CIS3333: Mathematics of Machine Learning Fall 2025

Lecture: Inner Products and Projections

Date: October 13th, 2025 *Author: Surbhi Goel*

1 Recap and Preview

Last time we covered change of basis, images and kernels, and affine subspaces. We saw that once we choose a basis B for a vector space V, we can represent vectors as coordinate vectors and linear mappings as transformation matrices.

Today we add geometric structure: inner products let us measure lengths, angles, and distances. This will let us solve the optimization problem from Lecture 1—finding the best parameters θ to fit data.

2 Inner Products

2.1 Definition and Properties

To add geometric structure to any vector space, we need inner products.

- Recall: a linear map is one that is linear with respect to addition and scalar multiplication: $\Phi(\lambda x + \psi y) = \lambda\Phi(x) + \psi\Phi(y)$
- A **bilinear map** $\Omega : V \times V \to \mathbb{R}$ is a function of two arguments that is linear in both arguments:
$$\Omega(\lambda x + \psi y, z) = \lambda\Omega(x, z) + \psi\Omega(y, z)$$
$$\Omega(x, \lambda y + \psi z) = \lambda\Omega(x, y) + \psi\Omega(x, z)$$
- A bilinear map Ω is **symmetric** if $\Omega(x, y) = \Omega(y, x)$
- A bilinear map is **positive definite** if
$$\forall x \in V \setminus \{0\} : \Omega(x, x) > 0, \quad \Omega(0, 0) = 0$$
- An **inner product** on V is a positive definite, symmetric bilinear map. We write $\Omega(x, y) = \langle x, y\rangle$. The pair $(V, \langle\cdot,\cdot\rangle)$ is called an **inner product space**.
- Inner products capture the notion of **alignment** between vectors—analogous to angles in Euclidean space.
- **Examples of inner products:**
  1. Standard dot product on $\mathbb{R}^n$: $\langle x, y\rangle = x^T y = \sum_i x_i y_i$
  2. Weighted inner product: $\langle x, y\rangle = \sum_i w_i x_i y_i$ for $w_i > 0$
  3. For polynomials $p, q \in \mathcal{P}_n$: $\langle p, q\rangle = \int_0^1 p(x)q(x)dx$

2.2 Norms Induced by Inner Products

Every inner product naturally gives us a way to measure the "size" of vectors.

- Any inner product induces a **norm**: $\|x\| = \sqrt{\langle x, x \rangle}$
- This satisfies the three axioms of a norm:
  1. Absolutely homogeneous: $\|\lambda x\| = |\lambda| \|x\|$
  2. Triangle inequality: $\|x + y\| \leq \|x\| + \|y\|$
  3. Positive definite: $\|x\| \geq 0$ and $\|x\| = 0 \Leftrightarrow x = 0$
- The induced norm measures the "length" or "size" of a vector
- **Examples for $\mathbb{R}^d$:**
  - From $\langle x, y \rangle = x^T y$ we get $\|x\|_2 = \sqrt{x^T x} = \sqrt{\sum_i x_i^2}$ (Euclidean norm)
  - From $\langle x, y \rangle = \sum_i w_i x_i y_i$ we get $\|x\| = \sqrt{\sum_i w_i x_i^2}$ (weighted norm)
- Note: Not all norms come from inner products. For example, $\|x\|_1 = \sum_i |x_i|$ and $\|x\|_\infty = \max_i |x_i|$ cannot be expressed as $\sqrt{\langle x, x \rangle}$ for any inner product.
- Cauchy-Schwarz inequality: For any inner product,

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

- The induced norm defines a **distance** (metric), $d(x, y) = \|x - y\|$, which satisfies:
  1. Non-negativity and identity: $d(x, y) \geq 0$ and $d(x, y) = 0 \iff x = y$
  2. Symmetry: $d(x, y) = d(y, x)$
  3. Triangle Inequality: $d(x, z) \leq d(x, y) + d(y, z)$

  For example, the Euclidean distance between $x, y \in \mathbb{R}^d$ is given by $\|x - y\|_2$.

2.3 Matrix Representation of Inner Products

Just as linear mappings become matrices when we choose a basis, inner products also have matrix representations.

- Let V be an inner product space with ordered basis $B = (b_1, \ldots, b_n)$. Any vectors $x, y \in V$ have coordinate vectors $\hat{x} = (\psi_1, \ldots, \psi_n)$ and $\hat{y} = (\lambda_1, \ldots, \lambda_n)$. By bilinearity,

$$\langle x, y \rangle = \left\langle \sum_i \psi_i b_i, \sum_j \lambda_j b_j \right\rangle = \sum_i \sum_j \psi_i \langle b_i, b_j \rangle \lambda_j = \hat{x}^T A \hat{y}$$

  where $A_{ij} = \langle b_i, b_j \rangle$.

- This parallels what we saw for linear mappings: just as a linear map Φ has a transformation matrix A_Φ, an inner product has a matrix A whose entries are inner products of basis vectors.
- Since the inner product is positive definite, we have $x^T Ax > 0$ for all $x \in V \setminus \{0\}$.
- **Definition:** A matrix A is **positive definite** if $x^T Ax > 0$ for all non-zero x, and **positive semi-definite** if $x^T Ax \geq 0$ for all x.
- **Theorem:** For a finite-dimensional vector space V with ordered basis B, $\langle \cdot, \cdot \rangle$ is an inner product if and only if there exists a symmetric positive definite matrix A such that $\langle x, y \rangle = \hat{x}^T A\hat{y}$.

3 Orthogonality

With inner products, we can formalize the notion of "perpendicularity."

3.1 Orthogonal and Orthonormal Vectors

- Two vectors x, y are **orthogonal** if $\langle x, y \rangle = 0$. We write $x \perp y$.
- If additionally $\|x\| = \|y\| = 1$, then x, y are **orthonormal**.
- Orthogonality generalizes perpendicularity to arbitrary inner products.
- **Example:** $(1, 1)$ and $(-1, 1)$ are orthogonal with respect to $\langle x, y \rangle = x^T y$, but not with respect to $\langle x, y \rangle = x^T \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix} y$.
- A basis $B = (b_1, \ldots, b_n)$ is an **orthonormal basis (ONB)** if $\langle b_i, b_j \rangle = 0$ (orthogonal) for $i \neq j$ and $\langle b_i, b_i \rangle = 1$ (normalized) for all i.
- Orthonormal bases have minimal redundancy and are standardized—ideal for computations.
- An **orthonormal matrix** $Q \in \mathbb{R}^{n \times n}$ has orthonormal columns. This means:
  - $Q^T Q = QQ^T = I$
  - $Q^{-1} = Q^T$ (inverse is just transpose!)
  - $\|Qx\|^2 = \|x\|^2$ (preserves lengths)

3.2 Orthogonal Complements

- Let V be a D-dimensional vector space and $U \subseteq V$ be an M-dimensional subspace.
- The **orthogonal complement** $U^\perp \subseteq V$ is the $(D - M)$-dimensional subspace:

$$U^\perp = \{v \in V : \langle u, v \rangle = 0 \text{ for all } u \in U\}$$

- Since $U \cap U^{\perp} = \{0\}$, every vector $x \in V$ decomposes uniquely as:

$$x = x_U + x_{U^{\perp}}$$

where $x_U \in U$ and $x_{U^{\perp}} \in U^{\perp}$.

- If $(b_1, \dots, b_M)$ is a basis for U and $(b_1^{\perp}, \dots, b_{D-M}^{\perp})$ is a basis for $U^{\perp}$, then:

$$x = \sum_{m=1}^{M} \lambda_m b_m + \sum_{j=1}^{D-M} \psi_j b_j^{\perp}$$

- **Example:** In $\mathbb{R}^3$, if U is a plane through the origin, then $U^{\perp}$ is the line perpendicular to that plane.

4 Projections

4.1 Definition and Properties

- Let $U \subseteq V$ be a subspace. A linear mapping $\pi : V \to U$ is a **projection** if $\pi^2 = \pi$.
- When we choose a basis, projection matrices P satisfy $P^2 = P$.
- An **orthogonal projection** $\pi_U : V \to U$ finds the closest point in U:

$$\pi_U(x) = \arg\min_{u \in U} \|x - u\|$$

- **Key insight:** If $\pi_U(x)$ is the closest point to x in U, then the error $x - \pi_U(x)$ must be orthogonal to U:

$$\langle x - \pi_U(x), u \rangle = 0 \quad \text{for all } u \in U$$

- This means $x - \pi_U(x) \in U^{\perp}$, so $x = \pi_U(x) + (x - \pi_U(x))$ is the orthogonal decomposition.
- **Computing the projection (General Case):** Let $(b_1, \dots, b_M)$ be a basis for U. The coordinates λ of the projection $\pi_U(x) = \sum_j \lambda_j b_j$ are found by solving the linear system derived from the orthogonality condition $\langle b_i, x - \pi_U(x) \rangle = 0$:

$$\sum_{j=1}^{M} \lambda_j \langle b_i, b_j \rangle = \langle b_i, x \rangle \quad \text{for each } i = 1, \dots, M$$

This is the system $G\lambda = c$, where $G_{ij} = \langle b_i, b_j \rangle$ is the invertible **Gram matrix** and $c_i = \langle b_i, x \rangle$. The unique solution is $\lambda = G^{-1}c$.

- **Formula for $\mathbb{R}^D$ with the Dot Product:** In the common case where our vector space is $V = \mathbb{R}^D$ and our inner product is the standard dot product, this abstract solution becomes concrete. Let B be the matrix whose columns are the basis vectors $b_i \in \mathbb{R}^D$. The system becomes:

$$B^{\top}(x - B\lambda) = 0 \Leftrightarrow B^{\top}x = B^{\top}B\lambda$$

Since the columns of B form a basis, they are linearly independent, so the Gram matrix $B^\top B$ is invertible. We can solve for the coordinates λ:

$$\lambda = (B^\top B)^{-1} B^\top x$$

Substituting back, the projection is given by the matrix formula:

$$\pi_U(x) = B\lambda = B(B^\top B)^{-1} B^\top x$$

- If the basis vectors in B are orthonormal with respect to the dot product, then $B^\top B = I$, and the formula simplifies to $\pi_U(x) = BB^\top x$. The coordinates are simply $\lambda = B^\top x$.

- **Projection onto affine spaces:** For $L = x_0 + U$, we have $\pi_L(x) = x_0 + \pi_U(x - x_0)$.

5 Application: Least Squares

We can now apply projections to solve the linear regression problem: given a data matrix $X \in \mathbb{R}^{N \times D}$ and targets $Y \in \mathbb{R}^N$, find parameters $\theta \in \mathbb{R}^D$ that best satisfy $X\theta \approx Y$.

- The system $X\theta = Y$ often has no exact solution, as this would require Y to be in the column space of X, denoted $\mathrm{col}(X)$.

- The **least squares** approach is to find the parameters $\hat{\theta}$ that minimize the squared ℓ_2 error:

$$\hat{\theta} = \arg\min_{\theta} \|Y - X\theta\|_2^2$$

- **Geometric interpretation:** The vector of predictions $X\theta$ is always in $\mathrm{col}(X)$. Minimizing the error is equivalent to finding the point in $\mathrm{col}(X)$ closest to Y. This is the orthogonal projection of Y onto $\mathrm{col}(X)$.

- Let $\hat{Y} = \pi_{\mathrm{col}(X)}(Y)$. Using the projection formula with the standard dot product (where the columns of X form the basis), we have:

$$\hat{Y} = X(X^T X)^{-1} X^T Y$$

- We are looking for the $\hat{\theta}$ such that $X\hat{\theta} = \hat{Y}$. By inspection, we arrive at the **least squares solution** (or **normal equations**):

$$\hat{\theta} = (X^T X)^{-1} X^T Y$$

- This solution is unique if the columns of X are linearly independent (i.e., $\mathrm{rk}(X) = D$), which ensures $X^T X$ is invertible. The matrix $(X^T X)^{-1} X^T$ is the **pseudoinverse** of X.

- **Beyond least squares:** This framework is extensible. To prevent overfitting, we often add a penalty to the loss function, known as regularization: $\min_\theta \|Y - X\theta\|^2 + \lambda\|\theta\|^2$. Here, the first norm measures fit to data, while the second norm measures model complexity. This same principle of using norms, inner products, and projections to find an optimal balance between fit and complexity is what will allow us to solve optimization problems in the infinite-dimensional function spaces central to the Representer Theorem.