Statistical Learning

Pierpaolo Brutti

Overview of Linear Algebra and Functional Analysis
Basic Concepts and Examples.

Theory recap (. . . if you need it!)

Vector spaces

We start from the basic concepts of **linear algebra**, which is to say the study of **vector spaces**. Additional details can be found in any textbook on linear algebra. Typical feature spaces characterizing most learning problems have a rich mathematical structure, which arises from the fact that they allow a number of useful operations to be carried out on their elements: *addition*, *multiplication with scalars*, and the product between the elements themselves, usually called *dot* or *inner product*.

Definition 1 (Real Vector Space)**.** A set $\mathbb{V}$ is called a vector space (or linear space) over $\mathbb{R}$ if we can define two binary operations $(+, \times)$ called *addition* and (scalar) *multiplication* respectively such that, for all $\boldsymbol{v}$, $\boldsymbol{v}_1$, $\boldsymbol{v}_2$, $\boldsymbol{v}_3$ in $\mathbb{V}$, and β, β_1, β_2 in $\mathbb{R}$, we have

1. $\boldsymbol{v}_1 + (\boldsymbol{v}_2 + \boldsymbol{v}_3) = (\boldsymbol{v}_1 + \boldsymbol{v}_2) + \boldsymbol{v}_3$ ⇝ $+$ is *associative*,
2. $(\boldsymbol{v}_1 + \boldsymbol{v}_2) = (\boldsymbol{v}_2 + \boldsymbol{v}_1) \in \mathbb{V}$ ⇝ $+$ is *commutative*,
3. we can find an element $\mathbf{0}_{\mathbb{V}} \in \mathbb{V}$, such that $\boldsymbol{v} + \mathbf{0}_{\mathbb{V}} = \boldsymbol{v}$ for every $\boldsymbol{v} \in \mathbb{V}$ ⇝ *null element*,
4. for every $\boldsymbol{v} \in \mathbb{V}$, we can find an element of $\mathbb{V}$ denoted by $-\boldsymbol{v}$ such that $\boldsymbol{v} + (-\boldsymbol{v}) = \mathbf{0}_{\mathbb{V}}$,
5. $\beta \times \boldsymbol{v} \in \mathbb{V}$,
6. $1 \times \boldsymbol{v} = \boldsymbol{v}$,
7. $\beta_1 \times (\beta_2 \times \boldsymbol{v}) = (\beta_1\,\beta_2) \times \boldsymbol{v}$,
8. $\beta \times (\boldsymbol{v}_1 + \boldsymbol{v}_2) = \beta \times \boldsymbol{v}_1 + \beta \times \boldsymbol{v}_2$.

The first four conditions amount to saying that $(\mathbb{V}, +)$ is an *abelian* or *commutative group*.

Among the things we can do in a vector space are **linear combinations**,

$$\sum_{i=1}^{d} \beta_i \times \boldsymbol{v}_i, \text{ where } \beta_i \in \mathbb{R}, \text{ and } \boldsymbol{v}_i \in \mathbb{V},$$

and **convex combinations**

$$\sum_{i=1}^{d} \beta_i \times \boldsymbol{v}_i, \text{ where } \beta_i \geqslant 0, \ \sum_{i=1}^{d} \beta_i = 1, \text{ and } \boldsymbol{v}_i \in \mathbb{V}.$$

The set $\left\{ \sum_{i=1}^{d} \beta_i \times \boldsymbol{v}_i \,:\, \beta_i \in \mathbb{R} \right\}$ is referred to as the **span** of the vectors $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_d\}$.

A set of vectors $\{\boldsymbol{v}_i\}_i$, chosen such that none of the $\boldsymbol{v}_i$ can be written as a linear combination of the others, is called **linearly independent**. A set of vectors $\{\boldsymbol{v}_i\}_i$ that allows us to uniquely write each element of $\mathbb{V}$ as a linear combination is called a **basis** of $\mathbb{V}$. For the uniqueness to hold, the vectors have to be linearly independent. All bases of a vector space $\mathbb{V}$ have the same number of elements, called the **dimension** of $\mathbb{V}$.

Example 1 (Euclidean spaces)**.** The standard example of a finite-dimensional vector space is the euclidean space $\mathbb{R}^d$. In $\mathbb{R}^d$ addition and scalar multiplication are defined element-wise. The *canonical basis* of $\mathbb{R}^d$ is

$$\mathscr{B}_d = \{\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_d\}, \text{ where } \boldsymbol{\phi}_j[i] = \delta_{i,j} \quad \Leftrightarrow \quad \boldsymbol{\phi}_j = \begin{bmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix} \leftarrow j^{\texttt{th}} \text{ entry.}$$

Here we used an R-like notation $\boldsymbol{\phi}_j[i]$ to denote the $i^{\texttt{th}}$ entry of the $j^{\texttt{th}}$ basis vector, whereas $\delta_{i,j}$ is the *Kronecker symbol*,

$$\delta_{i,j} = \begin{cases} 1, & \text{if } i = j, \\ 0, & \text{otherwise} \end{cases}.$$

Vector addition in $\mathbb{R}^2$ and $\mathbb{R}^3$ is easily visualized by using elementary geometry, in particular the **parallelogram law**, which states that for two vectors $\boldsymbol{v}$ and $\boldsymbol{w}$, the sum $\boldsymbol{v} + \boldsymbol{w}$ is the vector defined by the diagonal of the parallelogram as shown is Figure 1.

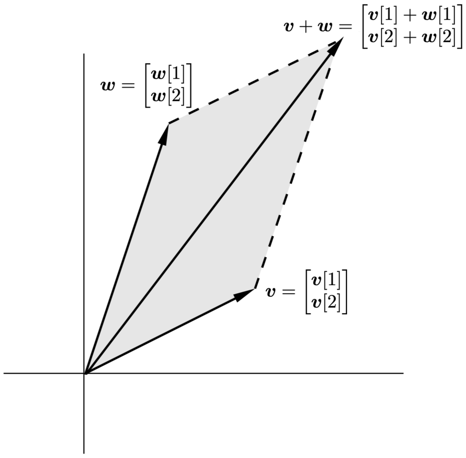

$\boldsymbol{v} + \boldsymbol{w} = \begin{bmatrix} \boldsymbol{v}[1] + \boldsymbol{w}[1] \\ \boldsymbol{v}[2] + \boldsymbol{w}[2] \end{bmatrix}$
$\boldsymbol{w} = \begin{bmatrix} \boldsymbol{w}[1] \\ \boldsymbol{w}[2] \end{bmatrix}$
$\boldsymbol{v} = \begin{bmatrix} \boldsymbol{v}[1] \\ \boldsymbol{v}[2] \end{bmatrix}$

Figure 1: Vector addition in $\mathbb{R}^2$ and the *parallelogram law*.

Linear algebra is the study of vector spaces and **linear maps** (sometimes called *operators*) between vector spaces. Given two real vector spaces $\mathbb{V}_1$ and $\mathbb{V}_2$, the latter are maps $L : \mathbb{V}_1 \mapsto \mathbb{V}_2$, such that, for all $\boldsymbol{v}_1$, $\boldsymbol{v}_2$ in $\mathbb{V}_1$, and β_1, β_2 in $\mathbb{R}$, satisfy

$$L\big(\beta_1 \boldsymbol{v}_1 + \beta_2 \boldsymbol{v}_2\big) = \beta_1\, L(\boldsymbol{v}_1) + \beta_2\, L(\boldsymbol{v}_2). \tag{1}$$

It is customary to omit the parentheses for linear maps; thus we normally write $L\,\boldsymbol{v}$ rather than $L(\boldsymbol{v})$. Let us go into more detail, using (for simplicity) the case where $\mathbb{V}_1$ and $\mathbb{V}_2$ are identical, have dimension d, and are written $\mathbb{V}$. Due to Equation (1), a linear map L is completely determined by the values it takes on a basis $\mathscr{B}_d = \big\{\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_d\big\}$ of $\mathbb{V}$. This can be seen by writing an arbitrary input $\boldsymbol{v} \in \mathbb{V}$ as a linear combination in terms of the basis vectors $\boldsymbol{v} = \sum_j \beta_j \boldsymbol{\phi}_j$, and then applying L;

$$L\,\boldsymbol{v} = L \sum_j \beta_j \cdot \boldsymbol{\phi}_j \overset{\texttt{Eq.}(1)}{=} \sum_j \beta_j \cdot L\,\boldsymbol{\phi}_j.$$

Now, since we assumed $L : \mathbb{V} \mapsto \mathbb{V}$, the image of each basis vector, $L\,\boldsymbol{\phi}_j$ still belongs to $\mathbb{V}$, and it must admit an expansion in terms of the same basis $\mathscr{B}_d$ so that it is in turn completely determined by its expansion coefficients $A[i,j]$, for $i \in \{1, \ldots, d\}$;

$$L\,\boldsymbol{\phi}_j = \sum_i A[i,j] \cdot \boldsymbol{\phi}_i.$$

The coefficients $A[i,j]$ form the matrix A of L with respect to the basis $\mathscr{B}_d$. We often think of linear maps as matrices in the first place, and use the same symbol to denote them. The *unit* (or *identity*) matrix is denoted by $\mathbb{I}_d$.

Norms and Dot Products

Thus far, we have explained the linear structure of spaces. We now move on to the *metric* structure. To this end, we introduce concepts of *length* (norms) and *angles* (inner products).

Definition 2 (Norm)**.** A function $\|\cdot\| : \mathbb{V} \mapsto \mathbb{R}_0^+$ such that for all $\boldsymbol{v}$, $\boldsymbol{v}_1$, $\boldsymbol{v}_2$ in $\mathbb{V}$ and β, β_1, β_2 in $\mathbb{R}$ satisfies

1. $\|\boldsymbol{v}_1 + \boldsymbol{v}_2\| \leqslant \|\boldsymbol{v}_1\| + \|\boldsymbol{v}_2\|$ $\rightsquigarrow$ *triangle inequality* or *subadditivity*,
2. $\|\beta\,\boldsymbol{v}\| = |\beta| \cdot \|\boldsymbol{v}\|$ $\rightsquigarrow$ *nonnegative homogeneity*
3. $\|\boldsymbol{v}\| \geqslant 0$ and $\|\boldsymbol{v}\| = 0$ if and only if $\boldsymbol{v} \neq \mathbf{0}_{\mathbb{V}}$ $\rightsquigarrow$ *positive definiteness.*

is called a **norm** on $\mathbb{V}$. If we drop the last condition in the third equation, we are left with what is called a **semi-norm**.

Any norm induces a **metric**, a **distance** $d(\cdot,\cdot)$ via

$$d(\boldsymbol{v}_1, \boldsymbol{v}_2) = \|\boldsymbol{v}_1 - \boldsymbol{v}_2\|; \tag{2}$$

likewise, any semi-norm defines a semi-metric. The (semi-)metric inherits certain properties from the (semi-)norm, in particular the *triangle inequality* and *positivity.* While every norm gives rise to a metric, the converse is not the case. In this sense, the concept of the norm is stronger. Similarly, every *inner product* (to be introduced next) gives rise to a norm, but not vice versa[1]. Before describing the dot product, we start with a more general concept.

Definition 3 (Bilinear Form)**.** A **bilinear form** on a vector space $\mathbb{V}$ is a function

$$\underset{(\boldsymbol{v}_1,\boldsymbol{v}_2)\mapsto Q(\boldsymbol{v}_1,\boldsymbol{v}_2)}{Q : \mathbb{V} \times \mathbb{V} \mapsto \mathbb{R}\,.}$$

with the property that for all $\boldsymbol{v}_1$, $\boldsymbol{v}_2$, $\boldsymbol{v}_3$ in $\mathbb{V}$, and all β_1, β_2 in $\mathbb{R}$ we have

1. $Q\big((\beta_1\,\boldsymbol{v}_1 + \beta_2\,\boldsymbol{v}_2), \boldsymbol{v}_3\big) = \beta_1\,Q(\boldsymbol{v}_1, \boldsymbol{v}_3) + \beta_2\,Q(\boldsymbol{v}_2, \boldsymbol{v}_3)$,
2. $Q\big(\boldsymbol{v}_3, (\beta_1\,\boldsymbol{v}_1 + \beta_2\,\boldsymbol{v}_2)\big) = \beta_1\,Q(\boldsymbol{v}_3, \boldsymbol{v}_1) + \beta_2\,Q(\boldsymbol{v}_3, \boldsymbol{v}_1)$.

If the bilinear form also satisfies $Q(\boldsymbol{v}_1, \boldsymbol{v}_2) = Q(\boldsymbol{v}_2, \boldsymbol{v}_1)$ for all $\boldsymbol{v}_1$, $\boldsymbol{v}_2$ in $\mathbb{V}$, it is called *symmetric.*

Definition 4 (Inner Product)**.** A **dot** or **inner product** on a vector space $\mathbb{V}$ is a symmetric bilinear form,

$$\underset{(\boldsymbol{v}_1,\boldsymbol{v}_2)\rightsquigarrow\langle\boldsymbol{v}_1,\boldsymbol{v}_2\rangle}{\langle\cdot,\cdot\rangle : \mathbb{V} \times \mathbb{V} \mapsto \mathbb{R}}$$

that is strictly positive definite; in other words, it has the property that for all $\boldsymbol{v} \in \mathbb{V}$,

$$\langle\boldsymbol{v}, \boldsymbol{v}\rangle \geqslant 0,$$

with equality only for $\boldsymbol{v} = \boldsymbol{0}_{\mathbb{V}}$.

Definition 5 (Normed Space and Dot Product Space)**.** A **normed space** is a vector space endowed with a norm; a **dot product space** (sometimes called *pre-Hilbert* space) is a vector space endowed with a dot product.

Any dot product defines a corresponding norm via

$$\|\boldsymbol{v}\| \overset{\texttt{def}}{=} \sqrt{\langle\boldsymbol{v}, \boldsymbol{v}\rangle}. \tag{3}$$

An important result that relates norms and inner product is the so called **Cauchy-Schwarz inequality**: for all $\boldsymbol{v}_1$ $\boldsymbol{v}_2$ in $\mathbb{V}$,

$$\big|\langle\boldsymbol{v}_1, \boldsymbol{v}_2\rangle\big| \leqslant \|\boldsymbol{v}_1\| \cdot \|\boldsymbol{v}_2\|. \tag{4}$$

Examining the proof of this result, it can be seen that the equality holds only if $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ are linearly dependent. In some instances, the left hand side can be much smaller than the right hand side. An extreme case is when $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ are **orthogonal**, and, by definition, $\langle\boldsymbol{v}_1, \boldsymbol{v}_2\rangle = 0$.

Example 2. (Euclidean norm) The standard example of a dot product space is again the Euclidean space $\mathbb{R}^d$. We usually employ the *canonical dot product*,

$$\langle\boldsymbol{v}_1, \boldsymbol{v}_2\rangle_{\mathbb{R}^d} = \boldsymbol{v}_1^{\mathsf{T}}\boldsymbol{v}_2 = \sum_{i=1}^{d} \boldsymbol{v}_1[i]\,\boldsymbol{v}_2[i].$$

In fact, a significant portion of linear algebra is geometric in nature since comes out of the need to **generalize** the basic geometry of the plane $\mathbb{R}^2$ and 3D space $\mathbb{R}^3$ to nonvisual higher-dimensional cases.

Hence, for example, the usual length of a vector $\boldsymbol{v}$ in $\mathbb{R}^2$ or $\mathbb{R}^3$ is obtained from elementary geometry via *Pythagorean Theorem* as the length of the hypotenuse of a right triangle as shown in Figure 2. This measure of length $\|\boldsymbol{v}\| = \sqrt{\boldsymbol{v}^{\mathsf{T}}\boldsymbol{v}}$ is then called *Euclidean norm* in $\mathbb{R}^d$ and its induced distance between vectors can be easily visualized in $\mathbb{R}^3$ with the aid of the parallelogram law (see Figure 2).

Simple algebra also shows that $\|\boldsymbol{v}\|$ complies with the desiderata listed in Definition 2 which, in turn, generalizes this intuitive notion of lenght and its fundamental geometric properties to more general vector spaces.

[1] It can be shown that a norm is induced by an inner product if and only if it satisfies the parallelogram law.

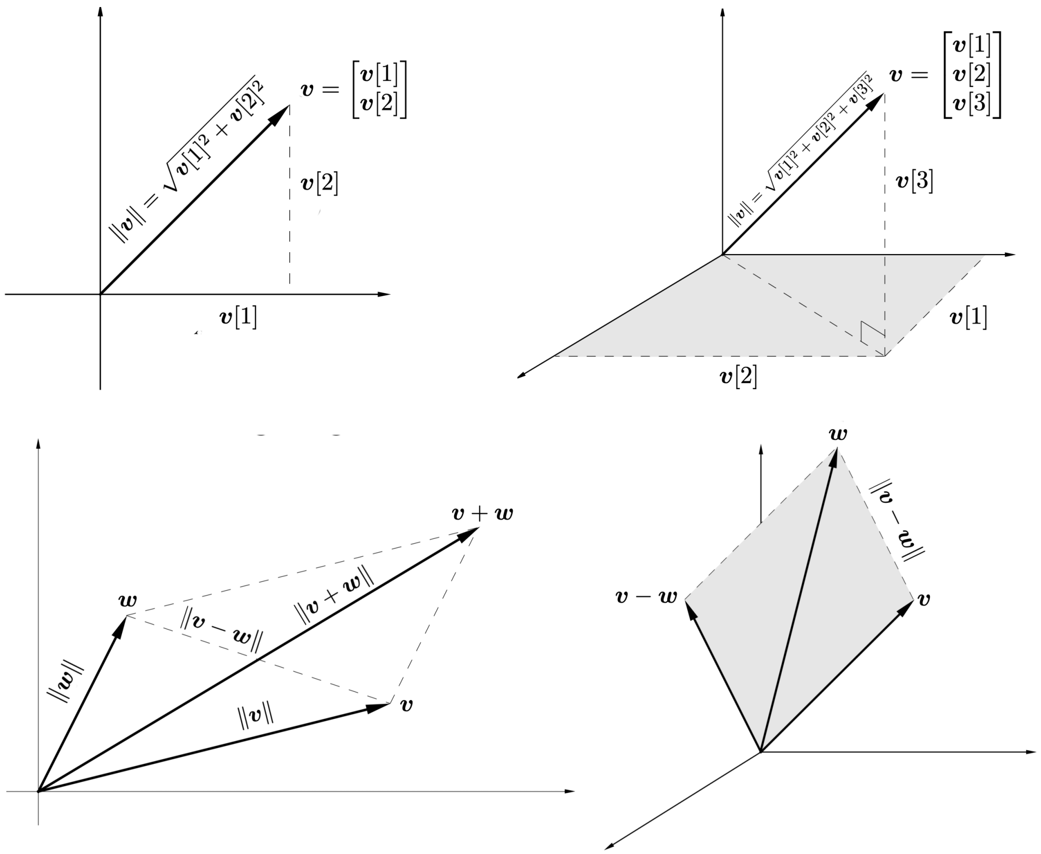

$\boldsymbol{v} = \begin{bmatrix} \boldsymbol{v}[1] \\ \boldsymbol{v}[2] \end{bmatrix}$
$\|\boldsymbol{v}\| = \sqrt{\boldsymbol{v}[1]^2 + \boldsymbol{v}[2]^2}$
$\boldsymbol{v}[2]$
$\boldsymbol{v}[1]$
$\boldsymbol{v} = \begin{bmatrix} \boldsymbol{v}[1] \\ \boldsymbol{v}[2] \\ \boldsymbol{v}[3] \end{bmatrix}$
$\|\boldsymbol{v}\| = \sqrt{\boldsymbol{v}[1]^2 + \boldsymbol{v}[2]^2 + \boldsymbol{v}[3]^2}$
$\boldsymbol{v}[3]$
$\boldsymbol{v}[1]$
$\boldsymbol{v}[2]$
$\boldsymbol{v} + \boldsymbol{w}$
$\boldsymbol{w}$
$\|\boldsymbol{v} + \boldsymbol{w}\|$
$\|\boldsymbol{v} - \boldsymbol{w}\|$
$\|\boldsymbol{w}\|$
$\|\boldsymbol{v}\|$
$\boldsymbol{v}$
$\boldsymbol{w}$
$\|\boldsymbol{v} - \boldsymbol{w}\|$
$\boldsymbol{v} - \boldsymbol{w}$
$\boldsymbol{v}$

Figure 2: Top: Length of a vector by Pythagorean Theorem. Bottom: *Parallelogram Law* and distance between vectors.

Clearly, there are notions of length on $\mathbb{R}^d$ other than the euclidean measure. Deliveroo riders, for example, navigate our city blocks following their topology including one way streets and so on, so they are prone to measure distances not as crow flies but rather in terms of lengths on a directed grid. This "grid norm" is better known as the *1-norm* being a special case of a more general class of norms defined below. Notice that the many property of this specific norm are central to the development of modern successful statistical techniques like the LASSO and its many variations.

Definition 6 (p-Norms)**.** For $p \geqslant 1$, the p-norm of $\boldsymbol{v} \in \mathbb{R}^d$ is defined as

$$\|\boldsymbol{v}\|_p^p = \sum_{i=1}^{d} |\boldsymbol{v}[i]|^p \quad \text{with} \quad \|\boldsymbol{v}\|_\infty = \lim_{p \to +\infty} \|\boldsymbol{v}\|_p = \max_i \big\{\boldsymbol{v}[i]\big\}.$$

Remark: In this family, **only** the 2-norm is induced by an inner product.

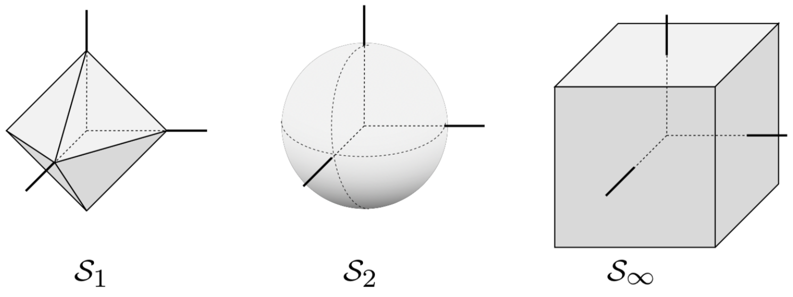


Figure 3: Unit p-spheres: $\mathcal{S}_p = \big\{\boldsymbol{v} : \|\boldsymbol{v}\|_p = 1\big\}$ for $p \in \{1, 2, +\infty\}$.

To get a feel of what's going on with this family of norms, it helps to look at the shape and relative sizes of their unit p-spheres $\mathcal{S}_p = \big\{\boldsymbol{v} : \|\boldsymbol{v}\|_p = 1\big\}$. As illustrated in Figure 3, the unit 1-, 2- and ∞-spheres in $\mathbb{R}^3$ are an octahedron, a ball, and a cube, respectively, and it's visually evident that $\mathcal{S}_1 \subset \mathcal{S}_2 \subset \mathcal{S}_\infty$. This means that $\|\boldsymbol{v}\|_1 \geqslant \|\boldsymbol{v}\|_2 \geqslant \|\boldsymbol{v}\|_\infty$ for all $\boldsymbol{v} \in \mathbb{R}^3$, and this is true more in general in $\mathbb{R}^d$.

Let's now review the notion of angle between vectors. *En passant*, we already mentioned that two vectors $\boldsymbol{v}$ and $\boldsymbol{w}$ in a generic vector space $\mathbb{V}$ are **orthogonal** if $\langle \boldsymbol{v}, \boldsymbol{w} \rangle = 0$. Why does this make sense? Let's go back to $\mathbb{R}^3$. Here two vectors are orthogonal (*perpendicular*) if the angle between them is a right angle (90°). But the visual concept of a right angle is not at our disposal in higher dimension, so we must dig a little deeper. The essence of perpendicularity in $\mathbb{R}^2$ and $\mathbb{R}^3$ is again embodied by the classical *Pythagorean Theorem*, which says that $\boldsymbol{v}$ and $\boldsymbol{w}$ are orthogonal if and only if $\|\boldsymbol{v}\|^2 + \|\boldsymbol{w}\|^2 = \|\boldsymbol{v} - \boldsymbol{w}\|$ where $\|\cdot\|$ here denotes the Euclidean norm (see Figure 4). So we can rewrite the Pythagorean statement as

$$0 = \|\boldsymbol{v}\|^2 + \|\boldsymbol{w}\|^2 - \|\boldsymbol{v} - \boldsymbol{w}\| = \boldsymbol{v}^{\mathsf{T}}\boldsymbol{v} + \boldsymbol{w}^{\mathsf{T}}\boldsymbol{w} - (\boldsymbol{v} - \boldsymbol{w})^{\mathsf{T}}(\boldsymbol{v} - \boldsymbol{w}) = \boldsymbol{v}^{\mathsf{T}}\boldsymbol{v} + \boldsymbol{w}^{\mathsf{T}}\boldsymbol{w} - \big(\boldsymbol{v}^{\mathsf{T}}\boldsymbol{v} - \boldsymbol{v}^{\mathsf{T}}\boldsymbol{w} - \boldsymbol{w}^{\mathsf{T}}\boldsymbol{v} + \boldsymbol{w}^{\mathsf{T}}\boldsymbol{w}\big) = 2\,\boldsymbol{v}^{\mathsf{T}}\boldsymbol{w}.$$

Therefore, $\boldsymbol{v}$ and $\boldsymbol{w}$ are orthogonal in $\mathbb{R}^3$ if and only if $\boldsymbol{v}^{\mathsf{T}}\boldsymbol{w} = \langle \boldsymbol{v}, \boldsymbol{w} \rangle = 0$. The natural extension of this to a generic dot product provides a definition of orthogonality in more general spaces.

Now that "right angle" in higher dimensions make sense, let's define more general angles. We will proceed as before but, rather than the Pythagorean Theorem, we will start from its most direct generalization in elementary geometry, the **Law of Cosines** in $\mathbb{R}^2$ and $\mathbb{R}^3$ which says that $\|\boldsymbol{v} - \boldsymbol{w}\|^2 = \|\boldsymbol{v}\|^2 + \|\boldsymbol{w}\|^2 - 2\,\|\boldsymbol{v}\|\,\|\boldsymbol{w}\|\cos(\theta)$. If $\boldsymbol{v}$ and $\boldsymbol{w}$ are orthogonal, then this reduces to the Pythagorean Theorem. In general,

$$\cos(\theta) = \frac{\|\boldsymbol{v}\|^2 + \|\boldsymbol{w}\|^2 - \|\boldsymbol{v} - \boldsymbol{w}\|^2}{2\,\|\boldsymbol{v}\|\,\|\boldsymbol{w}\|} = \frac{\boldsymbol{v}^{\mathsf{T}}\boldsymbol{v} + \boldsymbol{w}^{\mathsf{T}}\boldsymbol{w} - \big(\boldsymbol{v} - \boldsymbol{w}\big)^{\mathsf{T}}\big(\boldsymbol{v} - \boldsymbol{w}\big)}{2\,\|\boldsymbol{v}\|\,\|\boldsymbol{w}\|} = \frac{2\,\boldsymbol{v}^{\mathsf{T}}\boldsymbol{w}}{2\,\|\boldsymbol{v}\|\,\|\boldsymbol{w}\|} = \frac{\langle \boldsymbol{v}, \boldsymbol{w} \rangle}{\|\boldsymbol{v}\|\,\|\boldsymbol{w}\|}.$$

This is easily extended to generic inner product spaces since the *Cauchy-Schwarz inequality* guarantees that $\frac{\langle \boldsymbol{v}, \boldsymbol{w} \rangle}{\|\boldsymbol{v}\|\,\|\boldsymbol{w}\|} \in [-1, +1]$, hence we can define the **angle** between two nonzero vectors $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ as the unique value θ between 0 and π that satisfies

$$\langle \boldsymbol{v}_1, \boldsymbol{v}_2 \rangle = \|\boldsymbol{v}_1\| \cdot \|\boldsymbol{v}_2\| \cdot \cos(\theta) \quad \Leftrightarrow \quad \theta = \angle(\boldsymbol{v}_1, \boldsymbol{v}_2) = \arccos\left(\tfrac{\langle \boldsymbol{v}_1, \boldsymbol{v}_2 \rangle}{\|\boldsymbol{v}_1\| \cdot \|\boldsymbol{v}_2\|}\right),$$

where $\arccos(\cdot)$ denotes the *inverse cosine*, normalized to lie in the interval $[0, \pi]$.

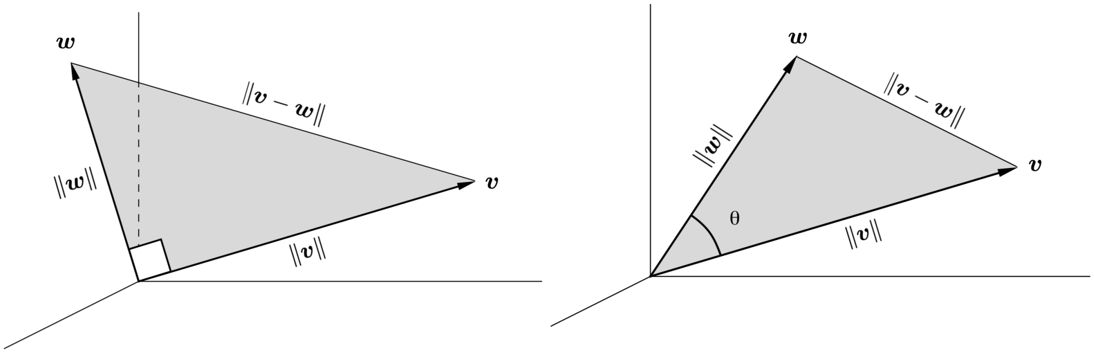


Figure 4: Angles between vectors in $\mathbb{R}^3$.

Example 3 (Correlation coefficient)**.** From a statistical point of view, if we look at $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ as centered (i.e. mean zero) n dimensional data vectors, then we can write

$$\mathbb{C}\mathrm{ov}(\boldsymbol{v}_1, \boldsymbol{v}_2) = \frac{1}{n}\sum_i \boldsymbol{v}_1[i] \cdot \boldsymbol{v}_2[i] = \frac{1}{n}\boldsymbol{v}_1^{\mathsf{T}}\boldsymbol{v}_2 = \frac{1}{n}\langle \boldsymbol{v}_1, \boldsymbol{v}_2 \rangle,$$
$$\mathbb{V}\mathrm{ar}(\boldsymbol{v}_j) = \frac{1}{n}\sum_i \boldsymbol{v}_j^2[i] = \frac{1}{n}\boldsymbol{v}_j^{\mathsf{T}}\boldsymbol{v}_j = \frac{1}{n}\|\boldsymbol{v}_j\|^2, \quad \forall\, j \in \{1, 2\}.$$

Hence, our beloved correlation coefficient can be easily rewritten as

$$\rho(\boldsymbol{v}_1, \boldsymbol{v}_2) = \frac{\mathbb{C}\mathrm{ov}(\boldsymbol{v}_1, \boldsymbol{v}_2)}{\sqrt{\mathbb{V}\mathrm{ar}(\boldsymbol{v}_1)} \cdot \sqrt{\mathbb{V}\mathrm{ar}(\boldsymbol{v}_2)}} = \frac{\langle \boldsymbol{v}_1, \boldsymbol{v}_2 \rangle}{\|\boldsymbol{v}_1\| \cdot \|\boldsymbol{v}_2\|},$$

thus $\rho = \cos(\theta)$ where $\theta = \angle(\boldsymbol{v}_1, \boldsymbol{v}_2)$. More in general, covariances *are* inner product, correlations *are* angles, and Cauchy-Schwarz is there to show that $|\rho| \leqslant 1$ with equality only if there's a perfect linear relationship between $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$.

Orthonormal Systems, Orthogonal Projections and Best Approximations.

One of the most useful constructions possible in dot product spaces are orthonormal basis expansions. Suppose that $\{\boldsymbol{\phi}_1, \dots, \boldsymbol{\phi}_d\}$ where $d \in \mathbb{N}$, form an **orthonormal set** or **system**; that is, they are mutually orthogonal and have norm 1,

$$\langle \boldsymbol{\phi}_i, \boldsymbol{\phi}_j \rangle = \delta_{i,j}.$$

If they also form a basis of $\mathbb{V}$, they are called an **orthonormal basis** (ONB). In this case, any $\boldsymbol{v} \in \mathbb{V}$ can be written as a linear combination,

$$\boldsymbol{v} = \sum_j \beta_j \cdot \boldsymbol{\phi}_j,$$

where this time the expansion coefficients $\{\beta_j\}_j$ have a *very* specific form, namely $\beta_j = \langle \boldsymbol{v}, \boldsymbol{\phi}_j \rangle$ and are called **generalized Fourier coefficients** for reasons that will become clear later. Now, to understand why this is actually the case, we may follow different paths. The simplest is just algebra. In fact, being $\{\boldsymbol{\phi}_1, \dots, \boldsymbol{\phi}_d\}$ a basis, we said that any vector $\boldsymbol{v}$ can be written as a linear combination of its elements $\boldsymbol{v} = \sum_j \beta_j \cdot \boldsymbol{\phi}_j$, now

$$\langle \boldsymbol{v}, \boldsymbol{\phi}_j \rangle = \Big\langle \sum_i \beta_i \, \boldsymbol{\phi}_i, \boldsymbol{\phi}_j \Big\rangle = \sum_i \beta_i \cdot \langle \boldsymbol{\phi}_i, \boldsymbol{\phi}_j \rangle = \beta_j \cdot \|\boldsymbol{\phi}_j\|^2 = \beta_j.$$

Definition 7 (Generalized Fourier Expansions)**.** If $\mathscr{B}_d = \{\boldsymbol{\phi}_1, \dots, \boldsymbol{\phi}_d\}$ is an *orthonormal basis* for an inner-product space $\mathbb{V}$, then each $\boldsymbol{v} \in \mathbb{V}$ can be expressed as

$$\boldsymbol{v} = \sum_j \langle \boldsymbol{v}, \boldsymbol{\phi}_j \rangle \cdot \boldsymbol{\phi}_j, \tag{5}$$

This is called the *(Generalized) Fourier Expansion*[a] of $\boldsymbol{v}$. The scalars $\beta_j = \langle \boldsymbol{v}, \boldsymbol{\phi}_j \rangle$ are the coordinates of $\boldsymbol{v}$ with respect to $\mathscr{B}_d$, and they are called the *(Generalized) Fourier Coefficients* (GFC).

[a] Jean Baptiste Joseph Fourier (1768–1830) was a French mathematician and physicist who, while studying heat flow, developed expansions similar to Equation 5. Fourier's work dealt with special infinite-dimensional inner-product spaces involving trigonometric functions.

Example 4. (Canonical basis in $\mathbb{R}^d$) Using as an example the canonical dot product in $\mathbb{R}^d$, we see that, for the *canonical basis* of $\mathbb{R}^d$, each coefficient $\beta_j = \langle \boldsymbol{v}, \boldsymbol{\phi}_j \rangle$ just picks out one entry from the column vector $\boldsymbol{v}$, thus $\boldsymbol{v} = \sum_j \boldsymbol{v}[j] \, \boldsymbol{\phi}_j$.

Geometrically, the Fourier expansion resolves $\boldsymbol{v}$ into d mutually orthogonal vectors, each of which represents the orthogonal projection of $\boldsymbol{v}$ onto the space (line) spanned by $\boldsymbol{\phi}_j$. Let's pursue this point of view a bit further in $\mathbb{R}^d$ – although it can be easily generalized to a generic vector space – by considering the matrix Φ having the basis elements $\{\boldsymbol{\phi}_j\}_j$ as column vectors

$$\underset{(d \times d)}{\Phi} = \big[\boldsymbol{\phi}_1 \cdots \boldsymbol{\phi}_d\big] \quad \rightsquigarrow \quad \text{take a generic vector of coefficients } \boldsymbol{\alpha} \in \mathbb{R}^d \text{ and consider } \widehat{\boldsymbol{v}}(\boldsymbol{\alpha}) = \sum_j \alpha_j \, \boldsymbol{\phi}_j = \Phi \, \boldsymbol{\alpha}.$$

Now, for a fixed target vector $\boldsymbol{v} \in \mathbb{R}^d$, we may ask: how do we have to choose $\boldsymbol{\alpha}$ in order to obtain the **best approximation** to $\boldsymbol{v}$? Of course we have to qualify better what do we mean by *best approximation*. Let's take a **Least Squares** perspective, and let's try to minimize the Euclidean distance between $\boldsymbol{v}$ and its approximation $\widehat{\boldsymbol{v}}(\boldsymbol{\alpha})$, that is

$$\boldsymbol{\alpha}_{\text{LS}} = \underset{\boldsymbol{\alpha} \in \mathbb{R}^d}{\operatorname{argmin}} \, \big\| \widehat{\boldsymbol{v}}(\boldsymbol{\alpha}) - \boldsymbol{v} \big\|^2 = \underset{\boldsymbol{\alpha} \in \mathbb{R}^d}{\operatorname{argmin}} \, \big(\Phi \, \boldsymbol{\alpha} - \boldsymbol{v}\big)^{\intercal} \big(\Phi \, \boldsymbol{\alpha} - \boldsymbol{v}\big) = \underset{\boldsymbol{\alpha} \in \mathbb{R}^d}{\operatorname{argmin}} \, \big\{ \boldsymbol{\alpha}^{\intercal} (\Phi^{\intercal} \Phi) \, \boldsymbol{\alpha} - 2 \, \boldsymbol{\alpha}^{\intercal} \Phi^{\intercal} \, \boldsymbol{v} + \boldsymbol{v}^{\intercal} \boldsymbol{v} \big\}.$$

Differentiating this last expression w.r.t. $\boldsymbol{\alpha}$, setting to zero and solving, we get

$$\frac{\partial}{\partial \boldsymbol{\alpha}} \boldsymbol{\alpha}^{\intercal} (\Phi^{\intercal} \Phi) \, \boldsymbol{\alpha} - 2 \, \boldsymbol{\alpha}^{\intercal} \Phi^{\intercal} \, \boldsymbol{v} + \boldsymbol{v}^{\intercal} \boldsymbol{v} = 0 \quad \Leftrightarrow \quad \Phi^{\intercal} \Phi \, \boldsymbol{\alpha} = \Phi^{\intercal} \boldsymbol{v} \quad \Leftrightarrow \quad \boldsymbol{\alpha}_{\text{LS}} = \big(\Phi^{\intercal} \Phi\big)^{-1} \Phi^{\intercal} \boldsymbol{v}. \tag{6}$$

Notice that Φ is an **orthogonal matrix** having *orthonormal* vector as columns, hence $\Phi^{-1} = \Phi^{\intercal}$ and consequently the **Gramian matrix** $(\Phi^{\intercal} \Phi) = \mathbb{I}_d$. As expected then, we finally get that the best approximation in the 2-norm sense is actually an **interpolation** because

$$\boldsymbol{\alpha}_{\text{LS}} = \Phi^{\intercal} \boldsymbol{v} = \begin{bmatrix} \boldsymbol{\phi}_1^{\intercal} \\ \vdots \\ \boldsymbol{\phi}_d^{\intercal} \end{bmatrix} \boldsymbol{v} = \begin{bmatrix} \boldsymbol{\phi}_1^{\intercal} \boldsymbol{v} \\ \vdots \\ \boldsymbol{\phi}_d^{\intercal} \boldsymbol{v} \end{bmatrix} = \begin{bmatrix} \langle \boldsymbol{v}, \boldsymbol{\phi}_1 \rangle \\ \vdots \\ \langle \boldsymbol{v}, \boldsymbol{\phi}_d \rangle \end{bmatrix} = \underset{(\text{GFC})}{\boldsymbol{\beta}} \quad \Rightarrow \quad \widehat{\boldsymbol{v}}(\boldsymbol{\alpha}_{\text{LS}}) = \Phi \, \boldsymbol{\alpha}_{\text{LS}} = (\Phi \, \Phi^{\intercal}) \, \boldsymbol{v} = \boldsymbol{v}. \tag{7}$$

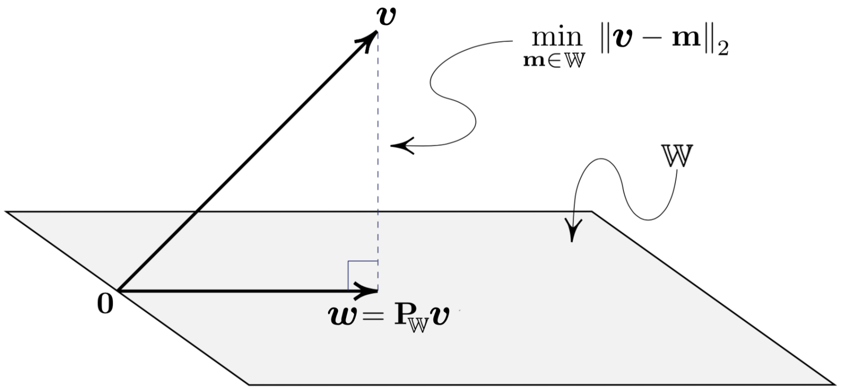
Nothing really changes if we look for a *reduced-rank* approximation based only on the first $k \leqslant d$ elements, for example, of our orthonormal basis $\mathscr{B}_d$. In fact, collecting these vectors in the $(d \times k)$ matrix $\Phi_k = \big[\boldsymbol{\phi}_1 \cdots \boldsymbol{\phi}_k\big]$, we get

$$(\Phi_k^{\intercal} \Phi_k) = \underset{(k \times d)}{\begin{bmatrix} \boldsymbol{\phi}_1^{\intercal} \\ \vdots \\ \boldsymbol{\phi}_d^{\intercal} \end{bmatrix}} \underset{(d \times k)}{\big[\boldsymbol{\phi}_1 \cdots \boldsymbol{\phi}_k\big]} = \underset{(k \times k)}{\begin{bmatrix} 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}} \quad \Rightarrow \quad \boldsymbol{\alpha}_{\text{LS}} = \big(\Phi_k^{\intercal} \Phi_k\big)^{-1} \Phi_k^{\intercal} \boldsymbol{v} = \Phi_k^{\intercal} \boldsymbol{v} = \underset{(\text{GFC})}{\boldsymbol{\beta}_k} \text{ and } \widehat{\boldsymbol{v}}_k(\boldsymbol{\alpha}_{\text{LS}}) = \Phi_k \, \boldsymbol{\beta}_k = \sum_{j=1}^{k} \underbrace{\langle \boldsymbol{v}, \boldsymbol{\phi}_j \rangle}_{\beta_j} \, \boldsymbol{\phi}_j,$$

where $\boldsymbol{\beta}_k = \boldsymbol{\beta}[1:k]$. Based on what we have seen so far, the *residuals* or *approximation error* we incur can be written as

$$\left\|\widehat{\boldsymbol{v}}_k(\boldsymbol{\alpha}_{\mathrm{LS}}) - \boldsymbol{v}\right\|^2 = \left\|\sum_{j=1}^{k} \beta_j\, \boldsymbol{\phi}_j - \sum_{j=1}^{d} \beta_j\, \boldsymbol{\phi}_j\right\|^2 = \left\|\sum_{j=k+1}^{d} \beta_j\, \boldsymbol{\phi}_j\right\|^2 = \left(\Phi_{-k}\boldsymbol{\beta}_{-k}\right)^{\mathtt{T}}\left(\Phi_{-k}\boldsymbol{\beta}_{-k}\right) = \boldsymbol{\beta}^{\mathtt{T}}_{-k}\left(\Phi^{\mathtt{T}}_{-k}\Phi_{-k}\right)\boldsymbol{\beta}_{-k} = \boldsymbol{\beta}^{\mathtt{T}}_{-k}\boldsymbol{\beta}_{-k} = \left\|\boldsymbol{\beta}_{-k}\right\|^2,$$

where, in `R` notation, $\boldsymbol{\beta}_{-k} = \boldsymbol{\beta}[(k+1):d]$ and $\Phi_{-k} = \left[\boldsymbol{\phi}_{k+1} \cdots \boldsymbol{\phi}_d\right]$. The notion of **orthogonal projection** in higher-dimensional spaces is consistent with the visual geometry in $\mathbb{R}^2$ and $\mathbb{R}^3$. In particular, it is visually evident from Figure 5 that if $\mathbb{W}$ is a generic subspace of $\mathbb{R}^3$ – for example the subspace spanned by the first 2 basis vectors $\{\boldsymbol{\phi}_1, \boldsymbol{\phi}_2\}$ – and if $\boldsymbol{v}$ is a vector outside of $\mathbb{W}$, then the point in $\mathbb{W}$ that is **closest** to $\boldsymbol{v}$ is $\boldsymbol{w} = P_{\mathbb{W}}\boldsymbol{v}$, the **orthogonal projection** of $\boldsymbol{v}$ onto $\mathbb{W}$.



$\boldsymbol{v}$
$\min_{\mathbf{m}\in\mathbb{W}} \|\boldsymbol{v} - \mathbf{m}\|_2$
$\mathbb{W}$
$\mathbf{0}$
$\boldsymbol{w} = \mathrm{P}_{\mathbb{W}}\boldsymbol{v}$

Figure 5: Best approximations as orthogonal projections.

The situation is exactly the same in higher dimensions. In fact we have the following result:

Theorem 1 (Closest Point Theorem). *Let $\mathbb{W}$ be a subspace of an inner-product space $\mathbb{V}$, and let $\boldsymbol{v}$ be a vector in $\mathbb{V}$. The unique vector in $\mathbb{W}$ that is closest to $\boldsymbol{v}$ is $\boldsymbol{w} = P_{\mathbb{W}}\,\boldsymbol{v}$, the orthogonal projection of $\boldsymbol{v}$ onto $\mathbb{W}$. In other words,*

$$\min_{\boldsymbol{m}\in\mathbb{W}} \left\|\boldsymbol{v} - \boldsymbol{m}\right\| = \left\|\boldsymbol{v} - P_{\mathbb{W}}\,\boldsymbol{v}\right\| \overset{\mathtt{def.}}{=} \mathtt{dist}\left(\boldsymbol{v}, \mathbb{W}\right).$$

This is called the orthogonal distance *between $\boldsymbol{v}$ and $\mathbb{W}$.*

Example 5. (Euclidean spaces) Consider $\mathbb{R}^d$ endowed with the *canonical* inner product. Our previous development shows that if $\mathbb{W} = \mathrm{span}\left(\{\boldsymbol{\phi}_j\}_{j=1}^{k}\right)$, then for any $\boldsymbol{v} \in \mathbb{R}^d$ its closest point, i.e. its best approximation w.r.t. the 2-norm, is

$$\sum_{j=1}^{k} \underbrace{\langle \boldsymbol{v}, \boldsymbol{\phi}_j\rangle}_{\beta_j}\, \boldsymbol{\phi}_j \overset{\mathtt{Th.}1}{=} P_{\mathbb{W}}\boldsymbol{v} \quad \text{and also} \quad \mathtt{dist}\left(\boldsymbol{v}, \mathbb{W}\right) = \left\|\boldsymbol{\beta}_{-k}\right\|.$$

<u>Remarks</u>: Viewing concepts from more than one perspective generally produces deeper understanding, and this is particularly true for the theory above, the theory of **Least Squares**. In our journey we went from the classical calculus-based solution to a vector space interpretation via Theorem 1. But we could also go the other way around, that is, by starting from a purely geometrical approach and landing on the same conclusion, a simple and, possibly, more direct path, that puts the entire picture in much sharper focus.

Before moving on, a couple of remarks:

<u>Remark</u>:

1. What are the relationships between orthogonality, orthonormality and linear independence? Working in $\mathbb{R}^d$, it is not difficult to find two vectors that are linear independent but <u>not</u> orthogonal. Similarly, every *orthogonal* set need <u>not</u> be linearly independent, as orthogonal sets can certainly include the zero vector, and any set which contains it is necessarily linearly dependent. On the other hand, two <u>nonzero</u> *orthogonal* vector <u>are</u> linearly independent. In fact, let $\{\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_k\}$ be a set of nonzero orthogonal vectors, and suppose there are coefficients $\{c_1, \ldots, c_k\}$ such that $c_1 \cdot \boldsymbol{\phi}_1 + \cdots + c_k \cdot \boldsymbol{\phi}_k = \mathbf{0}_{\mathbb{V}}$. Then, taking the inner product on both sides of this equality with any vector $\boldsymbol{\phi}_j$ for $j \in \{1, \ldots, k\}$ we see that

$$\langle c_1 \cdot \boldsymbol{\phi}_1 + \cdots + c_k \cdot \boldsymbol{\phi}_k, \boldsymbol{\phi}_j\rangle = 0 \quad \Leftrightarrow \quad \sum_{r=1}^{k} c_r \langle \boldsymbol{\phi}_r, \boldsymbol{\phi}_j\rangle = 0 \quad \overset{\text{nonzero}}{\Leftrightarrow} \quad \sum_{r=1}^{k} c_r \delta_{r,j} = 0 \quad \Leftrightarrow \quad c_j = 0 \quad \forall\, j \in \{1, \ldots, k\},$$

hence $\{\boldsymbol{\phi}_1, \dots, \boldsymbol{\phi}_k\}$ are also linearly independent. As a corollary of this result, every *orthonormal* set <u>is</u> linearly independent as, by definition, it is an orthogonal set consisting of nonzero vectors only.

2. Let's restate the ubiquitous **Pythagorean Theorem** in light of what we've seen so far. Referring to Figure 5, we see that in its general form, it reads as follows:

Theorem 2 (Pythagoras)**.** *If $\{\boldsymbol{\phi}_1, \dots, \boldsymbol{\phi}_k\}$ are orthonormal (they need <u>not</u> form a basis), and $\mathbb{W} = span\big(\{\boldsymbol{\phi}_j\}_{j=1}^k\big)$ then*

$$\|\boldsymbol{v}\|^2 = \|P_{\mathbb{W}}\boldsymbol{v}\|^2 + \|\boldsymbol{v} - P_{\mathbb{W}}\boldsymbol{v}\|^2 \quad \Leftrightarrow \quad \|\boldsymbol{v}\|^2 = \sum_{j=1}^{k} \langle \boldsymbol{v}, \boldsymbol{\phi}_j \rangle^2 + \left\| \boldsymbol{v} - \sum_{j=1}^{k} \langle \boldsymbol{v}, \boldsymbol{\phi}_j \rangle \, \boldsymbol{\phi}_j \right\|^2$$

3. There is a famous technique that goes under the name of **Gram-Schmidt algorithm** that produces an orthonormal collection of vectors $\{\boldsymbol{\phi}_1, \dots, \boldsymbol{\phi}_k\}$ out a linearly independent system $\{\boldsymbol{v}_1, \dots, \boldsymbol{v}_k\}$. The G-S algorithm is also able to find the first vector in the input system (if any) that is a linear combination of the previous ones and terminates.

Algorithm 1 (Gram-Schmidt algorithm)**.**

```
INPUT: $d$–vectors $\{\boldsymbol{v}_1, \dots, \boldsymbol{v}_k\}$.
INIT: $\boldsymbol{\phi}_1 = \frac{\boldsymbol{v}_1}{\|\boldsymbol{v}_1\|}$.

for($j$ in $2:k$)

    1. $\tilde{\boldsymbol{\phi}}_j = \boldsymbol{v}_j - P_{j-1}\boldsymbol{v}_j$      $\rightsquigarrow$  orthogonalization, where $P_{j-1}\boldsymbol{v}_j = \sum_{r=1}^{j-1} \langle \boldsymbol{v}_j, \boldsymbol{\phi}_r \rangle \boldsymbol{\phi}_j$,
    2. if $\tilde{\boldsymbol{\phi}}_j = \mathbf{0}$, quit          $\rightsquigarrow$  test for linear dependence,
    3. $\boldsymbol{\phi}_j = \frac{\tilde{\boldsymbol{\phi}}_j}{\|\tilde{\boldsymbol{\phi}}_j\|}$                  $\rightsquigarrow$  normalization.
```

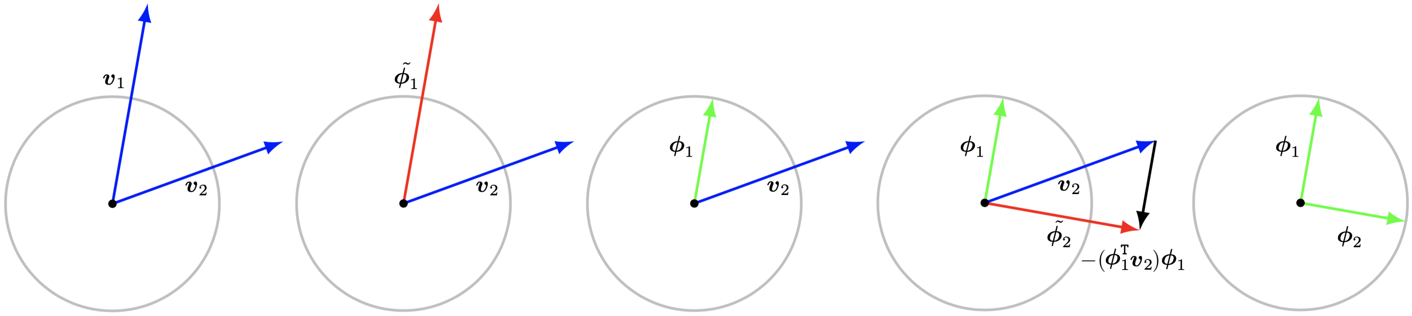


Figure 6: Gram-Schmidt algorithm applied to two 2-vectors.

Function as vectors

Consider $\mathscr{F}\big([a,b]\big)$, the set of all real (or complex) valued functions $m(\cdot)$ on the interval $[a,b]$. This is a **vector space** over the field of the real (or complex) numbers: given two functions $m_1(\cdot)$ and $m_2(\cdot)$, and two real numbers β_1 and β_2, we can form the sum $h(x) = \beta_1 \cdot m_1(x) + \beta_2 \cdot m_2(x)$ and the result is still a function on the same interval. Examination of the axioms listed in Definition 1 will show that $\mathscr{F}\big([a,b]\big)$ possesses all the other attributes of a vector space as well. We may think of the array of numbers $\big\{m(x)\big\}_{x\in[a,b]}$ as being the components of the vector. Intuitively, since there is a continuum of independent components - one for each point x - the space of functions is *infinite* dimensional.

The set of all functions is usually too large for us and we need to add some structure. Often a *nonparametric regression* function or *classifier* is chosen to lie in some function space, where the assumed structure is exploited by algorithms and theoretical analysis. As motivation, consider *nonparametric regression*. We observe $\{(\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_n, Y_n)\}$ and we want to estimate $m(\boldsymbol{x}) = \mathbb{E}(Y \mid \boldsymbol{X} = \boldsymbol{x})$. We cannot simply choose $m(\cdot)$ to minimize the training error $\sum_i \big(Y_i - m(\boldsymbol{X}_i)\big)^2$ as this will lead to interpolating the data. One approach is to minimize $\sum_i \big(Y_i - m(\boldsymbol{X}_i)\big)^2$ while restricting $m(\cdot)$ to be in a well behaved function space.

Hence, we will restrict ourselves to subspaces of functions with nice properties, such as being continuous or differentiable or with even more involved smoothness conditions (e.g. Lipschitz space, Holder spaces, Sobolev spaces, Besov spaces, etc). There is some fairly standard notation for these spaces: the space of functions which have k continuous derivatives is denoted by $\mathscr{C}^k\big([a,b]\big)$. For infinitely *smooth* functions, those with derivatives of *all* orders, we write $\mathscr{C}^\infty\big([a,b]\big)$. For the space of *analytic* functions, those whose Taylor expansion actually converges to the function, we write $\mathscr{C}^\omega\big([a,b]\big)$.

Norms and convergence

We can seldom write down an exact solution function to a real-world problem. We are usually forced to use numerical methods, or to expand as a power series in some small parameter. The result is a sequence of approximate solutions $m_n(x)$, which we hope will converge to the desired exact solution $m(x)$ as we make the numerical grid smaller, or take more terms in the power series, or get more and more data. Because as for generic vectors, there is more than one way to measure of the "size" of a function, the convergence of a sequence of functions to a limit function is not as simple a concept as the convergence of a sequence of numbers. Convergence means that the distance between the $m_n(\cdot)$ and the limit function $m(\cdot)$ gets smaller and smaller as n increases, so each different measure of this distance provides a new notion of what it means to converge. We are not going to make much use of formal "$\epsilon - \delta$" analysis, but you must realize that this distinction between different forms of convergence is not merely academic.

Here are some common forms of convergence depending on the notion of norm[2] we choose (and the induced distance!):

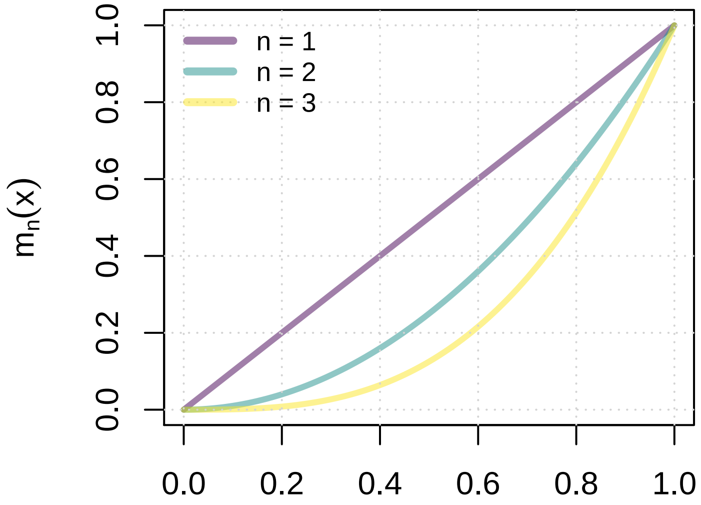
1. If, for each fixed $x \in \mathcal{X}$, its domain of definition, the set of numbers $\{m_n(x)\}_n$ converges to the number $m(x)$, then we say the sequence **converges pointwise**.
2. If the maximum separation
$$\|m\| = \sup_{x\in\mathcal{X}} \big|m(x)\big| \quad \rightsquigarrow \quad \sup_{x\in\mathcal{X}} \big|m_n(x) - m(x)\big|$$
goes to 0 as n diverges, then we say that $m_n(\cdot)$ **converges uniformly** to $m(\cdot)$ on $\mathcal{X}$.
3. If
$$\|m\| = \int_{\mathcal{X}} \big|m(x)\big|\,\mathrm{d}x \quad \rightsquigarrow \quad \int_{\mathcal{X}} \big|m_n(x) - m(x)\big|\,\mathrm{d}x$$
goes to zero as n diverges, then we say that $m_n(\cdot)$ **converges in mean** to $m(\cdot)$ on $\mathcal{X}$.

Uniform convergence implies pointwise convergence, but not vice versa. If $\mathcal{X}$ is a finite interval, then uniform convergence implies convergence in the mean, but convergence in the mean implies neither uniform nor pointwise convergence.

Example 6. Consider the sequence of function $m_n(x) = x^n$ on $\mathcal{X} = [0,1)$. To be clear, here, the round and square bracket notation means that the point $x = 0$ is included in the interval, but the point 1 is excluded.

As n becomes large we have $x^n \to 0$ pointwise in $\mathcal{X}$, but the convergence is not uniform because
$$\sup_{x\in[0,1)} \big|x^n - 0\big| = 1,$$
for all n. Replacing $\mathcal{X} = [0,1)$ with $\mathcal{X} = [0,1]$ we have neither uniform nor pointwise convergence of x^n to zero, but $x^n \to 0$ in mean.

[2] It can be shown that those are legit norms on the vector space of function on $\mathcal{X}$.



Norms from integrals: L^p spaces

The space $\mathtt{L}_p([a,b])$ for any $1 \leqslant p < \infty$, is defined to be $\mathscr{F}\big([a,b]\big)$ equipped with the natural extension of the p-norm (see Definition 6) to the continuum of indices that characterizes the function case, that is

$$\|m\|_p^p = \int_a^b \big|m(x)\big|^p \,\mathrm{d}x.$$

Hence, more formally, we define

$$\mathtt{L}_p([a,b]) = \{m : [a,b] \mapsto \mathbb{R} \text{ such that } \|m\|_p < \infty\}\,.$$

We say that $m_n \to m$ in $\mathtt{L}_p$ if the $\mathtt{L}_p$ distance $\|m(x) - m_n(x)\|_p$ tends to zero. We have already seen the $\mathtt{L}_1$ measure of distance in the definition of convergence in mean. As in that case, convergence in $\mathtt{L}_p$ says nothing about pointwise convergence.

Remarks:

- It seems intuitive to regard $\|\cdot\|_p$ as a norm. Notice however, that it is possible for a function to have $\|m\|_p = 0$ without $m(\cdot)$ being identically zero - a function that vanishes at all but a finite set of points, for example. This pathology violates the third requirement in our list for something to be called a norm (see Definition 2), but we circumvent the problem by simply declaring all such functions to be equivalent to the function that is identically zero.

  As awkward as it sounds, this means that elements of $\mathtt{L}_p$ spaces are not really functions, but only **equivalence classes** of functions - two functions being regarded as the same if they differ by a function of zero length, that is, by a function that is $\mathtt{L}_p$-equivalent to the identically zero function.

  Clearly these spaces are not for use when anything significant depends on the value of the function at any precise point. Consequently, these spaces have no hard-coded notion of **smoothness** inside: we may take any infinitely smooth function, wildly modify it on any set of x's with zero Lebesgue measure, and irreparably disrupt its smoothness still preserving its $\mathtt{L}_p$ nature.

An important property for any space to have is that of being **complete**. Roughly speaking, a space is complete if when some sequence of elements of the space "look" as if they are converging, then they are indeed converging and their limit is an element of the space. To make this concept precise, we need to say what we mean by the phrase "look as if they are converging". This we do by introducing the idea of a **Cauchy sequence**.

Definition 8 (Cauchy Sequence)**.** A sequence $\{m_n\}_n$ in a normed space is **Cauchy** if for any $\epsilon > 0$ we can find an $n_0 \in \mathbb{N}$ such that, for each $n_1, n_2 > n_0$ we have that $\|m_{n_1} - m_{n_2}\| < \epsilon$.

This definition can be loosely paraphrased to say that the elements of a Cauchy sequence get arbitrarily close to each other as n diverges.

Definition 9 (Banach space)**.** A normed vector space is **complete** with respect to its norm if every Cauchy sequence actually converges to some element in the space.

A complete normed vector space is called a **Banach space**.

If we interpret the norms as Lebesgue integrals then the $\mathtt{L}_p$ are complete, and therefore Banach spaces. The theory of Lebesgue integration is rather complicated, however, and is not really necessary.

$\mathtt{L}_2$ and Hilbert spaces

Remember what we said talking about p-norms: the 2-norm is the only one in this family induced by an inner product (see Definition 6): extending this result to our function setup, it makes $\mathtt{L}_2$ special!

The *Banach space* $\mathtt{L}_2$ in fact, is also a **Hilbert space**. This means that its norm is derived from an inner product[3]. If we define the inner product in the most natural way by extending the Euclidean finite dimensional case, we get

$$\langle m_1, m_2 \rangle_2 = \int_a^b m_1(x)\, m_2(x)\, \mathtt{d}x,$$

then the $\mathtt{L}_2$ norm can be written

$$\|m\|_2 = \sqrt{\langle m, m \rangle_2}.$$

When we omit the subscript on a norm, we mean it to be this one.

Being positive definite, we know that the inner product satisfies the Cauchy-Schwarz inequality

$$\big|\langle m_1, m_2 \rangle\big| \leqslant \|m_1\| \cdot \|m_2\|.$$

An interesting consequence that we did not mention in the finite dimensional case, is the following. Suppose that $m_n \to m$ in the $\mathtt{L}_2$ sense, that is $\|m_n - m\| \to 0$, then, for any other $h \in \mathtt{L}_2$, we get

$$\big|\langle m_n, h \rangle - \langle m, h \rangle\big| = \big|\langle (m_n - m), h \rangle\big| \overset{\mathtt{C-S}}{\leqslant} \|m_n - m\| \cdot \underset{<\infty}{\|h\|} \to 0,$$

and we can conclude that $\langle m_n, h \rangle \to \langle m, h \rangle$. In other words, the inner product is a **continuous functional** of its arguments. Please notice that this continuity hinges on $\|h\|$ being finite. It is for this reason that we do not permit $\|h\| = \infty$ functions to be elements of our Hilbert space.

Remark:

- Please note once again that $\mathtt{L}_2$ convergence $\|m_n - m\| \to 0$ does not imply pointwise convergence $m_n(x) \to m(x)$ for any fixed x. Remember that in $\mathtt{L}_2$, and $\mathtt{L}_p$ spaces more in general, it doesn't even make sense to talk about the value of a function in a single point x...these are classes of functions! For all this to be meaningful, we need the space to be what is called a **reproducing kernel Hilbert space** (RKHS) which we will discuss later.

Once we have an inner product, we know we can introduce the notion of *orthonormal set*. Of course a set of functions $\{\phi_j\}_j$ is orthonormal if $\langle \phi_j, \phi_k \rangle = \delta_{j,k}$. For example, after some algebra we can show that

$$2\int_0^2 \cos(j\pi x) \cdot \cos(k\pi x)\, \mathrm{d}x = \delta_{j,k}, \quad j, k \in \mathbb{N},$$

so the set of functions

$$\phi_0(x) = 1, \quad \phi_j(x) = \sqrt{2}\cos(j\,\pi\,x), \quad j \geqslant 1,$$

is **orthonormal** on $[0, 1]$.

This set of functions is also **complete** but in a different sense from our earlier use of this word (see Definition 9),

Definition 10 (Completeness)**.** A orthonormal set of functions $\{\phi_j\}_j$ in $\mathtt{L}_2$ is said to be **complete** if any $m \in \mathtt{L}_2$ admits a convergent (in $\mathtt{L}_2$ sense) expansion

$$m(x) = \sum_{j=0}^{\infty} \beta_j \cdot \phi_j(x),$$

for some set of coefficients $\{\beta_j\}_j$.

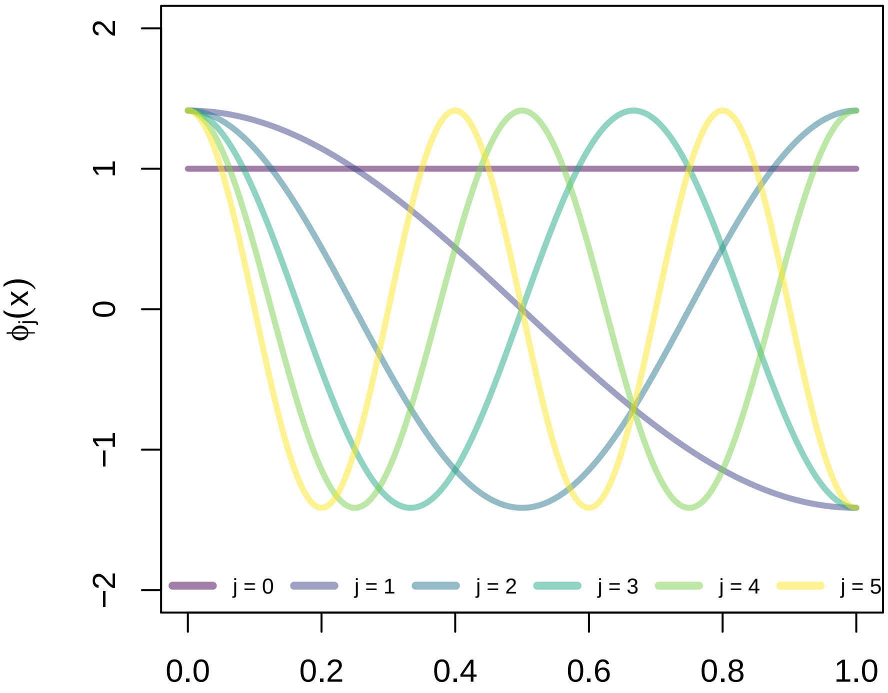
As before, if we assume that such an expansion exists, and that we can freely interchange the order of the sum and integral, we can multiply both sides of this expansion by $\phi_k(x)$, integrate over x, and use the orthonormality of the ϕ's to read off the expansion coefficients as

$$\beta_j = \langle m, \phi_j \rangle = \int m(x) \cdot \phi_j(x)\, \mathrm{d}x,$$

our beloved **Generalized Fourier Coefficients** (GFC), once again (see Definition 7).

[3] Every Hilbert space is a Banach space but the reverse is not true in general.

```r
# (Install if needed) and load some color palettes
suppressMessages(require("viridis", quietly = T))
# Cosine-basis
cos.basis = function(x, j = 4) 1*(j == 0) + sqrt(2)*cos(pi*j*x)*(j > 0)
# Plot the first 6 basis functions
j.max = 5
mycol = viridis(j.max + 1, alpha = .5)
# Open the graphical device
curve(cos.basis(x,0), n = 501,  ylim = c(-2,2), col = mycol[1], lwd = 3,
      main = " ", xlab = "", ylab = expression(phi[j](x)))
# Add the other basis functions
for (idx in 1:j.max) curve(cos.basis(x, j = idx), n = 501, add = T, col = mycol[idx + 1], lwd = 3)
legend("bottom", paste("j =", 0:j.max), col = mycol, lwd = 4, cex = .7, bty = "n", horiz = T)
```



Remarks:

1. In our developments we are focusing on L_2 but the results still apply to more general Hilbert spaces.
2. In particular, a useful feature of general Hilbert spaces is that they come with a generalization of the concept of a **basis**. Recall that a basis is a set of vectors that allows us to uniquely write each vector as a linear combination. In the context of infinite- dimensional Hilbert spaces, this is quite restrictive (note that linear combinations always involve finitely many terms) and leads to bases that are not countable. Therefore, as suggested by Definition 10, we usually work with a complete orthonormal system or an **orthonormal basis** (ONB).

   Formally, an ONB can be characterized with the property that no other nonzero vector is orthogonal to all of its elements.
3. *Every* Hilbert space admits an ONB. A Hilbert space is **separable** if there exists a countable orthonormal basis. As an example, L_2 with the inner product we have defined, is a separable Hilbert space.
4. Sticking to L_2, we may think to the basis elements $\phi_j(x)$ as function prototypes: simplified, even *cartoonish*, models that we compare to a function of interest to break down and better understand its complexity. Thinking statistically, the GFC's, being inner products, are simply measuring a form of correlation - or similarity more in general - between our function and the set of prototypes we have chosen to represent it: the larger a coefficient β_j is, the more our function resemble its associated prototype.
5. As one can imagine, for a given Hilbert space (e.g. L_2) there are many possible ONB's available. The way we select one system over another crucially depends on its ability to better "capture" a smoothness condition of interest to us (e.g. wavelets are typically optimal over Besov classes).

Example 7 (Other bases)**.**

Fourier basis on $[0,1]$:

$$\phi_1(x) = 1, \quad \phi_{2j}(x) = \frac{1}{\sqrt{2}}\cos(2\,j\,\pi\,x), \quad \phi_{2j}(x) = \frac{1}{\sqrt{2}}\sin(2\,j\,\pi\,x) \quad j \geqslant 1.$$

Fourier basis on $[0,2\pi]$:

$$\phi_1(x) = \frac{1}{\sqrt{2\pi}}\mathrm{e}^{ix}, \quad \phi_2(x) = \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-ix}, \quad \phi_3(x) = \frac{1}{\sqrt{2\pi}}\mathrm{e}^{2ix}, \quad \phi_3(x) = \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-2ix}, \cdots$$

Legendre basis on $[-1,+1]$:

$$P_0(x) = 1, \quad P_1(x) = x, \quad P_2(x) = \frac{1}{2}(3x^2-1), \quad P_3(x) = \frac{1}{2}(5x^3-3x), \cdots$$

These polynomials are defined by the relation

$$P_j(x) = \frac{1}{2^j\,j!}\frac{\mathrm{d}^j}{\mathrm{d}x^j}(x^2-1)^j.$$

The Legendre polynomials are orthogonal but not orthonormal, since

$$\int_{-1}^{+1} P_j^2(x)\,\mathrm{d}x = \frac{2}{2j+1}.$$

However, we can define modified Legendre polynomials as

$$\phi_j(x) = \sqrt{\frac{2j+1}{2}}\,P_j(x),$$

which then form an orthonormal basis for $\mathtt{L}_2([-1,+1])$.

Haar basis on $[0,1]$: Define the following two functions, usually called *father* and *mother* wavelet respectively

$$\phi(x) = \mathbb{I}_{[0,1]}(x), \qquad \psi(x) = \begin{cases} -1, & \text{if } x \in \left[0, \frac{1}{2}\right] \\ +1 & \text{if } x \in \left[\frac{1}{2}, 1\right] \end{cases}.$$

The Haar basis consists of the functions

$\left\{\phi(x), \psi_{j,k}(x) : j \in \mathbb{N}, k \in \{0,1,\ldots,2^j-1\}\right\}$, with $\psi_{j,k}(x) = 2^{j/2}\psi\big(2^j x - k\big)$, a rescaled and shifted mother wavelet.

This is a doubly indexed set of functions so, when $m(\cdot)$ is expanded in this basis, we write

$$m(x) = \alpha \cdot \phi(x) + \sum_{j=1}^{\infty}\sum_{k=1}^{2^j-1}\beta_{j,k}\cdot\psi_{j,k}(x),$$

where $\alpha = \langle m, \phi\rangle$ and $\beta_{j,k} = \langle m, \psi_{j,k}\rangle$. The Haar basis is an example of (first generation) **wavelet basis**.

Best approximation in Hilbert space

Let $\{\phi_j(x)\}_j$ be an orthonormal set of functions. The sum of the first J terms of the Fourier expansion of $m(\cdot)$, is the closest - measuring distance with the $\mathtt{L}_2$ norm - that one can get to $m(\cdot)$ whilst remaining in the (linear) space spanned by $\{\phi_1,\ldots,\phi_J\}$.

To see this, let $\{c_1, \dots, c_J\}$ be generic coefficients, and consider the square of the error-distance[4]

$$\begin{aligned}\Delta_J \overset{\texttt{def}}{=} \left\| m - \sum_{j=1}^J c_j\phi_j \right\|^2 &= \left\langle m - \sum_{j=1}^J c_j\phi_j, m - \sum_{j=1}^J c_j\phi_j \right\rangle = \|m\|^2 - 2\sum_j c_j \underbrace{\langle m,\phi_j\rangle}_{=\beta_j \ (\text{GFC})} + \sum_j c_j^2 \underbrace{\langle \phi_j,\phi_j\rangle}_{=1} \pm \sum_j \beta_j^2 \\ &= \|m\|^2 + \left[\sum_j c_j^2 - 2\sum_j c_j\cdot\beta_j + \sum_j \beta_j^2\right] - \sum_j\beta_j^2 = \|m\|^2 + \sum_j (c_j-\beta_j)^2 - \sum_j \beta_j^2.\end{aligned}$$

We seek to minimize Δ_J by a suitable choice of the coefficients c_j's, and clearly the best we can do is to set $c_j = \beta_j$ to make $\sum_j (c_j - \beta_j)^2 = 0$. Thus the Fourier coefficients are *the* optimal choice for c_j.

Remarks:

1. This "best approximation" result allows us to give an alternative definition of a *complete orthonormal set* and to obtain the formula $\beta_j = \langle \phi_j, m\rangle$ for the expansion coefficients without having to assume that we can integrate the infinite $\sum c_j\phi_j$ series term-by-term. Recall that a set of points $\mathcal{S}$ is a **dense subset** of a space $\mathcal{X}$ if any given point $x \in \mathcal{X}$ is the limit of a sequence of points in $\mathcal{S}$, i.e. there are elements of $\mathcal{S}$ lying arbitrarily close to x. For example, the set of rational numbers $\mathbb{Q}$ is a dense subset of $\mathbb{R}$.

   Using this language, we say that a set of orthonormal functions $\{\phi_j\}_j$ is **complete** if the set of all **finite** linear combinations of the basis elements ϕ_j is a **dense subset** of the entire Hilbert space. This assumption guarantees that, by taking J sufficiently large, our best approximation will be arbitrarily close - in $\mathtt{L}_2$ sense - to the target function $m(\cdot)$.

   Since the best approximation containing all the $\phi_j(\cdot)$ up to $\phi_J(\cdot)$ is the $J^{\mathtt{th}}$ partial sum of the (generalized) Fourier series, this shows that the Fourier series actually converges to $m(\cdot)$.

   We have therefore proved that, if we are given $\{\phi_j\}_j$ complete orthonormal set of functions on $[a, b]$, then any function $m(\cdot)$ having $\|m\|$ finite can be expanded as a convergent Fourier series

   $$m(x) = \sum_{j=1}^{\infty} \beta_j \cdot \phi_j(x), \quad \text{where} \quad \beta_j = \langle m, \phi_j\rangle = \int_a^b m(x)\phi_j(x)\,\mathrm{d}x,$$

   and the convergence is guaranteed only in the $\mathtt{L}_2$ sense, that is

   $$\lim_{J\to\infty} \Delta_J = \lim_{J\to\infty} \left\| m - \sum_{j=1}^J \beta_j \cdot \phi_j \right\| = 0.$$

2. We showed that $\Delta_J = \|m\|^2 - \sum_{j=1}^J \beta_j^2$, hence $\mathtt{L}_2$-convergence is equivalent to the statement that

   $$\lim_{J\to\infty} \Delta_J \to 0 \quad \Leftrightarrow \quad \|m\|^2 = \sum_{j=1}^{\infty} \beta_j^2. \tag{8}$$

   This last result is called **Parseval's Theorem**.

3. Suppose we have some non-orthogonal collection of functions $\{g_j\}_{j=1}^J$, and we have found the best approximation w.r.t. this system, $\sum_{j=1}^J \alpha_j g_j(x)$ say, to $m(\cdot)$. Now suppose also we are given a new $g_{J+1}(\cdot)$ element to add to our collection. We may then seek an improved $(J+1)$-term approximation $\sum_{j=1}^{J+1} \alpha'_j g_j(x)$ by including this new function, but finding this better fit will generally involve tweaking all the α_j's already computed, not just trying different values for α_{J+1}.

   One of the great advantage of approximating by orthogonal functions is that, given another member of an orthonormal family, we can improve the precision of the fit by adjusting only the coefficient of the new term. We do not have to perturb the previously obtained coefficients.

4. On the other hand, from a practical point of view, the use of orthonormal system is somewhat sub-optimal: injecting the right (small) amount of *redundancy* in the system, allows for better robustness of our representation w.r.t. the noise typical of many statistical applications (see Elad, 2010).

5. Let's denote by $m_J(x) = \sum_{j=1}^J \beta_j\, \phi_j(x)$ the best J-term approximation to $m(\cdot)$. By an extension of the Closest Point Theorem (see Theorem 1), we can geometrically characterize $m_J(x)$ as the *projection* of $m(\cdot)$ onto the *span* of $\{\phi_1, \dots, \phi_J\}$. We call $m_J(x)$ the J-term **linear approximation** of $m(\cdot)$. Now let $\mathcal{A}_J$ denote all functions of the form

[4] We consider the case of a real and not complex basis in order to avoid the annoyance of the complex conjugation.

$\sum_{j=1}^{\infty} c_j \phi_j(x)$ such that at most J of the c_j's are nonzero - note that $\mathcal{A}_J$ is not a linear space: if $g, h \in \mathcal{A}_J$ it does not follow that $g + h$ is in $\mathcal{A}_J$. The best approximation to a function $m(\cdot)$ in $\mathcal{A}_J$ is then

$$m^{\star}_J(x) = \sum_{j \in \Lambda_J} \beta_j \, \phi_j(x),$$

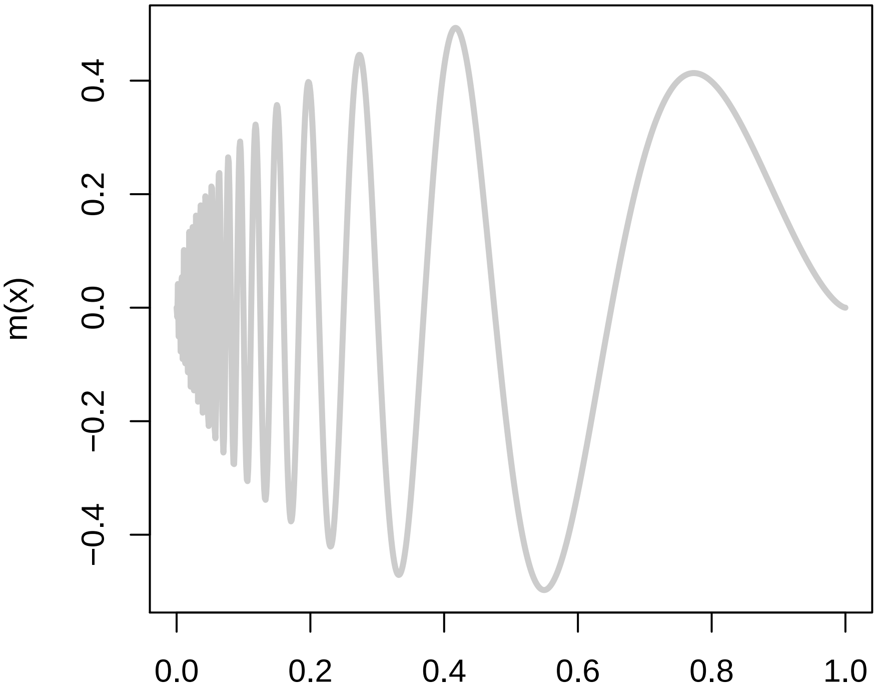
where Λ_J are the J indices corresponding to the J largest $|\beta_j|$'s. We call $m^{\star}_J(\cdot)$ the J–term **nonlinear approximation** of $m(\cdot)$ – a.k.a. greedy approximation.

Example 8 (Numerical example: Doppler function)**.** In this example we will use the cosine basis to build a *linear* approximation to the so called *Doppler function* scaled in $[0, 1]$[a]:

$$m(x) = \sqrt{x(1-x)} \sin\left(\frac{2.1\pi}{x + 0.05}\right), \quad x \in [0, 1].$$

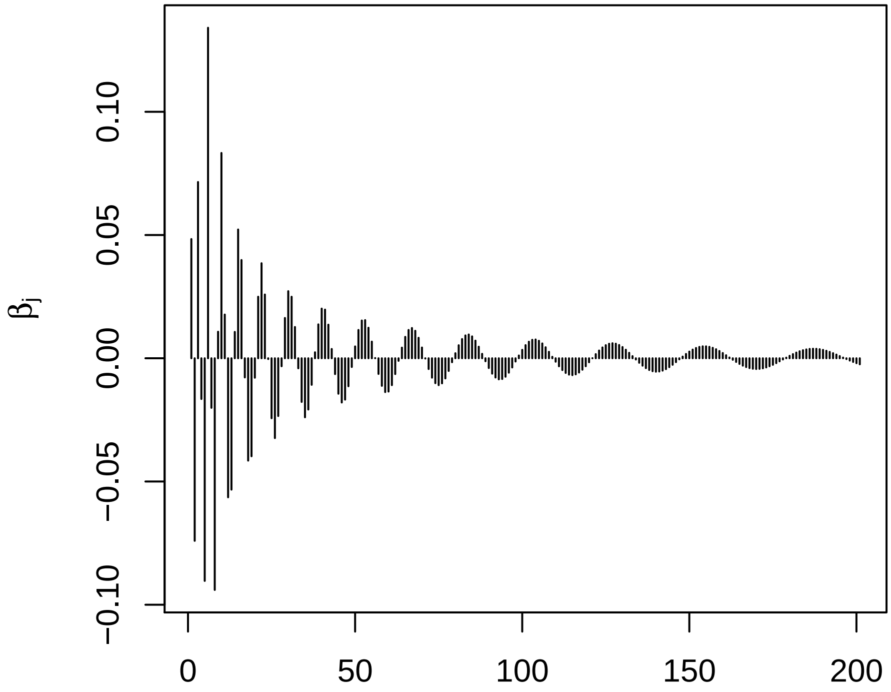
[a] Of course, if you are not dealing with a function defined over $[0, 1]$, it is sufficient to rescale its argument.

```r
# Doppler function scaled in [0,1]
doppler.fun <-  function(x) sqrt(x*(1 - x))*sin( (2.1*pi)/(x + 0.05) )
curve(doppler.fun(x), from = 0, to = 1, main = "", xlab = "", ylab = "m(x)",
        n = 1001, col = gray(.8), lwd = 3)
```



Let's now numerically evaluate the Fourier coefficients of the Doppler under our cosine-basis:

```r
j.max   <- 200
f.coeff <- rep(NA, j.max+1)
for (idx in 0:j.max){
  foo = tryCatch(
    integrate(function(x, j) doppler.fun(x) * cos.basis(x,j), lower = 0, upper = 1, j = idx)$value,
    error = function(e) NA
  )
  f.coeff[idx + 1] = foo
}
# Visualize the Fourier coefficients
plot(f.coeff, type = "h", ylab = expression(beta[j]), main = "", xlab = "")
```



<u>Remarks</u>:

1. Before moving on, let's stop for a second to ponder the plot above. What are we looking at? Oscillating Fourier coefficients β_j (w.r.t. the cosine basis), that taper off as j increases. This behavior is in essence the fingerprint of the Doppler function **smoothness** as captured by the cosine basis. More in general, the (speed of) decay of the Fourier coefficients in a particular basis, can be related to the smoothness of the function we are studying (again, as seen through *that* basis!).

   To see why, let's start from a very familiar notion of smoothness, and assume that the function $m(\cdot)$ has bounded derivatives up to order k. Thus we may expect that $\|m^{(k)}\|^2 < \infty$ where $m^{(k)}(\cdot)$ denotes the $k^{\texttt{th}}$ derivative of $m(\cdot)$. Now, considering once again the cosine-basis, let $m(x) = \sum_j \beta_j \phi_j(x)$, then it can be shown that

   $$\|m^{(k)}\|^2 = \int_0^1 |m^{(k)}(x)|^2 \, \mathrm{d}x = 2 \sum_{j=1}^{\infty} \beta_j^2 \cdot (\pi\, j)^{2k}.$$

   The only way that $\sum_{j=1}^{\infty} \beta_j^2 \cdot (\pi\, j)^{2k}$ can be finite is if the β_j's get small (at a suitable speed) when j gets large. To summarize

   If a function is smooth, then its Fourier coefficients β_j will be small for j large.

2. If we think again to this particular example involving the cosine-basis, it is definitely obvious what we've seen: in this case the basis elements $\{\phi_j\}_j$ are naturally ordered from lower to higher frequencies. Hence, by the very nature of this basis, ϕ_j's at low j capture the overall trend of $m(\cdot)$, its low resolution behavior, whereas ϕ_j's at very high j, will try to extract the fine details of $m(\cdot)$, those little tiny features "visible" only at very high resolution.

   Now, as we said, the Fourier coefficients quantify the similarity between the complicated function $m(\cdot)$ we are expanding, and each basis element, but the function $m(\cdot)$ is smooth! So, intuitively, beyond a given (possibly large) resolution level J, all we add to the representation should be irrelevant bits with associated necessarily small Fourier coefficients.

3. All this has also a clear statistical impact: even from a finite sample of size n, there's hope to decently estimate a (smooth enough) function. Roughly speaking, with n data-points the best we can hope for it to estimate n Fourier coefficients. If we focus on linear approximation in a cosine basis, this means that we are introducing a systematic bias in our estimates due the fact that we are throwing away (i.e. setting to zero) the β_j with $j > n$. But, if the function is smooth and n large enough, we also know that what we are ignoring should be small or, being in a statistical context, under the noise level (i.e. *statistically undetectable*) anyway.

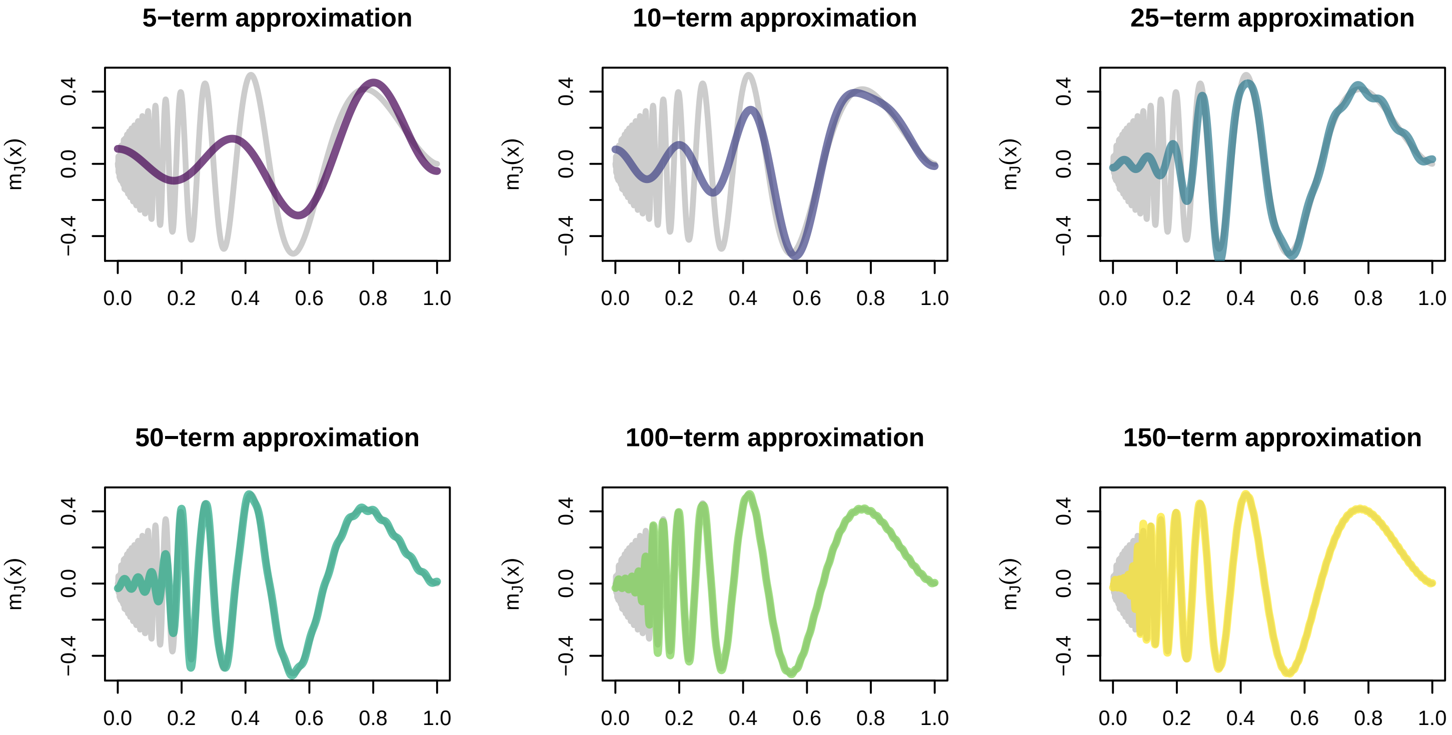
   It is exactly after this qualitative consideration that the importance of picking the best basis for the chosen notion of smoothness kicks in. Very briefly: for a given smoothness class of function, the speed at which the β_j's decay to zero varies depending on the ONB we choose. Of course we want to pick the ONB with associated the fastest rate, the one that squeezes the $\mathtt{L}_2$ energy of our target functions (i.e. their squared norm) in just a few coefficients (remember Parseval!) so that the *truncation bias* will be small and controllable even at modest sample sizes.

All this said, let's complete our work by looking at different linear approximations for the Doppler:

```
# Time to rebuild/approximate our Doppler with an n-term (linear) approximation.
# Let's make a function for this purpose...
proj.cos <- function(x, f.coeff, j.max = 10){
  out = rep(0, length(x))
  for(idx in 0:j.max){
    if ( !is.na(f.coeff[idx + 1]) ) out = out + f.coeff[idx + 1] * cos.basis(x, j = idx)
  }
  return(out)
}

# Visualize some n-terms approximations
j.seq = c(5, 10, 25, 50, 100, 150)
mycol = viridis(length(j.seq), alpha = .7)

par(mfrow = c(2,3))
for (idx in 1:length(j.seq)){
  # Original function
  curve(doppler.fun(x), from = 0, to = 1,
        main = paste(j.seq[idx], "-term approximation", sep = ""),
        xlab = "", ylab = expression(m[J](x)),
        n = 1001, col = gray(.8), lwd = 3)
  # Add approximation
  curve(proj.cos(x, f.coeff = f.coeff, j.seq[idx]),
        n = 1001, col = mycol[idx], lwd = 4,
        add = TRUE)
}
```



Remarks:

1. As expected, with only a few terms we start recovering the slowly varying parts of the function plus its overall trend, and then we incrementally add up details around the origin.
2. Notice that the Doppler is a typical example of function with inhomogeneous smoothness level along its domain: very wiggly around the origin, calm and "relaxed" as we move away from 0. For this very reason it was among the benchmark functions that in mid nineties David Donoho and coauthors used to show the superiority of wavelet bases in nonparametric function estimation (being totally "delocalized" in space, cosine basis are really sub-optimal here!).

3. The basis-expansion machinery used above can be readily extended to higher dimensions. So, suppose that $m(x_1, x_2)$ is a function of two variables, and assume that $0 \leqslant (x_1, x_2) \leqslant 1$. If $\{\phi_1(\cdot), \phi_1(\cdot), \dots\}$ is an orthonormal basis for $\mathtt{L}_2([0,1])$, then the functions

$$\big\{\phi_{j_1,j_2}(x_1,x_2) = \phi_{j_1}(x_1)\,\phi_{j_2}(x_2)\ :\ j_1, j_2 = 0, 1, \dots\big\},$$

form an orthonormal basis for $\mathtt{L}_2([0,1]\times[0,1])$, called the *tensor product basis*. The basis can be extended to d dimensions in the obvious way.

Suppose that $\phi_0(x) = 1$, then any function $m \in \mathtt{L}_2([0,1]\times[0,1])$ can be expanded in the tensor basis as

$$\begin{aligned} m(x_1,x_2) &= \sum_{j_1=0}^{\infty}\sum_{j_2=0}^{\infty}\beta_{j_1,j_2}\phi_{j_1,j_2}(x_1,x_2) = \sum_{j_1=0}^{\infty}\sum_{j_2=0}^{\infty}\beta_{j_1,j_2}\phi_{j_1}(x_1)\,\phi_{j_2}(x_2) = \\ &= \beta_{0,0} + \sum_{j=1}^{\infty}\beta_{j,0}\phi_j(x_1) + \sum_{j=1}^{\infty}\beta_{0,j}\phi_j(x_2) + \sum_{j_1=1}^{\infty}\sum_{j_2=1}^{\infty}\beta_{j_1,j_2}\phi_{j_1}(x_1)\,\phi_{j_2}(x_2), \end{aligned}$$

where β_{j_1,j_2} denotes the Fourier coefficient given by

$$\beta_{j_1,j_2} = \big\langle m(x_1,x_2), \phi_{j_1,j_2}(x_1,x_2)\big\rangle = \int_0^1\int_0^1 m(x_1,x_2)\,\phi_{j_1,j_2}(x_1,x_2)\,\mathrm{d}x_1\mathrm{d}x_2.$$

This expansion has an ANOVA-like structure consisting of a mean, main effects, and interactions. Statistically thinking, this structure *may* suggest a way to get better estimators by enforcing higher smoothness levels on higher–order interactions.

As an example, we can build a tensor basis $\{\phi_{j_1,j_2}(x_1,x_2)\}_{j_1,j_2}$ from the usual cosine basis

$$\phi_0(x) = 1, \quad \phi_j(x) = \sqrt{2}\cos(j\,\pi\,x), \quad j \geqslant 1,$$

and, consequently, jot down the linear (J_1, J_2)–approximation for a generic function $m(x_1,x_2)$:

$$m_{J_1,J_2}(x_1,x_2) = \sum_{j_1=0}^{J_1}\sum_{j_2=0}^{J_2}\beta_{j_1,j_2}\phi_{j_1,j_2}(x_1,x_2).$$