Mathematical Foundations of Machine Learning

Week 3: The dual space

Ayberk Zeytin

1. Linear transformations
2. Eigenvalues and eigenvectors
3. Diagonalization
4. Spectral theorem
5. Geometric interpretation
6. Schur decomposition

1 Recap — Week 2

In last week's lecture we continued our review of linear algebra by focusing on the geometry of inner product spaces. We introduced the Cauchy–Schwarz inequality as a fundamental bound, discussed subspaces, span, and dimension as tools for understanding structure, and examined orthogonality and projections, culminating in the Projection Theorem. The Gram–Schmidt process was presented as a systematic way to build orthonormal bases, and we closed with norms and their roles in measuring size and complexity of vectors.

Inner Products and Cauchy–Schwarz

- Inner product: $\langle \alpha x + y, z\rangle = \alpha\langle x, z\rangle + \langle y, z\rangle$, symmetry, positivity.
- Norm from inner product: $\|x\| = \sqrt{\langle x, x\rangle}$; angle via $\cos\theta = \frac{\langle x,y\rangle}{\|x\|\,\|y\|}$.
- **Cauchy–Schwarz:** $|\langle x, y\rangle| \leq \|x\|\,\|y\|$, with equality $\iff$ x, y are linearly dependent.

Example. In $\mathbb{R}^3$, let $x = (1, 2, -1)$ and $y = (2, 0, 1)$. Then

$$\langle x, y\rangle = 1 \cdot 2 + 2 \cdot 0 + (-1) \cdot 1 = 1, \qquad \|x\| = \sqrt{1 + 4 + 1} = \sqrt{6}, \quad \|y\| = \sqrt{4 + 0 + 1} = \sqrt{5},$$

so $|\langle x, y\rangle| = 1 \leq \sqrt{30} = \|x\|\,\|y\|$.

Subspaces, Span, Dimension

- $W \leq V$ if closed under addition and scalar multiplication.
- $\operatorname{span}(S) = \{\sum_{i=1}^{k} \alpha_i v_i\}$ is the smallest subspace containing S.
- Dimension: $\dim W = \min\{|\mathcal{B}| : W = \operatorname{span}(\mathcal{B})\}$.

Example. In $\mathbb{R}^4$, $S = \{(1, 0, 0, 0)^\top, (0, 1, 0, 0)^\top, (1, 1, 0, 0)^\top\}$ spans

$$\operatorname{span}(S) = \{(x_1, x_2, 0, 0)^\top : x_1, x_2 \in \mathbb{R}\},$$

so $\dim \operatorname{span}(S) = 2$ (a plane inside $\mathbb{R}^4$).

Orthogonality, Projections, Projection Theorem

- $u \perp v \iff \langle u, v\rangle = 0$; $W^\perp = \{x : \langle x, w\rangle = 0\ \forall w \in W\}$.
- Projection onto nonzero u: $\mathrm{proj}_u v = \dfrac{\langle v, u\rangle}{\langle u, u\rangle} u$.
- **Projection Theorem:** For finite-dim. $W \le V$, each x decomposes uniquely as $x = p+r$ with $p \in W$, $r \in W^\perp$.

Example. In $\mathbb{R}^3$, $u = (1, 1, 0)$, $v = (2, 1, 3)$. Then

$$\mathrm{proj}_u v = \frac{\langle v, u\rangle}{\langle u, u\rangle} u = \frac{2\cdot 1 + 1\cdot 1 + 3\cdot 0}{1^2 + 1^2 + 0^2}(1, 1, 0) = \frac{3}{2}(1, 1, 0) = (\tfrac{3}{2}, \tfrac{3}{2}, 0).$$

The residual $r = v - \mathrm{proj}_u v = (\frac{1}{2}, -\frac{1}{2}, 3)$ satisfies $\langle r, u\rangle = 0$, so $v = p + r$ with $p \in \mathrm{span}\{u\}$ and $r \in (\mathrm{span}\{u\})^\perp$.

Gram–Schmidt (to build an orthonormal basis)

Given independent $\{v_1, \ldots, v_k\}$, form

$$w_1 = v_1, \qquad w_j = v_j - \sum_{i=1}^{j-1} \frac{\langle v_j, w_i\rangle}{\langle w_i, w_i\rangle} w_i, \qquad u_j = \frac{w_j}{\|w_j\|}.$$

Then $\{u_1, \ldots, u_k\}$ is orthonormal and spans the same subspace.

Example. In $\mathbb{R}^2$, $v_1 = (3, 1)^\top$, $v_2 = (2, 2)^\top$.

$$w_1 = v_1, \quad w_2 = v_2 - \frac{\langle v_2, w_1\rangle}{\langle w_1, w_1\rangle} w_1 = (2, 2)^\top - \frac{8}{10}(3, 1)^\top = \Big(-\tfrac{2}{5}, \tfrac{6}{5}\Big)^\top,$$

$$u_1 = \frac{1}{\sqrt{10}}(3, 1)^\top, \qquad u_2 = \frac{1}{2\sqrt{10}}(-2, 6)^\top = \Big(-\tfrac{1}{\sqrt{10}}, \tfrac{3}{\sqrt{10}}\Big)^\top.$$

Norms (and where they matter)

- On $\mathbb{R}^n$: $\|x\|_1 = \sum_i |x_i|$, $\|x\|_2 = (\sum_i x_i^2)^{1/2}$, $\|x\|_\infty = \max_i |x_i|$.
- Geometry: unit balls differ (diamond/circle/square), shaping constraints and regularization.
- (Preview) Dual pairs: (ℓ_1, ℓ_∞), (ℓ_2, ℓ_2), and (ℓ_p, ℓ_q) with $\frac{1}{p} + \frac{1}{q} = 1$.

Example. For $x = (3, -4, 1)$,

$$\|x\|_1 = 8, \qquad \|x\|_2 = \sqrt{26}, \qquad \|x\|_\infty = 4.$$

Exercise. ▶ Show $W^\perp$ is a subspace.

▶ Verify $x = p + r$ with $p = A(A^\top A)^{-1}A^\top x$ when $W =$ span of columns of A.

▶ Run one Gram–Schmidt step for $v_1 = (1, 1, 0)$, $v_2 = (1, 0, 1)$.

2 Linear Transformations

Having studied vector spaces and their geometric structure, we now turn to the natural maps between them: *linear transformations*. These are the building blocks of much of linear algebra, and in machine learning they model how data is transformed through layers of an algorithm.

Definition 2.1 (Linear Transformation)**.** Let V, W be vector spaces over $\mathbb{R}$. A map $T : V \to W$ is called a **linear transformation** if for all $u, v \in V$ and $\alpha \in \mathbb{R}$,

$$T(u + v) = T(u) + T(v), \qquad T(\alpha v) = \alpha T(v).$$

Example. ▶ The map $T : \mathbb{R}^2 \to \mathbb{R}^2$ given by $T(x, y) = (2x, 3y)$ is linear.
▶ The derivative operator $D : C^1([0, 1]) \to C([0, 1])$, defined by $Df = f'$, is a linear transformation between function spaces.

Matrix Representation

Every linear transformation $T : \mathbb{R}^n \to \mathbb{R}^m$ is determined by its action on the standard basis vectors $e_1, \ldots, e_n$. If

$$T(e_j) = \begin{bmatrix} a_{1j} \\ a_{2j} \\ \vdots \\ a_{mj} \end{bmatrix},$$

then for $x = (x_1, \ldots, x_n)^\top \in \mathbb{R}^n$,

$$T(x) = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}.$$

Thus every linear transformation corresponds to multiplication by a matrix $A \in \mathbb{R}^{m \times n}$, and conversely, every such matrix defines a linear map. However, keep in mind that in order to talk about matrices representing linear transformation, one needs to fix bases.

Example. Let $T : \mathbb{R}^2 \to \mathbb{R}^2$ be given by $T(x, y) = (x + y,\, x - y)$. Then

$$T(e_1) = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \qquad T(e_2) = \begin{bmatrix} 1 \\ -1 \end{bmatrix},$$

so the associated matrix is

$$A = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}, \quad T(x, y)^\top = A \begin{bmatrix} x \\ y \end{bmatrix}.$$

Geometric View

Linear transformations capture familiar operations:

- **Scaling:** $T(x, y) = (2x, 2y)$ stretches vectors by a factor 2.
- **Rotation:** $T(x, y) = \big(\cos\theta\, x - \sin\theta\, y,\ \sin\theta\, x + \cos\theta\, y\big)$ rotates vectors by angle θ.
- **Projection:** $T(x, y) = (x, 0)$ projects $\mathbb{R}^2$ onto the x-axis.

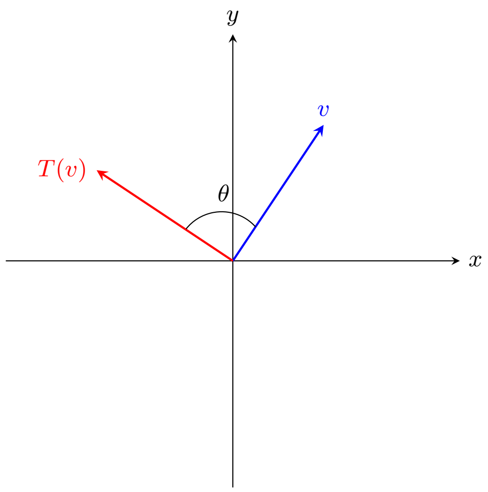

Figure 1: A linear transformation: rotation by angle θ.

Exercise. Find the matrix of the linear transformation $T : \mathbb{R}^2 \to \mathbb{R}^2$ that first reflects across the line $y = x$ and then rotates vectors by 90°.

Matrix Representation and Choice of Basis

So far, we have represented a linear transformation by a matrix relative to the *standard basis* of $\mathbb{R}^n$. More generally, every linear transformation admits a matrix representation once a basis of the domain and codomain are chosen. The entries of the matrix record how the images of the basis vectors are expressed in terms of the chosen basis.

Example. Let $T : \mathbb{R}^3 \to \mathbb{R}^3$ be defined by

$$T(x, y, z) = (x + 2y,\ y + z,\ x - z).$$

In the standard basis $\{e_1, e_2, e_3\}$, we compute:

$$T(e_1) = (1, 0, 1)^\top, \qquad T(e_2) = (2, 1, 0)^\top, \qquad T(e_3) = (0, 1, -1)^\top.$$

Thus the matrix representation of T in the standard basis is

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

Change of Basis

If we choose a different basis, the matrix of T changes. Suppose $\mathcal{B} = \{v_1, v_2, \dots, v_n\}$ is another basis of $\mathbb{R}^n$. To represent T in this basis, we express each $T(v_j)$ as a linear combination of $v_1, v_2, \dots, v_n$; the coefficients become the jth column of the new matrix. Indeed, given any vector $v \in V$, one can interpret v in terms of the basis $\mathcal{B}$ as, say, $v = \alpha_1 v_1 + \alpha_2 v_2 + \ldots + \alpha_n v_n$.

Formally, let P be the *change-of-basis matrix* from $\mathcal{B}$ to the standard basis, i.e.

$$P = \begin{bmatrix} v_1 & v_2 & v_3 \end{bmatrix}.$$

Then if A is the standard matrix of T, the matrix relative to $\mathcal{B}$ is

$$A_{\mathcal{B}} = P^{-1} A P.$$

Here one This may be interpreted as

Example. Consider again $T(x, y, z) = (x + 2y,\ y + z,\ x - z)$ with standard matrix

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

Let us choose a new basis

$$v_1 = (1, 0, 0)^\top, \quad v_2 = (1, 1, 0)^\top, \quad v_3 = (0, 0, 1)^\top.$$

Then

$$P = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \qquad P^{-1} = \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

The new representation of T is

$$A_{\mathcal{B}} = P^{-1} A P = \begin{bmatrix} 1 & 1 & -1 \\ 0 & 1 & 1 \\ 1 & -1 & -1 \end{bmatrix}.$$

Thus the same linear transformation acquires a different matrix when expressed in the new basis $\mathcal{B}$.

Geometric Note

Change of basis does not alter the transformation itself, only its *coordinates*. The relationship $A_{\mathcal{B}} = P^{-1} A P$ explains why eigenvalues are invariant under change of basis, while eigenvectors transform according to the chosen coordinates.

2.1 The Dual Space

Every vector space carries with it a natural companion: its **dual space**. If V is a vector space over $\mathbb{R}$, the *dual space* V^* is defined as

$$V^* = \{\, f : V \to \mathbb{R} \ \mid\ f \text{ is linear} \,\}.$$

The elements of V^* are called *(linear) functionals*. Just as vectors in V can be added and scaled, so too can functionals in V^*, making V^* itself into a vector space.

Example. On $V = \mathbb{R}^2$, the map $f(x, y) = 3x - 2y$ is a linear functional, since $f(\alpha(x_1, y_1) + (x_2, y_2)) = \alpha f(x_1, y_1) + f(x_2, y_2)$.

Exercise. Show, more generally, that if V and W are vector spaces over $\mathbb{R}$ (or any other field), then the set of all linear transformations fron V to W, that is the set :

$$\mathrm{Hom}(V, W) := \{T : V \to W \mid T \text{ is linear}\}$$

is itself a vector space over $\mathbb{R}$ (or any other field).

Dual Basis. If $\mathcal{B} = \{v_1, \dots, v_n\}$ is a basis of V, then there exists a unique *dual basis* $\mathcal{B}^* = \{f_1, \dots, f_n\} \subset V^*$ such that

$$f_i(v_j) = \delta_{ij} \quad (1 \le i, j \le n).$$

Here δ_{ij} is the Kronecker delta. The dual basis provides a convenient coordinate system for linear functionals, just as $\mathcal{B}$ does for vectors.

Example. Let $V = \mathbb{R}^3$ with the standard basis $e_1 = (1,0,0)$, $e_2 = (0,1,0)$, $e_3 = (0,0,1)$. The dual basis is given by the coordinate projection maps:

$$f_1(x,y,z) = x, \qquad f_2(x,y,z) = y, \qquad f_3(x,y,z) = z.$$

Relation to Matrices. Given a linear map $T : V \to W$, there is an induced *dual map* $T^* : W^* \to V^*$ defined by

$$(T^* g)(v) = g(T(v)).$$

If A is the matrix of T in some basis, then the matrix of T^* in the corresponding dual bases is $A^\top$, the transpose of A.

Example. Let $T : \mathbb{R}^2 \to \mathbb{R}^2$ be $T(x,y) = (x + 2y,\, y)$. Its matrix in the standard basis is

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

For the dual map $T^* : \mathbb{R}^{2*} \to \mathbb{R}^{2*}$, the matrix is

$$A^\top = \begin{bmatrix} 1 & 0 \\ 2 & 1 \end{bmatrix}.$$

Thus T and T^* carry the same information, but from the perspective of vectors versus linear functionals.

Geometric Note. While vectors in V represent "directions," functionals in V^* represent "measurements" or "projections" onto $\mathbb{R}$. The dual space thus encodes how our choice of basis and coordinates determines the evaluation of vectors, and it plays a key role in optimization, convexity, and regularization.