.1) Gradient Descent:
$\nabla f(\underline{x}^{(k)})$ is a gradient of an n-dimensional function $f(\underline{x})$ at iteration k.
The Gradient Descent Formula:
Gradient descent* at iteration $k + 1$:
$\underline{x}^{(k+1)} = \underline{x}^{(k)} - \alpha\nabla f(\underline{x}^{k})$, $\alpha > 0$ is a fixed step size.
(AKA, to compute the next point, we take the current point and subtract the gradient and some step size).
Steepest Descent is a faster version - as we minimise what we have left.
$\underline{x}^{(k+1)} = \underline{x}^{(k)} - \alpha_k\nabla f(\underline{x}^{k})$
$a_k = \operatorname{argmin} f(\underline{x}^{(k)} - \alpha\nabla f(\underline{x}^{k}))$
Graphically, we compute the gradient at our point, then go the opposite direction by step size a, and repeat.
For quadratic functions of the form
$f(x) = 1/2x^TQx - b^Tx$ where
Q^{nn} is real symmetric positive definite matrix, b^n is a real vector, and x^n is a real vector the gradient of the function is:
$\nabla f(x) = Qx - b$, Hessian = $\nabla^2 f(x) = Q$.
With this, the optimal step size a at iteration k is: $a_k = (g^k)^T(g^k) / ((g^k)^T)Hg^k$
$g^k = \nabla f(x)$ and $H = Q = \nabla^2 f(x)$.