

Inner products, operators

### Recap

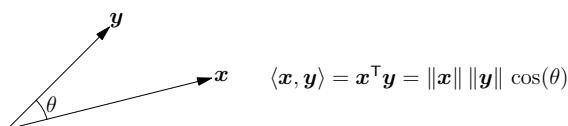
- We have looked at vector space (closed sets where we can add elements and multiply them by a scalar)
- Recall that vector spaces don't just apply to normal vectors ( $\mathbb{R}^n$ ), but to matrices, functions, sequences, random variables, . . .
- Proper distances or metrics,  $d(\mathbf{x}, \mathbf{y})$ , allow us to construct ideas about geometry of the vector space
- Norms,  $\|\mathbf{x}\|$ , that allow us to reason about the size of vector
- Norm induce a distance,  $d(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|$

### Axioms of Inner Products

- An inner product satisfies
  1.  $\langle \mathbf{x}, \mathbf{x} \rangle \geq 0$  for all  $\mathbf{x} \in \mathcal{V}$
  2.  $\langle \mathbf{x}, \mathbf{x} \rangle = 0$  if and only if  $\mathbf{x} = \mathbf{0}$
  3.  $\langle \alpha \mathbf{x}, \mathbf{y} \rangle = \alpha \langle \mathbf{x}, \mathbf{y} \rangle$
  4.  $\langle \mathbf{x}, \mathbf{y} + \mathbf{z} \rangle = \langle \mathbf{x}, \mathbf{y} \rangle + \langle \mathbf{x}, \mathbf{z} \rangle$
  5.  $\langle \mathbf{x}, \mathbf{y} \rangle = \langle \mathbf{y}, \mathbf{x} \rangle$
- We can show that  $\|\mathbf{x}\| = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}$  satisfies the axioms of a norm, so that an inner-product space is a normed space
- The norm associated with the inner-product for vectors in  $\mathbb{R}^n$  (i.e.  $\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^\top \mathbf{y}$ ) is the Euclidean norm  $\|\mathbf{x}\| = \sqrt{\mathbf{x}^\top \mathbf{x}}$

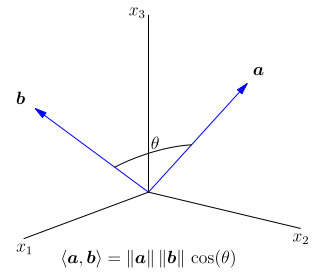
### Angles Between Vectors

- A natural interpretation of the inner product is in providing a measure of the angle between vectors



- Vectors are orthogonal if  $\langle \mathbf{x}, \mathbf{y} \rangle = 0$
- We can extend this idea to functions
 
$$\langle f(x), g(x) \rangle = \int_{x \in \mathcal{I}} f(x)g(x)dx = \|f(x)\| \|g(x)\| \cos(\theta)$$
- Note that  $\sin(x)$  and  $\cos(x)$  are orthogonal in the interval  $[0, 2\pi]$

1. Inner Products
2. Operators



### Inner Products

- We will often consider objects with an *inner product*
- For vectors in  $\mathbb{R}^n$

$$\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^\top \mathbf{y} = \sum_{i=1}^n x_i y_i$$

- For functions

$$\langle f, g \rangle = \int_{x \in \mathcal{I}} f(x)g(x)dx$$

- For  $m \times n$  matrices

$$\langle \mathbf{A}, \mathbf{B} \rangle = \text{Tr} \mathbf{A}^\top \mathbf{B} = \sum_{i=1}^m \sum_{j=1}^n A_{ij} B_{ij}$$

### Cauchy-Schwarz Inequality

- One of the most important results of inner-product spaces, known as the **Cauchy-Schwarz inequality** is that

$$\langle \mathbf{x}, \mathbf{y} \rangle^2 \leq \langle \mathbf{x}, \mathbf{x} \rangle \langle \mathbf{y}, \mathbf{y} \rangle = \|\mathbf{x}\|^2 \|\mathbf{y}\|^2$$

- Or

$$|\langle \mathbf{x}, \mathbf{y} \rangle| \leq \|\mathbf{x}\| \|\mathbf{y}\|$$

- This is a very general result so for example

$$\left| \int f(x)g(x)dx \right| \leq \sqrt{\left( \int f^2(x)dx \right) \left( \int g^2(x)dx \right)}$$

### Basis Functions

- Any set of vectors  $\{\mathbf{b}_i | i = 1, \dots\}$  that span the space can be used as a basis or coordinate system

- The simplest and most useful case is when the vectors are orthogonal and normalised (i.e.  $\|\mathbf{b}_i\| = 1$ )

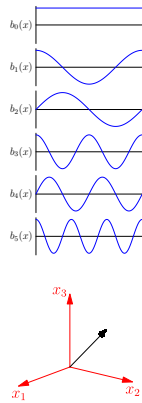
- In  $\mathbb{R}^3$  we could use  $\mathbf{b}_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$ ,  $\mathbf{b}_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$ ,  $\mathbf{b}_3 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$

- This is not unique as we can rotate our basis vectors

- For an orthogonal basis we can write any vector as  $\hat{\mathbf{x}} = \begin{pmatrix} \mathbf{x}^\top \mathbf{b}_1 \\ \mathbf{x}^\top \mathbf{b}_2 \\ \mathbf{x}^\top \mathbf{b}_3 \end{pmatrix}$

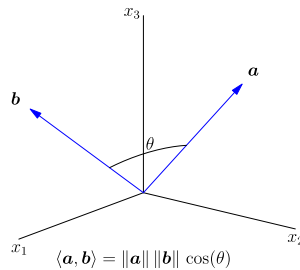
## Orthogonal Functions

- For functions we can use any ortho-normal set of functions as a basis
- The most familiar are the Fourier functions  $\sin(n\theta)$  and  $\cos(n\theta)$
- Any function in  $C(0, 2\pi)$  can be represented by a point  $f = \begin{pmatrix} \langle f(x), b_0(x) \rangle \\ \langle f(x), b_1(x) \rangle \\ \vdots \end{pmatrix}$
- There might be an infinite number of components
- This is analogous to points in  $\mathbb{R}^n$  (for large  $n$ )



## Outline

- Inner Products
- Operators



## Linear Operators

- Operators are in general very complicated, but a particular nice set of operators are linear operators
- $\mathcal{T}$  is a linear operator if
  - $\mathcal{T}[ax] = a\mathcal{T}[x]$
  - $\mathcal{T}[x + y] = \mathcal{T}[x] + \mathcal{T}[y]$
- For normal vectors ( $x \in \mathbb{R}^n$ ) the most general linear operation is

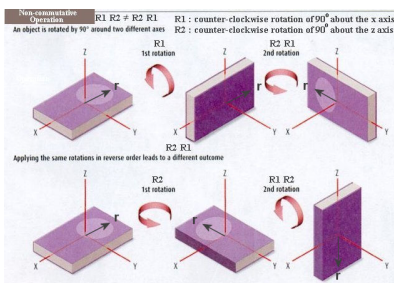
$$\mathcal{T}[x] = Mx$$

where  $M$  is a matrix

## Non-commutativity

- In general  $AB \neq BA$

$$\begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} \quad \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$$



## Algebraic Structure

- We have gone to these lengths as we want to show that many properties of vectors are shared by other objects (matrices, functions, etc.)
- The notions of distance (geometry), norms (size of vectors) and inner products (angles between vectors) provides a very rich set of concepts
- Vectors form the backbone of objects we will use repeated in machine learning
- The next piece of the jigsaw is to understand how we can transform these objects

## Operators

- In machine learning we are interested in transforming our input vectors into some output predictions
- To accomplish this we will apply some mapping or operators on the vector  $\mathcal{T} : \mathcal{V} \rightarrow \mathcal{V}'$
- This says that  $\mathcal{T}$  maps some object  $x \in \mathcal{V}$  to an object  $y = \mathcal{T}[x]$  in a new vector space  $\mathcal{V}'$
- This new vector space may or may not be the same as the original vector space
- Our objects may be any object in a vector space such as a function

## Matrix multiplication

- For an  $\ell \times m$  matrix  $A$  and an  $m \times n$  matrix  $B$  we can compute the  $(\ell \times n)$  product,  $C = AB$ , such that

$$C_{ij} = \sum_{k=1}^m A_{ik} B_{kj} \quad \left( \begin{pmatrix} \equiv \end{pmatrix} \right) \left( \begin{pmatrix} \equiv \end{pmatrix} \right) = \left( \begin{pmatrix} \equiv \end{pmatrix} \right)$$

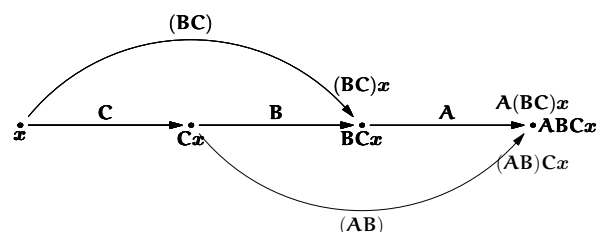
- Treating the vector  $x$  as a  $n \times 1$  matrix then

$$y = Ax \Rightarrow y_i = \sum_j M_{ij} x_j \quad \left( \begin{pmatrix} \equiv \end{pmatrix} \right) \left( \begin{pmatrix} \equiv \end{pmatrix} \right) = \left( \begin{pmatrix} \equiv \end{pmatrix} \right)$$

- Using the same matrix notation we can define the inner product as

$$\langle x, y \rangle = x^T y = \sum_{i=1}^n x_i y_i \quad \left( \begin{pmatrix} \equiv \end{pmatrix} \right) \left( \begin{pmatrix} \equiv \end{pmatrix} \right) = \left( \begin{pmatrix} \equiv \end{pmatrix} \right)$$

## Associativity of Mappings



- For all  $x$  we have  $A(BC)x = (AB)Cx$
- This implies  $A(BC) = (AB)C$

- The equivalent of a matrix for functions (i.e. a linear operator) is known as a kernel  $K(x, y)$

$$g(x) = \mathcal{T}[f] = \int_{y \in \mathcal{I}} K(x, y) f(y) dy$$

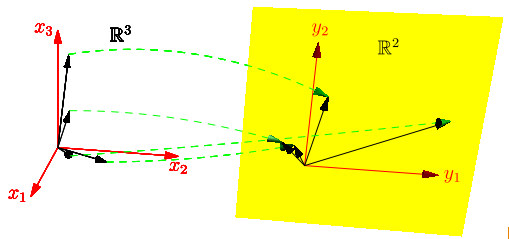
- Our domain does not need to be one dimensional, e.g.

$$g(x) = \mathcal{T}[f] = \int_{y \in \mathcal{I}} K(x, y) f(y) dy$$

- We shall soon see examples of high-dimensional kernels

## General Linear Mappings

- In general a linear operator will map vectors between different vector spaces
- E.g.  $\mathbb{R}^3 \rightarrow \mathbb{R}^2$



## Summary

- We haven't covered much machine learning as such—sorry
- But mathematics is the language of machine learning and you have to get used to it
- Mathematics is like programming, if you don't understand the syntax and you can't write it down then it's meaningless
- We've taken a high level view of inner product spaces and operator, this will pay us back later as we look at kernel methods

- Kernels are used heavily in machine learning
- In kernel methods such as SVM, SVR, Kernel-PCA
- They are also used in Gaussian Processes
- In all these cases we consider symmetric, positive semi-definite kernels
- Sometimes they can be interpreted as covariance between random functions

$$K(x, y) = \mathbb{E}_{f \sim \mathcal{P}}[(f(x) - \mu(x))(f(y) - \mu(y))]$$

## Square Matrices

- We will spend a lot of time on operators that map from a vector space onto itself  $\mathcal{T} : \mathcal{V} \rightarrow \mathcal{V}$
- For vectors in  $\mathbb{R}^n$  such linear operators are represented by square matrices
- When there is a one-to-one mapping then we have a unique inverse
- We will study such mappings in detail in the next lecture