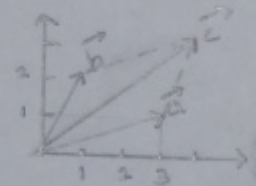
Recitation 1: Background.

I. Linear Algebra.

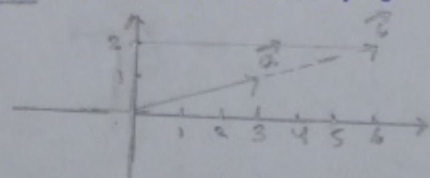
1. Vectors. consider the two vectors $\vec{a} = [3, 1]$ & $\vec{b} = [1, 2]$

- **Addition:** $\vec{c} = \vec{a} + \vec{b} = [3, 1] + [1, 2] = [3+1, 1+2] = [4, 3]$

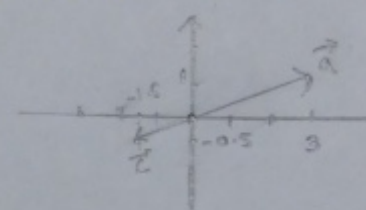


(complete parallelogram)

- **scalar multiplication:** $\vec{c} = \lambda \vec{a} = \lambda[3, 1] = [3\lambda, \lambda]$



$\lambda = 2$



$\lambda = -0.5$

- **Length of a vector:** $\|\vec{a}\| = \sqrt{\|\vec{a}\|^2}$
  $= \sqrt{3^2 + 1^2} = \sqrt{10}$.

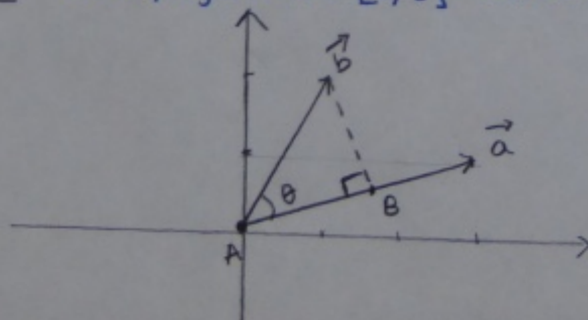
- **Dot product:** $\vec{a} = [a_1, a_2, a_3]$ $\vec{b} = [b_1, b_2, b_3]$
  $$\vec{a} \cdot \vec{b} = \langle \vec{a}, \vec{b} \rangle = \vec{a}\,\vec{b}^T = a_1 b_1 + a_2 b_2 + a_3 b_3$$
  $$= \|\vec{a}\|\,\|\vec{b}\| \cos(\theta)$$
  angle between $\vec{a}$ & $\vec{b}$.

  Ex: $\vec{a} = [3, 1]$ $\vec{b} = [1, 2]$
  $\vec{a} \cdot \vec{b} = 3 \times 1 + 1 \times 2 = 5$.

- **Angle between two vectors:** $\theta = \cos^{-1}\left[\dfrac{\vec{a} \cdot \vec{b}}{\|a\| \cdot \|b\|}\right]$

- **Orthogonal vectors:** $\theta = 90°$, $\cos\theta = 0$
  $\vec{a}$ & $\vec{b}$ are orthogonal iff $\boxed{\vec{a} \cdot \vec{b} = 0}$

- **Projection:** Let's project $\vec{b} = [1, 2]$ onto $\vec{a} = [3, 1]$.



- scalar proj $\vec{b}$ onto $\vec{a}$:
  signed = length of segment AB.
- vector projection $\vec{b}$ onto $\vec{a}$:
  = vector $\overrightarrow{AB}$

→ continued.

• Projection (cont.):

SCALAR
$$AB = \|\vec{b}\| \cdot \boxed{\cos\theta} \leftarrow \frac{\vec{a}\cdot\vec{b}}{\|a\|\cdot\|b\|}$$
$$= \frac{\vec{a}\cdot\vec{b}}{\|a\|}$$
← note that this can be < 0 if $\theta > 90°$

VECTOR
$$\vec{AB} = AB \cdot \frac{\vec{a}}{\|a\|} = \frac{\vec{a}\cdot\vec{b}}{\|a\|^2}\,\vec{a}$$
↑ vector of magnitude 1 in the direction of $\vec{a}$.

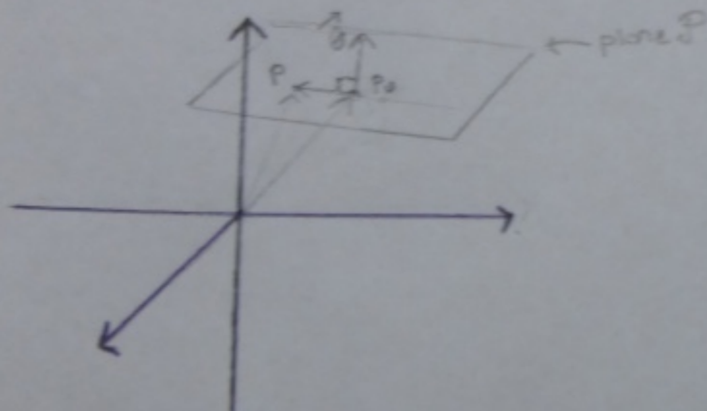
2. Planes.

A plane can be described by:

- A point P_0 belonging to the plane

AND

- A vector $\vec{\theta}$ orthogonal to the plane. (indicates the 'inclination' of the plane)



• Equation describing the plane:

Let p be a point on the plane.

$\vec{P}-\vec{P_0} \perp \vec{\theta}$: $(\vec{P}-\vec{P_0})\cdot\vec{\theta} = 0$.

Example in $\mathbb{R}^3$: $\vec{P} = [x, y, z]^T$

$\vec{P_0} = [x_0, y_0, z_0]^T$

$\vec{\theta} = [\theta_1, \theta_2, \theta_3]^T$

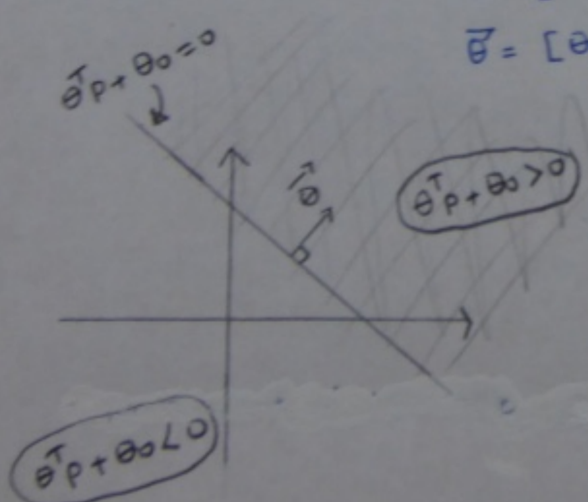
$$\begin{array}{ccc} \vec{P}-\vec{P_0} & & \vec{\theta} \end{array}$$
$$\begin{bmatrix} x - x_0 \\ y - y_0 \\ z - z_0 \end{bmatrix} \cdot \begin{bmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \end{bmatrix} = 0$$

expanding:

$$\theta_1(x - x_0) + \theta_2(y - y_0) + \theta_3(z - z_0) = 0$$

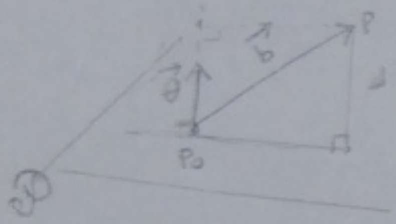
$$\underbrace{\theta_1 x + \theta_2 y + \theta_3 z}_{\vec{\theta}^T\cdot p} \underbrace{- \theta_1 x_0 - \theta_2 y_0 - \theta_3 z_0}_{-\theta^T P_0} = 0$$

↑ denote this by θ_0!



$$\Rightarrow \boxed{\text{Equation: } \theta^T \cdot P + \theta_0 = 0}$$

• Distance between a point & a plane.



• Find the signed distance from point p to plane $\mathcal{P}$ described by the normal vector $\vec{\theta}$ & point p_0.

The signed distance from p to $\mathcal{P}$ is simply the scalar projection of $\vec{b} = \vec{p} - \vec{p}_0$ onto $\vec{\theta}$.

(recall: scalar proj. of $\vec{b}$ onto $\vec{\theta}$ is >0 if $\measuredangle(\vec{b}, \vec{\theta}) < 90°$ i.e. if $\vec{b}$ & $\vec{\theta}$ point in the same direction & it is <0 if they point in opposite directions).

• signed distance $p \to \mathcal{P}$: $d = \dfrac{\vec{b} \cdot \vec{\theta}}{\|\theta\|} = \dfrac{(\vec{p} - \vec{p}_0) \cdot \vec{\theta}}{\|\theta\|} = \dfrac{\vec{p} \cdot \vec{\theta} + \vec{\theta_0}}{\|\theta\|}$

• unsigned distance $p \to \mathcal{P}$: $|d| = \dfrac{|\vec{p} \cdot \vec{\theta} + \vec{\theta_0}|}{\|\theta_0\|}$

II. Probability:

- **Discrete R.V.**
  - Numerical outcome of a random experiment.
  - "Discrete": can only take a countable number of values.
  - **pmf:** (probability mass function of X)

    $P_X(x)$ = Probability that r.v. X takes the value x.

  - Example: X = outcome of a roll of a fair six-sided die.

    $P_X(3) = 1/6$.

    $$P_X(x) = \begin{cases} 1/6 & \text{if } x = 1,2,\dots,6 \\ 0 & \text{o.w.} \end{cases}$$

  - $\sum_x P_X(x) = 1$
  - Find $P(X \geq 4)$: $P(X \geq 4) = P(X = 4 \text{ or } X = 5 \text{ or } X = 6)$

    $= P_X(4) + P_X(5) + P_X(6) = 3/6 = 1/2$.

  - Find $E[X]$: $E[X]$ = weighted average of values of X

    $= \sum_{x=1}^{6} x P_X(x) = 1 \cdot P_X(1) + \dots + 6 \cdot P_X(6) = \frac{1}{6}(1 + \dots + 6$

    $= 7/2$.

- **Joint pmf:** $P_{X,Y}(x,y) = P[X = x \text{ AND } Y = y]$.

- **Independence:** X & Y are independent if knowing the value of one of them gives no information about the value of the other.

  Formally, X & Y are independent iff $P_{X,Y}(x,y) = P_X(x) \cdot P_Y(y) \quad \forall x,y$

- **Example:** Let X be the result of the 1st die roll & Y be the result of the 2nd die roll.

  Assume that the results of the two die rolls are independent.

  $P_{X,Y}(3,6) = P_X(3) \cdot P_Y(6) = 1/6 \cdot 1/6 = 1/36$.

• Continuous random variable:

• Random variable with a continuous range of possible values.

Ex: time until next customer arrives at a store

• Formally, X is a continuous r.v. if there is a nonnegative function f_X s.t

$$P(X \in B) = \int_B f_X(x)\,dx$$

any segment of the real line

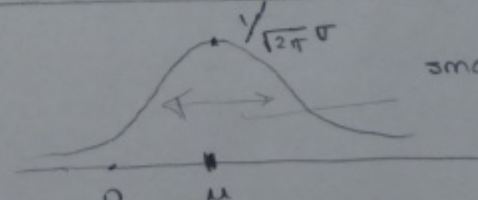
f_X is the probability density function (PDF) of X.

• PDF:
- $f_X(x) \geq 0 \quad \forall x$.
- $\int_{-\infty}^{\infty} f_X(x)\,dx = 1$
- $f_X(x)$ is NOT the probability that $X = x$. [$P(X=x)=0 \ \forall x$ for a continuous r.v.]
- $f_X(x)$ is a 'density' (roughly a probability per unit length).

• Expected value: • $E[X] = \int_{-\infty}^{\infty} x \cdot f_X(x)\,dx$.

• Independence: X & Y are independent iff $\underbrace{f_{X,Y}(x,y)}_{\text{Joint pdf of } X+Y} = f_X(x) \cdot f_Y(y) \quad \forall x, y$

Normal R.V: $X \sim N(\mu, \sigma^2)$



smaller σ $\Rightarrow$ smaller width (more concentration around the mean)

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma}}$$

$E[X] = \mu$.

$\text{var}(X) = \sigma^2$.

• At what value of x is $f_X(x)$ maximized? (call it $\hat{x}$)

- visually, from the picture of the pdf, mode = mean = $\mu = \hat{x}$
- $\hat{x} = \text{argmax}\, f_X(x)$ i.e. $f'(x)\Big|_{\hat{x}} = 0$

$$\frac{1}{\sqrt{2\pi\sigma^2}}\left[-\frac{2(\hat{x}-\mu)}{2\sigma}\right] \cdot e^{-\frac{(\hat{x}-\mu)^2}{2\sigma}} = 0$$

$$\Rightarrow \boxed{\hat{x} = \mu}$$

5c) Suppose a set of points $D = \{x_1, \dots, x_n\}$ are drawn independently from a given univariate gaussian $N(x; \mu, \sigma^2)$.
Write down an expression for the multivariate (joint) probability density function.

- Since the draws are independent, $f_{X_1, \dots, X_n}(x_1, \dots, x_n) = f_{X_1}(x_1) \cdot f_{X_2}(x_2) \cdots f_{X_n}(x_n)$
$$= \prod_{i=1}^{n} f_{X_i}(x_i).$$

- $X_i \sim N(\mu, \sigma^2)$
$$f_{X_i}(x_i) = \frac{1}{\sqrt{2\pi}\,\sigma} \cdot e^{-\frac{(x_i - \mu)^2}{2\sigma}}$$

- $$f_{X_1, \dots, X_n}(x_1, \dots, x_n) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{(x_1 - \mu)^2}{2\sigma}} \cdot \dots \cdot \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{(x_n - \mu)^2}{2\sigma}}$$

$$= \frac{1}{(2\pi)^{n/2}\,\sigma^n} \cdot \exp\left[-\frac{1}{2\sigma} \sum_{i=1}^{n} (x_i - \mu)^2\right]$$

III .. Gradient, optimization.

- Gradient = generalization of 'derivative' to functions of many variables.

  = a vector whose components are the partial derivatives of f

- $f(x_1, \dots, x_n)$ differentiable. $f: \mathbb{R}^n \to \mathbb{R}$
- $\nabla f = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{bmatrix}$

- "Slope" of the tangent to the graph of the function.
- Points in the direction of greatest increase of the function $\circledast$

Ex:

$$f(x, y, z) = x^2 + 3y + z^4$$

$$\nabla f = 2x\vec{i} + 3\vec{j} + 4z^3\vec{k} = \begin{bmatrix} 2x \\ 3 \\ 4z^3 \end{bmatrix}$$

- Local maximum/minimum $\vec{x_0}$ $\boxed{\nabla f(\vec{x_0}) = 0}$ (There is no direction that leads to increase)

problem 6: $\underset{\text{assumed to be given}}{x} = [x_1, x_2]^T \quad \theta = [\theta_1, \theta_2]^T$

$$L(x, \theta) = \log\Big[\overbrace{1 + \exp(\underbrace{-x_1\theta_1 - x_2\theta_2}_{-\vec{\theta}\cdot\vec{x}})}^{\gamma}\Big]$$

$$\frac{\partial L(x,\theta)}{\partial \theta_1} = \frac{\partial}{\partial \gamma}\log\gamma \cdot \frac{\partial \gamma}{\partial \theta_1}$$

$$= \frac{1}{1 + \exp(-x_1\theta_1 - x_2\theta_2)} \cdot \left[-x_1 \exp\{-x_1\theta_1 - x_2\theta_2\}\right]$$

$$= \frac{-x_1 \exp\{-\vec{x}\cdot\vec{\theta}\}}{1 + \exp\{-\vec{x}\cdot\vec{\theta}\}}$$

similarly $$\frac{\partial L(x,\theta)}{\partial \theta_2} = \frac{-x_2 \exp\{-\vec{x}\cdot\vec{\theta}\}}{1 + \exp\{-\vec{x}\cdot\vec{\theta}\}}.$$

so $$\nabla_\theta L(x,\theta) = \begin{bmatrix} -x_1 & -x_2 \end{bmatrix}^T \frac{\exp\{-\vec{x}\cdot\vec{\theta}\}}{1 + \exp\{-\vec{x}\cdot\vec{\theta}\}} = -\vec{x} \cdot \frac{\exp(-\vec{\theta}\cdot\vec{x})}{1 + \exp(-\vec{\theta}\cdot\vec{x})}$$

b) Into which direction does the gradient point?

Direction of greatest increase of $L(x, \theta)$.

$\theta' = \theta + \epsilon \nabla_\theta L(x, \theta)$

if $\epsilon > 0$, $L(x, \theta)$ becomes larger.

if $\epsilon < 0$, $L(x, \theta)$ becomes smaller.