Recap: math

Statistical Methods in NLP 2
ISCL-BA-08

Çağrı Çöltekin
ccoltekin@sfs.uni-tuebingen.de

University of Tübingen
Seminar für Sprachwissenschaft

Summer Semester 2025

version: none @2025-04-25

Linear algebra

Linear algebra is the field of mathematics that studies *vectors* and *matrices*.

- A vector is an ordered sequence of numbers

$$\mathbf{v} = (6, 17)$$

- A matrix is a rectangular arrangement of numbers

$$\mathbf{A} = \begin{bmatrix} 2 & 1 \\ 1 & 4 \end{bmatrix}$$

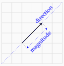
- A well-known application of linear algebra is solving a set of linear equations

$$\begin{aligned} 2x_1 + x_2 &= 6 \\ x_1 + 4x_2 &= 17 \end{aligned} \iff \begin{bmatrix} 2 & 1 \\ 1 & 4 \end{bmatrix} \times \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 6 \\ 17 \end{bmatrix}$$

Vectors

- Vectors are objects with a magnitude and a direction
- We represent vectors with an ordered list of numbers $\mathbf{v} = (v_1, v_2, \dots v_n)$
- The number n (the number of elements or entries of the vector) is its dimension
- We often call an n dimensional vector as n-vector
- The vector of n real numbers is said to be in $\mathbb{R}^n$ ($\mathbf{v} \in \mathbb{R}^n$)
- Typical notation for vectors:

$$\mathbf{v} = \vec{v} = (v_1, v_2, v_3) = \langle v_1, v_2, v_3 \rangle = \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix}$$

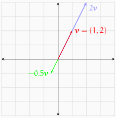


Multiplying a vector with a scalar

- For a vector $\mathbf{v} = (v_1, v_2)$ and a scalar a,

$$a\mathbf{v} = (av_1, av_2)$$

- multiplying with a scalar 'scales' the vector
- We can use the notation $a\mathbf{1}$ for a vector whose all entries are a



Vector addition and subtraction

For vectors $\mathbf{v} = (v_1, v_2)$ and $\mathbf{u} = (w_1, w_2)$

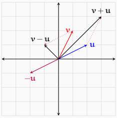
- $\mathbf{v} + \mathbf{u} = (v_1 + w_1, v_2 + w_2)$

$$(1, 2) + (2, 1) = (3, 3)$$

- $\mathbf{v} - \mathbf{u} = \mathbf{v} + (-\mathbf{u})$

$$(1, 2) - (2, 1) = (-1, 1)$$

- For any vector $\mathbf{v}$, $\mathbf{v} + \mathbf{0} = \mathbf{v}$



Properties of vector operations

- Vector addition and scalar multiplication is commutative

$$\mathbf{u} + \mathbf{v} = \mathbf{v} + \mathbf{u}$$

$$a\mathbf{u} = \mathbf{u}a$$

- Scalar multiplication and vector addition also show the following distributive properties

$$a(\mathbf{u} + \mathbf{v}) = a\mathbf{u} + a\mathbf{v}$$

$$(a + b)\mathbf{v} = a\mathbf{v} + b\mathbf{v}$$

Dot (inner) product

- Dot product is an operation between two vectors with same dimensions

$$\mathbf{u} \cdot \mathbf{v} = u_1 v_1 + u_2 v_2 + \dots + u_n v_n$$

- Calculate the dot products for the following vectors

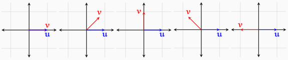
$$\begin{bmatrix} 4 \\ 3 \end{bmatrix} \cdot \begin{bmatrix} 3 \\ 4 \end{bmatrix} \quad \begin{bmatrix} 4 \\ -3 \end{bmatrix} \cdot \begin{bmatrix} -3 \\ 4 \end{bmatrix} \quad \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} \cdot \begin{bmatrix} -2 \\ -4 \\ -6 \end{bmatrix} \quad \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} \cdot \begin{bmatrix} 2 \\ 4 \\ 6 \end{bmatrix} \quad \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \cdot \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}$$

- Note that dot product is larger when the vectors are 'similar'

Properties of dot product

- *Commutativity* $\mathbf{u} \cdot \mathbf{v} = \mathbf{v} \cdot \mathbf{u}$
- *Distributivity with vector addition* $\mathbf{u} \cdot (\mathbf{v} + \mathbf{v}) = \mathbf{u} \cdot \mathbf{v} + \mathbf{u} \cdot \mathbf{u}$
- *Associativity with scalar multiplication* $(a\mathbf{u}) \cdot (b\mathbf{v}) = ab(\mathbf{u} \cdot \mathbf{v})$.
- Note that dot product is not associative, since the result of the dot product is not a vector, but a scalar

Dot product with unit vectors



$\mathbf{u} \cdot \mathbf{v} = 1$ $\quad \mathbf{u} \cdot \mathbf{v} = 0.7$ $\quad \mathbf{u} \cdot \mathbf{v} = 0$ $\quad \mathbf{u} \cdot \mathbf{v} = -0.7$ $\quad \mathbf{u} \cdot \mathbf{v} = -1$

- The dot product is larger if the vectors point to the similar directions

L2 norm

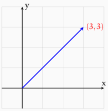
- Euclidean norm, or L2 (or L_2) norm is the most commonly used norm
- For $\mathbf{v} = (v_1, v_2, \dots v_n)$,

$$\begin{aligned} \|v\|_2 &= \sqrt{v_1^2 + v_2^2 + \dots v_n} \\ &= \sqrt{\mathbf{v} \cdot \mathbf{v}} \end{aligned}$$

- For example,

$$\|(3, 3)\|_2 = \sqrt{3^2 + 3^2} = \sqrt{18}$$

- L2 norm is the default, we often skip the subscript $\|\mathbf{v}\|$

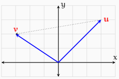


Euclidean distance

- Euclidean distance between two vectors is the L2 norm of their difference

$$D(\mathbf{u}, \mathbf{v}) = \|\mathbf{u} - \mathbf{v}\| = \sqrt{(-6)^2 + (-1)^2}$$

- Euclidean distance is a metric
  - symmetric $\|\mathbf{v} - \mathbf{u}\| = \|\mathbf{u} - \mathbf{v}\|$
  - non-negative
  - and obeys the triangle inequality $D(\mathbf{u}, \mathbf{v}) \leqslant D(\mathbf{u}, \mathbf{w}) + D(\mathbf{w}, \mathbf{v})$ for any $\mathbf{w}$



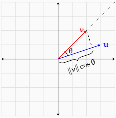
Cosine similarity

- The cosine of the angle between two vectors

$$\cos\theta = \frac{\mathbf{v} \cdot \mathbf{u}}{\|\mathbf{v}\| \cdot \|\mathbf{u}\|}$$

is called *cosine similarity*

- Unlike dot product, the cosine similarity is not sensitive to the magnitudes of the vectors
- The cosine similarity is bounded in range $[-1, +1]$



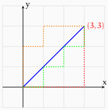
L1 norm

- Another norm we will often encounter is the L1 norm

$$\|v\|_1 = |v_1| + |v_2|$$

$$\|(3,3)\|_1 = |3| + |3| = 6$$

- L1 norm is related to Manhattan distance



Multiplying a matrix with a scalar

Similar to vectors, each element is multiplied by the scalar.

$$2\begin{bmatrix}2 & 1\\1 & 4\end{bmatrix} = \begin{bmatrix}2\times 2 & 2\times 1\\2\times 1 & 2\times 4\end{bmatrix} = \begin{bmatrix}4 & 2\\2 & 8\end{bmatrix}$$

Matrix addition and subtraction

Each element is added to (or subtracted from) the corresponding element

$$\begin{bmatrix}2 & 1\\1 & 4\end{bmatrix} + \begin{bmatrix}0 & 1\\1 & 0\end{bmatrix} = \begin{bmatrix}2 & 2\\2 & 4\end{bmatrix}$$

Note:

- Matrix addition and subtraction are defined on matrices of the same dimensions

Transpose of a matrix

Transpose of a $n \times m$ matrix is an $m \times n$ matrix whose rows are the columns of the original matrix.
Transpose of a matrix $\mathbf{A}$ is denoted with $\mathbf{A}^T$.

$$\text{If } \mathbf{A} = \begin{bmatrix}a & b\\c & d\\e & f\end{bmatrix}, \mathbf{A}^T = \begin{bmatrix}a & c & e\\b & d & f\end{bmatrix}.$$

Matrix–vector multiplication

- An $n \times m$ matrix can be multiplied with a m-vector to yield a n-vector
- Example

$$\begin{bmatrix}2 & 1 & 0\\1 & 0 & 1\end{bmatrix} \times \begin{bmatrix}0\\1\\0\end{bmatrix} = \begin{bmatrix}2\times 0 + 1\times 1 + 0\times 1\\1\times 0 + 0\times 1 + 1\times 0\end{bmatrix} = \begin{bmatrix}1\\0\end{bmatrix}$$

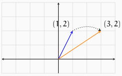
- One view of this operation: each entry in the resulting vector is a dot product (of rows of the matrix and the vector)
- Another: the result is a linear combination of the columns of the matrix (with the entries in the vector as coefficients)

$$0 \times \begin{bmatrix}2\\1\end{bmatrix} + 1 \times \begin{bmatrix}1\\0\end{bmatrix} + 0 \times \begin{bmatrix}0\\1\end{bmatrix} = \begin{bmatrix}1\\0\end{bmatrix}$$

Matrix multiplication transforms vectors

$$\begin{bmatrix}3 & 0\\0 & 1\end{bmatrix} \times \begin{bmatrix}1\\2\end{bmatrix} = \begin{bmatrix}3\\2\end{bmatrix}$$

- Matrices define a linear operator or function
- Linear transformations scale and/or rotate/reflect a vector



Transformations by non-square matrices

- Multiplying a vector with (compatible) rectangular matrix results in a vector with different dimensionality
- Example $\mathbb{R}^3 \to \mathbb{R}^2$

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

- Example $\mathbb{R}^3 \to \mathbb{R}^4$

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

- Multiplying a vector with a matrix trasnforms it into the 'column space' of the matrix

Dot product as matrix multiplication

In machine learning (and many other disciplines), we treat an n-vector as an $n \times 1$ matrix.
Then, the *dot product* of two vectors is

$$\mathbf{u}^T\mathbf{v}$$

For example, $\mathbf{u} = (2, 2)$ and $\mathbf{v} = (2, -2)$,

$$\begin{bmatrix}2 & 2\end{bmatrix} \times \begin{bmatrix}2\\-2\end{bmatrix} = 2\times 2 + 2\times -2 = 4 - 4 = 0$$

- This is a 1×1 matrix, but matrices and vectors with single entries are often treated as scalars

Question: What is the transformation performed by dot product?

Outer product

The *outer product* of two column vectors is defined as

$$\mathbf{v}\mathbf{u}^T$$

$$\begin{bmatrix}1\\2\end{bmatrix} \times \begin{bmatrix}1 & 2 & 3\end{bmatrix} = \begin{bmatrix}1 & 2 & 3\\2 & 4 & 6\end{bmatrix}$$

Note:

- The result is a matrix
- The vectors do not have to be the same length

Matrix multiplication

- if $\mathbf{A}$ is a $n \times k$ matrix, and $\mathbf{B}$ is a $k \times m$ matrix, their product $\mathbf{C}$ is a $n \times m$ matrix
- Elements of C, $c_{i,j}$, are defined as

$$c_{ij} = \sum_{\ell=1}^{k} a_{i\ell} b_{\ell j}$$

- Note: $c_{i,j}$ is the dot product of the i^{th} row of $\mathbf{A}$ and the j^{th} column of $\mathbf{B}$

Matrix multiplication

(demonstration)

$$\begin{pmatrix}a_{11} & a_{12} & \dots & a_{1k}\\a_{21} & a_{22} & \dots & a_{2k}\\\vdots & \vdots & \ddots & \vdots\\a_{n1} & a_{n2} & \dots & a_{nk}\end{pmatrix} \times \begin{pmatrix}b_{11} & b_{12} & \dots & b_{1m}\\b_{21} & b_{22} & \dots & b_{2m}\\\vdots & \vdots & \ddots & \vdots\\b_{k1} & b_{k2} & \dots & b_{km}\end{pmatrix}$$

$$c_{ij} = a_{i1}b_{1j} + a_{i2}b_{2j} + \dots a_{ik}b_{kj}$$

$$= \begin{pmatrix}c_{11} & c_{12} & \dots & c_{1m}\\c_{21} & c_{22} & \dots & c_{2m}\\\vdots & \vdots & \ddots & \vdots\\c_{n1} & c_{n2} & \dots & c_{nm}\end{pmatrix}$$

Alternative ways to think about matrix multiplication

If we have $\mathbf{AB} = \mathbf{C}$,

- Column vectors of $\mathbf{C}$, $\mathbf{c}_j = \mathbf{A}\mathbf{b}_j$
- Row vectors of $\mathbf{C}$, $\mathbf{c}_i^T = \mathbf{a}_i^T\mathbf{B}$
- $\mathbf{C}$ is also the sum of outer product of columns of $\mathbf{A}$ and rows of $\mathbf{B}$

$$C = \sum \mathbf{a}_i \mathbf{b}_i^T$$

Properties of matrix multiplication

- Associativity
$$(\mathbf{AB})\mathbf{C} = \mathbf{A}(\mathbf{BC})$$
- Distributivity
$$\mathbf{A}(\mathbf{B}+\mathbf{C}) = \mathbf{AB}+\mathbf{AC}$$
$$(\mathbf{A}+\mathbf{B})\mathbf{C} = \mathbf{AC}+\mathbf{BC}$$
- Multiplication by Identity
$$\mathbf{IA} = \mathbf{AI} = \mathbf{A}$$
- Matrix multiplication is not commutative $\mathbf{AB} \neq \mathbf{BA}$ (in general)
- Matrix multiplication and transpose
$$(\mathbf{AB})^T = \mathbf{B}^T\mathbf{A}^T$$

Row reduction and solving systmes of linear equations

$$\begin{matrix} x_1 & - & x_2 & = & -1 \\ 2x_1 & - & x_2 & = & 1 \end{matrix} \iff \begin{bmatrix} 1 & -1 \\ 2 & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$$

- We apply a set of *elementary row operations* to the *augmented matrix* to obtain an *upper triangle matrix*
$$\left[\begin{array}{cc|c} 1 & -1 & -1 \\ 2 & -1 & 1 \end{array}\right]$$
- Elementary row operations are
  - Multiply one of the rows with a non-zero scalar
  - Add (or subtract) a multiple of one row from another
  - Swap two rows

Solution with row reduction

$$\left[\begin{array}{cc|c} 1 & -1 & -1 \\ 2 & -1 & 1 \end{array}\right]$$

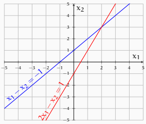
- Add $-2 \times$ row 1 to row 2
$$\left[\begin{array}{cc|c} 1 & -1 & -1 \\ 0 & 1 & 3 \end{array}\right]$$
- This corresponds to:
$$\begin{matrix} x_1 & - & x_2 & = & -1 \\ & & x_2 & = & 3 \end{matrix}$$

  where we already see $x_2 = 3$
- *Back-substituting* this in the first equation gives the same answer $x_1 = 2$

Solving systems of linear equations

Geometric interpretation (1)

- The solution is the intersection of the lines defined by the equations

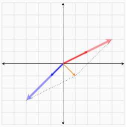


Solving systems of linear equations

Geometric interpretation (2)

- The solution satisfies the linear combination of the column vectors

$$2\begin{bmatrix} 2 \\ 1 \end{bmatrix} + 3\begin{bmatrix} -1 \\ -1 \end{bmatrix} = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$



Singular matrices and matrix rank

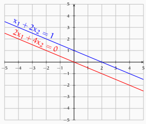
- If the elimination results in one or more rows with all zeros, the matrix is said to be *singular*
- This means – effectively – we have fewer equations than unknowns
- If a square matrix is not singular, we can find a unique solution for any right-hand side
- The systems of equations with a singular matrix results in either none or an infinite number of solutions
- The number of columns (or rows) with a pivot is called the *rank* of the matrix
- A non-singular square matrix is said to be full-rank

Systems of equations with singular matrices

- What is the rank of the following matrix?
$$\mathbf{A} = \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix}$$
- Can we solve $\mathbf{Ax} = \mathbf{b}$
  - for $\mathbf{b} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$?
  - for $\mathbf{b} = \begin{bmatrix} 3 \\ 6 \end{bmatrix}$?

Systems of equations with singular matrices

Demonstration of no solution

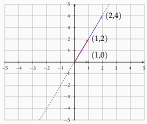


$$\begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$
$$\Rightarrow \begin{matrix} 2x_1 & + & x_2 & = & 1 \\ 4x_1 & + & 2x_2 & = & 0 \end{matrix}$$

- Lines are parallel to each other: no intersection, no solution

Systems of equations with singular matrices

Demonstration of no solution (another view)

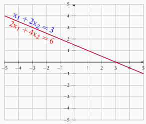


$$\begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$
$$\Rightarrow x_1 \begin{bmatrix} 1 \\ 2 \end{bmatrix} + x_2 \begin{bmatrix} 2 \\ 4 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

- All linear combinations of $\begin{bmatrix} 1 \\ 2 \end{bmatrix}$ and $\begin{bmatrix} 2 \\ 4 \end{bmatrix}$ bound to be on the dotted line: no linear combination can produce $\begin{bmatrix} 1 \\ 0 \end{bmatrix}$

Systems of equations with singular matrices

Demonstration of infinite number of solutions

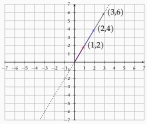


$$\begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 3 \\ 6 \end{bmatrix}$$
$$\Rightarrow \begin{matrix} 2x_1 & + & x_2 & = & 3 \\ 4x_1 & + & 2x_2 & = & 6 \end{matrix}$$

- Lines are identical: any point on the line is a solution

Systems of equations with singular matrices

Demonstration of infinite number of solutions (another view)



$$\begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 3 \\ 6 \end{bmatrix}$$
$$\Rightarrow x_1 \begin{bmatrix} 1 \\ 2 \end{bmatrix} + x_2 \begin{bmatrix} 2 \\ 4 \end{bmatrix} = \begin{bmatrix} 3 \\ 6 \end{bmatrix}$$

- There are many (x_1, x_2) combinations that satisfy the equation. An obvious one: $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$
- More?

Matrix inverse

- If we have a single linear equation with a single unknown: $ax = b$, the solution is
$$x = \frac{1}{a}b \quad \text{or} \quad x = a^{-1}b$$
- We can use an analogous method with systems of linear equations
$$\text{if} \quad \mathbf{Ax} = \mathbf{b} \quad \text{then,} \quad \mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$$
- Matrix inverse is only defined for square matrices (and not all square matrices are invertible)
- When it exists, $\mathbf{A}^{-1}\mathbf{A} = \mathbf{AA}^{-1} = \mathbf{I}$
- If a square matrix is invertible, a version of elimination can be used to find the inverse
  - Create the augmented matrix $[\mathbf{A}|\mathbf{I}]$
  - Use elementary row operations to obtain $[\mathbf{I}|\mathbf{B}]$
  - If successful, $\mathbf{B} = \mathbf{A}^{-1}$

Systems of equations with rectangular matrices
wide matrices (more columns than rows)

- This means $n \times m$ rectangular matrices with $n < m$,
- Note: the rank of such a matrix is always $\leqslant n$
- Exercise: solve

$$\begin{bmatrix} 4 & 2 & 4 \\ 2 & 2 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 10 \\ 4 \end{bmatrix}$$

- In this case we have
  - no solution if rank $r < n$ (number of rows)
  - infinitely many solution if rank $r = n$

Systems of equations with rectangular matrices
tall matrices (more rows than columns)

- This means $n \times m$ rectangular matrices with $m < n$,
- Note: the rank of such a matrix is always $\leqslant m$
- Exercise: solve

$$\begin{bmatrix} 4 & 2 \\ 2 & 2 \\ 4 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 10 \\ 4 \\ 4 \end{bmatrix}$$

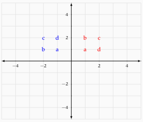
- In this case we have
  - a unique solution if the right-hand side is in the column space of the matrix
  - no solution otherwise
- We will work with this case more often

Determinant

- The determinant of a square matrix is a number that provides a lot of information about the matrix
  - Whether the matrix has an inverse or not
  - Calculating eigenvalues and eigenvectors
  - Solving systems of linear equations
  - Determining the (signed) 'change of volume' caused by the linear transformation defined by the matrix

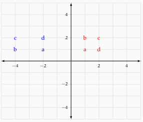
Determinant
example geometric interpretation (1)

- $A = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$
- $\det(A) = ?$



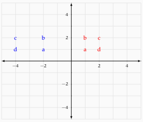
Determinant
example geometric interpretation (2)

- $A = \begin{bmatrix} 0 & -1 \\ 2 & 0 \end{bmatrix}$
- $\det(A) = ?$



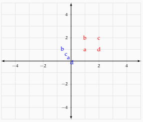
Determinant
example geometric interpretation (3)

- $A = \begin{bmatrix} -2 & 0 \\ 0 & 1 \end{bmatrix}$
- $\det(A) = ?$



Determinant
example geometric interpretation (3)

- $A = \begin{bmatrix} \cos 120 \\ \sin 120 \end{bmatrix} \times \begin{bmatrix} \cos 120 & \sin 120 \end{bmatrix} = \begin{bmatrix} 0.25 & -0.43 \\ -0.43 & 0.75 \end{bmatrix}$
- $\det(A) = ?$



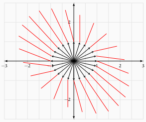
Eigenvalues and eigenvectors

- We can view any linear transformation as a combination of scaling and rotation (and reflection)
- The linear transformation defined by a matrix does not change the directions of some vectors, vectors in these directions are called the *eigenvectors*
- The scaling factor in these directions is called *eigenvalues*
- More formally, if v is an eigenvector of $\mathbf{A}$ with corresponding eigenvalue λ,

$$\mathbf{A}\boldsymbol{v} = \lambda\boldsymbol{v}$$

- Independent eigenvectors of a symmetric are orthogonal

Eigenvalues and eigenvectors
visualization



Diagonalization
(eigenvalue decomposition)

- An $n \times n$ with n independent eigenvalues can be *diagonalized* using eigenvalues and eigenvectors
- We take the matrix $\mathbf{S}$ whose columns are the eigenvalues of $\mathbf{A}$, and the diagonal matrix $\boldsymbol{\Lambda}$ with eigenvalues of $\mathbf{A}$, then

$$\mathbf{AS} = \mathbf{S}\boldsymbol{\Lambda}$$
$$\mathbf{A} = \mathbf{S}\boldsymbol{\Lambda}\mathbf{S}^{-1}$$
$$\mathbf{S}^{-1}\mathbf{AS} = \boldsymbol{\Lambda}$$

Matrix powers and matrix inverse

- Matrix powers can be easily calculated with diagonalization

$$\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$$
$$\mathbf{AA}\mathbf{x} = \lambda\mathbf{A}\mathbf{x}$$
$$\mathbf{A}^2\mathbf{x} = \lambda^2\mathbf{x}$$

- In general,

$$\mathbf{A}^2 = \mathbf{S}\boldsymbol{\Lambda}\mathbf{S}^{-1}\mathbf{S}\boldsymbol{\Lambda}\mathbf{S}^{-1}$$
$$= \mathbf{S}\boldsymbol{\Lambda}^2\mathbf{S}^{-1}$$
$$\mathbf{A}^k = \mathbf{S}\boldsymbol{\Lambda}^k\mathbf{S}^{-1}$$

- Inverse is also easy to obtain after eigendecomposition

$$\mathbf{A}^{-1} = \mathbf{S}\boldsymbol{\Lambda}^{-1}\mathbf{S}^{-1}$$

Singular Value Decomposition

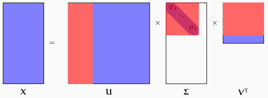
- Singular value decomposition (SVD) of an $n \times m$ matrix $\mathbf{X}$ is

$$\mathbf{X} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T$$

U is a $n \times n$ orthogonal matrix
Σ is a $n \times m$ diagonal matrix of singular values
V^T is a $m \times m$ orthogonal matrix.

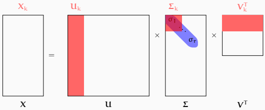
- Singular vectors in $\mathbf{U}$ are the eigenvalues of $\mathbf{XX}^T$
- Singular vectors in $\mathbf{V}^T$ are the eigenvalues of $\mathbf{X}^T\mathbf{X}$

Singular Value Decomposition



- Since $n - r$ rows and $m - r$ rows of Σ is 0, the decomposition does need the full matrices

Low rank estimation of a matrix



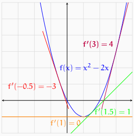
$X_k = U_k \Sigma_k V_k^T$ is the best rank k estimation of matrix X

Derivatives

- Derivative of a function $f(x)$ is another function $f'(x)$ indicating the rate of change in $f(x)$
- Alternatively: $f'(x) = \frac{df}{dx}(x)$
- When derivative exists, it determines the tangent line to the function at a given point
- Example from physics: velocity is the derivative of the position
- Our main interest:
  - the points where the derivative is 0.00 are the stationary points (maxima, minima, inflection points)
  - the derivative evaluated at other points indicate the direction and steepness of the curve defined by the function

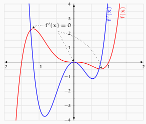
Example: derivatives

- $f'(x)$ is negative when $f(x)$ is decreasing, positive when it is increasing
- The absolute value of $f'(x)$ indicates how fast $f(x)$ changes when x changes
- $f'(x) = 0$ when at a *stationary point*
- $f'(a)$ is a (good) approximation to the $f(x)$ near the a



Derivatives and extrema

- Derivative of a function is 0 at minimum, maximum and inflection points
- Derivative is useful for optimization (minimization of maximization) problems
- We need additional tests to determine the type of critical points



Partial derivatives and gradient

- In ML, we are often interested in (error) functions of many variables
- A partial derivative is derivative of a multivariate function with respect to a single variable, noted $\frac{\partial f}{\partial x}$
- A very useful quantity, called *gradient*, is the vector of partial derivatives with respect to each variable

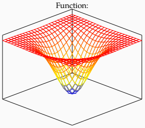
$$\nabla f(x_1, \ldots, x_n) = \left(\frac{\partial f}{\partial x_1}, \ldots, \frac{\partial f}{\partial x_n} \right)$$

- Gradient points to the direction of the steepest change
- Example: if $f(x, y) = x^3 + yx$

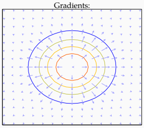
$$\nabla f(x, y) = (3x^2 + y, x)$$

Gradient visualization

Function:



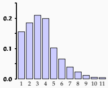
Gradients:



Probability mass function

Example: probabilities for sentence length in words

- *Probability mass function* (*PMF*) of a *discrete* random variable (X) maps every possible (x) value to its probability $(P(X = x))$.

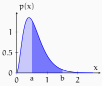


x	$P(X = x)$
1	0.155
2	0.185
3	0.210
4	0.194
5	0.102
6	0.066
7	0.039
8	0.023
9	0.012
10	0.005
11	0.004

Probability density function (PDF)

- Continuous variables have *probability density functions*
- $p(x)$ is not a probability (note the notation: we use lowercase p for PDF)
- Area under $p(x)$ sums to 1.00
- $P(X = x) = 0$
- Non zero probabilities are possible for ranges:

$$P(a \leqslant x \leqslant b) = \int_a^b p(x)dx$$



Joint and marginal probability

Two or more random variables form a *joint probability distribution*.

An example with letter bigrams:

	a	b	c	d	e	f	g	h	
a	0.04	0.02	0.02	0.03	0.05	0.01	0.02	0.06	0.23
b	0.01	0.00	0.00	0.00	0.01	0.00	0.00	0.01	0.04
c	0.02	0.00	0.00	0.00	0.01	0.00	0.00	0.01	0.05
d	0.02	0.00	0.00	0.01	0.02	0.00	0.01	0.02	0.08
e	0.06	0.02	0.01	0.03	0.08	0.01	0.01	0.07	0.29
f	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.01	0.02
g	0.01	0.00	0.00	0.01	0.02	0.00	0.01	0.02	0.07
h	0.08	0.00	0.00	0.01	0.10	0.00	0.01	0.02	0.22
	0.23	0.04	0.05	0.08	0.29	0.02	0.07	0.22	

Self information / surprisal

Self information (or *surprisal*) associated with an event x is

$$I(x) = \log \frac{1}{P(x)} = -\log P(x)$$

- If the event is certain, the information (or surprise) associated with it is 0.00
- Low probability (surprising) events have higher *information content*
- Base of the log determines the unit of information

2.00 bits
e nats
10.00 dit, ban, hartley

Entropy

Entropy is a measure of the uncertainty of a random variable:

$$H(X) = -\sum_x P(x) \log P(x)$$

- Entropy is the lower bound on the best average code length, given the distribution P that generates the data
- Entropy is average surprisal: $H(X) = E[-\log P(x)]$
- It generalizes to continuous distributions as well (replace sum with integral)

Entropy is about a distribution, while surprisal is about individual events

Pointwise mutual information

Pointwise mutual information (PMI) between two events is defined as

$$\mathrm{PMI}(x, y) = \log_2 \frac{P(x, y)}{P(x)P(y)}$$

- Reminder: $P(x, y) = P(x)P(y)$ if two events are independent PMI
  - 0 if the events are independent
  - + if events cooccur more than they would occur by chance
  - − if events cooccur less than they would occur by chance
- Pointwise mutual information is symmetric $\mathrm{PMI}(X, Y) = \mathrm{PMI}(Y, X)$
- PMI is often used as a measure of association (e.g., between words) in computational/corpus linguistics

Mutual information

Mutual information measures mutual dependence between two random variables

$$\mathrm{MI}(X, Y) = \sum_x \sum_y P(x, y) \log_2 \frac{P(x, y)}{P(x)P(y)}$$

- MI is the average (expected value of) PMI
- PMI is defined on events, MI is defined on distributions
- Note the similarity with the covariance (or correlation)
- Unlike correlation, mutual information is
  - also defined for discrete variables
  - also sensitive the non-linear dependence

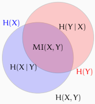
Conditional entropy

Conditional entropy is the entropy of a random variable conditioned on another random variable.

$$\begin{aligned} H(X \mid Y) &= \sum_{y \in Y} P(y) H(X \mid Y = y) \\ &= -\sum_{x \in X, y \in Y} P(x, y) \log P(x \mid y) \end{aligned}$$

- $H(X \mid Y) = H(X)$ if random variables are independent
- Conditional entropy is lower if random variables are dependent

Entropy, mutual information and conditional entropy



Cross entropy

Cross entropy measures entropy of a distribution P, under another distribution Q.

$$H(P, Q) = -\sum_x P(x) \log Q(x)$$

- It often arises in the context of approximation:
  - if we approximate the true distribution P with Q
- It is always larger than $H(P)$: it is the (non-optimum) average code-length of P coded using Q
- It is a common *error function* in ML for categorical distributions

Note: the notation $H(X, Y)$ is also used for *joint entropy*.

Perplexity

Perplexity is the exponential version of (cross) entropy:

$$\mathrm{PP}(X) = 2^{H(X)}$$

- Perplexity 'undoes' the logarithmic scaling
- Perplexity easier to interpret in some contexts
- Especially for language models, its interpretation is the average 'branching factor'

Predict the next word: ⟨S⟩ The perplexity of a random variable ⟨/S⟩

KL-divergence / relative entropy

For two distribution P and Q with same support, Kullback–Leibler divergence of Q from P (or relative entropy of P given Q) is defined as

$$D_{KL}(P \| Q) = \sum_x P(x) \log_2 \frac{P(x)}{Q(x)}$$

- D_{KL} measures the amount of extra bits needed when Q is used instead of P
- $D_{KL}(P \| Q) = H(P, Q) - H(P)$
- Used for measuring the difference between two distributions
- Note: it is not symmetric (not a distance measure)

Final remarks

- The knowledge most if these topics are assumed, and important for understanding modern methods in ML
- For math (and also for programming), it is difficult to master the concepts with passive participation. You need to practice

Next:

- Recap: regression
- Recap: classification

Some sources of information

On Linear algebra:

- A classic reference book in the field is Strang (2009)
- Alsto video lectures from the author:
  `https://www.youtube.com/playlist?list=PL0-GT3co4r2y2YErbmuJw2L5tW4Ew205B`
- A nice video series by 3Blue1Brown (also some calculus):
  `https://www.youtube.com/playlist?list=PLZHQObOWTQDMsr9K-rj53DwVRMYO3t5Yr`
- Shifrin and Adams (2011) and Farin and Hansford (2014) are textbooks with a more practical/graphical orientation.
- Cherney, Denton, and Waldron (2013) and Beezer (2014) are two textbooks that are freely available.

Some sources of information (cont.)

On probability theory:

- Please read, and follow the exercises in Goldwater (2018)
- See Grinstead and Snell (2012) a more conventional introduction to probability theory. This book is also freely available
- For an influential, but not quite conventional approach, see Jaynes (2007)

For information theory:

- MacKay (2003): a freely available textbook with further topics in ML, also includes probability theory,
- Shannon (1948)

In general for math:

- Many open books on math:
  `https://www.openculture.com/free-math-textbooks`

Some sources of information (cont.)

Beezer, Robert A. (2014). *A First Course in Linear Algebra*. version 3.40. Congruent Press. ISBN: 9780984417551. URL: http://linear.ups.edu/.

Cherney, David, Tom Denton, and Andrew Waldron (2013). *Linear algebra*. math.ucdavis.edu. URL: https://www.math.ucdavis.edu/~linear/.

Farin, Gerald E. and Dianne Hansford (2014). *Practical linear algebra: a geometry toolbox*. Third edition. CRC Press. ISBN: 978-1-4665-7958-3.

Goldwater, Sharon (2018). *Basic probability theory*. URL: https://homepages.inf.ed.ac.uk/sgwater/teaching/general/probability.%20pdf.

Grinstead, Charles Miller and James Laurie Snell (2012). *Introduction to probability*. American Mathematical Society. ISBN: 9780821894149. URL: http://www.dartmouth.edu/~chance/teaching_aids/books_articles/probability_book/book.html.

Jaynes, Edwin T (2007). *Probability Theory: The Logic of Science*. Ed. by G. Larry Bretthorst. Cambridge University Press. ISBN: 978-05-2159-271-0.

Some sources of information (cont.)

MacKay, David J. C. (2003). *Information Theory, Inference and Learning Algorithms*. Cambridge University Press. ISBN: 978-05-2164-298-9. URL: http://www.inference.phy.cam.ac.uk/itprnn/book.html.

Shannon, Claude E. (1948). "A mathematical theory of communication". In: *Bell Systems Technical Journal* 27, pp. 379–423, 623–656.

Shifrin, Theodore and Malcolm R Adams (2011). *Linear Algebra. A Geometric Approach*. 2nd. W. H. Freeman. ISBN: 978-1-4292-1521-3.

Strang, Gilbert (2009). *Introduction to Linear Algebra, Fourth Edition*. 4th ed. Wellesley Cambridge Press. ISBN: 9780980232714.