Lecture 3 Matrices

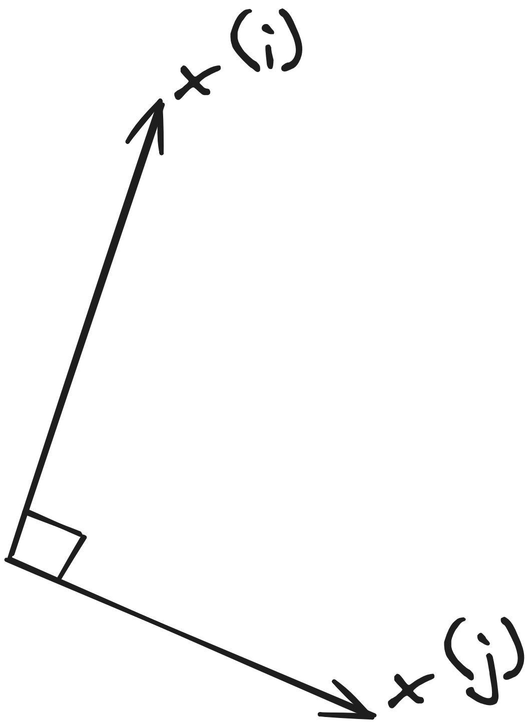
Note by Samion Suwito on 1/28/25

Vectors

Recall their interpretations as data points, directions and functions

Orthogonality

Vectors $x^{(1)}, \dots x^{(m)}$ are orthogonal if $\langle x^{(i)}, x^{(j)} \rangle = 0 \; \forall i \neq j$



Claim: Non-zero orthogonal vectors are LI.
Suppose

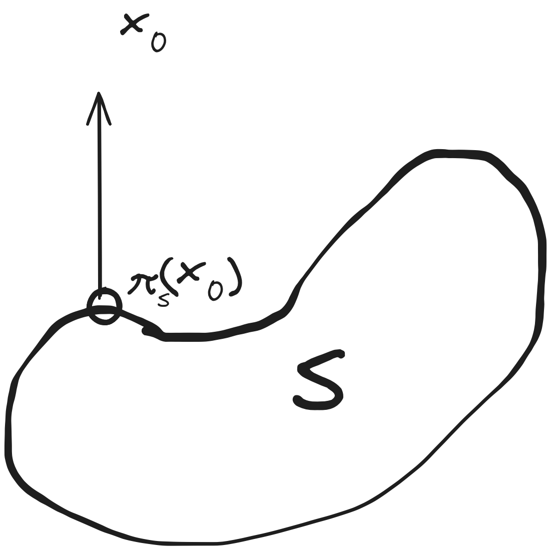
$$x^{(1)} = \sum_{j=2}^{m} \alpha_j x^{(j)} \implies$$

$$\langle x^{(i)}, x^{(j)} \rangle = \sum_{j=1}^{m} \alpha_j \langle x^{(j)}, x^{(i)} \rangle = 0$$

Projections

Let $\mathcal{X}$ be an IPS, $\mathcal{S} \subset \mathcal{X}$ if $x_0 \in \mathcal{X}$, define projection of x_0 onto $\mathcal{S}$

$$\pi_{\mathcal{S}}(x_0) = \arg\min_{s \in \mathcal{S}} \|x_0 - s\|$$



Hilbert Projection Theorem:
If $\mathcal{S}$ is a subspace, then $\pi_s(x_0)$ exists, is unique, and is uniquely characterised by "orthogonality principle":

$$\langle x_0 - \pi_{\mathcal{S}}(x_0), s\rangle = 0 \;\; \forall s \in \mathcal{S}$$

Example:
Let $x^{(1)}, \ldots, x^{(m)}$ be an orthonormal basis for $\mathcal{S}$. Claim that

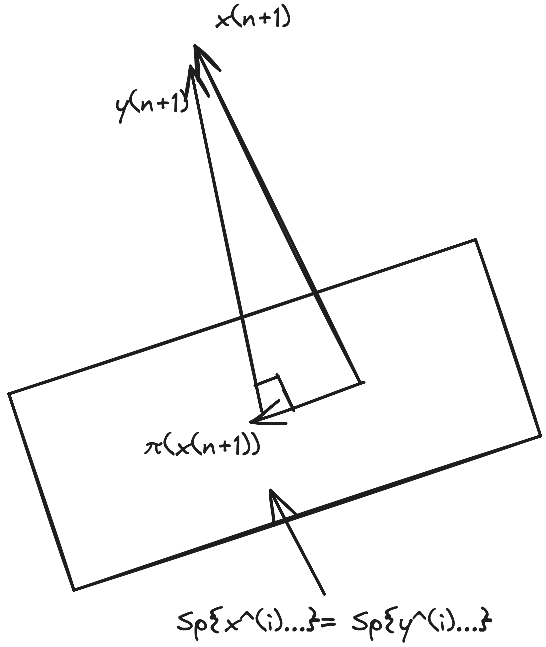
$$\pi_s(x) = \sum \langle x^{(i)}, x\rangle x^{(i)}$$

Check:
$\langle x - \Sigma\langle x^{(i)}, x\rangle x^{(i)}, s\rangle$ is equal to
$\langle x, s\rangle - \Sigma\langle x^{(i)}, x\rangle\langle x^{(i)}, s\rangle$, $s \in \mathcal{S}$, $s = \Sigma\alpha_i x^{(i)}$for some α_i

$$\sum \alpha_i \left\langle x, x^{(i)}\right\rangle - \sum \alpha_i \langle x^{(i)}, x\rangle = 0$$

Example 2:
If $\mathcal{S}^{\perp} = \{x \in \mathcal{X} : \langle x, s\rangle = 0 \;\; \forall s \in \mathcal{S}\}$ then $\mathcal{X} = \mathcal{S} \oplus \mathcal{S}^{\perp}$
Example: **Gram Schmidt**
Given collection $x^{(i)}, \ldots, x^{(k)}$, find orthogonal $y^{(1)}, \ldots, y^{k}$ such that $\text{Span}(x \ldots) = \text{Span}(y \ldots)$.

Start with $y^{(1)} = x^{(1)}$
$y^{(n+1)} = x^{(n+1)} - \pi_{\text{Span}(x^{(1)},...,x^{(n)})}(x^{(n+1)})$
Take component of y that is orthogonal to $\mathcal{S}$ essentially
$= x^{(n+1)} - \pi_{\text{Sp}(y^{(1)},...,y^{(n)})}(x^{(n+1)})$



Projection and orthogonality is different side of the same coin
Affine Space is a translated vector space

Gradients

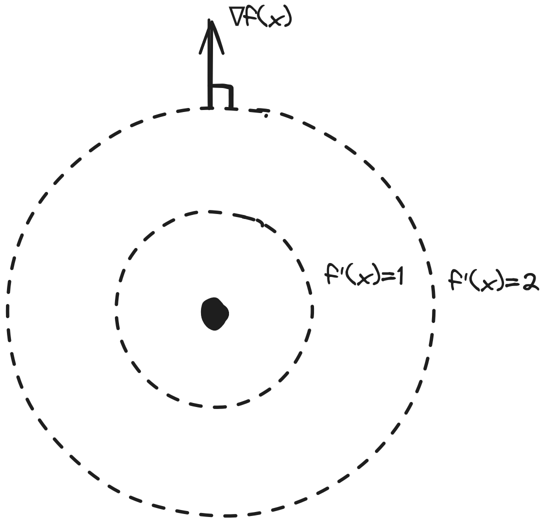
final example of vectors as directions interpretations is gradients

Given differentiable $f : \mathbb{R}^n \to \mathbb{R}$

$$\nabla f(x) = \begin{pmatrix} \frac{\partial}{\partial x_1} f(x) \\ \vdots \\ \frac{\partial}{\partial x_n} f(x) \end{pmatrix}$$

Key properties of gradients

1. $\nabla f(x)$ points in direction of steepest ascent (therefore negative ∇ points in direction of steepest descent; gradient descent)
2. $\nabla f(x)$ is perpendicular(not orthogonal, only in Euclidean sense (90º)) to the level set of f containing $f(x)$



Matrices

A matrix is representation of a linear map between two spaces
A linear map $A : \mathbb{R}^n \to \mathbb{R}^m$ can always be represented as a matrix. Indeed, if $(e_i)_{i=1}^n$ is a natural basis for $\mathbb{R}^n$

$$A(x) = A\left(\sum x_i e_i\right) = \sum x_i A(e_i)$$

where $A(e_i) \in \mathbb{R}^m =$

$$\begin{pmatrix} a_{1i} \\ \vdots \\ a_{mi} \end{pmatrix}$$

Therefore A can be represented as

$$\begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m_1} & \cdots & a_{mn} \end{pmatrix}$$

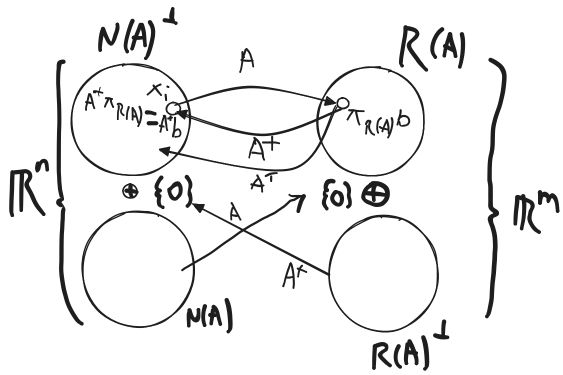
If $A : \mathbb{R}^n \to \mathbb{R}^m, B : \mathbb{R}^m \to \mathbb{R}^k$ then composition BA is a linear map from $\mathbb{R}^n \to \mathbb{R}$ with matrix representation:

$$BA_{ij} = \sum_{l=1}^{m} B_{il} A_{lj} = \text{matrix multiplication}$$

$BA(x) = B(A(x))$

$BA \in \mathbb{R}^{k \times m}$
A_{lj} is inner product between ith row of B and jth col of A.

Associated to every matrix $A \in \mathbb{R}^{m \times n}$ are two important subspaces:
Range of A = $R(A) := \{Ax : x \in \mathbb{R}^n\} \subset \mathbb{R}^m$ it's a subspace by the linearity of A
Nullspace of A = $\mathcal{N}(A) := \{x \in \mathbb{R}^n : Ax = 0\} \subset \mathbb{R}^n$
$\mathrm{rank}(A) := \dim(R(A))$



Important idea:
$R(A) = N(A^\top)^\perp$

This decomposition can be useful in simplifying/transforming problems
Example: Many problems in Learning can be represented by:

$$\min_{w \in \mathbb{R}^n} \mathcal{L}(Aw) \quad A \in \mathbb{R}^{m \times n}$$

$\mathcal{L}$ is a generic loss function
$w = w_0 + w_1$
Where $w_0 \in \mathcal{N}(A)^\perp$ and $w_1 \in \mathcal{N}(A)$.
$\mathcal{L}(A(w_0 + w_1) = \mathcal{L}(Aw_0) = \mathcal{L}\left(\mathcal{A}\mathcal{A}^\top v\right)$
Which therefore turns the min

$$\min_{v \in \mathbb{R}^m} \mathcal{L}\left(AA^\top v\right)$$

Turning the problem from $\mathbb{R}^n$ to $\mathbb{R}^m$ which can be useful if $n \gg m$

Consider a "regularised" regression problem.

$$\min_{w \in \mathbb{R}^n} \mathcal{L}(Aw) + \lambda \|w\|_2$$

Write $w = w_0 + w_1$ where $w_0 \in \mathcal{N}(A)^\perp$ and $w_1 \in \mathcal{N}(A)$

$$\min_{w \in \mathbb{R}^n} \mathcal{L}(Aw_0) + \lambda \|w_0 + w_1\|_2$$

Let $\|w_0 + w_1\|^2 = \|w_0\|^2 + \|w_1\|^2 \geq \|w_0^2\|$
Pythagorean theorem since
Therefore min statement

$$\geq \min_{w_0 \in \mathcal{N}(A)^\perp} \mathcal{L}(Aw_0) + \lambda \|w_0\|$$

actually equality by restricting to $w \in \mathcal{N}(A)^\perp$ making $w_1 = 0$. Also the dimension of that is $\operatorname{rank}(A)$ because of FTLA

Matrix Inverse

For $A \in \mathbb{R}^{n \times n}$, we can say A is invertible if it its 1-1 and onto, $Ax \neq Ay \forall x \neq y \in \mathbb{R}^n \iff Ax \neq 0 \forall x \neq 0$
In this case the inverse transformation is denoted by A^{-1}
A is invertible $\iff \mathcal{N}(A) = \{0\} \iff \operatorname{rank}(A) = n$

Weaker definition of **pseudo-inverse**
for general matrices $A \in \mathbb{R}^{m \times n}$
A^{Pi} should satisfy $AA^{Pi}A = A$
$A^{Pi} \in \mathbb{R}^{n \times m}$

Common terminology:
Square matrix: $m = n$
Symmetric matrix: $A = A^\intercal$
Orthogonal matrix: Where $A^{-1} = A^\intercal$
Rank-one (dyad) matrix: $A = uV^\intercal \; u \in \mathbb{R}^m, v \in \mathbb{R}^n$