COE548: LARGE LANGUAGE MODELS

Topic: Linear Algebra Review

Outline

Matrices and Vectors

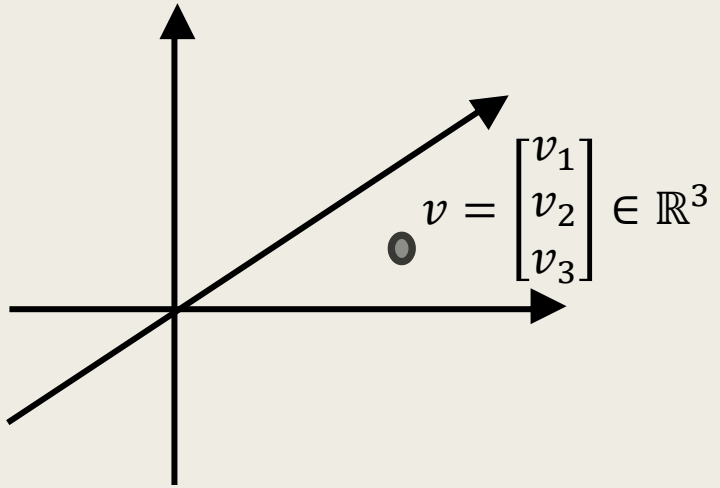
- Addition and Scalar Multiplication
- Matrix-Vector and Matrix Matrix Operations
- Inverse and Transpose
- Geometric Interpretations: Projections
- Eigenvalues and Eigenvectors
- Quadratic Forms and Definitiveness
- Matrix Decomposition

Matrix Calculus

- Calculating gradients
- Calculating Hessians
- Product Rule

Matrices and Vectors

- Vector: $v \in \mathbb{R}^d$ is a point that lives in a d-dimensional space.
- Example of $v \in \mathbb{R}^3$



- Thus a vector $v \in \mathbb{R}^d$ has the shape $v = \begin{bmatrix} v_1 \\ \vdots \\ v_d \end{bmatrix}$

Entries: v_i is the i^{th} element in the vector, where $i \in \{1, \dots, d\}$.

Matrices and Vectors

- Matrix: Rectangular (2D) array of numbers

$$A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{bmatrix} \in \mathbb{R}^{3\times 2}, \qquad A = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} \in \mathbb{R}^{2\times 3}, \qquad A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \in \mathbb{R}^{2\times 2}$$

Where the dimension is written as the number of rows x the number of columns.

Entries: A_{ij} is the entry in the i^{th} row and j^{th} column of matrix A.

- Vector: A special case of a matrix, an $n \times 1$ dimensional matrix:

$$y = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} \in \mathbb{R}^3$$

Matrices and Vectors

- Matrix transpose:

$$\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix}^T = \begin{bmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{bmatrix}$$

- If B is the transpose of A, then $B_{ij} = A_{ji}$ for all entries of A.
  - *In other words if $A \in \mathbb{R}^{n \times m}$ then $B = A^T \in \mathbb{R}^{m \times n}$*
- Vector transpose: For vector $v = \begin{bmatrix} v_1 \\ \vdots \\ v_d \end{bmatrix} \in \mathbb{R}^{d \times 1}$, then its transpose is

$$v^T = \begin{bmatrix} v_1 & \dots & v_d \end{bmatrix} \in \mathbb{R}^{1 \times d}$$

Matrices and Vectors

More definitions (matrices):

- Diagonal matrix, having non-zero diagonal entries:

$$A_{n\times m} = \begin{bmatrix} a_{11} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & a_{nm} \end{bmatrix}$$

- Symmetric matrix: $AA^T = A^T A$
- Trace of a matrix: $Tr(A_{m\times m}) = \sum_{i=1}^{m} A_{ii}$ is the sum of the diagonal entries.
- Rank of a matrix: the maximum number of linearly independent columns of the matrix (also, number of non-zero eigenvalues).

Matrices and Vectors

- Matrix addition (matrices must be same dimensions)

$$A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{bmatrix}, \qquad B = \begin{bmatrix} 7 & 8 \\ 9 & 10 \\ 11 & 12 \end{bmatrix} \rightarrow A + B = \begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{bmatrix} + \begin{bmatrix} 7 & 8 \\ 9 & 10 \\ 11 & 12 \end{bmatrix} = \begin{bmatrix} 8 & 10 \\ 12 & 14 \\ 16 & 18 \end{bmatrix}$$

- Vector addition (vectors must be same dimensions)

$$x = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}, \qquad y = \begin{bmatrix} 4 \\ 5 \\ 6 \end{bmatrix} \rightarrow x + y = \begin{bmatrix} 5 \\ 7 \\ 9 \end{bmatrix}$$

Matrices and Vectors

- Scalar multiplication

$$3A = 3 \times \begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{bmatrix} \rightarrow \begin{bmatrix} 3 & 6 \\ 9 & 12 \\ 15 & 18 \end{bmatrix} = \begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{bmatrix} \times 3$$

$$\frac{1}{2}A = \frac{1}{2} \times \begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{bmatrix} \rightarrow \begin{bmatrix} 0.5 & 1 \\ 1.5 & 2 \\ 2.5 & 3 \end{bmatrix}$$

- Similarly done for vectors

$$3y = 3 \times \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} \rightarrow \begin{bmatrix} 3 \\ 6 \\ 9 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} \times 3$$

Matrices and Vectors

- Combination of operands

$$3y - z = 3 \times \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} - \begin{bmatrix} 0.5 \\ 1 \\ 2 \end{bmatrix} \rightarrow \begin{bmatrix} 2.5 \\ 5 \\ 7 \end{bmatrix}$$

The same is applicable to matrices.

Matrices and Vectors

- Matrix-vector multiplication

$$A_{2\times 3}y_{3\times 1} = \begin{bmatrix}1 & 2 & 3\\ 4 & 5 & 6\end{bmatrix}\begin{bmatrix}1\\ 2\\ 3\end{bmatrix} \rightarrow \begin{bmatrix}1\times 1+2\times 2+3\times 3\\ 4\times 1+5\times 2+6\times 3\end{bmatrix} = \begin{bmatrix}14\\ 32\end{bmatrix} \in \mathbb{R}^{2\times 1}$$

Notice that the dimensions need to align. The numbers of columns of the first matrix must be the same as the number of rows of the second matrix (or vector).

- Matrix-matrix multiplication

$$A_{n\times m}B_{m\times o} = C_{n\times o}$$

$$\begin{bmatrix}a_{11} & a_{12} & a_{13}\\ a_{21} & a_{22} & a_{23}\end{bmatrix}\begin{bmatrix}b_{11} & b_{12}\\ b_{21} & b_{22}\\ b_{31} & b_{32}\end{bmatrix} = \begin{bmatrix}a_{11}b_{11}+a_{12}b_{21}+a_{13}b_{31} & a_{11}b_{12}+a_{12}b_{22}+a_{13}b_{32}\\ a_{21}b_{11}+a_{22}b_{21}+a_{23}b_{31} & a_{21}b_{12}+a_{22}b_{22}+a_{23}b_{32}\end{bmatrix} = \begin{bmatrix}c_{11} & c_{12}\\ c_{21} & c_{22}\end{bmatrix}$$

Matrices and Vectors

Matrix properties:

- Matrices are not commutative $AB \neq BA$

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

Also, consider $A_{n\times m}B_{m\times o}$, then$B_{m\times o}A_{n\times m}$ does not align along the dimensions.

- Matrices are associative $ABC = (AB)C = A(BC)$
- Identity matrix

$$I = \begin{bmatrix}1 & 0 & 0\\0 & \ddots & 0\\0 & 0 & 1\end{bmatrix}$$

Such that $AI = IA$.

Matrices and Vectors

- Matrix inverse:
- If A is an m× m matrix (i.e., square matrix) and has an inverse, then

$$AA^{-1} = A^{-1}A = I$$

Example:

$$\begin{bmatrix} 3 & 4 \\ 2 & 16 \end{bmatrix} \begin{bmatrix} 0.4 & -0.1 \\ -0.05 & 0.075 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

Matrices and Vectors

More definitions (vectors):

- Inner product (as far as this course is concerned is the same as dot product): For two same-dimensional vectors $x, y \in \mathbb{R}^d$ the dot product is

$$\sum_{i=1}^{d} x_i y_i = x^T y = y^T x = \sum_{i=1}^{d} y_i x_i \in \mathbb{R}$$

Which returns a scalar.

- Outer product: Let $x \in \mathbb{R}^d$ and $y \in \mathbb{R}^p$, then the outer product is $xy^T \in \mathbb{R}^{d \times p}$ (not commutative like inner product).
  - *Returns a rank-1 matrix.*

What's the Connection to The Course?

- Data representation (a way to represent the data for modeling)

$$X = \begin{bmatrix} f_{11} & \cdots & f_{1d} \\ \vdots & \ddots & \vdots \\ f_{k1} & \cdots & f_{kd} \end{bmatrix}, \qquad y = \begin{bmatrix} y_1 \\ \vdots \\ y_k \end{bmatrix}$$

Here there are k rows in X of different examples of the data, and each column represents a feature. And the elements of the vector y are the supervision values corresponding to each of the data examples.

We will also use operations on these matrices and vectors to manipulate the data for our modeling needs.

Example: Houses represented by features such as size, number of bedrooms, number of bathrooms, construction date, etc. and supervision values with the house prices.

What's the Connection to The Course?

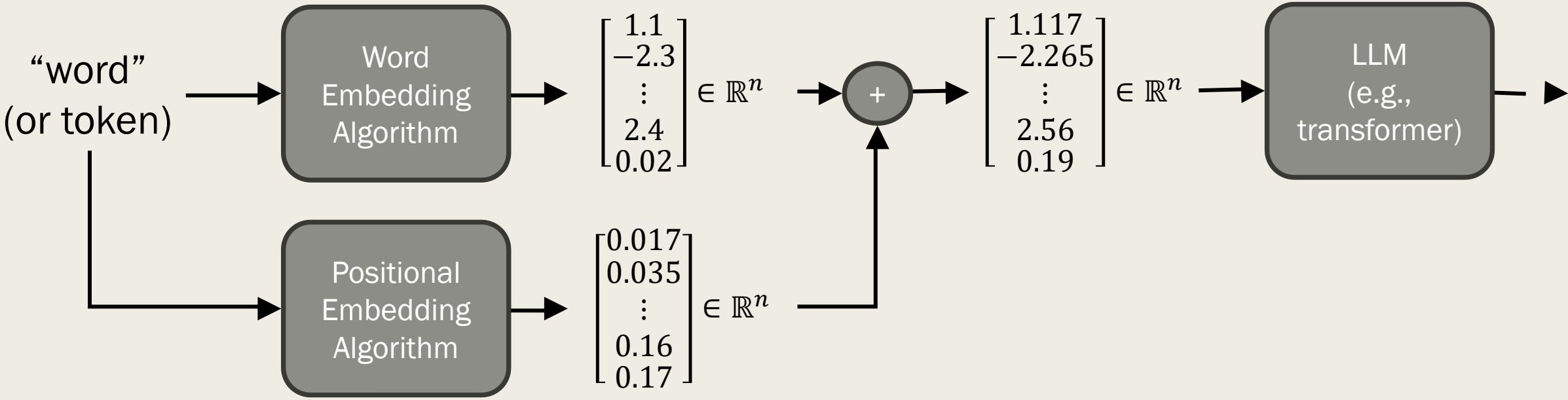
- Covariance matrices (will be used in probabilistic/statistical theory)

$$S \in \mathbb{R}^{d \times d}$$

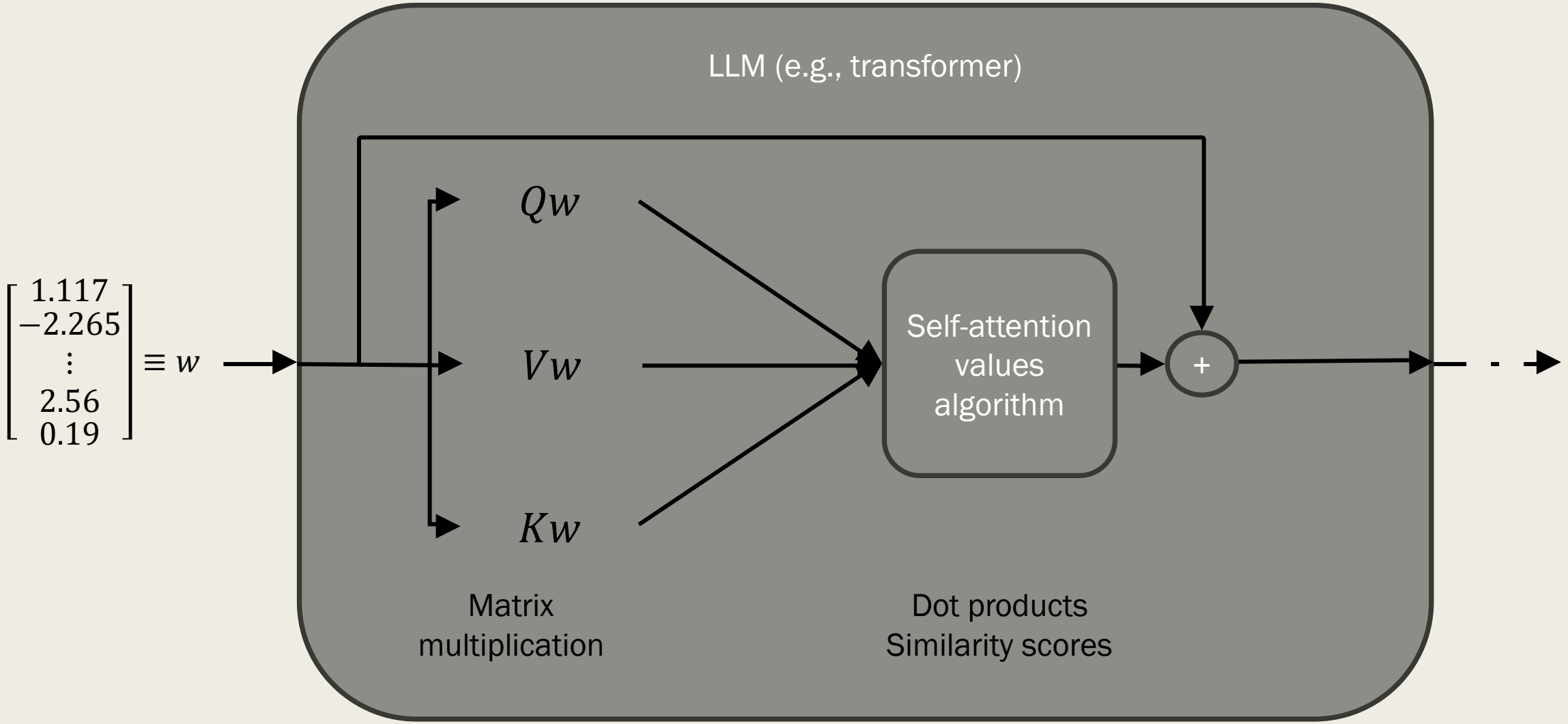
Which are symmetric matrices (will cover covariance matrices next lecture).

- Calculus (mainly in optimization where trying to minimize a loss function)
  - *Gradients (tend to be vectors containing derivatives)*
  - *Jacobians (tend to be matrices containing derivatives)*
  - *Hessians (tend to be second derivative matrices in the multivariate setting)*
- Kernel methods (not necessarily covered in the course but important to know)

Example: Linear Algebra in LLMs



Example: Linear Algebra in LLMs



Geometric Interpretations

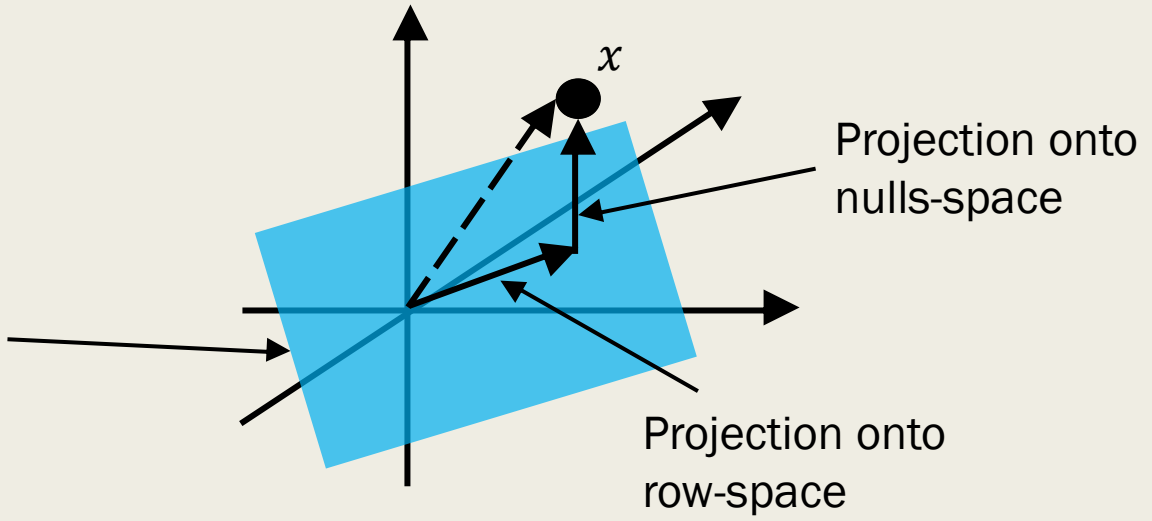
- Let A $\in \mathbb{R}^{m \times n}$ and x $\in \mathbb{R}^{n}$, then $Ax \in \mathbb{R}^{m}$.
  - *Rather than think of A as a matrix, think of it as a function that transforms x into a different m-dimensional vector.*
  - *Thus, A is a function that maps points from some input-space to an output-space.*
  - *If A is full rank (e.g., rank r=3 for 3x3 A), then every point in the input-space gets uniquely mapped to a point in the output-space (one-to-one mapping).*
  - *If there is a matrix B that reverse maps the points from the output-space to the input-space, given that A is full-rank, then $B = A^{-1}$.*
  - *If A is not full-rank (e.g., rank r<3 for 3x3 A – rank deficient), then there exist a r-dimensional subspace in the input-space and another in the output-space with a one-to-one map of all the points in those subspaces corresponding to A.*

Geometric Interpretations

- Let A ∈ $\mathbb{R}^{m\times n}$ and x ∈ $\mathbb{R}^{n}$, then $Ax \in \mathbb{R}^{m}$.
  - *Any vector x can be transformed by A nonetheless.*
  - *x can be decomposed into a summation between its projection on the subspace (aka row-space) and the null-space. In other words,*

$$x = Proj(x; row - space) + Proj(x; null - space) \equiv x_R + x_N$$

The sub-space (or row-space, the set of all points that can always be represented as some linear combination of all the rows of A) always passes through the origin



Geometric Interpretations

- Let A ∈ $\mathbb{R}^{m\times n}$ and x ∈ $\mathbb{R}^{n}$, then $Ax \in \mathbb{R}^{m}$.
  - *Any vector x can be transformed by A nonetheless.*
  - *x can be decomposed into a summation between its projection on the subspace (aka row-space) and the null-space. In other words,*

$$x = Proj(x; row - space) + Proj(x; null - space) \equiv x_R + x_N$$

And any point in the null-space of A multiplied by A gets mapped to the origin. Thus,

$$A(x) = A(x_R + x_N) = A(x_R) + A(x_N) = Ax_R$$

All in all, the function, or transformation matrix, A with rank r will project any vector x onto a r-dimensional row-space then map it to another r-dimensional subspace in the output (also referred to as the column-space).

Geometric Interpretations

- Let v and b be two vectors where we want to project b onto the subspace that v lies on.
- The projection matrix of v is $\frac{vv^T}{v^t v}$. Then if you multiply b by this projection matrix, you will get the projection of b onto the subspace spanned by v.

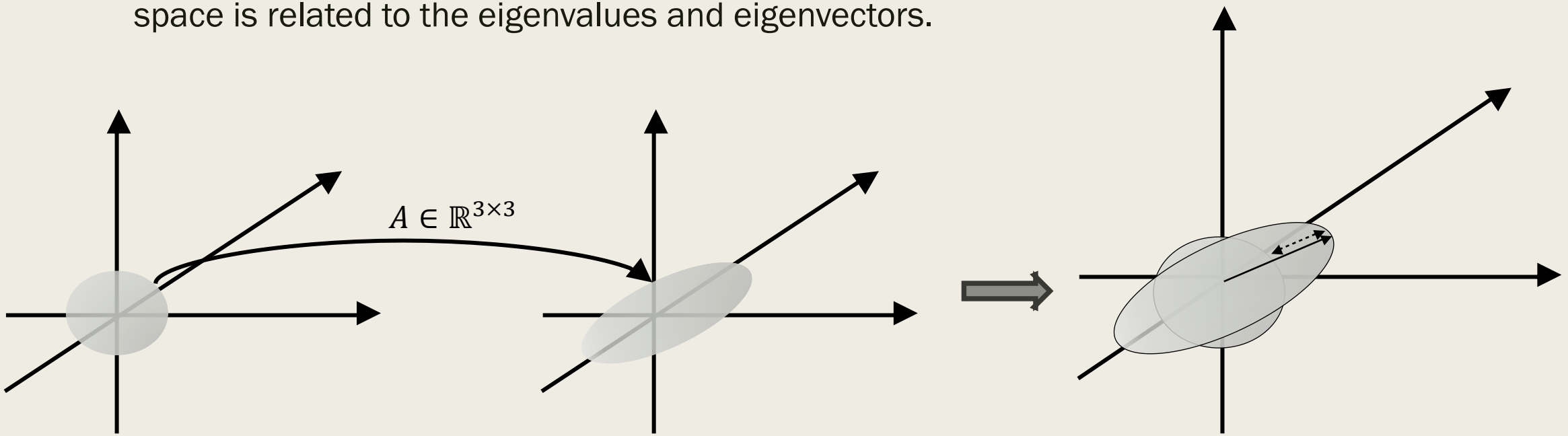
$$\frac{vv^T}{v^t v} b = \frac{v}{\|v\|}\left(\frac{v}{\|v\|}\right)^{\mathrm{T}} \mathrm{b} \equiv \tilde{v}\tilde{v}^T b = \tilde{v}(\tilde{v}^T b)$$

If instead of a vector v we want to use matrices instead, so rather than a subspace spanned by one vector we want to project b onto a subspace spanned by multiple vectors, then we have the projection matrix as:

$$X(X^T X)^{-1} X^T$$

Eigenvalues and Eigenvectors

- Consider a square matrix $A \in \mathbb{R}^{m \times m}$. To help envision consider $m = 3$.
- Take a unit sphere around the origin in $\mathbb{R}^3$, then how each point on the sphere gets compressed or expanded in each dimension from the input space to the output space is related to the eigenvalues and eigenvectors.



Eigenvalues and Eigenvectors

- To calculate the eigenvalues and eigenvectors of A:
$$Av = \lambda v \rightarrow (A - \lambda I)v = 0$$
- For a non-trivial solution (i.e., $v \neq 0$), the determinant of $(A - \lambda I)$ must be zero:
$$det(A - \lambda I) = 0$$
- This equation is called the **characteristic equation** of A which can be solved to find the eigenvalues.
- Once the eigenvalues are found, the corresponding eigenvectors can be determined by substituting each λ back into the equation $(A - \lambda I)v = 0$ and solving for v.
- Note: The determinant of a matrix is the product of all the eigenvalues. Or in terms of projections, the determinant is the volume of the output shape divided by the volume of the input shape. (Non full-rank equal zero).

Eigenvalues and Eigenvectors

Example: Find the eigenvalues and eigenvectors of $A = \begin{bmatrix} 4 & 1 \\ 2 & 3 \end{bmatrix}$.

- First - find eigenvalue by calculating the characteristic equation:

$$det(A - \lambda I) = \det\left(\begin{bmatrix} 4 & 1 \\ 2 & 3 \end{bmatrix} - \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix}\right) = \det\left(\begin{bmatrix} 4-\lambda & 1 \\ 2 & 3-\lambda \end{bmatrix}\right) = 0$$

Which gives us: $(4 - \lambda)(3 - \lambda) - 2 = \lambda^2 - 7\lambda + 10 = 0$

Thus, $\lambda_1 = 5$ and $\lambda_2 = 2$.

- Second - Calculate eigenvectors using eigenvalues:

$$\left(\begin{bmatrix} 4 & 1 \\ 2 & 3 \end{bmatrix} - \begin{bmatrix} 5 & 0 \\ 0 & 5 \end{bmatrix}\right)\begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} -1 & 1 \\ 2 & -2 \end{bmatrix}\begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \rightarrow v_1 = v_2$$

$$\left(\begin{bmatrix} 4 & 1 \\ 2 & 3 \end{bmatrix} - \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}\right)\begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} 2 & 1 \\ 2 & 1 \end{bmatrix}\begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \rightarrow v_2 = -2v_1$$

Eigenvalues and Eigenvectors

Definitions:

- Spectrum: Collection of all eigenvalues of a matrix A.
- Spectral Theorem: Every symmetric square matrix, $A \in \mathbb{R}^{d \times d}$ and $A = A^T$, has real-valued eigenvalues and orthonormal eigenvectors.
  - *Important because it covers matrices we are interested in:*
    - Hessians (matrix of second derivatives)
    - Covariance matrices
    - Kernel matrices

Quadratic Forms

- Given a square matrix $A \in \mathbb{R}^{d \times d}$ and vector $x \in \mathbb{R}^{d}$, then

$$x^T A x$$

Is the quadratic form.

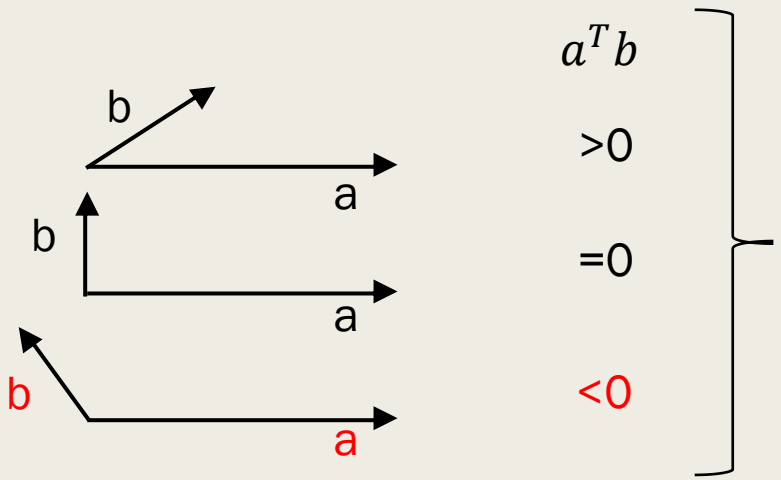
- In general we assume A is symmetric as well because there will always exist another matrix B such that

$$x^T B x = x^T A x, \qquad B = B^T = \frac{1}{2} A + \frac{1}{2} A^T$$

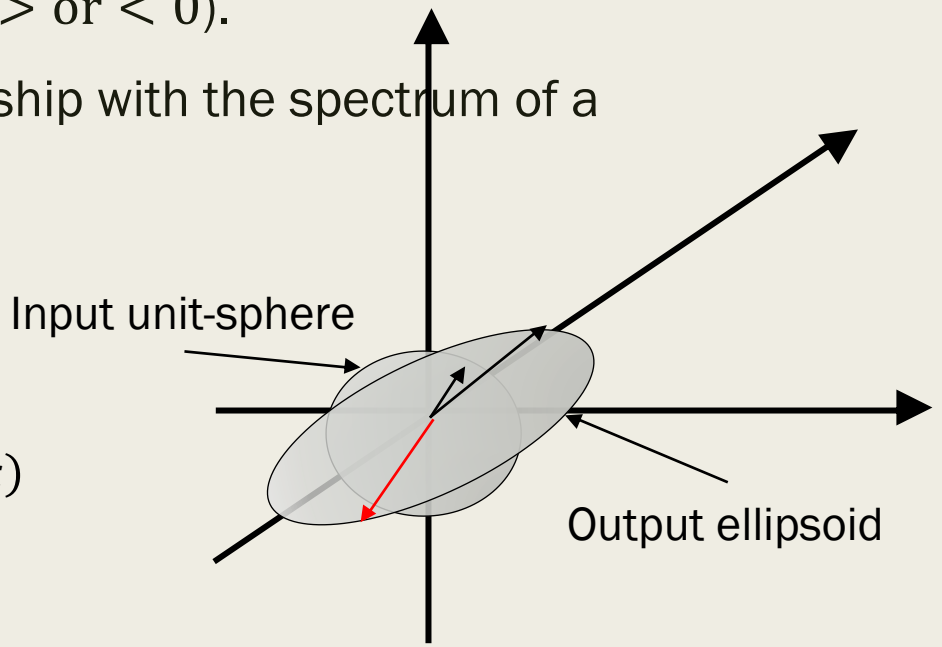
Definitiveness

- If $x^T Ax > 0,\ \forall x \neq 0$, then A is positive definite (all $\lambda(A) > 0$).
- If $x^T Ax \geq 0,\ \forall x \neq 0$, then A is positive semi-definite (all $\lambda(A) \geq 0$).
- If $x^T Ax < 0,\ \forall x \neq 0$, then A is negative definite (all $\lambda(A) < 0$).
- If $x^T Ax \leq 0,\ \forall x \neq 0$, then A is negative semi-definite (all $\lambda(A) \leq 0$).
- If $x^T Ax > \text{or} < 0,\ \forall x \neq 0$, then A is indefinite ($\lambda(A) > \text{or} < 0$).

Note: The definitiveness of a matrix has a 1-to-1 relationship with the spectrum of a matrix.



Let $a^T b \equiv x^T (Ax)$



Matrix Decomposition

- Singular value decomposition (SVD)
  - *For any matrix $A \in \mathbb{R}^{m \times n}$, we can decompose it into $A = USV^T$*
  - $A(x) = U\big(S(V^T x)\big)$
  - *U, V are orthonormal matrices, and S is a diagonal matrix of singular values of A.*
- Eigenvalue decomposition
  - *For square matrix $A \in \mathbb{R}^{m \times m}$, we can decompose it into $A = UDU^{-1}$*
  - $A(x) = U\big(D(U^{-1}x)\big)$
  - *U is matrix whose columns are the eigenvectors of A, and D is a diagonal matrix of the eigenvalues of A.*

Matrix Decomposition

	Step 1	Step 2	Step 3
SVD	$V^T(x)$ (rotation)	$S(V^T(x))$ (Scale along each dimension by diagonal entry value, real-valued scaling)	$U(S(V^T(x)))$ (Rotation 2)
EvD	$U^{-1}(x)$ (rotation)	$D\left(U^{-1}(x)\right)$ (Scale along each dimension by diagonal entry value, could be complex eigenvalues so scaling could act as rotation)	$U(D\left(U^{-1}(x)\right))$ (Rotation, inverse of step 1)

Matrix Calculus

- $f: \mathbb{R} \rightarrow \mathbb{R}$ A Function that maps a real-valued number from the real-line to another real-valued number on the real line.
  - *Value: real-valued*
  - *First derivative: real-valued in* $\mathbb{R}$
  - *Second derivative: real-valued in* $\mathbb{R}$
  - *Example:* x^2
- $f: \mathbb{R}^d \rightarrow \mathbb{R}$ A Function that maps a real-valued d-dimensional vector to another real-valued number on the real line.
  - *Value: real-valued in* $\mathbb{R}$
  - *First derivative: real-valued in* $\mathbb{R}^d$ *(gradient)*
  - *Second derivative: real-valued in* $\mathbb{R}^{d \times d}$ *(Hessian)*
  - *Example: loss function*

Matrix Calculus

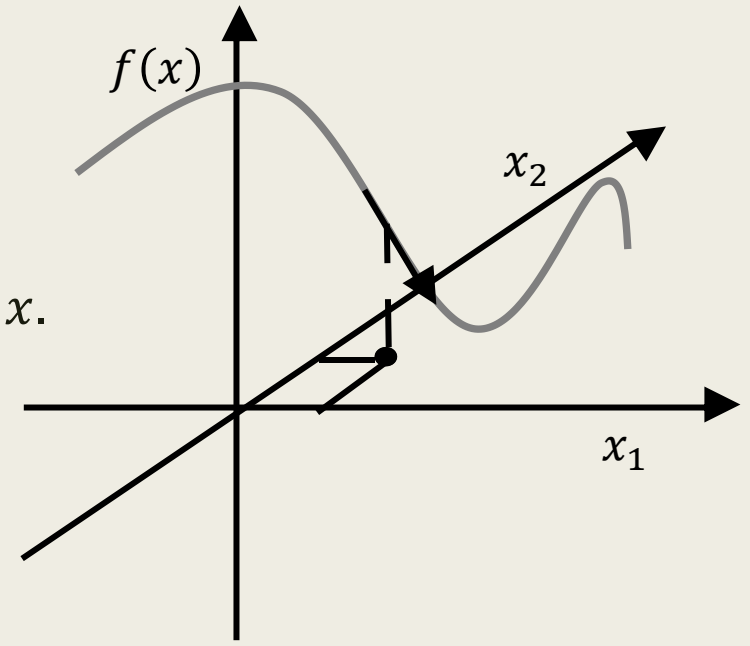
- $f: \mathbb{R}^d \rightarrow \mathbb{R}^p$ A Function that maps a real-valued d-dimensional vector to another real-valued number on the real line (we will see these in terms of neural network layers).
  - *Value: real-valued in* $\mathbb{R}^p$
  - *First derivative: real-valued in* $\mathbb{R}^{d \times p}$ *(Jacobian)*
  - *Second derivative: real-valued in* $\mathbb{R}^{d \times p \times p}$ *(Higher-order tensor)*
  - *Example: neural network layer*

Calculating Gradients

- Notation: $\nabla_x f(x)$.
- For some $x \in \mathbb{R}^d$ and $f: \mathbb{R}^d \to \mathbb{R}$

$$\nabla_x f(x) = \begin{bmatrix} \frac{\partial f(x)}{\partial x_1} \\ \vdots \\ \frac{\partial f(x)}{\partial x_d} \end{bmatrix}$$

This gives the direction of steepest descent in each dimension of x.



Calculating Gradients

- Example: Calculate $\nabla_x b^T x$ for some $x \in \mathbb{R}^d$

$$\nabla_x b^T x = \begin{bmatrix} \vdots \\ \dfrac{\partial b^T x}{\partial x_i} \\ \vdots \end{bmatrix} = \begin{bmatrix} \vdots \\ \dfrac{\partial [b_1 x_1 + \cdots b_d x_d]}{\partial x_i} \\ \vdots \end{bmatrix} = \begin{bmatrix} b_1 \\ \vdots \\ b_d \end{bmatrix} = b$$

Calculating Gradients

- For some $A \in \mathbb{R}^{m \times n}$ and $f: \mathbb{R}^{m \times n} \to \mathbb{R}$

$$\nabla_A f(A) = \begin{bmatrix} \frac{\partial f(A)}{\partial A_{11}} & \cdots & \frac{\partial f(A)}{\partial A_{1n}} \\ \vdots & \ddots & \vdots \\ \frac{\partial f(A)}{\partial A_{n1}} & \cdots & \frac{\partial f(A)}{\partial A_{nn}} \end{bmatrix}$$

- $\nabla_A \log |A| = A^{-1}$

Calculating the Hessian

- For some x ∈ $\mathbb{R}^d$ and $f: \mathbb{R}^d \to \mathbb{R}$

$$\nabla_x^2 f(x) = \begin{bmatrix} \frac{\partial^2 f(x)}{\partial x_1 \partial x_1} & \cdots & \frac{\partial^2 f(x)}{\partial x_1 \partial x_d} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f(x)}{\partial x_d \partial x_1} & \cdots & \frac{\partial^2 f(x)}{\partial x_d \partial x_d} \end{bmatrix}$$

Product Rule

- Product rule: $\frac{d}{dx} f(x)g(x) = \left[\frac{d}{dx} f(x)\right] g(x) + f(x)\left[\frac{d}{dx} g(x)\right]$

- If we want to calculate $\nabla_x x^T Ax$, then we need to break it into the chain of products:

$$\nabla_x x^T Ax = \nabla_x x^T Ax + \nabla_x x^T Ax$$
$$\nabla_x x^T Ax = Ax + A^T x = (A + A^T)x$$

And if A symmetric: $\nabla_x x^T Ax = 2Ax$

References

- **Stanford CS229 – Lecture 1:** https://www.youtube.com/watch?v=KzH1ovd4Ots&list=PLoROMvodv4rNH7qL6-efu_q2_bPuy0adh
- **Stanford CS229 – Lecture 2:** https://www.youtube.com/watch?v=b0HvwszmqcQ&list=PLoROMvodv4rNH7qL6-efu_q2_bPuy0adh&index=3