CEE 616: Probabilistic Machine Learning

Lecture 1e: Foundations—Linear Algera

Jimi Oke

UMass Amherst

College of Engineering

Feb 16, 2025

Outline

Scalars, vectors and matrices

- **Scalar:** a single number, e.g $a \in \mathbb{R}$
- **Vector:** ordered array of numbers, e.g.$\boldsymbol{x} \in \mathbb{R}^n$

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

- **Matrix:** two-dimensional array of numbers, e.g. $\boldsymbol{X} \in \mathbb{R}^{n \times p}$

$$\boldsymbol{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

Each element $x_{ij} = [\boldsymbol{X}]_{ij} \in \mathbb{R}, \forall i \in \{1:n\}, j \in \{1:p\}$

Tensors

- **Tensor:** generalization of a matrix to arbitrary number of indices/dimensions, e.g. 3 (x_{ijk})
- Number of dimensions is called **order**/**rank**
- A common application is the representation of an RGB image, e.g. a square 256-pixel image can be denoted by $\boldsymbol{A} \in \mathbb{R}^{256 \times 256 \times 3}$

Vector and matrix multiplication

Dot/inner product of two vectors

$$\boldsymbol{x} \cdot \boldsymbol{y} = \boldsymbol{x}^\top \boldsymbol{y} = \begin{bmatrix} x_1 & \cdots & x_N \end{bmatrix} \times \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix} = \sum_{n=1}^{N} x_n y_n \tag{1}$$

Matrix product

$$\boldsymbol{A}_{M \times N} \boldsymbol{B}_{N \times D} = \boldsymbol{C}_{M \times D} \tag{2}$$

Each element $[\boldsymbol{C}]_{md}$ is obtained as the dot product between the mth row of $\boldsymbol{A}$ and the d-th column of $\boldsymbol{B}$.

Matrix multiplication properties

- Distributivity

$$\boldsymbol{A}(\boldsymbol{B}+\boldsymbol{C}) = \boldsymbol{AB} + \boldsymbol{AC} \tag{3}$$

- Associativity

$$\boldsymbol{A}(\boldsymbol{BC}) = (\boldsymbol{AB})\boldsymbol{C} \tag{4}$$

- Conjugate transposability

$$(\boldsymbol{AB})^\top = \boldsymbol{B}^\top \boldsymbol{A}^\top \tag{5}$$

Non-commutativity

Unlike the inner product of two vectors, where $\boldsymbol{x}^\top \boldsymbol{y} = \boldsymbol{y}^\top \boldsymbol{x}$, matrix multiplication is not commutative:

$$\boldsymbol{AB} \neq \boldsymbol{BA} \tag{6}$$

Linear independence

A set of vectors $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\}$ is said to be linearly independent (LI) if no vector in the set can be expressed as a linear combination of the others.

- If a vector $\boldsymbol{x}_j$ in the set can be written as:

$$\boldsymbol{x}_j = \sum_{i \neq j}^{n} \alpha_i \boldsymbol{x}_i \tag{7}$$

  Then $\boldsymbol{x}_j$ is dependent on the others, so the set is not LI

- Another way to view this:

$$\nexists \boldsymbol{\alpha} \in \mathbb{R}^n \quad \text{s.t.} \quad \alpha_1 \boldsymbol{x}_1 + \cdots + \alpha_n \boldsymbol{x}_n = \mathbf{0} \tag{8}$$

Span

The **span** of a set of vectors $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\}$ is the set of all vectors that can be expressed as a linear combination of the set:

$$\text{span}(\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\}) := \left\{\boldsymbol{v} : \boldsymbol{v} = \sum_{i=1}^{n} \alpha_i \boldsymbol{x}_i, \quad \alpha_i \in \mathbb{R}\right\} \tag{9}$$

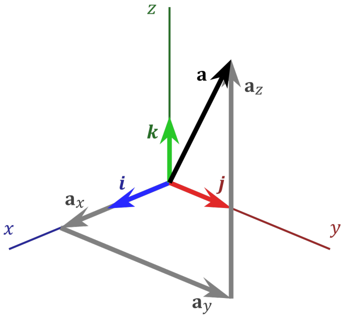
Result

If a set of vectors is LI, then any vector $\boldsymbol{v} \in \mathbb{R}^n$ can be written as a linear combination of $\boldsymbol{x}_1$ through $\boldsymbol{x}_n$

Basis

A set of LI vectors that span an entire vector space $\mathcal{V}$ is called a **basis** $\mathcal{B}$.

- Every element of $\boldsymbol{v} \in \mathcal{V}$ can be expressed as a linear combination of corresponding elements in $\mathcal{B}$.



Standard basis vectors in $\mathbb{R}^3$: $\boldsymbol{i}, \boldsymbol{j}, \boldsymbol{k}$ or $\boldsymbol{e}_1, \boldsymbol{e}_2, \boldsymbol{e}_3$.

Source: `https://en.wikipedia.org/wiki/Standard_basis`

Vector norms

Norm of a vector (p-norm)

$$||\boldsymbol{x}||_D = \left(\sum_n |x_n|^D \right)^{\frac{1}{D}} \tag{10}$$

- Euclidean, ℓ^2, or 2-norm:

$$||\boldsymbol{x}||_2 = \sqrt{\sum_{n=1}^{N} x_n^2} \tag{11}$$

  $||\boldsymbol{x}||_2^2 = \boldsymbol{x}^\top \boldsymbol{x}$
- ℓ^1 or 1-norm:

$$||\boldsymbol{x}||_1 = \sum_{n=1}^{N} |x_n| \tag{12}$$

- Max-norm:

$$||\boldsymbol{x}||_\infty = \max_n |x_n| \tag{13}$$

- Unit vector: one whose Euclidean norm is 1: $||\boldsymbol{x}||_2 = 1$

Special vectors

- The vector of ones: $\mathbf{1}$.
- Vector of zeros: $\mathbf{0}$.
- Unit (or one-hot) vector: all zeros except for entry i with a value of 1.

$$\boldsymbol{e}_4 \in \mathbb{R}^5 = (0, 0, 0, 1, 0) \tag{14}$$

Transpose and trace

- The transpose $\boldsymbol{A}^\top$ of a matrix $\boldsymbol{A}$ is obtained by converting row elements to column elements and vice versa:

$$[\boldsymbol{A}^\top]_{ij} = [\boldsymbol{A}]_{ji} \tag{15}$$

  If $\boldsymbol{A}$ is an $n \times p$ matrix, then $\boldsymbol{A}^\top$ is $p \times n$.
- A matrix $\boldsymbol{A} \in \mathbb{R}^{n \times p}$ is **square** if $n = p$.
- If a square matrix $\boldsymbol{A}$ is equal to it transpose ($\boldsymbol{A} = \boldsymbol{A}^\top$), then $\boldsymbol{A}$ is **symmetric**.
- The trace $\mathrm{tr}(\boldsymbol{A})$ of a square matrix $\boldsymbol{A}$ is the sum of its diagonal elements:

$$\mathrm{tr}(\boldsymbol{A}) = \sum_{i=1}^{n} a_{ii} \tag{16}$$

Matrix norms

Given a matrix $\boldsymbol{A} \in \mathbb{R}^{M \times N}$ defining a function $f(\boldsymbol{x}) = \boldsymbol{A}\boldsymbol{x}$, the **induced p-norm** of $\boldsymbol{A}$ is given by:

$$||\boldsymbol{A}||_D = \max_{||\boldsymbol{x}||=1} ||\boldsymbol{A}\boldsymbol{x}||_D \tag{17}$$

For $D = 2$:

$$||\boldsymbol{A}||_2 = \sqrt{\lambda_{\max}(\boldsymbol{A}^\top \boldsymbol{A})} = \max_n \sigma_n \tag{18}$$

where $\lambda_{\max}$ is the greatest eigenvalue and σ_n is n-th singular value.

Determinants

Geometrically, the determinant, $\det(\boldsymbol{A})$ or $|\boldsymbol{A}|$, of a square matrix is the [directional] scaling factor of a unit area/volume transformed by the matrix.

- If $\boldsymbol{A} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$, then:

$$|\boldsymbol{A}| = ad - bc \tag{19}$$

- The determinant of a singular (noninvertible) matrix is 0.
- The determinant of matrix is equal to the product of its eigenvalues:

$$|\boldsymbol{A}| = \prod_{i=1}^{n} \lambda_i \tag{20}$$

Range and nullspace

The range (column space) of a matrix $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ is the set of all the vectors that can be expressed as a linear combination of the column vectors of $\boldsymbol{A}$.

$$\text{range}(\boldsymbol{A}) := \{\boldsymbol{v} \in \mathbb{R}^m : \boldsymbol{v} = \boldsymbol{A}\boldsymbol{x}, \boldsymbol{x} \in \mathbb{R}^n\} \tag{21}$$

The nullspace of a matrix is the set of all vectors $\boldsymbol{v}$ such that $\boldsymbol{A}\boldsymbol{v} = \boldsymbol{0}$:

$$\text{nullspace}(\boldsymbol{A}) := \{\boldsymbol{v} \in \mathbb{R}^n : \boldsymbol{A}\boldsymbol{v} = \boldsymbol{0}\} \tag{22}$$

Rank

The rank of a matrix is the greatest number of its LI column vectors (or row vectors)

- This is equivalent to the dimension of the vector space spanned by its columns (or by its rows)
- Given $\boldsymbol{A} \in \mathbb{R}^{m \times n}$:

$$\text{rank}(\boldsymbol{A}) \leq \min(m, n) \tag{23}$$

- A matrix $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ is full rank if $\text{rank}(\boldsymbol{A}) = \min(m, n)$.
- Any matrix that is not full rank is **rank deficient**
- A square matrix is invertible if and only if it is full rank
- $\text{rank}(\boldsymbol{A}) = \text{rank}(\boldsymbol{A}^\top) = \text{rank}(\boldsymbol{A}^\top \boldsymbol{A}) = \text{rank}(\boldsymbol{A}\boldsymbol{A}^\top)$

Condition number

The condition number κ of a matrix measures how numerically stable it is under computation:

$$\kappa(\boldsymbol{A}) = ||\boldsymbol{A}|| \cdot ||\boldsymbol{A}^{-1}|| \tag{24}$$

where $||\boldsymbol{A}||$ is typically taken as the ℓ_2 norm of $\boldsymbol{A}$.

- $\boldsymbol{A}$ is well-conditioned when $\kappa(\boldsymbol{A})$ is close to 1
- $\boldsymbol{A}$ is ill-conditioned when $\kappa(\boldsymbol{A})$ is large (nearly singular/noninvertible)

Special matrices

- **Diagonal matrix:** square matrix; all elements 0 except on the main diagonal

$$\boldsymbol{A} = \text{diag}(\boldsymbol{a}) = \begin{bmatrix} a_1 & 0 & \cdots & 0 \\ 0 & a_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & a_n \end{bmatrix}$$

- **Identity matrix:** diagonal matrix whose non-zero elements are 1

$$\boldsymbol{I}_n = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}_{n \times n}$$

- **Block diagonal matrix:** concatenates matrices onto the main diagonal of a single one

$$\boldsymbol{Z} = \text{blkdiag}(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}) = \begin{bmatrix} \boldsymbol{A} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{B} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{C} \end{bmatrix}$$

Triangular matrices

- Upper triangular matrix: all non-zero entries are either on or above the diagonal:

$$\begin{pmatrix} 1 & 4 & 7 & -2 \\ 0 & -3 & 1 & 10 \\ 0 & 0 & 7 & 2 \\ 0 & 0 & 0 & -1 \end{pmatrix}$$

- Lower triangular matrix: all non-zero entries are either on or below the diagonal:

$$\begin{pmatrix} 3 & 0 & 0 \\ -1 & -2 & 0 \\ 9 & 4 & 1 \end{pmatrix}$$

- The diagonal elements A_{ii} of a triangular matrix $\boldsymbol{A}$ are its eigenvalues.

Positive definite matrices

Given a symmetric matrix $\boldsymbol{A} \in \mathbb{S}^n$ and a non-zero vector $\boldsymbol{x} \in \mathbb{R}^n$, $\boldsymbol{A}$ is:

- positive definite if $\forall \boldsymbol{x}$: $\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x} > 0$
- positive semidefinite (psd) if $\forall \boldsymbol{x}$: $\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x} \geq 0$
- negative definite if $\forall \boldsymbol{x}$: $\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x} < 0$
- negative semidefinite if $\forall \boldsymbol{x}$: $\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x} \leq 0$
- indefinite if neither psd nor nsd
- **Gram matrix $\boldsymbol{G}$** is always psd:

$$\boldsymbol{G} = \boldsymbol{X}^\top \boldsymbol{X} \tag{25}$$

  given any matrix $\boldsymbol{X} \in \mathbb{R}^{m \times n}$

Orthogonal matrix

- Two vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ are orthogonal if $\boldsymbol{x}^\top \boldsymbol{y} = 0$.
- A normalized vector is one whose 2-norm is 1.
- Thus, a set of vectors that is pairwise orthogonal and normalized is **orthonormal**
- A square matrix $\boldsymbol{U} \in \mathbb{R}^{n \times n}$ is **orthogonal** if all its columns are orthonormal
- The inverse of $\boldsymbol{U}$ is its transpose:

$$\boldsymbol{U}^\top \boldsymbol{U} = \boldsymbol{U}\boldsymbol{U}^\top = \boldsymbol{I} \tag{26}$$

which implies: $\boldsymbol{U}^{-1} = \boldsymbol{U}^\top$

Nonsingular matrices

A matrix is nonsingular (invertible) only if it is square and if its columns are linearly independent.
The inverse of a matrix $\boldsymbol{X}$ is given by $\boldsymbol{X}^{-1}$ and satisfies the property:

$$\boldsymbol{X}^{-1}\boldsymbol{X} = \boldsymbol{X}\boldsymbol{X}^{-1} = \boldsymbol{I}_n \tag{27}$$

- $\boldsymbol{X}^{-1}$ exists $\iff |\boldsymbol{X}| \neq 0$
- $(\boldsymbol{X}^{-1})^{-1} = \boldsymbol{X}$
- $(\boldsymbol{W}\boldsymbol{X})^{-1} = \boldsymbol{X}^{-1}\boldsymbol{W}^{-1}$
- For 2D case:

$$\boldsymbol{A} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}, \quad \boldsymbol{A}^{-1} = \frac{1}{|\boldsymbol{A}|}\begin{pmatrix} d & -b \\ -c & a \end{pmatrix} \tag{28}$$

- For block diagonal matrix:

$$\begin{pmatrix} \boldsymbol{A} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{B} \end{pmatrix}^{-1} = \begin{pmatrix} \boldsymbol{A}^{-1} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{B}^{-1} \end{pmatrix} \tag{29}$$

Other important results

Read **PMLI** 7.3 for details

- Schur complement: Given a partitioned matrix $\boldsymbol{M} = \begin{pmatrix} \boldsymbol{E} & \boldsymbol{F} \\ \boldsymbol{G} & \boldsymbol{H} \end{pmatrix}$ then:

$$\boldsymbol{M}/\boldsymbol{H} = \boldsymbol{E} - \boldsymbol{F}\boldsymbol{H}^{-1}\boldsymbol{G} \quad \text{(Schur complement of } \boldsymbol{M} \text{ wrt } \boldsymbol{H}\text{)} \tag{30}$$

$$\boldsymbol{M}/\boldsymbol{E} = \boldsymbol{H} - \boldsymbol{G}\boldsymbol{E}^{-1}\boldsymbol{F} \tag{31}$$

- Matrix inversion lemma (Sherman-Morrison formula)
- Matrix determinant lemma:

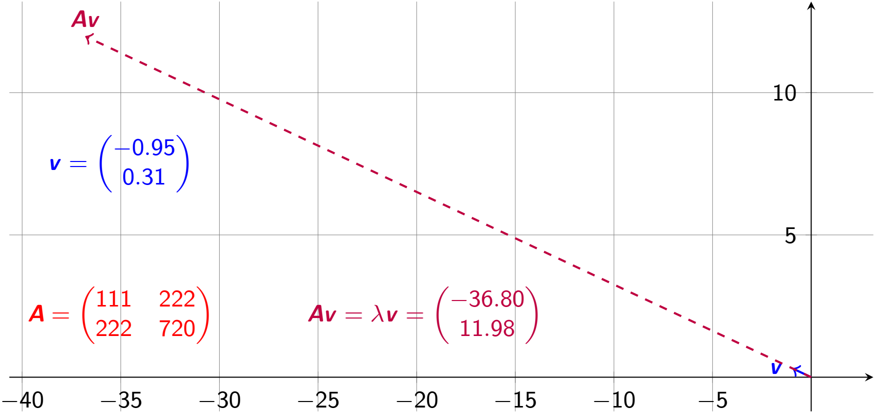
$$|\boldsymbol{A} + \boldsymbol{u}\boldsymbol{v}^\top| = (1 + \boldsymbol{v}^\top \boldsymbol{A}^{-1})\boldsymbol{u}|\boldsymbol{A}| \tag{32}$$

Eigenvectors and eigenvalues (review)

A vector $\boldsymbol{v}$ is an eigenvector of a matrix A if its transformation by A scales rather than changes the direction of $\boldsymbol{v}$:

$$\boldsymbol{A}\boldsymbol{v} = \lambda \boldsymbol{v} \tag{33}$$

where λ is the **eigenvalue** (a scalar).



Eigenvalue-related properties

- The solutions of the characteristic equation

$$\det(\lambda \boldsymbol{I} - \boldsymbol{A})\boldsymbol{v} = \boldsymbol{0}, \quad \boldsymbol{v} \neq \boldsymbol{0} \tag{34}$$

  are the eigenvalues λ_i of $\boldsymbol{A}$.
- The trace of a matrix is the sum of its eigenvalues

$$\mathrm{tr}(\boldsymbol{A}) = \sum_{i=1}^{n} \lambda_i \tag{35}$$

- The determinant of a matrix is the product of its eigenvalues

$$\det(\boldsymbol{A}) = \prod_{i=1}^{n} \lambda_i \tag{36}$$

- The rank of a matrix is equal to the number of its non-zero eigenvalues

Eigenvalue decomposition (EVD)

We can write:

$$\boldsymbol{AU} = \boldsymbol{U\Lambda} \tag{37}$$

where the columns of $\boldsymbol{U} \in \mathbb{R}^{n \times n}$ are the eigenvectors of $\boldsymbol{A}$

$$\boldsymbol{U} \in \mathbb{R}^{n \times n} = \begin{pmatrix} | & | & & | \\ \boldsymbol{u}_1 & \boldsymbol{u}_2 & \ldots & \boldsymbol{u}_n \\ | & | & & | \end{pmatrix} \tag{38}$$

and $\boldsymbol{\Lambda}$ is a diagonal matrix of the eigenvalues:

$$\boldsymbol{\Lambda} = \text{diag}(\lambda_1, \ldots, \lambda_n) \tag{39}$$

- $\boldsymbol{A}$ is said to be diagonalizable
- EVD: If $\boldsymbol{U}$ is invertible, then we can also write:

$$\boldsymbol{A} = \boldsymbol{U\Lambda U}^{-1} \tag{40}$$

Symmetric matrices

When $\boldsymbol{A}$ is real and symmetric, then $\boldsymbol{U}$ is orthogonal:

$$\boldsymbol{A} = \boldsymbol{U}\Lambda\boldsymbol{U}^{-1} = \boldsymbol{U}\Lambda\boldsymbol{U}^{\top} = \sum_{i=1}^{n} \lambda_i \boldsymbol{u}_i \boldsymbol{u}_i^{\top} \tag{41}$$

And

$$\boldsymbol{A}^{-1} = \boldsymbol{U}\Lambda^{-1}\boldsymbol{U}^{\top} = \sum_{i=1}^{d} \frac{1}{\lambda_i} \boldsymbol{u}_i \boldsymbol{u}_i^{\top} \tag{42}$$

Singular value decomposition (SVD)

The singular value decomposition of $\boldsymbol{X}$ generalizes EVD to rectangular matrices:

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Gamma}\boldsymbol{V}^T \tag{43}$$

where:

- $\boldsymbol{X}$ is an $n \times p$ data matrix
- $\boldsymbol{U}$ is an $n \times p$ orthogonal[1] matrix. The columns of $\boldsymbol{U}$ are called *left singular vectors*
- $\boldsymbol{\Gamma}$ is a $p \times p$ diagonal matrix (whose elements are called *singular values*)
- $\boldsymbol{V}$ is an $p \times p$ orthogonal[2] matrix. The columns of $\boldsymbol{V}$ are called *right singular vectors*
- The columns of $\boldsymbol{U}\boldsymbol{\Gamma}$ are called the **principal components** of $\boldsymbol{X}$.

[1] i.e. $\boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{I}$ and $\boldsymbol{U}^T = \boldsymbol{U}^{-1}$
[2] i.e. $\boldsymbol{V}^T\boldsymbol{V} = \boldsymbol{I}$ and $\boldsymbol{V}^T = \boldsymbol{V}^{-1}$

SVD (cont.)

$$\underbrace{\begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix}}_{\boldsymbol{X}} = \underbrace{\begin{pmatrix} u_{11} & \cdots & u_{1p} \\ \vdots & \ddots & \vdots \\ u_{n1} & \cdots & u_{np} \end{pmatrix}}_{\boldsymbol{U}\text{: eigenvectors of } \boldsymbol{XX}^T} \underbrace{\begin{pmatrix} \sqrt{\lambda_1} & \cdots & 0 \\ 0 & \ddots & 0 \\ 0 & \cdots & \sqrt{\lambda_p} \end{pmatrix}}_{\substack{\boldsymbol{\Gamma}\text{: } \sqrt{\text{eigenvalues}} \text{ of } \boldsymbol{XX}^T \\ \text{also singular values of } \boldsymbol{X}}} \underbrace{\begin{pmatrix} v_{11} & \cdots & v_{1p} \\ \vdots & \ddots & \vdots \\ v_{p1} & \cdots & v_{pp} \end{pmatrix}}_{\boldsymbol{V}\text{: eigenvectors of } \boldsymbol{X}^T\boldsymbol{X}}$$

- The columns $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_p$ are the left singular vectors of $\boldsymbol{X}$
- The columns $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_p$ are the right singular vectors of $\boldsymbol{X}$
- The elements $\sqrt{\lambda_1} \geq \ldots \geq \sqrt{\lambda_p} = 0$ are the singular values of $\boldsymbol{X}$
- $\lambda_1 \geq \ldots \geq \lambda_p = 0$ are the eigenvalues of $\boldsymbol{XX}^T$ and also of $\boldsymbol{X}^T\boldsymbol{X}$

Solving a system of equations

A system of linear equations can be represented and solved as follows:

$$\begin{aligned} \boldsymbol{A}\boldsymbol{x} &= \boldsymbol{b} \\ \boldsymbol{A}^{-1}\boldsymbol{A}\boldsymbol{x} &= \boldsymbol{A}^{-1}\boldsymbol{b} \\ \boldsymbol{I}_n\boldsymbol{x} &= \boldsymbol{A}^{-1}\boldsymbol{b} \\ \boldsymbol{x} &= \boldsymbol{A}^{-1}\boldsymbol{b} \end{aligned}$$

- If $\boldsymbol{A} \in \mathbb{R}^{m \times n}$, then there are m equations and n unknowns
- If $m = n$, then $\boldsymbol{A}$ is full rank, and there is a unique solution
- If $m < n$, the system is **underdetermined** (no unique solution)
- If $m > n$, the system is **overdetermined** (no exact solution; solve via least squares)

Least squares

Given a system $\boldsymbol{Ax} = \boldsymbol{b}$, the least squares objective is:

$$\min_{\boldsymbol{x}} f(\boldsymbol{x}) \equiv \min_{\boldsymbol{x}} \frac{1}{2}||\boldsymbol{Ax} - \boldsymbol{b}||_2^2 \tag{44}$$

The gradient is given by:

$$\boldsymbol{g}(\boldsymbol{x}) = \frac{\partial}{\partial \boldsymbol{x}} f(\boldsymbol{x}) = \boldsymbol{A}^\top \boldsymbol{Ax} - \boldsymbol{A}^\top \boldsymbol{b} \tag{45}$$

The optimal $\hat{\boldsymbol{x}}$ is found by solving for $\boldsymbol{g}(\boldsymbol{x}) = \boldsymbol{0}$:

$$\boldsymbol{A}^\top \boldsymbol{Ax} = \boldsymbol{A}^\top \boldsymbol{b} \tag{46}$$

And thus,

$$\hat{\boldsymbol{x}} = (\boldsymbol{A}^\top \boldsymbol{A})^{-1} \boldsymbol{A}^\top \boldsymbol{b} \tag{47}$$

is the ordinary least squares (OLS) solution. Checking that the Hessian $\boldsymbol{H}(\boldsymbol{x}) = \boldsymbol{A}^\top \boldsymbol{A}$ is pd confirms the solution is unique.

Reading assignments

- **PMLI** 7
- **PMLCE** 2