[1]

LECTURE NOTES ON LINEAR ALGEBRA

A. A. ERGÜR, O. ESKEW, AND D.M. GANDHI

Abstract. Lecture Notes on Linear Algebra given by Dr. A. A. Ergür on 23 January 2025 and 26 January 2025 respectively.

1. Vector Space over $\mathbb{R}$.

A vector space over $\mathbb{R}$ is a collection of objects that can be:

- Added to each other.
- Multiplied by a real number.

For example, in $\mathbb{R}^2$, let $a = (a_1, a_2)$, $b = (b_1, b_2)$, and $3a = (3a_1, 3a_2)$. Then:

$$a + b = (a_1, a_2) + (b_1, b_2) = (a_1 + b_1, a_2 + b_2)$$

Example (see figure 1): Shapes in $\mathbb{R}^2$ that include the origin $(0, 0)$.
A diagram illustrates:

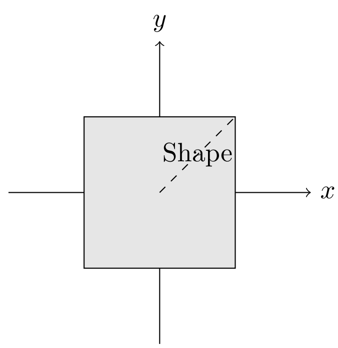

FIG. 1. *Shapes in $\mathbb{R}^2$ including the origin[3]*

- Shape K (a square with a dot at $(0, 0)$) plus shape L (a circle with a dot at $(0, 0)$) results in shape $K + L$ (a rounded square with a dot at $(0, 0)$).
- Scalar multiplication: $2 \times K$ (a square with a dot at $(0, 0)$) results in $2K$ (a larger square with a dot at $(0, 0)$).

1.1. Basis of a Vector Space.

In $\mathbb{R}^2$, consider the standard basis:

$$e_1 = (1, 0), \quad e_2 = (0, 1)$$

Any vector $x = (x_1, x_2) \in \mathbb{R}^2$ can be written as:

$$x = x_1 e_1 + x_2 e_2$$

2. Norms.

A norm is a function that attaches a number to each element x of a vector space, intended to measure its size. For $x \in \mathbb{R}^n$, where $x = (x_1, \ldots, x_n)$:

*We thank Robbins family for supporting the Algorithmic Foundations of Data Science Course

- Euclidean norm (ℓ_2-norm):
$$\|x\|_2 = \left(x_1^2 + x_2^2 + \cdots + x_n^2\right)^{1/2}$$
- ℓ_1-norm:
$$\|x\|_1 = |x_1| + |x_2| + \cdots + |x_n|$$
- ℓ_p-norm ($1 \leq p < \infty$):
$$\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p}$$
- Infinity norm (ℓ_∞-norm):
$$\|x\|_\infty = \max_{1 \leq i \leq n} |x_i|$$
- Example: Consider $x = \left(\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n}}, \ldots, \frac{1}{\sqrt{n}}\right) \in \mathbb{R}^n$.
$$\|x\|_2 = \left(\frac{1}{n} + \frac{1}{n} + \cdots + \frac{1}{n}\right)^{1/2} = \left(n \cdot \frac{1}{n}\right)^{1/2} = 1$$
$$\|x\|_1 = \frac{1}{\sqrt{n}} + \frac{1}{\sqrt{n}} + \cdots + \frac{1}{\sqrt{n}} = n \cdot \frac{1}{\sqrt{n}} = \sqrt{n}$$
$$\|x\|_\infty = \frac{1}{\sqrt{n}}$$

Note: Images of ℓ_p-unit balls (Fig 2 and 3) are provided by Kayden Mimmace, with code available on GitHub.

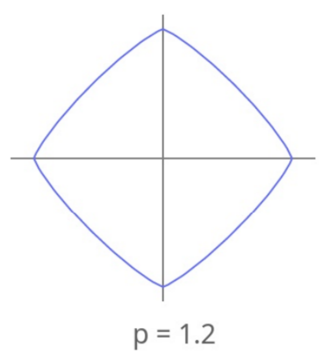

FIG. 2. *ℓ_p-unit ball in $\mathbb{R}^2$ for $p = 1.2$ generated using the code [2].*

2.1. ℓ_p-Unit Balls.

The ℓ_p-unit ball in $\mathbb{R}^n$ is defined as:
$$S_p := \{x \in \mathbb{R}^n : \|x\|_p \leq 1\}$$

Diagrams illustrate:
- For $p = 1$: A diamond shape in $\mathbb{R}^2$.
- For $p = 1.2$: A rounded diamond in $\mathbb{R}^2$ and a 3D plot in $\mathbb{R}^3$.

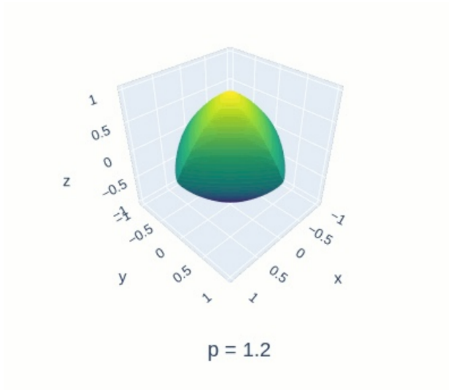

FIG. 3. *3D ℓ_p-unit ball for $p = 1.2$ [2].*

2.2. Hölder Inequality. For $x \in \mathbb{R}^n$, and $1 \leq p \leq q \leq \infty$:

$$\|x\|_q \leq \|x\|_p \leq n^{1/q-1/p}\|x\|_q$$

For $p = 1$, $q = 2$, and x as in the example above:

$$\|x\|_2 \leq \|x\|_1 \leq n^{1/2-1}\|x\|_2$$

3. Exercise. Show that for $x \in \mathbb{R}^n$, if $p > 2\log n$, then:

$$\|x\|_\infty \leq \|x\|_p \leq c\|x\|_\infty$$

4. Inner Product. For a real vector space V, an inner product $\langle \cdot, \cdot \rangle$ satisfies:
- Symmetry: $\langle x, y \rangle = \langle y, x \rangle$
- Linearity: $\langle ax + by, z \rangle = a\langle x, z \rangle + b\langle y, z \rangle$
- Positive definiteness: $\langle x, x \rangle \geq 0$, and $\langle x, x \rangle = 0$ if and only if $x = 0$.

These properties hold for any $x, y, z \in V$, $a, b \in \mathbb{R}$. The inner product induces a norm:

$$\|x\| = \sqrt{\langle x, x \rangle}$$

4.1. Cauchy-Schwarz Inequality.

$$|\langle x, y \rangle| \leq \|x\| \cdot \|y\|$$

4.2. Angles. The angle θ between vectors x and y is given by:

$$\frac{\langle x, y \rangle}{\|x\| \cdot \|y\|} = \cos\theta$$

A diagram shows vectors x and y with an angle θ between them.

5. Linear Maps and Matrices. Every linear map is represented by a matrix. A linear map $f : V \to \mathbb{R}$ satisfies:

$$f(ax + by) = af(x) + bf(y)$$

For example, consider:

$$f(x) = 3x_1 + 2x_2, \quad x \in \mathbb{R}^2$$

$$g(x) = 3x_1 + 2x_2 + 5$$

$$h(x) = x_1^2 + 3x_2^2$$

A linear map $A : x \to Ax$ can be represented by a matrix. For instance:

$$A = \begin{bmatrix} 3 & 1 \\ 5 & 2 \end{bmatrix}, \quad A \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 3x_1 + x_2 \\ 5x_1 + 2x_2 \end{bmatrix}$$

Define basis vectors:

$$e_1 = \begin{bmatrix} 5 \\ 3 \end{bmatrix}, \quad e_2 = \begin{bmatrix} 3 \\ 2 \end{bmatrix}$$

The discussion involves the inner product $\langle \cdot, \cdot \rangle$. If A^T is the transpose of A, then:

$$\langle Ax, y \rangle = \langle x, A^T y \rangle$$

for all x, y.
Question: How to see or hear what a matrix does to a vector?

5.1. Eigenvalues and Eigenvectors. For $A \in \mathbb{R}^{n \times n}$, $x \in \mathbb{R}^n$, and $\lambda \in \mathbb{R}$:

$$Ax = \lambda x$$

Here, x is an eigenvector, and λ is an eigenvalue.

- Every $n \times n$ matrix A has n complex eigenvalues.

5.2. Singular Value Decomposition (SVD). Not every matrix A is diagonalizable. Consider using the eigenvalues of X, where:

$$A^T A = X, \quad X^T A^T A = X X^T$$

Theorem 4.22 (SVD Theorem): Let $A \in \mathbb{R}^{m \times n}$ be a rectangular matrix of rank $r \in [0, \min(m, n)]$. The SVD of A is a decomposition of the form:

$$A = U \Sigma V^T$$

where:

- $U \in \mathbb{R}^{m \times m}$ is an orthogonal matrix with column vectors u_i, $i = 1, \ldots, m$, satisfying $U^T U = I_m$.
- $V \in \mathbb{R}^{n \times n}$ is an orthogonal matrix with column vectors v_j, $j = 1, \ldots, n$, satisfying $V^T V = I_n$.
- $\Sigma \in \mathbb{R}^{m \times n}$ is a diagonal matrix with $\Sigma_{ii} = \sigma_i \geq 0$, and $\Sigma_{ij} = 0$ for $i \neq j$.

Thus:

$$A = U\Sigma V^T$$

The σ_i are eigenvalues of A^TA.

$$A \in \mathbb{R}^{m\times n}, \quad U \in \mathbb{R}^{m\times m}, \quad V \in \mathbb{R}^{n\times n}, \quad \Sigma \in \mathbb{R}^{m\times n}$$

$$\Sigma = \begin{bmatrix} \sigma_1 & & & & & & \\ & \sigma_2 & & & & & \\ & & \ddots & & & & \\ & & & \sigma_r & & & \\ & & & & 0 & & \\ & & & & & \ddots & \\ & & & & & & 0 \end{bmatrix}$$

5.3. Diagonalization. In general, for any matrix A, if we have:

$$P^TAP = D \quad \Rightarrow \quad A = PDP^{-1}$$

where $P \in \mathbb{R}^{n\times n}$ is invertible, and D is diagonal:

$$D = \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{bmatrix}$$

we say A is diagonalizable.
Symmetric Matrices: If $A^T = A$, then A has real eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$, and eigenvectors $u_1, u_2, \ldots, u_n$, with:

$$U = \begin{bmatrix} u_1 & u_2 & \cdots & u_n \end{bmatrix}, \quad \langle u_i, u_j \rangle = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$$

If $U^TU = I_n$, then for all $x \in \mathbb{R}^n$:

$$\|Ux\|_2 = \|x\|_2, \quad \langle Ux, Uy \rangle = \langle x, y \rangle$$

$$x = yz, \quad x = z^Ty^T$$

5.4. Eigenbasis. For $x = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$:

$$x = x_1u_1 + x_2u_2 + \cdots + x_nu_n \quad (\text{in } U\text{-basis, } x = (x_1, x_2, \ldots, x_n))$$

$$Ax = (Ax_1u_1 + Ax_2u_2 + \cdots + Ax_nu_n)$$

$$Ax = \lambda_1x_1u_1 + \lambda_2x_2u_2 + \cdots + \lambda_nx_nu_n$$

In the eigenbasis, A becomes:

$$\Lambda = \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{bmatrix}$$

$$A = U\Lambda U^{-1}$$

This is undoing the change of basis to a diagonal form.

$$U^T = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}, \quad U = \begin{bmatrix} u_1 & u_2 & \cdots & u_n \end{bmatrix}$$

$$UU^T = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} = I_n$$

5.5. Symmetric Matrices and Eigenvalues.

- Some eigenvalues may be repeating.
- Example: For $A = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$, what is λ? Compare with $\begin{bmatrix} 2 & \\ & 2 \end{bmatrix}$.
- If $A^T = A$, i.e., A is symmetric, then all eigenvalues are real.

Let λ_1, λ_2 be distinct eigenvalues of A with eigenvectors x and y:

$$Ax = \lambda_1 x, \quad Ay = \lambda_2 y$$

Suppose A is symmetric. Then:

$$\langle Ax, y\rangle = \langle x, A^T y\rangle = \langle x, Ay\rangle$$

$$\lambda_1\langle x, y\rangle = \lambda_2\langle x, y\rangle$$

Since $\lambda_1 \neq \lambda_2$:

$$\lambda_1\langle x, y\rangle = \lambda_2\langle x, y\rangle \quad \Rightarrow \quad \langle x, y\rangle = 0$$

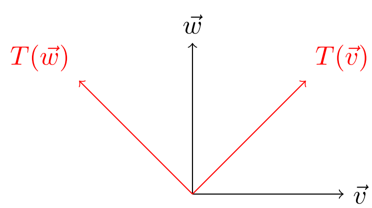

FIG. 4. *Orthogonality and linear transformations[3]*

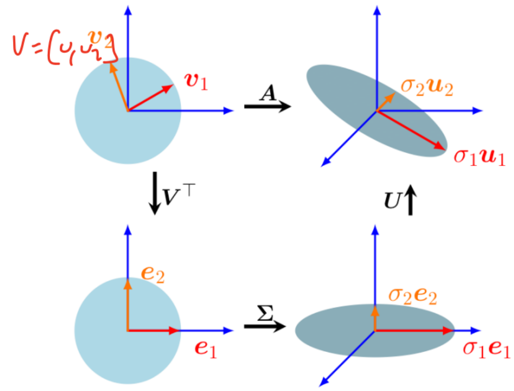

FIG. 5. *SVD transformation of a unit sphere, adapted from [1]*

5.6. SVD and Projections. If A is $n \times n$ and no x exists such that $Ax = 0$, consider the SVD:

$$A = U\Sigma V^T$$

Diagrams illustrate the transformation of a unit sphere under A:

- $V = \{v_1, v_2, v_3\}$, a sphere in $\mathbb{R}^3$, transforms via Σ to an ellipsoid with axes $\sigma_1 v_1, \sigma_2 v_2, \sigma_3 v_3$, and then via U^T.
- V^T maps the ellipsoid back to a sphere with axes $\sigma_1 e_1, \sigma_2 e_2$, and U rotates it.

Question: How about projections?

5.7. Projections. Define a projection:

$$A^2 = A \quad \Leftrightarrow \quad A \cdot (Ax) = Ax \quad \text{for all } x \in \mathbb{R}^n$$

Then A is a projection.
Example (Main): Let $y_1, y_2 \in \mathbb{R}^3$, and let H be the span of y_1, y_2. A represents the projection of x onto H.

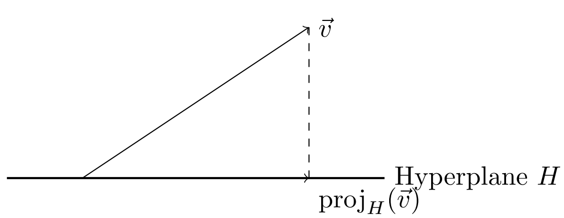

FIG. 6. *Projection onto a hyperplane H [3]*

A diagram shows $x \in \mathbb{R}^3$, H as a plane, and Ax as the projection of x onto H.

- Ax is the closest point to x in H.
- For $t \in H$, $x - Ax \perp t$, i.e., $\langle x - Ax, t\rangle = 0$.

For example:

$$y_1 = \begin{bmatrix}1\\0\\0\end{bmatrix}, \quad y_2 = \begin{bmatrix}0\\1\\0\end{bmatrix}$$

$$Y = \begin{bmatrix}1 & 0\\0 & 1\\0 & 0\end{bmatrix} \in \mathbb{R}^{3\times 2}$$

$$Y = QR$$

Find A:

$$A = QQ^T$$

$$Q^TQ = I \quad \Rightarrow \quad QQ^T = I$$

For $A = U\Sigma V^T$:

$$\langle Ax, Ax\rangle = \|Ax\|_2^2$$

$$\|Ax\|_2^2 = \langle Ax, Ax\rangle = \langle x, A^TAx\rangle$$

$$\|Ax\|_2^2 = \lambda\langle x, x\rangle = \lambda\|x\|_2^2$$

So, all σ_i are non-negative.

$$\Sigma = \begin{bmatrix}\sigma_1 &&&&&& \\ & \sigma_2 &&&&& \\ && \ddots &&&& \\ &&& \sigma_r &&& \\ &&&& 0 && \\ &&&&& \ddots & \\ &&&&&& 0\end{bmatrix} \in \mathbb{R}^{m\times n}$$

$$\|Ax\|_2^2 = 0 \quad \Rightarrow \quad Ux\|_2 = 0 \quad \Rightarrow \quad Ax = 0$$

- The number of zeros is related to the kernel of A, i.e., $\{x : Ax = 0\}$.

$$A = U\Sigma V^T$$

$$Ax = U\Sigma V^Tx = U\begin{bmatrix}\sigma_1 &&& \\ & \sigma_2 && \\ && \ddots & \\ &&& 0\end{bmatrix}V^Tx$$

6. Tensors. An order-d tensor with n variables is an $n \times n \times \cdots \times n$ (d times) data array.

- $d = 3$: Very common, e.g., $\mathbb{R}^{n\times n\times n}$.
- $d = 2$: Matrix, $\mathbb{R}^{n\times n}$.
- $d = 1$: Vector, $\mathbb{R}^n$.

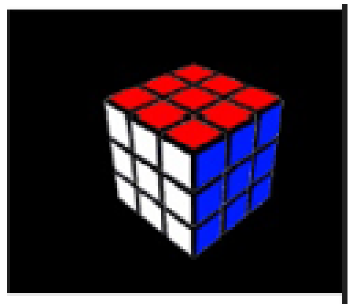

FIG. 7. *A 3 x 3 x 3 tensor [3]*

7. QR Decomposition. For $Y \in \mathbb{R}^{m\times n}$, $m \geq n$:

$$Y = QR, \quad Q \in \mathbb{R}^{m\times n}, \quad R \in \mathbb{R}^{n\times n}$$

where $Q^TQ = I_n$, Q is orthogonal, and R is upper triangular.
A diagram illustrates:

$$Y \in \mathbb{R}^{m\times n}, \quad Q \in \mathbb{R}^{m\times n}, \quad R \in \mathbb{R}^{n\times n}$$

with R having $n - \text{rank}(Y)$ zero rows.

- Suppose $\text{rank}(Y) = n$, $m \geq n$. Then $Y = QR$, $Y \in \mathbb{R}^{m\times n}$, $R \in \mathbb{R}^{n\times n}$ is invertible.

Goal: Given $x \in \mathbb{R}^m$, find $w \in \mathbb{R}^n$ such that:

$$\|x - Yw\|_2^2 = \min_{t\in\text{span}(Y)} \|x - t\|_2^2$$

$$Yw = QQ^Tx, \quad w = R^{-1}Q^Tx$$

$$Y \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{bmatrix} = w_1y_1 + w_2y_2 + \cdots + w_ny_n$$

Using Singular Values to Analyze A. Let A be an $m \times n$ matrix. The singular value decomposition (SVD) of A is given by:

$$A = U \begin{bmatrix} \delta_1 & & \\ & \ddots & \\ & & \delta_n \end{bmatrix} V^T$$

where U is $m \times m$, $U^TU = I_m$, V is $n \times n$, $V^TV = I_n$, and $\delta_1, \delta_2, \ldots, \delta_n$ are the singular values of A.

- For $x \in \mathbb{R}^n$,

$$\|x\|_2 \cdot \delta_n(A) \leq \|Ax\|_2 \leq \delta_1(A) \cdot \|x\|_2$$

- $\delta_1(A)$ is called the **operator norm** of A, denoted by $\|A\|_2$ or $\|A\|_{\text{op}}$.

- The ratio $\frac{\delta_1(A)}{\delta_n(A)}$ is called the **condition number** of A, denoted by $\kappa(A)$. This is used in LAPACK.
- The sum $\delta_1(A) + \delta_2(A) + \cdots + \delta_n(A)$ is called the **nuclear norm**, denoted by $\|A\|_*$.
- The sum $\delta_1(A)^2+\delta_2(A)^2+\cdots+\delta_n(A)^2$ is called the **Frobenius norm** or **Hilbert-Schmidt norm**, denoted by $\|A\|_F$ or $\|A\|_{\mathrm{HS}}$.

Trace Norm or $\|A\|_2$. This has a specific meaning. Consider a matrix A:

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}$$

Then the Frobenius norm squared is:

$$\|A\|_F^2 = \delta_1(A)^2 + \cdots + \delta_n(A)^2 = \sum_{1\leq i,j\leq n} a_{ij}^2$$

What's up with trace-norm naming?. The trace of a matrix X is defined as:

$$\mathrm{Tr}(X) = x_{11} + x_{22} + \cdots + x_{nn} \quad \text{for} \quad X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nn} \end{bmatrix}$$

For the inner product on matrices $\langle A, B\rangle = \mathrm{Tr}(B^T A)$, the Frobenius norm can be expressed as:

$$\|A\|_F = \sqrt{\langle A, A\rangle} = \sqrt{\mathrm{Tr}(A^T A)}$$

When Do Eigenvalues and Singular Values Coincide?. Let A be a symmetric matrix with singular value decomposition (SVD) $A = U\Sigma V^T$. Then:

$$A^T = V\Sigma U^T$$

Since A is symmetric, $A^T = A$, so:

$$V\Sigma U^T = U\Sigma V^T$$

This implies $V = U$. Thus, the SVD becomes:

$$A = U\Sigma U^T$$

Now consider the eigenvalue decomposition of A:

$$A = V\Lambda V^T$$

Since $V = U$, we have:

$$\Lambda = \Sigma$$

Thus, $\Lambda = \Sigma$, meaning the eigenvalues of A must coincide with the singular values. Additionally, since Σ contains non-negative values, all eigenvalues of A are non-negative.

- If all eigenvalues are non-negative, A is positive semidefinite (PSD).
- If all eigenvalues are positive, A is positive definite (PD).

Cholesky Decomposition. If A is an $n \times n$ positive definite matrix, then it has a Cholesky decomposition:

$$A = RR^T$$

where R is a real, upper triangular matrix with positive diagonal entries.
If A is PD and has the form $A = RR^T$, then for any $x, y \in \mathbb{R}^n$:

$$\langle Ax, y \rangle = \langle x, Ay \rangle = \langle x, RR^T y \rangle$$

$$\langle Ax, y \rangle = \langle Rx, Ry \rangle$$

Thus, if we define a new inner product $\langle \cdot, \cdot \rangle_R$ such that:

$$\langle x, y \rangle_R = \langle Rx, Ry \rangle$$

we have:

$$\langle Ax, y \rangle = \langle x, y \rangle_R$$

This implies:
- All possible inner products on $\mathbb{R}^n \leftrightarrow$ all PD matrices A.
- All similarity measures using angles.

Linear Regression. Input: Labeled vectors (x_i, y_i), $i = 1, 2, \ldots, N$, where $x_i \in \mathbb{R}^n$ (vector with n coordinates) and $y_i \in \mathbb{R}$.
Goal: Develop a linear model to predict the output value y given $x = (x_1, \ldots, x_n)$.
In other words, find a linear function $f : \mathbb{R}^n \to \mathbb{R}$ that best fits the data.
What does "best" mean?
For now, define the residual sum of squares (RSS):

$$\mathrm{RSS}(f) = \sum_{i=1}^{N} (y_i - f(x_i))^2$$

The goal is to minimize $\mathrm{RSS}(f)$ among all linear functions f.
In this context, "linear" is used in a restrictive way to mean:

$$f(x) = w_1 x_1 + w_2 x_2 + \cdots + w_n x_n + w_0$$

So the model is:

$$y_i \approx w_1 x_{i1} + w_2 x_{i2} + \cdots + w_n x_{in} + w_0 \quad (\star)$$

Let $w = (w_1, w_2, \ldots, w_n, w_0)$, and define:

$$\tilde{x}_i = (x_{i1}, x_{i2}, \ldots, x_{in}, 1)$$

$$(\star) \quad \text{becomes} \quad \langle w, \tilde{x}_i \rangle \approx w^T \tilde{x}_i$$

The RSS can be written as:

$$\mathrm{RSS}(f) = \sum_{i=1}^{N} (y_i - w^T \tilde{x}_i)^2$$

More concisely, define:

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}, \quad X = \begin{bmatrix} x_{11} & x_{21} & \cdots & x_{N1} \\ x_{12} & x_{22} & \cdots & x_{N2} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 1 \end{bmatrix}$$

Then:

$$\mathrm{RSS}(f) = \|y - X^T w\|^2$$

This means projecting $y = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}$ into the span of the rows of X.

If $X^T = QR$, where Q is $N \times (n+1)$ and R is $(n+1) \times (n+1)$, then:

$$X^T w = QRw \quad \text{and} \quad w = R^{-1} Q^T y$$

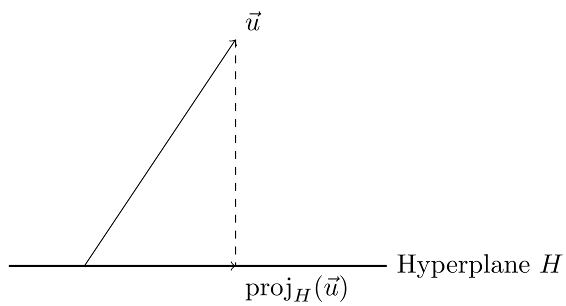

Fig. 8. *Second example of a projection onto a hyperplane[3]*

A diagram illustrates this: y (real labels) is projected onto the column span of X, with $X^T w$ being the best linear approximation.

How do you compute that Q?

PCA (Principal Component Analysis). For $d = 1$:

Input: $x_1, x_2, \ldots, x_N \in \mathbb{R}^n$.

Goal: Find a d-dimensional vector space $L \subset \mathbb{R}^n$ such that:

$$\sum_{i=1}^{N} \|x_i - P_L(x_i)\|^2$$

is minimized, where $P_L(x_i)$ is the projection of x_i onto L.

- **Simplification 1:** Define the mean of the data:

$$\mu = \frac{1}{N} \sum_{i=1}^{N} x_i$$

So the centered data is:

$$(0, 0, \dots, 0) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)$$

We'll assume $\mu = (0, \dots, 0)$.

- **Simplification 2:** If $v_1, v_2, \dots, v_d \in \mathbb{R}^n$ is an orthonormal basis for the vector space L, and we set:
$$W = \begin{bmatrix} v_1 & v_2 & \cdots & v_d \end{bmatrix} \quad \text{(an } n \times d \text{ matrix)}$$
then the projection is:
$$P_L(x) = WW^T x$$

Goal: Find $W = \begin{bmatrix} v_1 & v_2 & \cdots & v_d \end{bmatrix}$, an $n \times d$ matrix, such that $W^T W = I_d$, and:

$$\sum_{i=1}^{N} \|x_i - WW^T x_i\|^2 = \min_{L} \sum_{i=1}^{N} \|x_i - P_L(x_i)\|^2$$

Objective Function Restated.

$$\sum_{i=1}^{N} \|x_i - WW^T x_i\|^2 = \sum_{i=1}^{N} \tilde{x}_i^T (I - WW^T)\tilde{x}_i$$

where:

$$\tilde{x}_i = (x_i - \mu) \quad \text{(but we assumed } \mu = 0 \text{, so } \tilde{x}_i = x_i\text{).}$$

Define the data matrix:

$$X = \begin{bmatrix} x_1 & x_2 & \cdots & x_N \end{bmatrix} \quad \text{(an } n \times N \text{ matrix).}$$

Then:

$$X^T (I - WW^T) X \quad \text{(an } N \times N \text{ matrix).}$$

The trace of the objective function is:

$$\text{Trace}(X^T (I - WW^T) X) = \sum_{i=1}^{N} \|x_i - WW^T x_i\|^2$$

So the optimization problem becomes:

$$\min_{W \in \mathbb{R}^{n \times d},\, W^T W = I_d} \text{Trace}(X^T (I - WW^T) X)$$

This is equivalent to:

$$\min_{W \in \mathbb{R}^{n \times d},\, W^T W = I_d} \text{Trace}(X^T X) - \text{Trace}(X^T WW^T X)$$

which is equivalent to:

$$\max_{W \in \mathbb{R}^{n \times d},\, W^T W = I_d} \text{Trace}(X^T WW^T X) \quad \langle A, A \rangle \quad \text{where} \quad A^T = W^T X, \quad A = X^T W$$

Alternatively:

$$\max_{W \in \mathbb{R}^{n\times d},\, W^T W = I_d} \|W^T X\|_F^2$$

where $W^T X$ is a $d \times N$ matrix, and X is an $n \times N$ matrix with d singular values. This can also be written as:

$$\max_{W \in \mathbb{R}^{n\times d},\, W^T W = I_d} \sum_{i=1}^{d} \delta_i (W^T X)^2$$

Given the SVD of X:

$$X = U \Sigma V^T \quad \text{(where } U \text{ is } n \times n,\ \Sigma \text{ is } n \times N,\ V \text{ is } N \times N\text{)},$$

$$\Sigma = \begin{bmatrix} \delta_1 & & & & & \\ & \delta_2 & & & & \\ & & \ddots & & & \\ & & & \delta_r & & \\ & & & 0 & & \\ & & & & \ddots & \\ & & & & & 0 \end{bmatrix},$$

$$U = \begin{bmatrix} u_1 & u_2 & \cdots & u_n \end{bmatrix},$$

we pick the best d column vectors from U, i.e., $W^T = \begin{bmatrix} u_1 & u_2 & \cdots & u_d \end{bmatrix}$.

Recap: Linear Regression.

$$x_1, x_2, \ldots, x_N \in \mathbb{R}^n, \quad y_1, y_2, \ldots, y_N \in \mathbb{R}, \quad (y_1, y_2, \ldots, y_N) \in \mathbb{R}^N$$

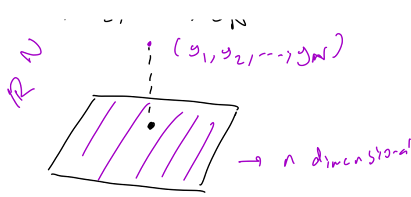

FIG. 9. *Basis of a subspace and projection onto it[3].*

A diagram illustrates $\mathbb{R}^N$ with the n-dimensional row span of X:

$$X = \begin{bmatrix} x_1 & x_2 & \cdots & x_N \end{bmatrix} \quad \text{(an } n \times N \text{ matrix).}$$

Another diagram shows $x_1, x_2, \ldots, x_N \in \mathbb{R}^n$ projected onto an n-dimensional subspace (PCA):

$$X = \begin{bmatrix} x_1 & x_2 & \cdots & x_N \end{bmatrix} \quad \text{(an } n \times N \text{ matrix).}$$

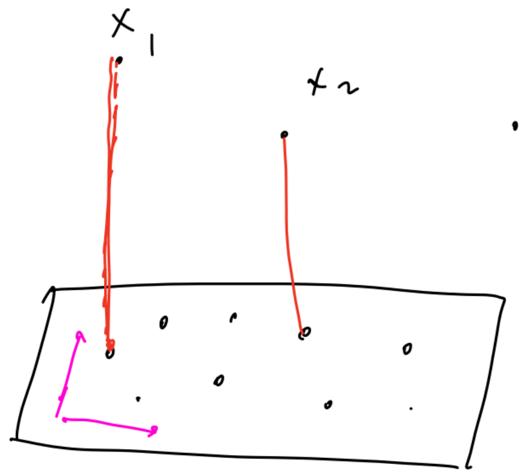

FIG. 10. *Projection of vectors onto a plane[3].*

The SVD of X:

$$X = U \begin{bmatrix} \delta_1 & & & & & & \\ & \delta_2 & & & & & \\ & & \ddots & & & & \\ & & & \delta_r & & & \\ & & & & 0 & & \\ & & & & & \ddots & \\ & & & & & & 0 \end{bmatrix} V^T$$

where U is $n \times n$, and we pick the first r columns corresponding to non-zero singular values.

Exercises.

Problem 1. Let A be a symmetric matrix ($A^T = A$), and let $\langle \cdot, \cdot \rangle$ be an inner product. Show that $\langle x, Ay \rangle = \langle Ax, y \rangle$.
Since A is symmetric, $A^T = A$. Using the standard inner product $\langle x, y \rangle = x^T y$, we have:

$$\langle x, Ay \rangle = x^T (Ay) = x^T Ay$$

$$\langle Ax, y \rangle = (Ax)^T y = x^T A^T y = x^T Ay \quad (\text{since } A^T = A)$$

Thus:

$$\langle x, Ay \rangle = x^T Ay = \langle Ax, y \rangle$$

Problem 2. Define the function $f : \mathbb{R}^2 \to \mathbb{R}^2$, where $f(x)$ is obtained by turning x counter-clockwise by 45°. Find the matrix that represents this function using the standard basis $e_1 = (1, 0)$ and $e_2 = (0, 1)$.
A counter-clockwise rotation by 45° in $\mathbb{R}^2$ is represented by the matrix:

$$R = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$$

For $\theta = 45°$, we have $\cos 45° = \sin 45° = \frac{\sqrt{2}}{2}$, so:

$$R = \begin{pmatrix} \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \end{pmatrix}$$

Problem 3. Let $Q \in \mathbb{R}^{n\times n}$ be a matrix such that $Q^T Q = I_n$. Show that for any $x, y \in \mathbb{R}^n$, the angle between x and y is the same as the angle between Qx and Qy.
The angle θ between x and y is given by:

$$\cos\theta = \frac{\langle x, y\rangle}{\|x\|\|y\|} = \frac{x^T y}{\|x\|\|y\|}$$

For Qx and Qy, compute the inner product:

$$\langle Qx, Qy\rangle = (Qx)^T(Qy) = x^T Q^T Qy = x^T I_n y = x^T y = \langle x, y\rangle$$

The norms are:

$$\|Qx\| = \sqrt{(Qx)^T(Qx)} = \sqrt{x^T Q^T Qx} = \sqrt{x^T x} = \|x\|$$

Similarly, $\|Qy\| = \|y\|$. Thus:

$$\cos\theta' = \frac{\langle Qx, Qy\rangle}{\|Qx\|\|Qy\|} = \frac{x^T y}{\|x\|\|y\|} = \cos\theta$$

The angles are the same.

Problem 4. Recall that for $1 \le p < \infty$ and $x \in \mathbb{R}^n$, we define $\|x\|_p = \left(\sum_{i=1}^n |x_i|^p\right)^{1/p}$.

Part a. Let $x \in \mathbb{R}^n$ be a vector with 100 non-zero entries. Show that $\frac{1}{10} \le \frac{\|x\|_1}{\|x\|_\infty}$.
We have $\|x\|_1 = \sum_{i=1}^n |x_i|$ and $\|x\|_\infty = \max_i |x_i|$. Let x have 100 non-zero entries, say $|x_i| = a_i$ for $i = 1$ to 100, and $x_i = 0$ otherwise. Then:

$$\|x\|_1 = \sum_{i=1}^{100} a_i, \quad \|x\|_\infty = \max_{i=1}^{100} a_i = M$$

If all non-zero entries are equal to M, then:

$$\|x\|_1 = 100M, \quad \|x\|_\infty = M \implies \frac{\|x\|_1}{\|x\|_\infty} = 100 \ge \frac{1}{10}$$

In the minimal case (e.g., one entry is M, others smaller), $\|x\|_1 \ge M$, so:

$$\frac{\|x\|_1}{\|x\|_\infty} \ge 1 \ge \frac{1}{10}$$

The inequality holds.

Part b. Let $x \in \mathbb{R}^{8000}$. Show that $\ell \cdot \|x\|_q \leq \|x\|_\infty \leq \|x\|_q$, where e denotes the natural base. (Note: Assuming ℓ is a typo or constant; interpreting as a norm comparison.) For $\|x\|_\infty = \max_i |x_i|$ and $\|x\|_q = \left(\sum_{i=1}^{8000} |x_i|^q\right)^{1/q}$, we have:

$$\|x\|_\infty \leq \|x\|_q \leq 8000^{1/q}\|x\|_\infty$$

The exact role of ℓ or e is unclear, but the standard norm comparison holds as shown.

Exercises (First Set).

Exercise 1. Let L be the line spanned by the vector $(-1, 1, 0) \in \mathbb{R}^3$. Let A be the matrix that represents the projection onto this line. Compute A.
The vector $v = (-1, 1, 0)$ spans the line. The projection matrix is:

$$A = \frac{vv^T}{v^T v}$$

$$v^T v = (-1)^2 + 1^2 + 0^2 = 2$$

$$vv^T = \begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix} \begin{pmatrix} -1 & 1 & 0 \end{pmatrix} = \begin{pmatrix} 1 & -1 & 0 \\ -1 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

$$A = \frac{1}{2} \begin{pmatrix} 1 & -1 & 0 \\ -1 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix} = \begin{pmatrix} \frac{1}{2} & -\frac{1}{2} & 0 \\ -\frac{1}{2} & \frac{1}{2} & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

Exercise 2. Generate 100 random Gaussian five-dimensional vectors. Compute the matrix that represents the projection onto the span of $x_1, x_2, \ldots, x_{100} \in \mathbb{R}^5$ using QR decomposition.
Using Python with NumPy:

```
import numpy as np
np.random.seed(42)
X = np.random.randn(5, 100)  # 5x100 matrix
Q, R = np.linalg.qr(X)
A = Q @ Q.T  # Projection matrix
```

The matrix A is 5×5 and projects onto the span of the columns of X.

Exercises (Second Set).

Exercise 2. We define the Hilbert-Schmidt norm as $\|A\|_{HS} = \left(\sum_{i,j} a_{ij}^2\right)^{1/2}$, and the trace inner product as $\langle A, B \rangle = \mathrm{Trace}(B^T A)$, with norm $\|A\|_2 = \sqrt{\langle A, A \rangle}$. Show $\|A\|_{HS} = \|A\|_2$.

$$\|A\|_{HS} = \left(\sum_{i,j} a_{ij}^2\right)^{1/2}$$

$$\|A\|_2 = \sqrt{\mathrm{Trace}(A^T A)}$$

The (i,i)-th entry of A^TA is $\sum_j a_{ji}^2$, so:

$$\text{Trace}(A^TA) = \sum_i \sum_j a_{ji}^2 = \sum_{i,j} a_{ij}^2$$

$$\|A\|_2 = \sqrt{\text{Trace}(A^TA)} = \sqrt{\sum_{i,j} a_{ij}^2} = \|A\|_{HS}$$

For further study, we recommend the textbooks by Murphy [3] and Streil [4].

REFERENCES

[1] M. P. Deisenroth, A. A. Faisal, and C. S. Ong, *Mathematics for machine learning*, Cambridge University Press, 2020.

[2] K. Mimmace, *Unit ball visualizations*. https://github.com/kaydenmimmace/unit-ball-visualizations, 2025. Accessed: 2025-05-14.

[3] K. P. Murphy, *Probabilistic Machine Learning: An Introduction*, 2023, https://probml.github.io/pml-book/book1.html. Available online.

[4] S. Streil, *Linear algebra done wrong*, 2023, https://www.math.brown.edu/streil/papers/LADW/LADW.html. Available online.