Optimisation

Objective → Minimizing a function

Or, $\min f(\bar{x})$; $\bar{x} \in (x_1, x_2, x_3 + \dots + x_n)$
Minimizing a function also means reducing an error.
Optimising can mean ① Reducing Time ② Power ③ Error etc.
Maximizing is also a minimization problem.

*Prerequisites

Matrix & Vector

$$\bar{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \underset{\text{tuple}}{(x_1, x_2, x_3)}$$

$n \times 1$ column matrix = vector

These are vectors.

① Inner Product (Generalisation of Dot Product)

→ Tells us about how close two vectors are.

$(\bar{x} \cdot \bar{y}) = \langle \bar{x}, \bar{y} \rangle = \bar{x}^T y$ → Vector multiplication of $\bar{x}^T$ & y.

(x^T → Transpose → Row ⇄ Column)

$\bar{x}^T \bar{y}$ → Returns a scalar

*** Properties of Inner Product**

① $\langle \bar{x}_1, \bar{y}_1 + \bar{y}_2 \rangle = \langle \bar{x}_1, \bar{y}_1 \rangle + \langle \bar{x}_1, \bar{y}_2 \rangle$ = **Additivity**

② $\langle \bar{x}_1, \bar{y}_1 \rangle = \langle \bar{y}_1, \bar{x}_1 \rangle$ = **Symmetry**

③ $\langle \lambda \bar{x}_1, \bar{y}_1 \rangle = \langle \bar{x}_1, \lambda \bar{y}_1 \rangle = \lambda \langle \bar{x}_1, \bar{y}_1 \rangle$ = **Magnification**

④ $\langle \bar{x}, \bar{x} \rangle = 0$ if $\bar{x} = 0$

* Assuming two vectors (column matrices)

$$\bar{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \quad \bar{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \quad \& \quad \bar{x}^T = (x_1 \; x_2 \; .. \; x_n), \quad \bar{y}^T = (y_1 \; y_2 \; .. \; y_n)$$

$$\& \quad \langle \bar{x}, \bar{y} \rangle = \langle \bar{y}, \bar{x} \rangle = (\bar{x}^T \bar{y}) = \bar{y}^T \bar{x}$$

Orthogonality = Perpendicularity => Inner Product = 0
(General) (Specific)

Norm → Size → How big the vector is → Same as modulus but higher dimension.

* Eucledian Norm: $||\bar{x}||$

$$\bar{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \quad \therefore ||\bar{x}|| = \sqrt{x_1^2 + x_2^2 + \dots + x_n^2}$$

↳ 2 norm

{Those norms that don't fulfill triangle law are psuedo norms}

Definiteness

→ **Conditions of Norm**

① $||\bar{x}|| > 0$;

② $||\bar{x}|| = 0$; if $\bar{x} = 0$

→ Positive semi definiteness (Can be 0 or more)

③ $||\lambda \bar{a}|| = \lambda ||\bar{a}||$

④ Triangle Inequality

$$||\bar{a} + \bar{b}|| \leq ||\bar{a}|| + ||\bar{b}||$$

↳ (Kind of like triangle law)

$$\text{* P norm} = ||\bar{x}||_p = \left(\sum |x_i|^p \right)^{\frac{1}{p}}$$

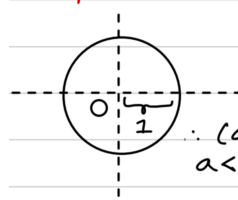
$$\text{1 Norm} = ||\bar{x}||_1 = \left(\sum |x_i| \right) \rightarrow \text{Manhattan Norm}$$

$$\infty \text{ Norm} = ||\bar{x}||_\infty = \max_i |x_i|$$

Q: Find all vectors with 2-norm = 1 (in 2D space = 2 variables)

let, $\bar{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$ & $\sqrt{x_1^2 + x_2^2} = 1$

$$\Rightarrow x_1^2 + x_2^2 = 1$$



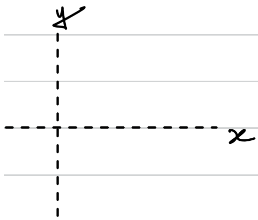
$\therefore (a, b) \; || \; a, b \in \mathbb{R}, \; a < 1, b < 1$

1 norm = 1

$|x_1| + |x_2| = 1$

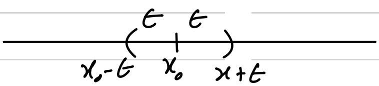
But,

$x_1 + x_2 = 1$
$-x_1 - x_2 = 1$
$x_1 - x_2 = 1$
$-x_1 + x_2 = 1$



In 1D → we get a line
2D → We " disc
3D → " " " Sphere
4D → !?

* Interval = Area around a point



; Interval $\boxed{|x - x_0| < \epsilon}$

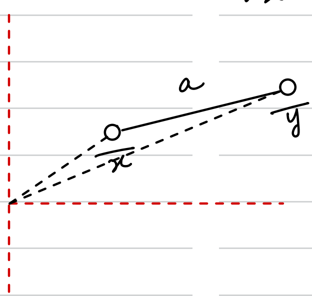
$S = \{ x ; \|x - a\| \in \mathbb{Z} \}$ → ϵ neigh

↳ But this is in 1D.

In higher dimensions, this is called a neighborhood.

*Distance (Metric)

↳ Scalar



$d(\overline{x}, \overline{y}) = \|\overline{y} - \overline{x}\|$

* Eigenvalues & Eigenvectors

$$\underbrace{\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}}_{\bar{A}} \underset{\bar{u}}{\begin{pmatrix} 1 \\ 1 \end{pmatrix}} = \begin{pmatrix} 1 \\ 1 \end{pmatrix} = (1)\bar{u}$$

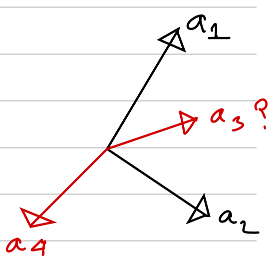
$$\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} 1 \\ -1 \end{pmatrix} = \begin{pmatrix} -1 \\ 1 \end{pmatrix} = (-1)\begin{pmatrix} 1 \\ -1 \end{pmatrix}$$

$A\bar{u} = \lambda\bar{u}$; $\bar{u} \neq 0$

* Independence :

→ Two vectors will be dependent if they are on each other.

$$a_1 = \begin{pmatrix} 2 \\ 1 \\ 5 \end{pmatrix} \quad a_2 = \begin{pmatrix} 4 \\ 2 \\ 0 \end{pmatrix} \quad a_3 = a_1 + a_2 = \begin{pmatrix} 6 \\ 3 \\ 5 \end{pmatrix}$$



* a_1, a_2 are independent vectors and span the plane $a_1 a_2$ & another vector $a_4 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$ is also independent & spans to a 3D space along with a_1, a_2.

*** Basis:** Independent vectors that 'span' the subspace.

*** Dimension:** Number of basis vectors for the subspace.

Basis of xy plane $\hat{i}$ & $\hat{j}$

Any linear combination $\hat{i}$ & $\hat{j}$ form any vector in the xy plane. $(a\hat{i} + b\hat{j})$. (The entire 2D space)

Span: How far you can reach with addition & vector multiplication.

A linear combination of 3 vectors $a\hat{i} + b\hat{j} + c\hat{k}$

***Basis:** The minimum number of vectors required to represent a particular vector space.

* In general basises do not need to be orthogonal.
* If basises are orthogonal, co-efs in the linear combination can be calculated easily.

Example: $\{u_1, u_2, u_3, \dots, u_n\}$

$$\overline{V} = \alpha_1 \overline{u}_1 + \alpha_2 \overline{u}_2 + \dots + \alpha_n \overline{u}_n \quad - (1)$$

$\overline{u}_1 \cdot (1)$, then

$$\overline{u}_1 \cdot \overline{V} = \alpha_1 \cdot \overline{u}_1 \cdot \overline{u}_1 \; ; \text{ so, } \alpha_1 = \frac{\overline{u}_1 \cdot \overline{V}}{\overline{u}_1 \cdot \overline{u}_1}$$

$$\boxed{\text{so, } \alpha_n = \frac{\overline{u}_n \overline{V}}{\overline{u} \cdot \overline{u}_n}}$$

* A vector is normalised $\|\overline{u}\| = 1$
* A basis of a vector space is orthogonal if other dot products of basises of that vector space yields 0.

* Orthonormal Basis: When $\overline{u_1} \cdot \overline{u_1} = 1$ but $\overline{u_1} \cdot \overline{u_x} = 0$

$$\overline{u_i} \cdot \overline{u_j} = \begin{Bmatrix} 0 & i \neq j \\ 1 & i = j \end{Bmatrix}$$

And if vectors are orthonormal the co-ef calculation gets easier because $\alpha_n = \frac{\overline{u_n} \cdot \overline{v}}{\overline{u_n} \cdot \overline{u_n}} = (\overline{u_n} \cdot \overline{V})/1$

Q: ① Show that $\left\{ \overset{\overline{a}}{\frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \end{pmatrix}}, \overset{\overline{b}}{1/\sqrt{2} \begin{pmatrix} 1 \\ -1 \end{pmatrix}} \right\}$ is a orthonormal set

1+>

1->

$$\rightarrow \overline{a} \cdot \overline{a} = (1/\sqrt{2}) \cdot (1/\sqrt{2}) + (1/\sqrt{2}) \cdot (1/\sqrt{2})$$
$$= 1/2 + 1/2 = 1$$

&

$$\overline{a} \cdot \overline{b} = (1/\sqrt{2}) \cdot (1/\sqrt{2}) + (1/\sqrt{2}) \cdot (-1/\sqrt{2})$$
$$= 1/2 - 1/2 = 0$$

② Represent $\begin{pmatrix} 3 \\ 9 \end{pmatrix}$ as a linear combination of $\overline{a}$ & $\overline{b}$

$$\overline{a} \qquad\qquad \overline{b}$$

$$\rightarrow \begin{pmatrix}3\\4\end{pmatrix} = \alpha\left(\frac{1}{\sqrt{2}}\begin{pmatrix}1\\1\end{pmatrix}\right) + \beta\left(\frac{1}{\sqrt{2}}\begin{pmatrix}1\\-1\end{pmatrix}\right) \quad \text{— (e)}$$

Now, $3 = \alpha \, 1/\sqrt{2} + \beta \, 1/\sqrt{2}$ — (I)   Solve for (I) & (II)

$4 = \alpha \, 1/\sqrt{2} - \beta \, 1/\sqrt{2}$ — (II)

or,

$\overline{a}$ & $\overline{b}$ are orthonormal so,

$$\overline{a}\cdot (e) \Rightarrow \begin{pmatrix}3\\4\end{pmatrix}\begin{pmatrix}1/\sqrt{2}\\1/\sqrt{2}\end{pmatrix} = \alpha\begin{pmatrix}1/\sqrt{2}\\1/\sqrt{2}\end{pmatrix}\cdot\begin{pmatrix}1/\sqrt{2}\\1/\sqrt{2}\end{pmatrix} + 0$$

$$= \alpha \times 1$$

Eigenvalue, Eigenvectors

Normalising a vector $= \dfrac{\overline{u}}{\|u\|}$ (Normalising is a generalisation of of finding a unit vector)

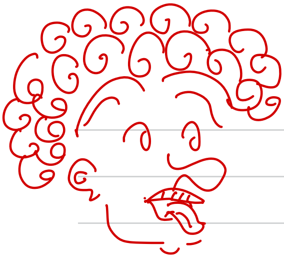
$A\overline{u} = \lambda\overline{u}$ ($\overline{u}$ is a normalised eigen vector)

$\Rightarrow \overline{u}^T A \overline{u} = \overline{u}^T(\lambda\overline{u})$

$= \lambda(\overline{u}^T\overline{u}) = \lambda \rightarrow$ Eigenvalue

$\therefore \lambda = \overline{u}^T A \overline{u}$

* Symmetric Matrix $= \overline{a}^T = \overline{a}$
  - ↳ Eigen value of a symmetric matrix is a real number.
  - ↳ If A is symmetric $\overline{u}_1 \cdot \overline{u}_2$



* Finding Eigen-values & Eigen-vectors

$A = \begin{pmatrix} 0 & 1 \\ -2 & -3 \end{pmatrix}$ eigen vector $\bar{u}$ will satisfy

$Ax = \lambda x$

Now,

$$Ax = \lambda x = \lambda I x \quad \left(I = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \right)$$

$$Ax = \lambda I x \Rightarrow A\bar{x} - \lambda I \bar{x} = 0 \Rightarrow \underbrace{(A - \lambda I)}_{\text{Not invertible}} x = 0$$

& $\|A - \lambda I\| = 0$

So,

$$\det\left(\begin{pmatrix} 0 & 1 \\ -2 & -3 \end{pmatrix} - \lambda \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \right) = 0$$

$$\Rightarrow \det\left(\begin{pmatrix} 0 & 1 \\ -2 & -3 \end{pmatrix} - \begin{pmatrix} \lambda & 0 \\ 0 & \lambda \end{pmatrix} \right) = 0$$

$$\Rightarrow \det \begin{pmatrix} -\lambda & 1 \\ -2 & \lambda - 3 \end{pmatrix} = 0$$

$$\Rightarrow \left(-\lambda(\lambda - 3) - (-2)(1) \right) = 0$$

$$\Rightarrow \lambda = -2, -1$$

So eigen-vectors of A are,

$$A\bar{x} = \lambda \bar{x} \Rightarrow \begin{pmatrix} 0 & 1 \\ -2 & -3 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \lambda \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

for $\lambda = -1$,

so, $\begin{pmatrix} 0 & 1 \\ -2 & -3 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = -1 \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} -x_1 \\ -x_2 \end{pmatrix}$

Now, $0 \cdot x_1 + 1 x_2 = -x_1$

$\Rightarrow x_2 = -x_1$ & $-2x_1 - 3x_2 = -x_2$

$\Rightarrow -2x_1 = +2x_2$

or $x_1 = -x_2$

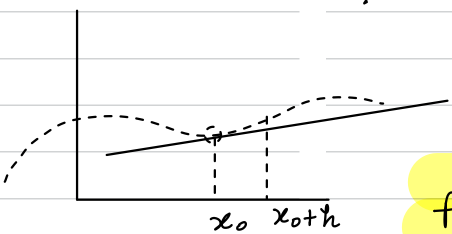
Now, as $x_1 = -x_2$; so an eigen vector is $\begin{pmatrix} a \\ -a \end{pmatrix}$ when $x = a$.

* Derivatives & Optimality Conditions

(i) Multivariable Calculus

$f(x) = f'(x) = df/dx$

→ Derivatives give us the information of how fast a function is decreasing or increasing.

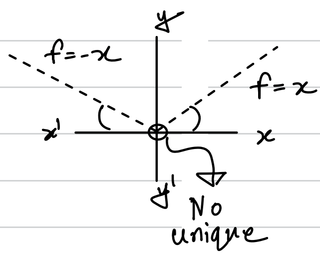


$$f'(x) = \lim_{h \to 0} \frac{f(x_0+h) - f(x_0)}{h}$$

$\Delta x = (x_0 + h) - x_0$
$= h$

$\Delta f = f(x_0+h) - f(x)$

→ Derivatives don't exist on all points of the function.



$f(x) = |x|$

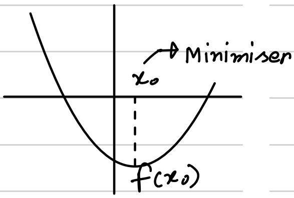
Slope changes over time & by time means as function expands.

* Conditions of a differentiable function
  1. Function must be continuous.
  2.

* Maxima is the maximum value reachable with the function

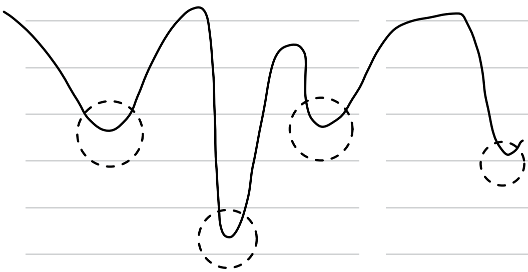
* Not all functions have minima or maxima. ($f(x) = x$ or $f(x) = 1/x$)

x_0 is the minimiser of $f(x)$ if $f(x_0)$ is the lowest value for $f(x)$.



Global Minima : $f(x_0) \leq f(x) \ \forall \ x$

Local Minima : $f(x_0) \leq f(x)$ for a small neighborhood around x_0.



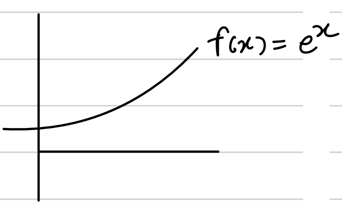
* Local Minima = 0 ; in a continuous function (Because it is a turning point ; but, this dosen't mean at some place if we get $f(x) = 0$, it'll not necessarily indicate a local minima.

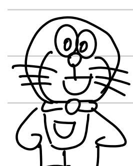
Taylor Series:

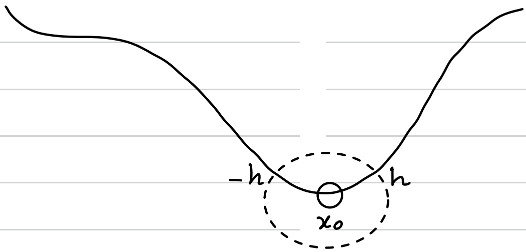
$$f(x+h) = f(x_0) + h f'(x_0) + \frac{h^2}{2!} f''(x_0) + \frac{h^3}{3!} f^3(x_0) + \dots + \frac{h^k}{k!} f^k(x_0)$$

+ error

& error $\leq \frac{h^k}{k!} f^k(x_0)$ if h is small enough.







$f(x^0+h) = f(x_0) + f'(x_0)\cdot h$

$f(x-h) = 1-2h$

if $f'(x^0) > 0$; then $f(x^0+h) < f(x_0)$ for -ve h

if $f'(x^0) < 0$; " $f(x^0+h) < f(x_0)$ for +ve h

* Necessary Condition: $f'(x) = 0$ at local minima / maxima.

* Sufficient Condition: 2nd Order

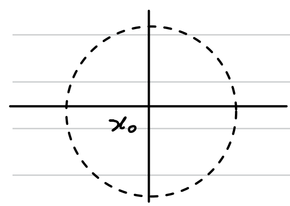
Min: $f''(x_0) > 0$

Max: $f''(x_0) < 0$

* We call something a minima

Neighborhood

① $f(\bar{x}_0) \leq f(\bar{x}) \ \forall \ \|x - x_0\| < \epsilon$ $\quad \mathcal{B} = \{ \bar{x} \mid \|\bar{x} - x_0\| < \epsilon \}$



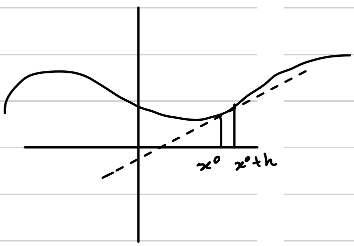
* Optimality Condition

① Necessary Condition $\longrightarrow$ $f'(x_0) = 0$

② Sufficient Condition $\longrightarrow$ $f''(x_0) > 0 \to$ min

$f''(x_0) < 0 \to$ max

} For local min/max



Didn't understand the things about

$f(x_0+h) = f(x_0) + \underbrace{h}f'(x_0) + \text{error}$

why h should be/not a little negative?

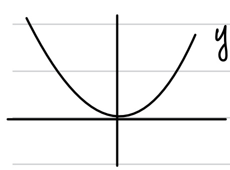
Adding up all of these

$$f(x_0+h) = f(x_0) + \underbrace{hf'(x^0)}_{\text{Becomes } 0} + \frac{h^2}{2!}f''(x^0)$$

Becomes 0 if it is a critical point

$f(x^0+h) > f(x^0) \to$ will always hold true regradless of the value of h.

Ex: $f(x) = x^4$; Find it's optimal point



NC: $f'(x) = 0$

$\Rightarrow 4x^3 = 0$

$\therefore x = 0$

$f''(x) = 12x^2 = 0$

$f^3(x) = 24x = 0$ } For $x = 0$

$f^4(x) = 24 \; = 0$ $\to$ One even derivitive is greater than 0 $\to$ Local Minima

$\cancel{f^5(x) = 0}$

But for $f(x) = x^5 - x$

$f'(x) = 5x^4 - 1 = 0 \Rightarrow 5x^4 = 1 \Rightarrow x^4 = 1/5 \Rightarrow x = \left(1/5\right)^{1/4}$

$f''(x) = 20x^3$ & $x = (\cdot 2)^{1/4} \Rightarrow 20 \times (\cdot 2^{1/4})^3 = 20 \times (\cdot 2)^{3/4} > 0$

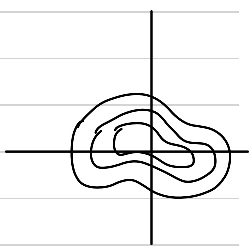
we'll get local minima

* Multivariable Calculus

$f(x,y)$ & $\delta_x = \frac{d}{dx} f(x,y) = \lim_{h \to 0} \frac{f(x+h,y) - f(x,y)}{h}$

$\delta_y = \frac{d}{dy} f(x,y) = \lim_{h \to 0} \frac{f(x,y+h) - f(x,y)}{h}$

* 3D Graphs are hard to draw but can be represented using a contour line. Contour lines represent the points in which the function does not change.



Partial Derivative Example

$f(x,y) = (x^2+y^2)^2 xy$

$\delta_x = (x^2+y^2)^2 y + xy \times 2(x^2+y^2) 2x$

$= (x^2+y^2)^2 y + 4x^2 y (x^2+y^2)$

$= (x^2+y^2) y (x^2+y^2+4x^2)$

$= (x^2+y^2) y (5x^2+y^2)$

$\delta_y = (x^2+y^2) x + xy \{2(x^2+y^2) 2y\}$

$= x(x^2+y^2) + 4xy^2 (x^2+y^2)$

*Directional Derivatives

We can write a multivariable function as
$f(x_1, x_2, x_3, \ldots\ldots x_n)$ or $f(\bar{x})$ here $\bar{x} = \begin{pmatrix} x_1 + h \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{pmatrix}$
& it's basises are $\underset{e_1}{\begin{pmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}} \underset{e_2}{\begin{pmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}} \cdots \underset{e_n}{\begin{pmatrix} \\ \end{pmatrix}}$ or standard basises.

$$Sx_1 = \lim_{h \to 0} \frac{f(x_1 + h) - f(x_1)}{h}$$; but in vector format it'll be

$$= \lim_{h \to 0} \frac{f(\bar{x} + he_1) - f(\bar{x})}{h}$$

$$f_d(\bar{x}) = \lim_{d \to 0} \frac{f(\bar{x} + h\bar{d}) - f(\bar{x})}{h}$$; d is a directional vector

* Multivariable Taylor Series

$f(x_1, x_2, \ldots, x_n)$

$\nabla \rightarrow$ The partial derivative matrix

$\rightarrow \begin{bmatrix} \frac{\partial}{\partial x_1} \\ \frac{\partial}{\partial x_2} \\ \vdots \\ \partial / \partial x_n \end{bmatrix}$

$$\frac{\partial}{\partial x_1} f = \lim_{x \to 0} \frac{f(x+h, x_2, x_3, \ldots) - f(x_1, x_2 \ldots)}{h}$$

$$= \lim_{h \to 0} \frac{f\left(\begin{pmatrix} x_1 + h \\ x_2 \\ \vdots \end{pmatrix}\right) - f\left(\begin{pmatrix} x_1 \\ x_2 \\ \vdots \end{pmatrix}\right)}{h}$$

$$= \lim_{h \to 0} \frac{f(\bar{x} + he_1) - f(\bar{x})}{h}$$

$$= \lim_{h \to 0} \frac{f(\bar{x} + h\bar{d}) - f(\bar{x})}{h}$$; $d \rightarrow$ directional vector

$$= \bar{d} \cdot \nabla f = f_d$$

Now,

$$f(x_0+h) = f(x_0) + hf'(x_0) + \frac{h^2}{2!}f''(x_0) + \dots + \frac{h^n}{n!}f^n(x_0) + \text{error}$$

$\rightarrow$ This was the taylor series of single variables

And,

$$f(\bar{x}_0+\bar{h}) = f(\bar{x_0}) + \bar{h}\cdot\bar{\nabla} f\Big|_{x_0} + \frac{1}{2!}(\bar{h}\cdot\bar{\nabla})^2 f\Big|_{x_0} + \frac{1}{n!}(\bar{h}\cdot\bar{\nabla})^n f\Big|_{x_0} + \text{error}$$

Now, $\bar{h} = \begin{pmatrix} h_1 \\ h_2 \\ \vdots \\ h_n \end{pmatrix}$ $\quad \vec{\nabla} = \begin{pmatrix} \frac{\delta}{\delta x_1} \\ \delta/\delta x_2 \\ \vdots \\ \delta/\delta x_n \end{pmatrix}$

so, $\bar{h}\cdot\bar{\nabla} = \begin{pmatrix} h_1\,\delta/\delta x_1 \\ h_2\,\delta/\delta x_1 \\ \vdots \\ h_n\,\delta/\delta x_n \end{pmatrix}$ $\quad$ so, $(\bar{h}\cdot\bar{\nabla})f = h_1\frac{\delta f}{\delta x_1} + h_2\frac{\delta f}{\delta x_2} + \dots + h_n\,\delta f/\delta x_n$

* Now, let's look into an example,

$\bar{x} = (x_1, x_2)$

$\bar{h} = (h_1, h_2)$

$\bar{\nabla} = (\delta/\delta x_1, \delta/\delta x_2)$

so,

$$\bar{h}\cdot\bar{\nabla} = h_1\frac{\delta}{\delta x_1} + h_2\,\delta/\delta x_2$$

Remember

$$(\bar{A}+\bar{B})^2 = \bar{A}^2 + \bar{B}^2 + \bar{A}\bar{B} + \bar{B}\bar{A}$$

Now,

$$(\bar{h}\cdot\bar{\nabla})^2 = \left(h_1\,\delta/\delta x_1 + h_2\,\delta/\delta x_2\right)^2$$

$$= h_1^2(\delta/\delta x_1)^2 + (h_2\,\delta/\delta x_2)^2 + h_1h_2\frac{\delta f}{\delta x_1\delta x_2} + h_2h_2\,\frac{\delta f}{\delta x_2\delta x_1}$$

These parts are equal in our functions

Mixed Partials

$$[h_1 \quad h_2] \begin{pmatrix} \frac{\delta^2}{\delta x_1^2} f & \frac{\delta^2}{\delta x_1 \delta x_2} f \\ \frac{\delta^2}{\delta x_2 \delta x_1} & \frac{\delta^2}{\delta x_2^2} f \end{pmatrix} \begin{pmatrix} h_1 \\ h_2 \end{pmatrix} = (\bar{h} \cdot \bar{\nabla})^2$$

Hessian → Matrix of all 2nd Derivatives

So, $\frac{1}{2!}(\bar{h}\cdot\bar{\nabla})f\Big|_{x_0} = \frac{1}{2!}\bar{h}^T H(\bar{x}_0) h$

* Optimality Condition of Multivariable

1. All 1st Order derivatives must be 0 {1st Order Optimality condition}
   $\vec{\nabla} \cdot f = 0$
2. If $\bar{x}_0$ is a local minimiser then $\bar{h}^T H h$ is +ve for all $\bar{h}$ eigs$(H(x_0)) > 0$ {And vice versa}

eigs$(H(x_0)) > 0$ → Local Min
eigs$(H(x_0)) < 0$ → Local Max } Conclusive

eigs$(H(x_0)) > 0$ → Local Min
eigs$(H(x_0)) < 0$ → Local Max } Inconclusive

& eigs$(H(x_0)) > 0 \ \forall \ \bar{x}_0$ → These are convex functions

* Odd Derivative = 0 $\rightsquigarrow$ Local Minima / Maxima

* Example 1: $x^3 - 3x$

$\rightarrow f(x) = x^3 - 3x$

$f'(x) \Rightarrow 3x^2 - 3 = 0$ $\quad$ $f''(x) = 6x$

$\Rightarrow x = \pm 1$ $\quad$ $f''(+1) = 6$ (Upward curve so; +ve)

$f''(-1) = 6$ (Downward curve so: −ve)

* Multivariable Taylor Series

$$f(x_0 + h) = f(x_0) + (\bar{h} \cdot \bar{\nabla}) f \Big|_{x_0} + \frac{1}{2!} (h \cdot \bar{\nabla})^2 f \Big|_{x_0}$$

$$+ \cdots + \frac{1}{k!} (\bar{h} \cdot \bar{\nabla})^k f \Big|_{x_0} + e$$

$$= f(x_0) + \underbrace{\bar{h} \cdot \bar{\nabla} f(x_0)}_{\text{Gradient}} + \frac{1}{2} \bar{h}^T \mathcal{H} \bar{h}$$

Suppose, $f(x_1, x_2) = x_1^2 + 2x_1 x_2$

$$\therefore \nabla f(x_1, x_2) = \begin{pmatrix} 2x_1 + x_2 \\ 2x_2 \end{pmatrix}$$

* $f(x_1, x_2) = (x_1 + x_2)^2 + x_1 x_2$

$$\rightarrow \quad \nabla f(x_1, x_2) = \begin{pmatrix} 2(x_1+x_2)(1) + x_2 \\ 2(x_1+x_2)(1) + x_1 \end{pmatrix} = 0$$

$$\text{or,} \quad \left.\begin{aligned} 2x_1 + 3x_2 = 0 \\ 3x_1 + 2x_2 = 0 \end{aligned}\right\} \quad x_1 = 0 \quad \& \quad x_2 = 0$$

$$\mathcal{H} = \begin{pmatrix} 2 & 3 \\ 3 & 2 \end{pmatrix}$$

Now, $\det(\mathcal{H} - \lambda I) = 0$; so, $(2-\lambda)^2 - 9 = 0$

$\lambda = 5, -1 \rightarrow$ Saddle Point

* Quadratic Function

$f(x) = ax^2 + bx + c$

$f(x_1, x_2) = 2x_1^2 + x_2^2 + 3x_1 x_2 + x_1 + x_2 + 1$

$f(x_1, x_2, x_3) = x_1^2 + x_2^2 + x_3^2 + 2(x_1 x_2 + x_2 x_3 + x_3 x_1) + C$

Any quadratic form can be represented as

$$f(\bar{x}) = \bar{x}^T A \bar{x} + \bar{b} x + c$$

Example

$$[x_1, x_2] \begin{bmatrix} 2 & 3/2 \\ 3/2 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 2x_1^2 + \tfrac{3}{2} x_1 x_2 + \tfrac{3}{2} x_1 x_2$$

* Few Formulas

$f(\bar{x}) = \bar{x} A \bar{x}$; $\nabla f = 2(A + A^T)\bar{x}$

$f(\bar{x}) = \bar{b}^T \bar{x}$; $\nabla f = \bar{b}$

Now,

---> Quadratic Form

$f(\bar{x}) = \bar{x}^T A \bar{x} + \bar{b}^T \bar{x} + c$

$\nabla f = 2A\bar{x} + b$ → (Necessary condition)

if $\det(A) \neq 0$; $\bar{A}$ is invertible so,

$\bar{x} = \frac{1}{2} A^{-1} B^{-1}$ & if $\bar{A}$ is non-invertible

$H(x) = 2A$

* Multivariable Quadratic Function

$f(\bar{x}) = \bar{x}^T A x + b^T \bar{x} + c$

$f(\bar{x}) = [x_1 \ \ x_2] \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} x_1 & x_2 \end{bmatrix}$

$= [x_1 \ \ x_2] \begin{bmatrix} x_1 a_{11} + a_{12} x_2 \\ x_1 a_{21} + a_{22} x_2 \end{bmatrix}$

$= a_{11} x_1^2 + a_{12} x_2 x_1 + a_{21} x_1 x_2$
$\quad + a_{22} x_2^2)$

$\delta_1 f = 2a_{11} x_1 + (a_{12} + a_{21}) x_2$

?

Few Identities

* $f(\bar{x}) = b^T \bar{x}$

$\nabla f = \bar{b}$ → Gradient

$\nabla^2 f = \bar{0}$ → Hickesian

* $f(\bar{x}) = \bar{x}^T A \bar{x}$

$\nabla f = (A + A^T)\bar{x} = 2A_s \bar{X}$

$\nabla^2 f = (A + A^T) = 2A_s$

Now, $f(\bar{x}) = \bar{x}^T A \bar{x} + b^T \bar{x} + c$

$\nabla f = 2A_s \bar{x} + \bar{b} = 0$; Grad $= 0$ $\quad \therefore \bar{x}^* = -\frac{1}{2} A_s^{-1} \bar{b}$

$\nabla^2 f = 2A_s$

$$f(x) = \bar{x}^T A \bar{x} + \bar{b}^T A \bar{x} + c$$
$$= \bar{x}^T A \bar{x} + \hat{b}^T \bar{x} + c$$

$$\nabla f = (A + A^T)\bar{x} + \bar{b} = 0$$ → Stationary points = For the xs, $\Delta f == 0$.

$$\nabla^2 f = (A + A^T)$$ → Minima & Maxima depends on eigen values of this vector.

$e.v > 0$ → Minima
$e.v < 0$ → Maxima
$e.v = 0$ → Could be saddle point

Some $e.v < 0$ & some $e.v > 0$ → Definitely saddle points.

* Least Squares Problem:

Basically says, "Solve $Ax = b$";

$$\begin{pmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 7 \\ 8 \\ 9 \end{pmatrix}$$; This can be solved easily, so we try to find $A\bar{x} \approx b$

Or, we try to minimise $\|A\bar{x} - \bar{b}\|$

Now,

$$f(x) = \|Ax - b\|^2 = (A\bar{x} - \bar{b})^T (A\bar{x} - \bar{b}) = (\bar{x}^T A^T - \bar{b}^T)(A\bar{x} - \bar{b})$$
$$= \bar{x}^T A^T A \bar{x} \underbrace{- \bar{x}^T A^T \bar{b} - b^T A \bar{x}}_{\text{Equal Parts}} + \bar{b}^T \bar{b}$$
$$= \bar{x}^T A^T A \bar{x} - 2 b^T A x + \bar{b}^T \bar{b} =$$

$$\nabla f = 2A^T A x - 2A^T b$$ Grad

$$\nabla^2 f = 2A^T A$$ Hessian

Overdetermined System: # of variables < # of eqⁿs.

Full Rank Matrix: If A is full rank than A^TA is invertible

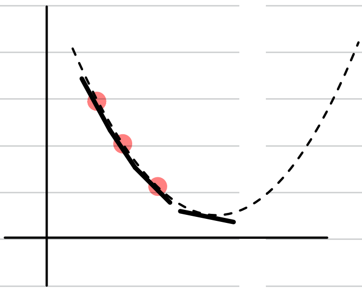
* 2nd Order Condition

$A^TA \succcurlyeq 0 \qquad \lambda \cdot \bar{u}$

$$\underbrace{\bar{u}^T (A^TA) \bar{u}}_{\geq 0} = \lambda \underbrace{(\bar{u}^T \bar{u})}_{\geq 0}$$

so, $\lambda \geq 0$

* Gradient Descent



* Initial point: x_0
* Descent direction: $\hat{d}_0$

$\bar{x}_1 = \bar{x}_0 + \hat{d}_0 s_0$

$\bar{x}_2 = \bar{x}_1 + \hat{d}_1 \cdot s_1$

$s_i \rightarrow$ Step size

But, Gradient Descent

$$\hat{d}_i = \frac{-\nabla f(x_i)}{\|\nabla f(x_i)\|}$$

$$s_i = \|\nabla f(x_i)\| \times \underbrace{\eta}_{\text{learning rate}}$$

* Least Squares

$A\bar{x} \approx \bar{b}$ $\quad \min_{x} \|A\bar{x}-\bar{b}\|^2 \rightsquigarrow$ Least Squares problem

Now,

$$\|A\bar{x}-\bar{b}\|^2 = (A\bar{x}-\bar{b})^T \cdot (A\bar{x}-\bar{b})$$
$$= (\bar{x}^T A^T - \bar{b}^T)(A\bar{x}-\bar{b})$$
$$= x^T (A^T A)\bar{x} - x^T A^T \bar{b} - b^T A\bar{x} + b^T b$$
$$= x^T (A^T A)\bar{x} - 2(A^T b)^T + b^T b$$

$$x^T A^T \bar{b} = b^T A \bar{x}$$

$$\underbrace{x^T A x}_{\downarrow\, A^T A} + \underbrace{b^T x}_{\downarrow\, A^T b} + \underbrace{c}_{\downarrow\, b^T b} \qquad \nabla f = 2(A^T A)\bar{x} - 2(A^T b)^T \bar{x} + \bar{b}^T \bar{b}$$

$$\boxed{\bar{x} = (A^T A)^{-1}(A^T b)} \rightarrow \text{Least Square Solution}$$

* Example:

Fit this data with a linear function

(x_1, x_2)	Y
1, 1	0
1, -1	-1
-1, 2	3
3, 4	5

$$f(x_1, x_2) = a_1 x_1 + a_2 x_2 + b$$

$$a_1(1) + a_2(1) + b = 0$$
$$a_1(1) + a_2(-1) + b = -1$$
$$a_1(-1) + a_2(2) + b = 3$$
$$a_1(3) + a_2(4) + b = 5$$

$$\begin{pmatrix} 1 & 1 & 1 \\ 1 & -1 & 1 \\ -1 & 2 & 1 \\ 3 & 4 & 1 \end{pmatrix} \begin{pmatrix} a_1 \\ \\ \\ \end{pmatrix}$$

Picture on phone

Q: Fit a quadratic function

x	y
-1	1
1	2
3	3
4	5

$$f(x) = ax^2 + bx + c$$

$$a(-1)^2 + b(-1) + c = 1$$
$$a(1)^2 + b(1) + c = 2$$
$$a(3)^2 + b(3) + c = 3$$
$$(4)^2 + b(4) + c = 4$$

$$\underbrace{\begin{pmatrix} -1^2 & -1 & 1 \\ 1^2 & 1 & 1 \\ 3^2 & 3 & 1 \\ 4^2 & 4 & 1 \end{pmatrix}}_{\overline{A}} \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \underbrace{\begin{pmatrix} 1 \\ 2 \\ 3 \\ 5 \end{pmatrix}}_{\overline{b}}$$

* Variation of Least square (Regularised Least Square) → Check Online

$$\sum_{\text{data}} (a\cdot 1 + b\cdot 1 + c - 2)^2 + \underline{\text{penalty}}$$

$$\|A\bar{x} - \bar{b}\|^2 + \alpha_1 x_1 + \alpha_2 x_2^2 + \cdots + \alpha_n x_n^2$$

$$[x_1 \; x_2 \; \cdots \; x_n] \begin{bmatrix} \alpha_1 & & & \\ & \alpha_2 & & \\ & & \alpha_3 & \\ & & & \alpha_n \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{bmatrix}$$

$$\min \left(\|A\bar{x} - \bar{b}\|^2 + \bar{x}^T D \bar{X} \right)$$

$$x^T (A^T A + D) x - 2(A^T b)^T \bar{x} + b^T b +$$

$f(x) = ae^x + b\sin x + c\log x$

Ex:

x	y
-1	1
1	2
3	3
4	5

$f(x) = ax^2 + bx + c$

$$\begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 9 & 3 & 1 \\ 16 & 4 & 1 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 5 \end{bmatrix}$$

$$\bar{x} = (A^T A)^{-1} A^T \bar{b}$$

$$\bar{a} = (x^T x + D)^{-1} x^T y$$

Regularisation function

$10 a_1^2 + a_2^2 + a_3^2$

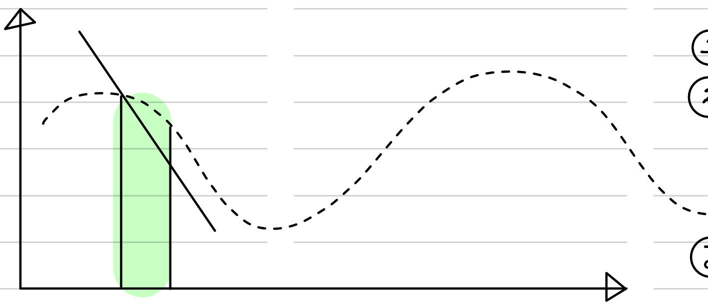
*Calculating Gradients

Example: Find stationary points of $f(x) = x^6 - x^4 + x^3 + x + 1$

$f'(x) = 6x^5 - 4x^3 + 3x^2 = 0$

No general formula exist for degree 4 & above.

We'll use gradient descent for solving for gradients of higher order functions.



1. Calculate slope.
2. Go to the opposite point where f(x) increases.
3. Slope decreases as we move towards minima. (Reduce step size as we go closer)

*Gradient Descent will be touched later on but Gradient calculation is an important thing. We'll discuss about a few techniques of calculating gradients.

1 1st derivative approach

→ $f'(x)$ is the gradient but it becomes complex to calculate. Ex:

$f(x) = e^x$

② Numerical Differentiation

→ Approximates derivatives $f'(x_0) = \frac{f(x_0+h) - f(x_0)}{h}$

↳ Forward difference method

$f'(x_0) = \frac{f(x_0) - f(x_0+h)}{h}$

↳ Backward difference.

$f'(x_0) = \frac{f(x_0+h) - f(x_0-h)}{h}$

* Approximation gets more accurate, the lower h value we take.

But, as these are approximations, these have errors & we can use the taylor series to calculate these errors.

$$f(x^0+h) = f(x_0) + hf'(x_0) + \frac{h^2}{2!} f''(x_0) + \cdots$$

$$\Rightarrow f(x^0+h) - f(x_0) = hf'(x_0) + \frac{h^2}{2!} f''(x_0) + \cdots$$

$$\Rightarrow \frac{f(x^0+h) - f(x_0)}{h} = f'(x_0) + \underbrace{\frac{h}{2!} f''(x_0)}_{\text{Dominant error term } O(h)} + \text{-----} \longrightarrow ①$$

↳ For forward difference.

Similarly, for bacward difference,

$$-\frac{f(x^0-h)}{h} + \frac{f(x_0)}{h} = f'(x_0) - \underbrace{\frac{h}{2!} f''(x_0)}_{\text{Dominant error term.}} + \text{-----} \longrightarrow ②$$

① + ⑪

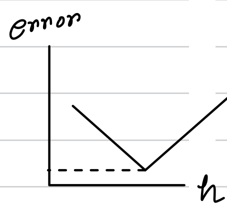
$$\frac{f(x_0+h) - \cancel{f(x_0)} - f(x_0+h) + \cancel{f(x_0)}}{2h} = f(x_0) + f(x_0) + \cancel{\frac{h}{2!} f''(x_0)} - \cancel{\frac{h}{2!} f''(x_0)} + \frac{h^2}{3!} f^3(x_0) + \frac{h^2}{3!} f^3(x_0) + \cdots$$

$$= \frac{\cancel{2} f(x_0)}{\cancel{2}} + \frac{\cancel{2} h^2}{3!} \frac{f^3(x_0) + \cdots}{\cancel{2}}$$

* Dominant term $O(h^2)$

& $O(h^2) < O(h)$ as $0 < h < 1$; so central difference is better.

* But h vs error graph



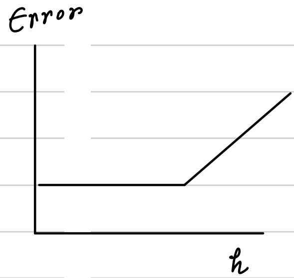
+ Reducing h till a certain point works.
+ That depends on the precision rate.

③ **Complex Diff** : $f(x_0 + ih) = f(x_0) + ihf'(x_0) + \frac{(ih)^2}{2!} f''(x_0) + \cdots\cdots$

$$\frac{Im(f(x_0 + ih))}{h} = \frac{\cancel{h} f'(x_0)}{\cancel{h}} - i\frac{h^{\cancel{3}\,2}}{3!} \frac{f'''(x_0) + \cdots\cdots}{\cancel{h}}$$

Error: $O(h^2)$

Error Graph:



*** Multivariable Case for the given techniques**

$f(\bar{x})$

$$\frac{\delta}{\delta x_1} f \approx \lim \frac{f\,[f((x_1+ih), x_2, x_3)]}{h}$$

⑨ Stochastic Grad Desc

$$f(x_1, x_2, x_3) = x_1 x_2 + x_3^2 \qquad u_1 = \tfrac{1}{\sqrt{2}}\begin{pmatrix}1\\1\\0\end{pmatrix} \quad u_2 = \tfrac{1}{\sqrt{2}}\begin{pmatrix}1\\-1\\0\end{pmatrix}$$

$$h = 0.001$$

Directional derivative,

$$\alpha_1 = \frac{f\left(\begin{pmatrix}1\\1\\1\end{pmatrix} + h\,\tfrac{1}{\sqrt{2}}\begin{pmatrix}1\\1\\0\end{pmatrix}\right) - f\begin{pmatrix}1\\1\\1\end{pmatrix}}{h}$$

$$= \frac{f(1.0007, 1.0007, 1) - f(1,1,1)}{0.001}$$

$$\nabla f \approx \alpha_1 \begin{pmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \\ 0 \end{pmatrix} + \alpha_2 \begin{pmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \\ 0 \end{pmatrix}$$

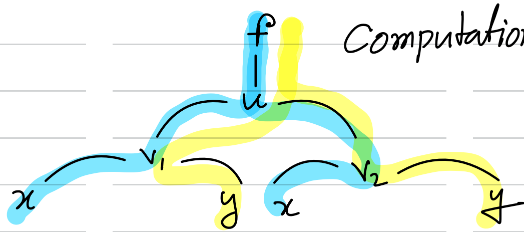
④ Automatic Difference

State of the art method

* Example: $f(x) = e^{\sin(x+y^2)}$

$$\frac{\delta f}{\delta x} = \frac{\delta}{\delta x} e^{(\dots)} \frac{\delta}{\delta x} \sin(\dots) \frac{\delta}{\delta x}(x)$$

$$\frac{\delta f}{\delta y} = \frac{\delta}{\delta y} e^{(\dots)} \frac{\delta}{\delta y} \sin(\dots) \frac{\delta}{\delta y}(y^2)$$



Computational Graph

*Blue parts get calculated once

$$\frac{\delta f}{\delta x} = \frac{\delta f}{\delta u} \times \frac{\delta u}{\delta v_1} \times \frac{\delta v_1}{\delta x} + \frac{\delta f}{\delta u} \times \frac{\delta u}{\delta v_2} \times \frac{\delta v_2}{\delta x}$$

$$= \frac{\delta f}{\delta u} \left(\frac{\delta u}{\delta v_1} \times \frac{\delta v_1}{\delta x} + \frac{\delta u}{\delta v_2} \times \frac{\delta v_2}{\delta x} \right)$$

&, $$\frac{\delta f}{\delta y} = \frac{\delta f}{\delta u} \left(\frac{\delta u}{\delta v_1} \times \frac{\delta v_1}{\delta y} + \frac{\delta u}{\delta v_2} \times \frac{\delta v_2}{\delta y} \right)$$

* How computer does it?

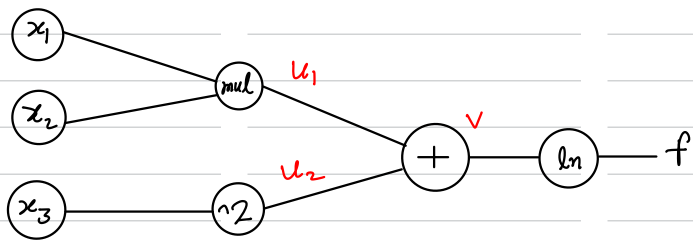
$f(x_1, x_2, x_3) = \ln(x_1 x_2 + x_3^2)$; find grad at $(1,1,1)$

Symbolic

$$\begin{pmatrix} \delta f/\delta x_1 \\ \delta f/\delta x_2 \\ \delta f/\delta x_3 \end{pmatrix} = \begin{pmatrix} \frac{1}{x_1x_2+x_3^2} \times x_2 \\ \frac{1}{x_1x_2+x_3^2} \times x_1 \\ \frac{1}{x_1x_2+x_3^2} \times 2x_3 \end{pmatrix} \text{ at } (1,1,1) = \begin{pmatrix} 1/2 \\ 1/2 \\ 1 \end{pmatrix}$$

Computational Process Steps:

① Computational Graph

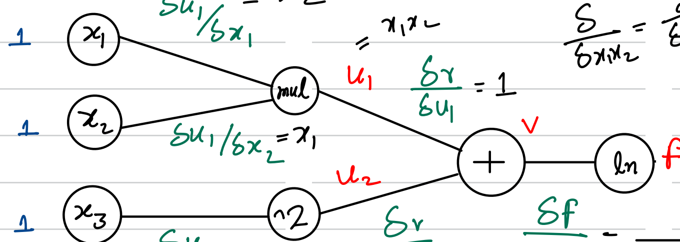


② Name intermediate variables (u_1, u_2, v etc above)

③ Calculate intermediate values

$$u_1 = 1, \quad u_2 = 1, \quad v = 2, \quad f = \ln(2)$$

④ Write intermediate derivatives



$\delta u_1/\delta x_1 = x_2$

$\delta u_1/\delta x_2 = x_1$

$\delta u_2/\delta x_3 = 2x_3$

$= x_1x_2$

$$\frac{\delta}{\delta x_1x_2} = \frac{\delta}{\delta x_1x_2}(x_1x_2 + x_3^2) = 1$$

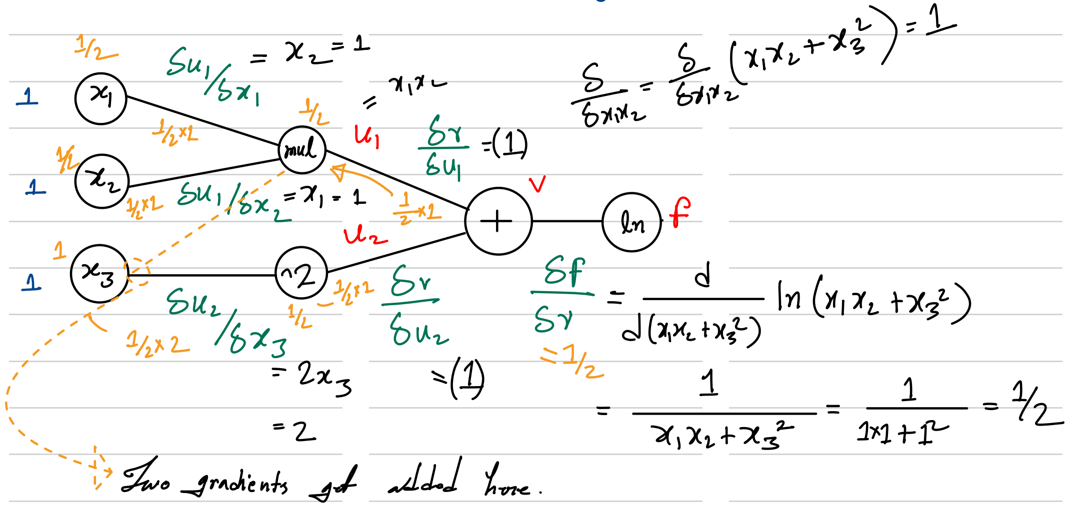
$$\frac{\delta v}{\delta u_1} = 1$$

$$\frac{\delta v}{\delta u_2} = 1$$

$$\frac{\delta f}{\delta v} = \frac{d}{d(x_1x_2+x_3^2)} \ln(x_1x_2+x_3^2)$$

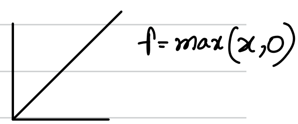
$$= \frac{1}{x_1x_2+x_3^2}$$

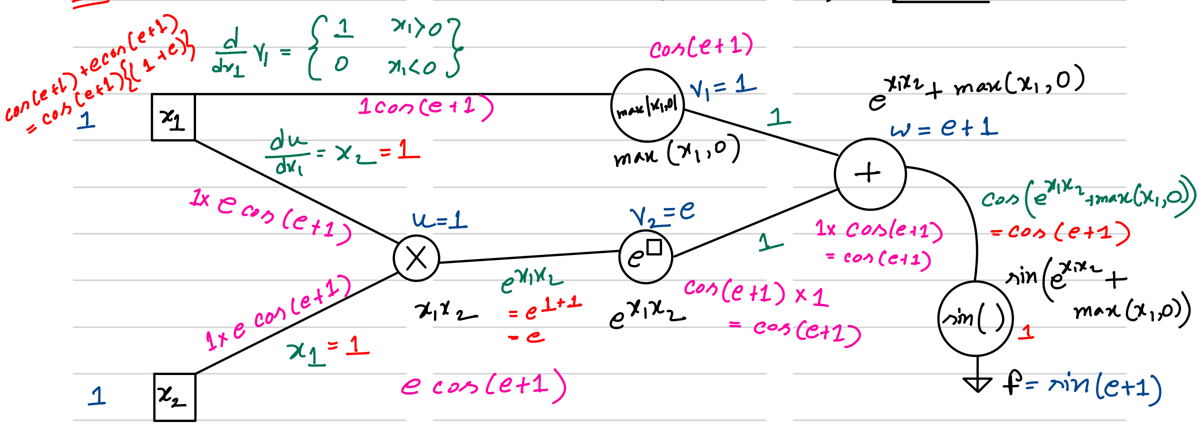
⑤ Calculate the values of the edges



* We have completed reverse autodiff *

Q: $f(x_1, x_2) = \sin\left(e^{x_1 x_2} + \max(x_1, 0)\right)$ at $(1,1)$





* Review

① Vector Space, $\mathbb{R}^n$ → Inner product

$$\overline{x}^T\overline{y} = \langle \overline{x}, \overline{y} \rangle \rightsquigarrow \text{Standard Inner Product}$$

* P-Norm: $\|\overline{x}\|_p$; $x = (x_1, x_2, x_3, \ldots, x_n)$

$$= (|x_1|^p + |x_2|^p + |x_3|^p + \ldots + |x_n|^p)^{1/p}$$

max norm: $\|\overline{x_2}\|, \|\overline{x_1}\|, \|\overline{x_3}\| = \max |x|_i$

* Distance: $\overline{x}, \overline{y}$

$$d(\overline{x}, \overline{y}) = \| x_2 - x_1 \|_2$$

* Neighborhood: $\overline{P}$

$$Ball(\overline{P}, \epsilon) = \{ \overline{x} \mid \|\overline{x} - \overline{P}\| < \epsilon \}$$

② Matrix:

* Any Matrix = Symmetric + Anti Symmetric

$$M = S + A$$

$$S = (M + MT)/2$$

$$A = (M - MT)/2$$

* Eigenvalues, Eigenvectors

$$A\overline{e} = \lambda \overline{e}$$

Mechanism: Caley-Hamilton Eqⁿ : $\det(A - \lambda I) = 0$

Symmetric Matrice : $A = AT$

In symmetric matrix:

(i) Eigenvalue : Always real.
(ii) Eigenvector: Orthogonal
(iii) Spectal Decomposition: $S = Q^T D Q$

Q → Matrix of Normalised eigenvector.

D → Diagonal Matric with eigen values in the diameter.

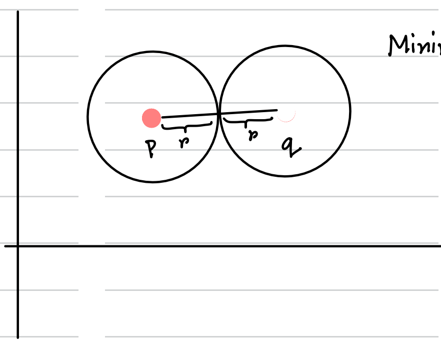
* Classification of Matrix

① All $e > 0$; Positive Definite
All $e \geq 0$; " Semi "
All $e < 0$; Negative Definite
All $e \leq 0$; Negative Semi "
All $e \gtrless 0$; Indefinite

$e = \lambda =$ Eigen value

$\det(M) = \prod e$
$\text{trace}(M) = \sum \lambda_i$

Q: Pset1



Minimum Distance : $2r$

Pset1 (Q4)

$\bar{u} = 3 \quad \bar{v} = 4 \quad w = 2\bar{u} + 3\bar{v}$ Find $\|5\bar{w}\|$

$\|5\bar{w}\| = 25\bar{u} + 35\bar{v} \qquad 5\bar{u} = 3\bar{u}$

$= 6\bar{u} + 12\bar{v} \qquad 5\bar{v} = 4\bar{v}$

$\left.\begin{aligned}\bar{u}^T\bar{u} = 1 \\ \bar{v}^T\bar{v} = 1\end{aligned}\right\}$ Normalised

$\|5\bar{w}\|_2 = \left((6\bar{u}+12\bar{v})^T(6\bar{u}+12\bar{v})\right)^{1/2}$

$= (36 + 144)^{1/2} = (180)^{1/2}$

2 Differentiation

$$f'(x) = \lim_{x \to 0} \frac{f(x_0+h) - f(x_0)}{h}$$

$$f'(x,y) \Rightarrow f'_x = \frac{f(x+h, y) - f(x,y)}{h}$$

$$f'_y = \frac{f(x, y+h) - f(x,y)}{h}$$

What we did here is,

$$f(x,y) \Rightarrow f_x = \frac{f\left(\begin{pmatrix} x \\ y \end{pmatrix} + h\begin{pmatrix} 1 \\ 0 \end{pmatrix}\right) - f\left(\begin{pmatrix} x \\ y \end{pmatrix}\right)}{h}$$

$$f_u(x,y) = \frac{f\left(\begin{pmatrix} x \\ y \end{pmatrix} + h\begin{pmatrix} u_1 \\ u_2 \end{pmatrix}\right) - f\left(\begin{pmatrix} x \\ y \end{pmatrix}\right)}{h}$$

$$= \bar{u} \cdot \nabla f$$

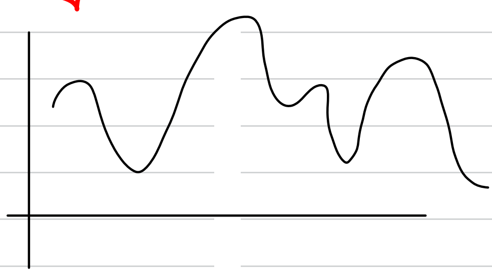
* Taylor Series:

$$f(x_0+h) = f(x_0) + h f'(x_0) + \frac{h^2}{2!} f''(x_0) + \dots + \frac{h^k}{k!} f^k(x_0) + \text{Error}$$

$\text{error}_k < h^k$ if h is small enough.

$$f(\bar{x}_0+\bar{h}) = f(\bar{x}_0) + (\bar{h}\cdot\bar{\nabla})f\Big|_{x_0} + \left\{\bar{h}\cdot\bar{\nabla}f\right\}$$
$$+ \frac{1}{2!}(\bar{h}\cdot\bar{\nabla})^2 f\Big|_{\bar{x}_0} + \left\{\frac{1}{2} h^T (\mathcal{H}) h\right\}$$
$$\text{-------} + \frac{1}{k!}(\bar{h}\cdot\bar{\nabla})^k f\Big|_{x_0}$$

③ Optimality Cond^n:



At x^*

① NC: $f'(x^*) = 0$ {Find critical points}

② SC: $f'(x^*) = 0$, $f''(x^*) > 0$, local min

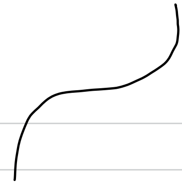
$f'(x^*) = 0$, $f''(x^*) < 0$, local max

if $f''(x^*) = 0$ then Odd derivative $= 0$

(If simultaneous derivatives lead to a 1st odd derivative $= 0$; function has a local optimum)

Even derivative $\gtrless 0$

Growth Rate Changing point



& 1st Odd derivative $\neq 0$ inflextion point. {Curvature changing point}

Saddle point = Hessian has mixed value

* Multivariable Optimality Cond^n

1. $\nabla f = 0$; Find critical values
2. $\text{eigs}(H) > 0$; local min
   $\text{eigs}(H) < 0$; local max
   $\text{eigs}(H) \gtrless 0$; saddle (Negative Positive Mix)

≥ 0 or $\leq 0 \to$ Inconclusive

→ All terms have a degree 2

* Quadratic Forms & Functions

$$f(x_1, x_2, \text{-----}, x_n) = x_1^2 + x_1 x_2$$

$$\begin{aligned} \bar{x}^T A \bar{x} &= \bar{x}^T (A_s + A_a) \bar{x} \\ &= \bar{x}^T A_s \bar{x} + \underbrace{\bar{x}^T A_a \bar{x}}_{0 \ \{\text{Rule!}\}} \\ &= \bar{x}^T A_s \bar{x} \end{aligned}$$

$f(\bar{x}) = \bar{x}^T A \bar{x} + \bar{b}^T \bar{x} + c$ {Comparable to $ax^2 + bx + c$}

$\nabla f = (A + A^T) \bar{x} + b = 2 A_s \bar{x} + b$; $x^* = -\frac{1}{2} (A_s)^{-1} \bar{b}$

$H = 2 A_s = (A + A^T)$

* Least Square Problem uses quadratic equations:

→ $A\bar{x} = b$ { Exact solution dosen't exist for

① A might not be invertible.

② A might not be a square matrix.

So we look for, $A\bar{x} \approx \bar{b}$ the $A\bar{x} - \bar{b} \approx 0$

$$\min \|A\bar{x} - \bar{b}\|^2 \approx 0^2$$

$$L.H.S = (A\bar{x} - \bar{b})^T (A\bar{x} - \bar{b})$$

$$= (A^T \bar{x}^T - \bar{b}^T)(A\bar{x} - \bar{b})$$

$$= A^T \bar{x}^T A\bar{x} - A^T x^T \bar{b} - \bar{b}^T A\bar{x} + b^T b$$

$$= x^T (A^T A)\bar{x} - 2(A^T \bar{b})^T \bar{x} + \bar{b}^T \bar{b}$$

* Regularised Least Square

* Gradient Descent :

*** PSet1**

1. Consider the following matrices:

A

ⓐ $A = \begin{bmatrix} 4 & 0 \\ 0 & 3 \end{bmatrix}$

$$A - \lambda I = \begin{pmatrix} 4 & 0 \\ 0 & 3 \end{pmatrix} - \begin{pmatrix} \lambda & 0 \\ 0 & \lambda \end{pmatrix}$$

$$= \begin{pmatrix} 4-\lambda & 0 \\ 0 & 3-\lambda \end{pmatrix}$$

1. Consider the following matrices:

$$\boldsymbol{A} = \begin{bmatrix} 4 & 0 \\ 0 & 3 \end{bmatrix}, \quad \boldsymbol{B} = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & -1 \\ 0 & -1 & 0 \end{bmatrix}$$

For each matrix,

(a) Find the eigenvalues and eigenvectors of the matrix.

(b) Show that, the eigenvectors corresponding to different eigenvalues are orthogonal.

(c) Find out whether the matrix is Positive (Semi)-Definite, Negative (Semi)-Definite or Indefinite.

Now,

$(4-\lambda)(3-\lambda) = 0$

$\lambda = 3 \;\&\; 4$ {Values} $\therefore$ Vectors

0 1

for $\lambda = 4$,

$(A - 4I)v = 0$

$\Rightarrow \left\{ \begin{pmatrix} 4 & 0 \\ 0 & 3 \end{pmatrix} - \begin{pmatrix} 4 & 0 \\ 0 & 4 \end{pmatrix} \right\} v = 0$

$\Rightarrow \begin{pmatrix} 0 & 0 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$

$0 \times x + y \times 0 = 0$ &

$0 \times x + y \times (-1) = 0$

$y = 0$ & $x = a$ $(a \in \mathbb{R})$

$\therefore v_1 = \begin{pmatrix} a \\ 0 \end{pmatrix}$

& $v_2 = \begin{pmatrix} 0 \\ a \end{pmatrix}$

ⓑ

1 0

$\langle v_1, v_2 \rangle = a \times 0 + 0 \times a$

$= 0 + 0$

$= 0$

1 0

$\therefore v_1$ & v_2 are orthogonal

ⓒ For

$\lambda = 3, 4 > 0 \rightarrow$ Positive Definite

e

ch matrix,

(a) Find the eigenvalues and e

genvectors

f

the matrix.

(b) Show that, the eigenvectors corresponding to different

$$A = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & -1 \\ 0 & -1 & 0 \end{pmatrix} \qquad \therefore A - \lambda I = \begin{pmatrix} -\lambda & 1 & 0 \\ 1 & -\lambda & -1 \\ 0 & -1 & -\lambda \end{pmatrix}$$

$$0 = \det(A - \lambda I) \Rightarrow -\lambda(\lambda^2 - 1) - 1(-\lambda - 0) + 0(-1 + \lambda) = 0$$

$$\Rightarrow -\lambda^3 + \lambda + \lambda = 0$$

$$\Rightarrow \lambda^3 - 2\lambda = 0 \Rightarrow \lambda(\lambda^2 - 2) = 0$$

$$\lambda = 0, +\sqrt{2}, -\sqrt{2}$$

Find all three eigen vectors and any combination of $\langle v_1, v_2 \rangle$ must be equal to 0.

$\lambda = 0$, $\lambda = -\sqrt{2} < 0$ & $\lambda = +\sqrt{2} > 0 \rightarrow$ Indefinite

2. Suppose, $\boldsymbol{u}$ and $\mathbf{v}$ are two orthonormal eigenvectors of a matrix $\boldsymbol{S}$ (i.e, $\|\boldsymbol{u}\| = \|\mathbf{v}\| = 1$ and $\langle \boldsymbol{u}, \mathbf{v} \rangle = 0$). The eigenvalues corresponding to $\boldsymbol{u}, \mathbf{v}$ are $3, 4$ respectively. Consider a vector $\mathbf{w}$ defined as -

$$\mathbf{w} = 2\boldsymbol{u} + 3\mathbf{v}$$

What is the value of $\|\boldsymbol{S}\mathbf{w}\|$?

$\rightarrow$

$u = 3$ $\quad v = 4$ $\quad$ so, $Su = 3u$ $\quad Sv = 4v$

$w = 2u + 3v$

$$\|Sw\|^2 = \|2Su + 3Sv\|$$

$$= \|2 \times (3u) + 3 \times (4v)\|$$

$$= \|6u + 12v\|$$

$$= 36 + 144 = 180$$

$$\therefore \|Sw\| = \sqrt{180} = 6\sqrt{5}$$

*** PSet 2**

1. Find the critical points of the following functions, and, classify them (local minima, maxima, or saddle point).

   (a) $f(x) = x^3 - 3x^2 + 3x$

   (b) $f(x_1, x_2) = x_1^2 + 2x_1x_2 + 3x_2^3$

   (c) $f(x_1, x_2) = x_1^4 + 2x_1^2x_2 + x_2^2 - 4x_1^2 - 8(x_1 + x_2)$

   (d) $f(x_1, x_2) = x_1^2 + 4x_1x_2 + x_2^2 + x_1 - x_2$

   (e) $f(x_1, x_2, x_3) = x_1x_2 + x_2x_3 + x_3x_1$

(a) $f(x) = x^3 - 3x^2 + 3x$ $f'(x) \Rightarrow 3x^2 - 6x + 3 = 0$

$\therefore x = 1, 1$

$f''(x) = 6x - 6$ & $f''(1) = 6\times 6 - 6 = 0 \rightarrow$ Further

$f'''(1) = 6 > 0 \rightarrow \neq 0$ so inflection point

$\therefore$ 1 is a critical inflection point.

(b) $f(x_1, x_2) = x_1^2 + 2x_1x_2 + 3x_2^3$

$f_{x_1x_1} = \frac{\delta f}{\delta x_1} \Rightarrow 2x_1 + 2x_2 = 0$ — (I)

$f_{x_2x_2} = \frac{\delta f}{\delta x_2} \Rightarrow 0 + 2x_1 + 9x_2^2 = 0$ — (II)

$x_1^* = 0$ & $x_2^* = 2/9$

$$\mathcal{H} = \begin{pmatrix} f_{x_1x_1} & f_{x_1x_2} \\ f_{x_2x_1} & f_{x_2x_2} \end{pmatrix} = \begin{pmatrix} 2 & 2 \\ 2 & 18x_2 \end{pmatrix} = \begin{pmatrix} 2 & 2 \\ 2 & 0 \end{pmatrix}$$

$$\det\begin{pmatrix} 2-\lambda & 2 \\ 2 & -\lambda \end{pmatrix} = 0; \quad \{-2\lambda + \lambda^2 - 4\} = 0$$

$\Rightarrow \lambda^2 - 2\lambda - 4 = 0$

$\Rightarrow \lambda = 1 \pm \sqrt{5}$

(e) $f(x_1, x_2, x_3) = x_1x_2 + x_2x_3 + x_3x_1$

$f_{x_1} = x_2 + 0 + x_3 = 0 + x_2 + x_3$

$f_{x_2} = x_1 + 0 + x_3$

$f_{x_3} = 0 + x_2 + 0$

$x_1 = 0, x_2 = 0, x_3 = 0$

$$H = \begin{pmatrix} f_{x_1x_1} & f_{x_1x_2} & f_{x_1x_3} \\ f_{x_2x_1} & f_{x_2x_2} & f_{x_2x_3} \\ f_{x_3x_1} & f_{x_3}f_{x_2} & f_{x_3x_3} \end{pmatrix} = \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix}$$

$\lambda_1 = -1, \lambda_2 = -1, \lambda_3 = 2 \rightarrow$ Mixed = Saddle point

3. Consider the function $f(x_1, x_2) = x_1^2 + 2x_1x_2 + 3x_2^3$. What is the directional derivative of this function along the unit vector $\frac{1}{\sqrt{2}}\begin{bmatrix}1\\1\end{bmatrix}$ at the point (3, 4) ?

Given, $f(x_1, x_2) = x_1^2 + 2x_1x_2 + 3x_2^2$

$$\nabla f = \begin{pmatrix} f_{x_1} \\ f_{x_2} \end{pmatrix} = \begin{pmatrix} 2x_1 + 2x_2 + 0 \\ 0 + x_1 + 9x_2^2 \end{pmatrix} = \begin{pmatrix} 2x_1 + 2x_2 \\ x_1 + 9x_2^2 \end{pmatrix}$$

$$= \begin{pmatrix} 6+8 \\ 3 + 9\times 16 \end{pmatrix} = \begin{pmatrix} 14 \\ 150 \end{pmatrix}$$

Direction Vector : $\nabla f \cdot u = \begin{pmatrix} 14 \\ 150 \end{pmatrix} \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \end{pmatrix}$

4. Consider the following function:

$$f(x_1, x_2) = x_1^2 - 6x_1x_2 + 8x_2^2 + 3x_1$$

This function can be written in the following form:

$$f(\mathbf{x}) = \mathbf{x}^\top \boldsymbol{A}\mathbf{x} + \mathbf{b}^\top \mathbf{x} + c$$

for some matrix $\boldsymbol{A}$, vector $\boldsymbol{b}$ and constant c. Here, $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$.

(a) What is $\boldsymbol{A}, \mathbf{b}, c$?

(b) Find the critical point of f.

(c) Find the Hessian $(\nabla^2 f)$ at the critical point.

(d) Determine whether the critical point is a minima, maxima or saddle point.

a) $f(x_1, x_2) = x_1^2 - 6x_1x_2 + 8x_2^2 + 3x_1$

$f(x) = x^T A x + b^T x + c$

$$A = \begin{pmatrix} 1 & -6/2 \\ -6/2 & 8 \end{pmatrix} = \begin{pmatrix} 1 & -3 \\ -3 & 8 \end{pmatrix}$$

$$b = \begin{pmatrix} 3 \\ 0 \end{pmatrix} \qquad c = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

* $f(\bar{x}) = b^T \bar{x}$

$\nabla f = \bar{b} \rightarrow$ Gradient

$\nabla^2 f = \bar{0} \rightarrow$ Hicksian

Pset3

1. For calculating numerical gradients, the formulas for forward difference, central difference and complex difference are -

$$[\tilde{f}'(x_0)]_{\text{forward}} = \frac{f(x_0+h)-f(x_0)}{h}$$
$$[\tilde{f}'(x_0)]_{\text{central}} = \frac{f(x_0+h)-f(x_0-h)}{2h}$$
$$[\tilde{f}'(x_0)]_{\text{complex}} = \frac{\mathsf{Im}(f(x_0+ih))}{h}$$

for some finite step-size h. For multi-variable functions, these formulas may be applied to each variable seperately while keeping the others fixed.

For the following functions, calculate the values of these three numerical gradients at $x = 1$ for a step size of $h = 0.001$.
Also, find the actual value of the derivative at $x = 1$, and calculate the absolute value of the error $\left|\widetilde{f}'(x_0) - f(x_0)\right|$. Write which formula gave the least amount of error.

(a) $f(x) = x^2$

(b) $f(x) = e^x \cdot x^3$

(c) $f(x) = x\ln(x)$ $\rightarrow$ $a + ib = re^{i\theta}$

2. For the following multi-variable functions, calculate the numerical directional directives in the direction $\mathbf{u} = \frac{1}{\sqrt{2}}\begin{bmatrix}1\\1\end{bmatrix}$ at the point $(x_1, x_2) = (2, 3)$, using the forward, central and complex difference formulas. Use a step-size of $h = 0.001$.

(a) $f(x_1, x_2) = x_1 \ln(x_2)$

(b) $f(x_1, x_2) = x_1^2/x_2$

(c) $f(x_1, x_2) = x_1^2 + 2x_1x_2 + 3x_2^3$

For the above functions, find the actual value of the directional derivatives at $(x_1, x_2) = (2, 3)$, and calculate the absolute error $\left|\tilde{f}_{\mathbf{u}} - f_{\mathbf{u}}\right|$. Write which formula gave the least amount of error.

(a) $f(x_1,x_2) = x_1 \ln(x_2)$ & $(x_1,x_2) = (2,3)$ $h = 0.001$ $u = \begin{pmatrix}1/\sqrt{2}\\1/\sqrt{2}\end{pmatrix}$

$$\Rightarrow f'(x_1,x_2) = \frac{f\left(\begin{pmatrix}x_1\\x_2\end{pmatrix} + \begin{pmatrix}1/\sqrt{2}\\1/\sqrt{2}\end{pmatrix}\right) - f\begin{pmatrix}x_1\\x_2\end{pmatrix}}{h}$$

$$= \frac{f(2+1/\sqrt{2},\ 3+1/\sqrt{2}) - f(2,3)}{0.001}$$

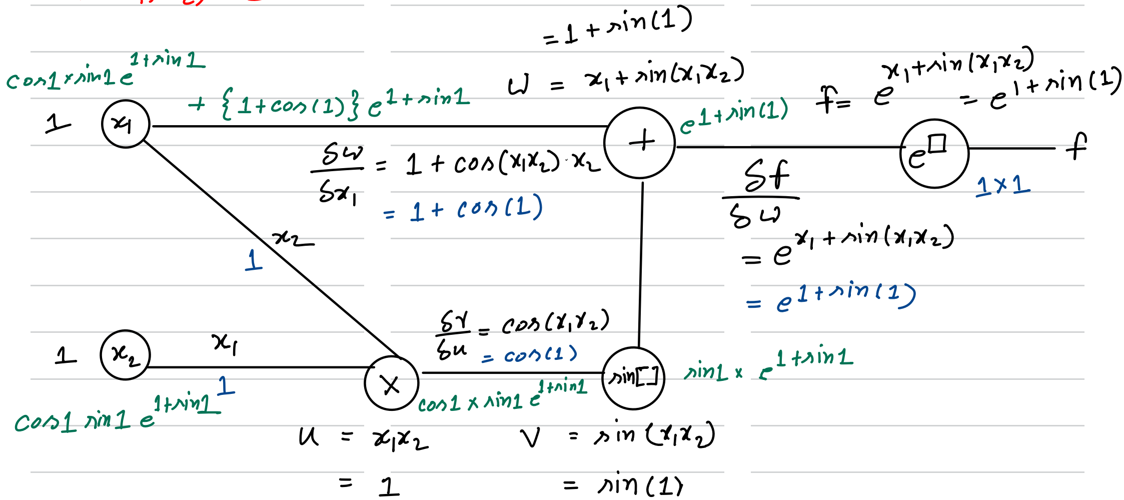
$$= -3099.68$$

(b) $f(x_1, x_2) = x_1^2 / x_2$

$$f'(x_1, x_2) = \frac{f\left(\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + h\begin{pmatrix} u_1 \\ u_2 \end{pmatrix}\right) - f\left(\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} - h\begin{pmatrix} u_1 \\ u_2 \end{pmatrix}\right)}{2h}$$

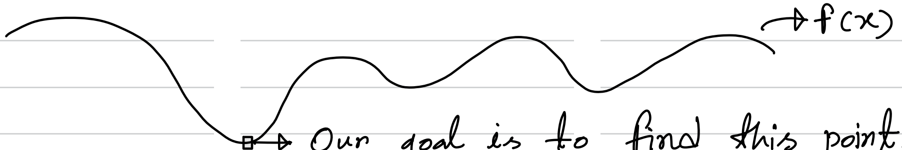
FWAD Ex:

$f(x_1, x_2) = e^{x_1 + \sin(x_1 x_2)}$



Gradient Descent

→ General solutions don't work for more than degree 4 polynomial.



→ Our goal is to find this point.

The steps we could take are:

1. We calculate gradient.
2. We go opposite of it.

$\Big\}$ Vanilla gradient descent

Function → $f(x)$, $\bar{x}^{(0)}$

$\bar{g}^{(0)} = \nabla f \big|_{\bar{x}_0}$; $\bar{x}^{(1)} = \bar{x}^{(0)} - \boxed{\alpha \cdot \bar{g}^{(0)}}$

↳ Step size

α → Step factor / Learning rate

$\bar{g}^{(1)} = \nabla f \big|_{\bar{x}_1}$; $\bar{x}_2 = \bar{x}_1 - \alpha \bar{g}_1$

→ We take longer steps when we're far from the gradient & shorter steps when we're near the gradient.

Step factor selection is an important factor & the lower the better but it'll take a lot of time but irl we don't have that much time, but; we can't just take a larger step size because it might overshoot.

We can use learning rate scheduling in V.G.D

$$\alpha^{(i+1)} = 0.9 \cdot \alpha^{(i)}$$

* AdaGrad → Aaptive Gradient

$$\overline{x}^{(i+1)} = \overline{x}^{(i)} - \alpha \overline{g}^{(i)}$$

$$\begin{pmatrix} x_1^{(i+1)} \\ x_2^{(i+1)} \\ \vdots \\ x_n^{(i+1)} \end{pmatrix} = \begin{pmatrix} x_1^{i} & - & \alpha_1 g_1^{i} \\ x_2^{i} & - & \alpha_2 g_2^{i} \\ \vdots & & \vdots \\ x_n^{(i)} & - & \alpha_3 g_n^{i} \end{pmatrix}$$

In adagrad each component has its own step factor. because we basically we want the larger variable to reduce faster.

→ We take one extra variable for each step

$$S_1 = g_1^{(0)2} + g_1^{(1)2} + \dots + g_1^{(k)2} \quad , \quad S_2 = g_2^{(0)2} + g_2^{(1)2} + \dots + g_2^{(2)2}$$

$$\alpha_1 = \frac{c}{\sqrt{S_1^{(i)}}}$$

Algorithm

$$S_k^{(i+1)} = S_k^{(i)} + g_k^{(i)2}$$

$g_k^{(i)} \rightarrow$

$$\alpha_k^{(i+1)} = \frac{c}{\sqrt{S_k^{(i+1)}}}$$

$$x_k^{(i+1)} = x_k^{(i)} - \alpha_k^{(i+1)} g_k^{(i)}$$

$$\therefore x_k^{(i+1)} = x_k^{(i)} - \frac{c}{\sqrt{S_k^{(i)} + g_k^{(i+1)}}} \times g_k^{(i)}$$

But $S_k^{(i+1)} = S_k^{(i)} + g_k^{(i)2}$ is monotonically increasing that can lead to problems & to address that issue, we have RMS Prop. ↳ Early Stopping

* RMS Prop

Update rule: $s_k^{(i+1)} = \gamma \cdot s_k^{(i)} + (1-\gamma)\, g_k^{(i)2}$

$0 \leq \gamma \leq 1$

→ Only this part changes.