Math for DS. Review

Emin Mammadov

RESOURCES USED
mml-book

COVERED
TO DO

LINEAR ALGEBRA
LA is the study of vectors and certain rules to manipulate vectors. We use x and y to represent vectors. We can write vectors as the tuples of n real numbers ($\mathbb{R}^n$): $a = [1, 2, 3]^T \in \mathbb{R}^n$. Linear algebra focuses on

Vector Spaces

Also known as linear spaces. Collection of vectors that can be added and multiplied ("scaled") by scalars. Essentially, there are three kinds of mathematical structures: groups, fields, vector spaces. Groups are charactersized by one kind of element, one operation; fields, one kind of element, two operations ("addition" and multiplication); vector spaces have two kinds of elements (vectors and scalars); scalars form a field, and operations that apply to (vector, vector) pairs and to (vector,scalar) pairs.

Group: Mathematically group can be characterized as $\mathcal{G}$ and an operation $\otimes$: $\mathcal{G} \otimes \mathcal{G} -> \mathcal{G}$. $\mathcal{G}$" $= (\mathcal{G})\otimes$ can be called a group if:

- Closure of group under tensor product ($\otimes$): for all $x, y \in \mathcal{G} : x \otimes y \in \mathcal{G}$
- Associativity: for all $x, y, z \in \mathcal{G} : (x \otimes y) \otimes z = x \otimes (y \otimes z)$
- Neutral element (identity element): $\exists e \in \mathcal{G} \quad \forall x \in \mathcal{G} : x \otimes e = x$ and $e \otimes x = x$
- Inverse elements: $\forall x \quad \exists e \in \mathcal{G} \in \mathcal{G} : x \otimes e = e$ and $e \otimes x = e$. Denoted by x^{-1}. It's with respect to the tensor product and not always mean $\frac{1}{x}$

A commutative group $x \otimes y = y \otimes x$ is Abelian Group

Vector Spaces: In groups, we looked at inner operations on sets, that operate on elements in the set $\mathcal{G}$. Vector spaces do inner operation (addition) and outer operation (scalar multiplication). Inner/outer operations have nothing to do with inner/outer products. Vector spaces are defined as a set $\mathcal{V}$ with 2 operations:

$$+ : \mathcal{V} \times \mathcal{V} -> \mathcal{V}$$
$$\cdot : R \times \mathcal{V} -> \mathcal{V}$$

A vector space is a set V of elements on which we have two operations + and $\cdot$ defined by the following properties:

- If u and v are any elements in V, then $u + v$ is in V and we say that V is closed under addition: $u + v = v + u; (u + v) + w = u + (v + w); \exists e = 0 \quad such \quad that \quad u + 0 = 0 + u$; for every u in V there exists an element $-u$ such that $u + -u = +u + u = 0$. This element $-u$ is known as additive inverse of u.
- If u is any element in V and λ is any real number, then $\lambda \cdot u \in V$: $c \cdot (u + v) = c \cdot u + c \cdot v; (c + d) \cdot u = c \cdot u + d \cdot u; c \cdot (d \cdot u) = (cd) \cdot u; 1 \cdot u = u \forall u \in V$

The elements $x \in V$ are vectors. The neutral element is zero vector and inner operation + is vector addition. The outer operation $\cdot$ is known as multiplication by scalars.

Vector subspaces A non-empty subset S of a vector space V is a subspace of V if it is also a vector space with respect to the same vector addition and scalar multiplication as V. If we assume that S is a non-empty subset of V, then it satisfies the following properties:

- $S \neq 0$;
- If u and v are vectors in S, then $u + v \in S$;
- If u is a vector in S, then for every λ, then $\lambda x \in S$

Linear Independence After doing all addition and multiplications, we end up in the Vector space. It is possible to find a set of vectors, known as basis, that can be used to represent every vector in the vector space by adding them together and scaling them. However, before basis, it is important to understand the idea of linear independence.

If vector x can be represented as $v = k_1 v_1 + k_2 v_2 + ... + k_n v_n$, then x is a linear combination of vectors $v_1, ..., v_k$

We say that vectors $v_1, v_2, v_3, ..., v_n \in V$ are linearly independent if the only scalars that satisfy the equation $k_1 \cdot v_1 + k_2 \cdot v_2 + ... + k_n \cdot v_n = 0$ are $k_1 = k_2 = k_3 = ...k_4 = 0$. Linearly independent vectors have no redundancy.

We say that vectors $v_1, v_2, v_3, ..., v_n \in V$ are linearly dependent if scalars that satisfy the equation $k_1 \cdot v_1 + k_2 \cdot v_2 + ... + k_n \cdot v_n = 0$ are $k_1, k_2, k_3, ..., k_n$ are not all zero (at least one $k_n \neq 0$.

The following properties can help to find if vectors are l. independent faster:

- If at least one of the vectors $x_1, x_2, ..., x_k$ is 0 then they are linearly independent. The same holds if two vectors are identical
- The vectors $x_1, x_2, ..., x_k : x_i \neq 0, i = 1, ..., k, k >= 2$ are linearly dependent iff one of them is a linear combination of others. For example, if one vector is a multiple of another vector; $x_i = \lambda x_j$, then the set of vectors $x_1, x_2, ..., x_k : x_i \neq 0, i = 1, ..., k$ is linearly dependent
- A practical way of checking whether vectors $x_1, ..., x_k \in V$ are linearly independent is to use Gaussian elimination until we get to REF/RREF. The pivot columns indicate the vectors, which are linearly independent of the vectors on the left; non-pivot columns can be expressed as linear combinations of the pivot columns. *All columns vectors are linearly independent iff all columns are pivot columns. If at least one column is non-pivot, then vectors are linearly dependent.*

In vector space V, m linear combinations of k vectors $x_1, ..., x_k$ are linearly dependent if $m > k$.

Basis and Rank

Generating Set and Span: If *every* vector in a space vector V can be written as a linear combination of $v_1, v_2, ..., v_k$, we say that V is spanned or generated by $v_1, v_2, ..., v_k$ and call the set of vectors $v_1, v_2, ..., v_k$ a spanning set for V. If you write out a linear combination and you find scalars, then the given vectors satisfy the spanning requirement.

Every spanning set for V that is linearly independent is a minimal spanning set. Such a set is called a basis of V. To formalize the definition of basis: A set of vectors $v_1, v_2, ..., v_k$ in a vector space V is called a basis for V if:

- Vectors are linearly independent
- Vectors span V
- It's a maximal linearly independent set of vectors; adding any other vector to this set will make it linearly dependent

The standard basis are combination of $1s, 0s$. So for R^2, a standard basis: $\mathcal{B} = [1, 0], [0, 1]$.

If we consider finite-dimensional vector spaces V. In this case, the dimension of V is the number of basis vectors of V, and we write $dim(V)$. If $U \subseteq V$ is a subspace of V, then $dim(U) = dim(V)$ iff $U = V$. The dimension of a vector space can be though of as the number of vectors in a basis for V. A basis of a subspace $U = span x_1, x_2, ..., x_m \subseteq R_n$ can be found by executing the following steps:

- Write the spanning vectors as columns of a matrix A
- Determine the REF of A
- Spanning vectors associated with the pivot columns are a basis of U

Rank

The number of linearly independent columns of a matrix $A \in R^{m \times n}$ equals the number of linearly independent rows and is called the rank of A and is denoted by $rank(A)$.

Rank has the following properties:

- $rank(A) = rank(A^T)$ the column rank is equal to row rank
- The columns of $A \in R^{m \times n}$ span a subspace $U \subseteq R^m$ with $dim(U) = rank(A)$. This subspace is known as image or range. A basis of U can be found after Gaussian elimination
- The rows of $A \in R^{m \times n}$ span a subspace $W \subseteq R^n$ with $dim(W) = rank(A)$. The basis of W can be found by applying Gaussian elimination to A^T
- For $A \in R^{n \times n}$, A is invertible iff $rank(A) = n$
- For $A \in R^{m \times n}$ and $b \in R^m$, it holds that the linear equation system $Ax = b$ can be solved iff $rank(A) = rank(A|b)$, where $A|b$ denotes the augmented system
- For $A \in R^{m \times n}$ the subspace of solutions for $Ax = 0$ possesses dimension $n - rank(A)$. This subspace is known as kernel or null space

- Matrix $A \in R^{m \times n}$ is full row rank when each of the rows of the matrix are linearly independent and full column rank when each of the columns of the matrix are linearly independent.

.

Linear Mappings

A linear transformation between vector spaces is a special transformation (mapping) which preserves the fundamental linear algebra operations - scalar multiplications and vector addition.

A mapping $T : V \longrightarrow W$ is called a linear transformation if for all vectors u, v in the vector space V and for any scalar k we have:

- $T(u + v) = T(u) + T(v)$; T preserves vector addition
- $T(ku) = kT(u)$; T preserves scalar multiplication

Linear maps are also known as homomorphism.

We can represent (and usually do) linear mappings as matrices. There are three linear transformations that need to be mentioned:

- One-to-one (injective) transformation; It's a linear transformation T in which every vector in the domain arrives at a different vector in the range under T. Mathematically, it is defined as: Let $T : V \longleftarrow W$ be a linear transform, u, v be in the domain V. The transform T is $1 - to - 1$ if: $u \neq v$ and $T(u) \neq T(v)$. At the same time, we can say that $T(u) = T(v)$ and $u = v$ because in that case, we start with the same vectors
- Surjective: Also known as onto transformation. Onto transformation is when all the information carried over by T fills the whole arrival vector space W. Mathematically, it is defined as: Let $T : V \longleftarrow W$ be a linear transform. The tranform T is onto iff for every w in the arrival vector space W there exists at least one v in the start vector space V such that: $w = T(v)$. If we show via range, then $range(T) = W$. This means that arriving vector of T fits all of W; $range(T) = W$
- If transformation is both 1-to-1 and onto, then it's known as bijective.

For inverse transformations to exist, they need to be bijective. There are more special cases of linear transformations:

- Isomorphism: If linear transformation is invertible, then vector spaces $V \longrightarrow W$ are isomorphic. Such a transformation is known as *isomorphism*. An isomorphism between vector spaces mean that these spaces are identical from a mathematical viewpoint, even though they are different spaces
- Endomorphism: A vector space that maps V to itself is known as endomorphism of V. It is a linear mapping (homomorphism) from an object to itself. $T : V \longrightarrow V$
- Automorphism: A biijective endomorphism

Finite-dimensional vector spaces V and W are isomorphic iff $dim(V) = dim(W)$. This means that there exists a linear, bijective mapping between two vector spaces of the same dimension.

consider vector spaces V, W, X. Then

- For linear mapping $T : V \longrightarrow W$ and $\Phi : W \longrightarrow X$, the mapping $T \circ \Phi : V \longrightarrow X$ is also linear
- If $T : V \longrightarrow W$ is an isomorphism, then $T^{-1} : W \longrightarrow V$ is an isomorphism, too
- If : $V \longrightarrow W$ and $\Phi : W \longrightarrow X$ are linear, then $T + \Phi$ and $\lambda\Phi$ are linear

Matrix Representation of Linear Mappings: We can write a basis of $n-$dimensional vector space V. Therefore, we define $B = (b_1, ..., b_n)$ and call this $n - tuple$ of an ordered basis of V.

Coordinates: Consider a vector space V and an ordered basis B. For any $x \in V$, we obtain a unique representation (linear combination)

$$x = \alpha_1 b_1 + ... + \alpha_n b_n$$

where α_n are the coordinates of x with respect to B, and vector α is a coordinate vector of x with respect to the ordered basis B.

Basis vectors are coordinate axis. Cartesian coordinate system in two dimensions is spanned by canonical basis vectors e_1, e_2.

Transformation matrix: Let A be an $m \times n$ matrix. The matrix transformation associated to A is the transformation: $T : R^n \longrightarrow R^m$ defined by $T(x) = Ax$. This is the transformation that takes a vector x in R^n to the vector Ax in R^m. A is a transformation matrix of T.

If $\hat{x}$ is the coordinate vector of $x \in V$ with respect to B and $\hat{y}$ is the coordinate vector of $y = T(x) \in W$ with respect to C (basis in W), then$\hat{y} = A_T \cdot \hat{x}$. This means that the transformation matrix can be used to map coordinates with respect to an ordered basis in V to coordinates with respect to an orded basis in W

Change of basis

Image and Kernel The image and kernel of a linear mapping are vector subspaces with certain properties. Kernel is also known as null space.

One of the most important results in linear algebra is :$dim(ker(T)) + dim(range(T)) = dim(V)$, which is valid for linear transformation $T : V \longrightarrow W$ A matrix, A can be used to represent a linear transformation, so finding the kernel means finding all of the vectors x such that $Ax = 0$. The linear system $Ax = b$ has a solution iff vector b is in the range (image) of linear transformation.

The set of vectors in the starting vector space which are transformed to the zero vector is called the **kernel** of the transformation and is denoted by ker(T). $ker(T) = 0$ is a null space of matrix A. If the kernel is just the zero vector, then linear transform has an inverse and all of the information was carried over (all of vectors in V is transferred to W). The official definition is:

Let $T : V \longrightarrow W$ be a linear transform (map). The set of all vectors v in V that are transformed to the zero vector in W is called the kernel of T, denoted by $ker(T)$. It is the set of vectors v in V such that $T(v) = 0$. $dim(ker(T))$ is known as nullity of T and denoted by nullity(T). If $T(u) = Au$ is a linear transformation then $ker(T)$ is the general solution u such that $Au = 0$.

Kernel of linear mapping T can also be expressed as: $ker(T) := T^{-1}(0_W) = v \in V : T(v) = 0_W$. Kernel is the general solution to the homogeneous system of linear equations $Ax = 0$ and captures all possible linear combinations of the elements in R^n that produce $0 \in R^m$. Kernel is a subspace of V. Kernel focuses on the relationship among the columns, and we can use it to determine how we can express a column as a linear combination of other columns.

The image/range is defined as: Let $T : V \longrightarrow W$ be a linear transform. The range of the linear transform is the set of all output vectors $w \in W$ for which there are inputs vectors $v \in V$ such that $T(v) = w$. We can write range(T) as $range(t) = T(v)|v \in V$.

We call V, W domain and codomain of T. The difference between kernel and range is that kernel of transformation lies in domain (start) and range lies in co-domain (arrival)

Both kernel and range compose an important theorem; rank-nullity theorem. It can be defined as: For vectors spaces V, W and a linear mapping T it holds that $dim(ker(T)) + dim(range(T)) = dim(V)$. The following properties are consequences of rank-nullity theorem

- If dim(range(T)) $<$ dim (V), then kernel(T) is non-trivial ($\neq 0$)
- If A is a transformation matrix with respect to an ordered basis and $dim(Im(T)) < dim(V)$, then the system of linear equations $Ax = 0$ has infinitely many solutions
- if $dim(V) = dim(W)$m then the following is true, since $Im(T) \subseteq W$: T is injective; T is surjective; T is bijective

ANALYTIC GEOMETRY

Norms

The length of a vector is the distance of the end of this directed line segment from the origin. Formally, a norm on a vector space V is a function: $|| \cdot || : V \longrightarrow R, x \longrightarrow ||x||$, which assigns each vector x its length $||x|| \in R$. The following properties hold for all scalars and $x, y \in R$:

- Absolutely homogeneous: $||\lambda x|| = |lambda|||x||$
- Triangle inequality: $||x + y|| \leq ||x|| + ||y||$
- Positive definite: $||x|| \geq 0$ and $||x|| = 0$, iff $x = 0$

L_1 *norm.* Also known as Manhattan norm. $||x||_1 = \sum_{i=1}^{n} |x_i|$ L_2 *norm.* Also known as Euclidean norm. $||x||_2 = \sqrt{\sum_{i=1}^{n} |x_i^2|}$

Inner products

A major purpose is to determine if vectors are orthogonal to each other. An answer is a scalar. An inner product is denoted as $< u, v >$, where u, v are vectors. An inner product on a real vector space V is an operation which assigns to each pair of vectors, u, v, a unique real number $< u, v >$ which satisfies the following axioms for all vectors u, v, w in V and all scalars c, k

- $< u, v >=< v, u >$. Known as commutative law (also symmetry)

- $< u + v, w >=< u, w > + < v, w >$. Known as distributive law
- $< ku, v >= k < u, v >$
- $< u, u > \geq 0$ and we have $< u, u >= 0$, iff $u = 0$
- Bilinearity: $< cu + dv; w >= c < u; w > +d < v; w >$

Symmetric, Positive Definite Matrices

Positive definite matrices are symmetric matrices whose eigenvalues are all positive. Positive semidefinite matrices are matrices whose eigenvalues are non-negative. The idea of symmetric positive semidefinite matrices is key in the definition of kernels.

It is easy to check if matrix is pd. If it is, the relationship $\forall x \in V \setminus 0 : x^T Ax > 0$. So, for example if $A = [9, 6; 6, 5]$, then $x^T Ax = [x_1, x_2] \times [A] \times [x_1; x_2] = 9x_1^2 + 12x_1x_2 + 5x_2^2$. For any $x \in V \setminus 0$ the condition of positivity is valid.

Length and Distances

Some but not all of the norms are based on inner products. Manhattan one is not. But Cauchy-Schwarz inequality is: For an inner product vector space, the induced norm satisfied the CS inequality:

$$|u \cdot v| \leq ||u||||v||$$

Distance and Metric

$$d(x, y) := ||x - y|| = \sqrt{< x - y, x - y >}$$

If we use the dot product as the inner product, then the distance is called Euclidean distance. The mapping:

$$d : V \times V \longrightarrow R$$
$$(x, y) \longmapsto d(x, y)$$

is called a metric.

A metric d satisfies the following:

- d is positive definite: $d(x, y) > 0$ and $d(x, y) = 0$ iff $x = y$
- d is symmetric; $d(x, y) = d(y, x)$
- triangle inequality: $d(x, z) \leq d(x, y) + d(y, z)$

Angles and Orthogonality

$$cos\omega = \frac{< x, y >}{||x||||y||}$$

This angle tells us how similar their orientations are.

Two vectors x and y are orthogonal iff $< x, y >= 0$ and in that case $x \perp y$.

Vector can be normalized via $\hat{u} = \frac{u}{||u||}$. Normalized vectors that are orthogonal are orthonormal; $||x|| = 1 = ||y||$.

$0 - vector$ is orthogonal to every vector in vector space.

Orthogonal matrix A square matrix $A \in R^{n \times n}$ is an orthogonal matrix iff its columns are orthonormal so that:

$$AA^T = I = A^T A$$

From this equation, it follows that $A^{-1} = A^T$

Orthonormal basis

Basis vectors are orthogonal to each other and length of each basis is 1. $B = e_1, e_2$, where $e_1 = [1, 0]; e_2 = [0, 1]$ is an example of orthonormal basis.

We can create a orthogonal basis for any finite dimensional vector space, given any arbitrary basis. The process is known as Gram-Schmidt process.

Orthogonal Complement

Orthogonal complements can be used to describe hyperplances in $n-$dimensional vector and affine spaces.

Consider $D-$ dimensional vector space V and $M-$dimensional subspace $U \subseteq V$. Then its orthogonal complement $U^{\perp}$ is a $(D - M)$-dimensional subspace of V and contains all vectors in V that are orthogonal to every vector in U.
COMPLETE!!!

Orthogonal projections

In ML, we often deal with high-dimensional data. However, it's quite often the case that only a small number of dimensions possess the most information needed. By projecting the original high-dimensional data onto a lower-dimensional feature space, we can work in this lower-dimensional feature space to learn more about dataset. Project is defined as: Let V be a vector space and $U \subseteq V$ a subspace of V. A linear mapping π: $V \longrightarrow U$ is called a projection if $\pi^2 = \pi$. Since linear mappings can be expressed by transformation matrices, this definition applies equally to a special kind of transformation matrices, the *projection matrices* P_π, which exhibit the property that $P_\pi^2 = P_\pi$

COMPLETE!!!

Rotations

A rotation is a linear mapping that rotates a plane by an angle θ about the origin. For a positive angle $\theta > 0$, we rotate in a counterclockwise direction. Important applications are robotics and computer graphics.

If we consider a standard basis in R^2, then we can rotate this coordinate system by angle θ. Rotated vectors are still linearly independent and still are basis of R^2. Rotations Φ are linear mappings and we can express them by a

rotation matrix $R(\theta)$. If we rotate ccw, then we can define a new linear mapping as: $\Phi(e_1) = [cos(\theta); sin(\theta)]; \Phi(e_2) = [-sin(\theta); cos(\theta)]$. Rotation matrix is $R(\theta) = [\Phi(e_1) \quad \Phi(e_2)]$.

MATRIX DECOMPOSITION

Mapping and transformations of vectors can be conveniently described as operations performed by matrices. Determinants and eigenvalues can characterize the overall properties of matrices. This chapter also discusses the decomposition of matrices and how these decomposition are useful for matrix approximations.

Determinant and Trace

Determinant is a mathematical object in the analysis and solution of systems of linear equations. Only defined for square matrices. Determinant is a real number.

- For any square matrix $A \in R^{n \times n}$ it holds that A is invertible iff $det(A) \neq 0$
- A square matrix T an upper-triangular matrix if matrix is 0 below the diagonal (and if 0 above diagonal, then lower-triangular). For a triangular matrix, determinant is a product of diagonal elements $det(T) = \prod_{i=1}^{n} T_{ii}$
- If A is invertible, then $det(A^{-1}) = \frac{1}{det(A)}$
- A square matrix $A \in R^{n \times n}$ has $det(A) \neq 0$ only if $rank(A) = n$. In other words, A is invertible only if it has a full rank (rows/columns are linearly independent).
- Trace of square matrix is defined as $tr(A) = \sum_{i=1}^{n} a_{ii}$.

Using determinants and traces, we can describe matrix A in terms of a polynomial, using the principle of characteristic equation: $p_A(\lambda) = det(A - \lambda I = c_0 + c_1\lambda + c_2\lambda^2 + c_{n-1}\lambda^{n-1} + (-1)^n\lambda^n$, where $c_0, ...$ is the characteristic polynomial of A. It allows us to compute eigenvalues and eigenvectors.

Eigenvectors and eigenvalues

Let A be a square matrix. Then λ is the eigenvalue of A and x is the corresponding eigenvector of A, if $Ax = \lambda x$. This equation is known as eigenvalue equation.

Almost all vectors change direction, when they are multiplied by A. Eigenvectors don't. Eigenvectors don't change orientation of original matrix A, just shrinks or expands the matrix. We can find eigenvalues from the characteristic polynomial, where roots are eigenvalues via $det(A - \lambda I) = 0$.

Eigenvectors are linearly independent. If a square matrix $A \in R^{n \times n}$ has less than n linearly independent eigenvectors.

The determinant of a square matrix $A \in R^{n \times n}$ is the product of its eigenvalues: $det(A) = \prod_{i=1}^{n} \lambda_i$.

Collinear vectors are vectors that parallel to one line or lying on one line. Two collinear vectors are co-directed if they have the same direction.

Cholesky Decomposition

We can use CD to decompose a symmetric, positive definite matrix. A symmetric, positive definite matrix is the one with all positive eigenvalues that is symmetric matrix (square matrix that is equal to its transpose). These types of matrices can be factorized into a product $A = LL^T$, where L is a lower-triangular matrix with positive

diagonal elements. L is a Cholesky of A and its unique. Cholesky factorization allows us to perform a linear transformation of random variables, which is heavily exploited when computing gradients in such models as variational auto-encoders.

Eigendecomposition and Diagonalization

A square matrix $A \in R^{n \times n}$ can be factored into $A = PDP^{-1}$, where $P \in R^{n \times n}$ and D is a diagonal matrix, whose diagonal entries are eigenvalues of A. P^{-1} transforms a basis change from the standard basis into the eigenbasis. This identifies the eigenvectors p_i onto the standard basis vectors e_i. D scales the vectors along these axes by the eigenvalues λ_i. P transforms these scaled vectors back into the standard coordinated yielding $\lambda_i p_i$.

This approach can also be used for raising a matrix into the power: $A^k = (PDP^{-1})^k = PD^kP^{-1}$

A symmetric matrix $S \in R^{n \times n}$ can always be diagnolized.

Singular Value Decomposition

Can be applied to all matrices, not only square ones. $A = U \sum V^T$. Compared to eigendecomposition:

- Vectors in eigendecomposition matrix P are not necessarily orthogonal; change of basis is not a simple rotation and scaling. Vectors in U and V are orthornormal, so they do represent rotation
- Both eigendecomp. and SVD are compositions of three linear mapping: change of basis in domain; independent scaling of each new basis vector and mapping from domain to codomain; change of basis in the codomain.

VECTOR CALCULUS

Differentiation

Difference quotient computes the slope of secant line through two points on the graph of f: $\frac{dy}{dx} = \frac{f(x+h)-f(x)}{h}$.

We can identify directions as unit vectors, those vectors whose length is 1; $||u||$. In that case we have $\frac{dy}{dx} = \frac{f(a+hu)-f(a)}{h}$, which means that we have a rate of change of $x \longrightarrow a$ in the direction of u. We can also rewrite the directional derivative as the dot product of the gradient and direction: $D_u f(a) = \nabla f(a)u$. The gradient $\nabla f(a)$ is a vector in a certain direction. Let u be any direction, and let θ be the angle between the vectors $\nabla f(a)$ and u. As the result, the formula above can be rewritten as: $\nabla f(a) \cdot u = ||\nabla f(a)|| \cdot cos(\theta)$. As the result, the largest directional derivative $\nabla f(a) \cdot u$ happens when $\theta = 0$, or in other words when u is in direction of the gradient $\nabla f(x)$. That is known as direction of greatest ascent. If we $\theta = 180$, then we get the greatest descent. Essentially, there are three major cases:

- $\theta = 0$. u points in the direction of $\nabla f(a)$. The rate of change $D_u f(a)$ is assumed to be maximum $||\nabla f(a)|| \geq 0$. This is the direction of steepest ascent.
- $\theta = \pi$. u points in the direction of $-\nabla f(a)$. The rate of change $D_u f(a)$ is assumed to be minimum $-||\nabla f(a)|| \leq 0$. This is the direction of steepest descent.
- $\theta = \pi/2| - \pi/2$. u points in the direction of that is perpendicular$\nabla f(a)$. The rate of change $D_u f(a)$ 0.

The direction of steepest descent is the vector $-\nabla f(x_0)$. The goal is to find t that we can use to find $x_{k+1} = x_k - t\nabla f(x_0)$. To do that:

- Find $\nabla f(x)$ at the initial point x_0

- Minimize $g(t) = f(x(t))$
- Find t and calculate $x_{k+1} = x_k - t_k \nabla f(x_0)$. This t_k is a stepsize.

If we want to know how f changes with respect to several variables, then unit vector becomes $u =< a, b >$ and directional derivative becomes: $D_u f(x_0, y_0) = lim_{h \to 0} \frac{f(x_0+ah, y_0+bh) - f(x_0, y_0)}{h} = \nabla f(x_0, y_0) \cdot u$.

Taylor Series

A Taylor polynomila of degree n is an approximation of a function, which does not need to be a polynomial. It is useful in ML because for some functions working directly with $f(x)$ is challenging and it's easier to work with Taylor series approximation.

$T(x) = \sum_{k=0}^{\inf} \frac{f^(k) \cdot x_0}{k!}(x - x_0)^k$. For $x_0 = 0$, we have Maclaurin series.

Jacobian matrix is the collection of first-order partial derivatives. Jacobian determinant is useful in the context of training DNNs using reparametrization trick, also called infinite perturbation analysis.

Higher-order Derivatives

Hessian is the collection of all second-order partial derivatives.

Linearization and Multivariate Taylor Series

Linearization is the process of taking the gradient of a nonlinear function with respect to all variables and creating a linear representation at that point. It is required for certain types of analysis such as stability analysis, solution with a Laplace transform, and to put the model into linear state-space form. Linearization is important because linear functions are easier to deal with. Using linearization, one can estimate function values near known points.

If f is differentiable at $x = a$ (center of the approximation), then the equation of the tangent line is:

$$L(x) = f(a) + f'(a)(x - a)$$

In ML, we often need to compute expectation: $E_x[f(x)] = \int f(x)p(x)dx$. Even if $p(x)$ is in a convenient form, this integral cannot be solved analytically and Taylor series expansion is one way of finding an approximate solution: Assuming $p(x) = \mathcal{N}(\mu, \sum)$ is Gaussian, then the first-order Taylor series expansion around μ locally linearizes the nonlinear function f