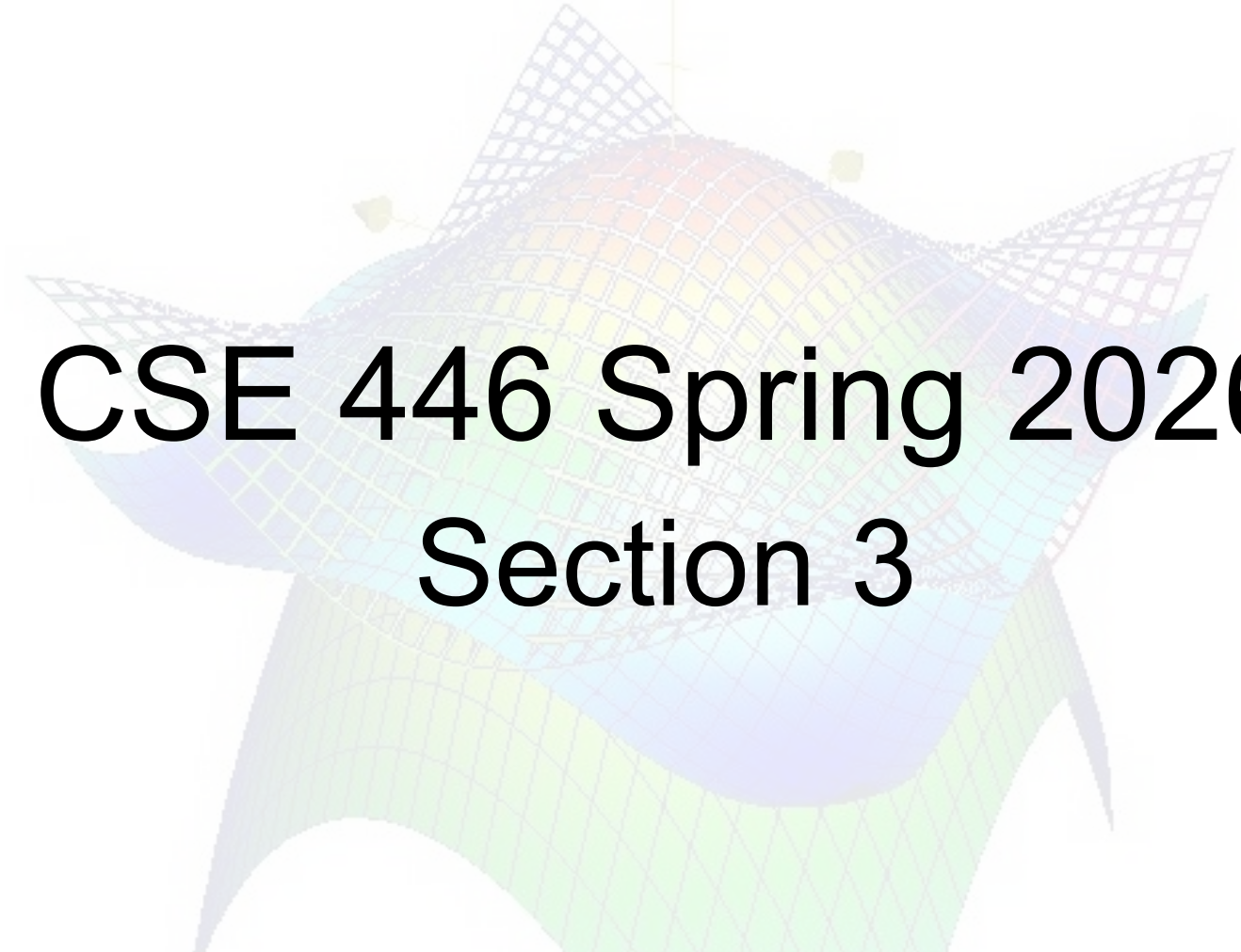


CSE 446 Spring 2026
Section 3

Announcements

HW1 is due next Wednesday, April 22

Midterm May 1 (details on class website)

Remember to email your section handout attempt to your TA if you missed a section

Today

1. Vector calculus
2. Approximations

Scalar derivative			Vector derivative		
$f(x)$	$\rightarrow$	$\frac{\mathrm{d}f}{\mathrm{d}x}$	$f(\mathbf{x})$	$\rightarrow$	$\frac{\mathrm{d}f}{\mathrm{d}\mathbf{x}}$
bx	$\rightarrow$	b	$\mathbf{x}^T\mathbf{B}$	$\rightarrow$	$\mathbf{B}$
bx	$\rightarrow$	b	$\mathbf{x}^T\mathbf{b}$	$\rightarrow$	$\mathbf{b}$
x^2	$\rightarrow$	$2x$	$\mathbf{x}^T\mathbf{x}$	$\rightarrow$	$2\mathbf{x}$
bx^2	$\rightarrow$	$2bx$	$\mathbf{x}^T\mathbf{B}\mathbf{x}$	$\rightarrow$	$2\mathbf{B}\mathbf{x}$

Derivatives: how small change in input changes output

1st derivative, second derivative (1D)

$$\frac{dy}{dx} = 2x - 1$$

$$\frac{d^2y}{dx^2} = 2$$

1. **Extend** to more variables
2. **Approximate** the gradients

Jacobian, Hessian (generalize to n input and m output variables)

$$\mathbf{J_f}(x, y) = \begin{bmatrix} \frac{\partial f_1}{\partial x} & \frac{\partial f_1}{\partial y} \\ \frac{\partial f_2}{\partial x} & \frac{\partial f_2}{\partial y} \end{bmatrix}$$

$$H_x = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} \\ \frac{\partial^2 f}{\partial x_1 x_2} & \frac{\partial^2 f}{\partial x_2^2} \end{bmatrix}$$

Gradient

How do we calculate the gradient of a function with a vector input?

$$x = \begin{pmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{pmatrix}$$

$$f(x) : \mathbb{R}^n \rightarrow \mathbb{R}$$

This is normal function that outputs a real number

Slides by Marco D.

We simply do partial derivatives n times

$$\nabla_x f(x) = \begin{bmatrix} \frac{\partial f(x)}{\partial x_1} \\ \frac{\partial f(x)}{\partial x_2} \\ \vdots \\ \frac{\partial f(x)}{\partial x_n} \end{bmatrix}$$

$$x = \begin{pmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{pmatrix}$$

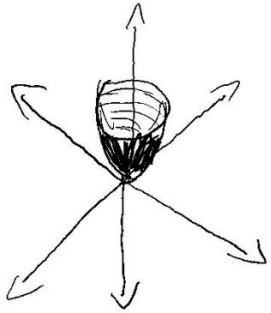
Slides by Marco D.

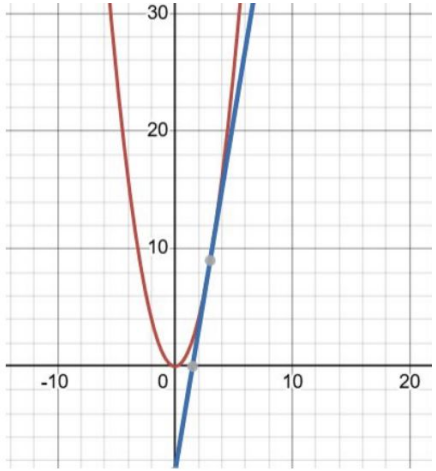
Visualizing the Gradient

$f : \mathbb{R}^n \rightarrow \mathbb{R}$ ← i.e: scalar value

Example: $z = x_1^2 + x_2^2$:

can also be thought of as x and y

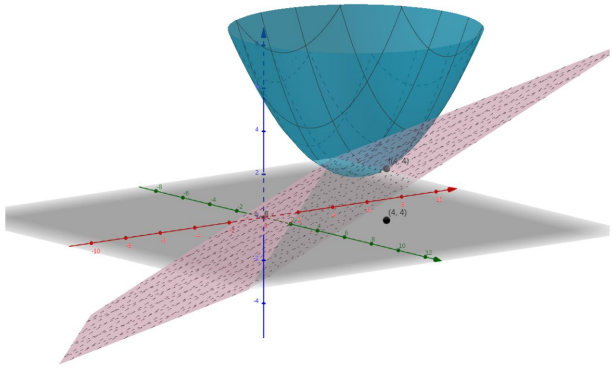




$$\nabla_x f(x) = \begin{bmatrix} \frac{\partial f(x)}{\partial x_1} \\ \frac{\partial f(x)}{\partial x_2} \end{bmatrix} = \begin{bmatrix} 2x_1 \\ 2x_2 \end{bmatrix}$$

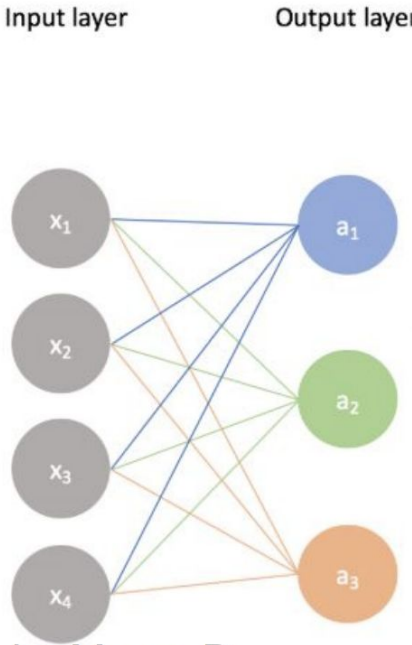
all x's

Tells you the slope of the tangent plane in the x_1 and x_2 directions



Jacobian

In machine learning, we don't usually have the privilege of having a function that outputs a real number. Usually, the function will output a vector. For example:



A simple neural network

$$\begin{bmatrix} w_1 & w_2 & w_3 & w_4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} + \begin{bmatrix} b \end{bmatrix} = \begin{bmatrix} w_1x_1 + w_2x_2 + w_3x_3 + w_4x_4 + b \end{bmatrix} \xrightarrow{activation} \begin{bmatrix} a_1 \end{bmatrix}$$

What do we do now???

Slides by Marco D.

$$g : \mathbb{R}^n \rightarrow \mathbb{R}^m$$

Example: $g(x) = Wx$

where: $W \in \mathbb{R}^{m \times n}$, $x \in \mathbb{R}^n$

$$m\begin{bmatrix} W \\ \\ \\ n \end{bmatrix} \quad n\begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}_1 = m\begin{bmatrix} \\ \\ \\ 1 \end{bmatrix}$$

Now what is $\nabla_x g(x)$?

$$m\begin{bmatrix} W \\ \\ \\ n \end{bmatrix} \quad n\begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = m\begin{bmatrix} w_1 \cdot x \\ w_2 \cdot x \\ \vdots \\ w_m \cdot x \end{bmatrix}_1$$

(dot product)

$$\nabla_x g(x) = m\begin{bmatrix} \nabla_x w_1 \cdot x \\ \nabla_x w_2 \cdot x \\ \vdots \\ \nabla_x w_m \cdot x \end{bmatrix}_1 \quad \nabla_x (w_1^1 x_1 + w_1^2 x_2 \dots)$$

Gradient!

$$= \begin{bmatrix} w_1 x_1 + w_2 x_2 + w_3 x_3 + w_4 x_4 + b \\ \\ \end{bmatrix} \xrightarrow[\text{activation}]{} \begin{bmatrix} a_1 \\ \\ \end{bmatrix}$$

$$\nabla_x g(x) = \begin{bmatrix} \nabla_x w_1 \cdot x \\ \nabla_x w_2 \cdot x \\ \vdots \\ \nabla_x w_m \cdot x \end{bmatrix}_{m \times 1}$$

$\nabla_x (w_1^1 x_1 + w_1^2 x_2 \ldots)$

Gradient!

Each row becomes an *n* element gradient

(For notational purposes, g_i = computation for row *i* of the output)

$$\begin{bmatrix} \frac{\partial g_1(x)}{\partial x_1} & \cdots & \frac{\partial g_1(x)}{\partial x_n} \\ \vdots & & \\ \frac{\partial g_m(x)}{\partial x_1} & \cdots & \frac{\partial g_m(x)}{\partial x_n} \end{bmatrix}_{m \times n}$$

$g_1(x) = w_1^1 x_1 + w_1^2 x_2 \ldots$

$g_2(x) = w_2^1 x_1 + w_2^2 x_2 \ldots$

$\vdots$

$g_m(x) = w_m^1 x_1 + w_m^2 x_2 \ldots$

This is the Jacobian of $g(x)$

Interpreting the Jacobian

How do we interpret the jacobian matrix?

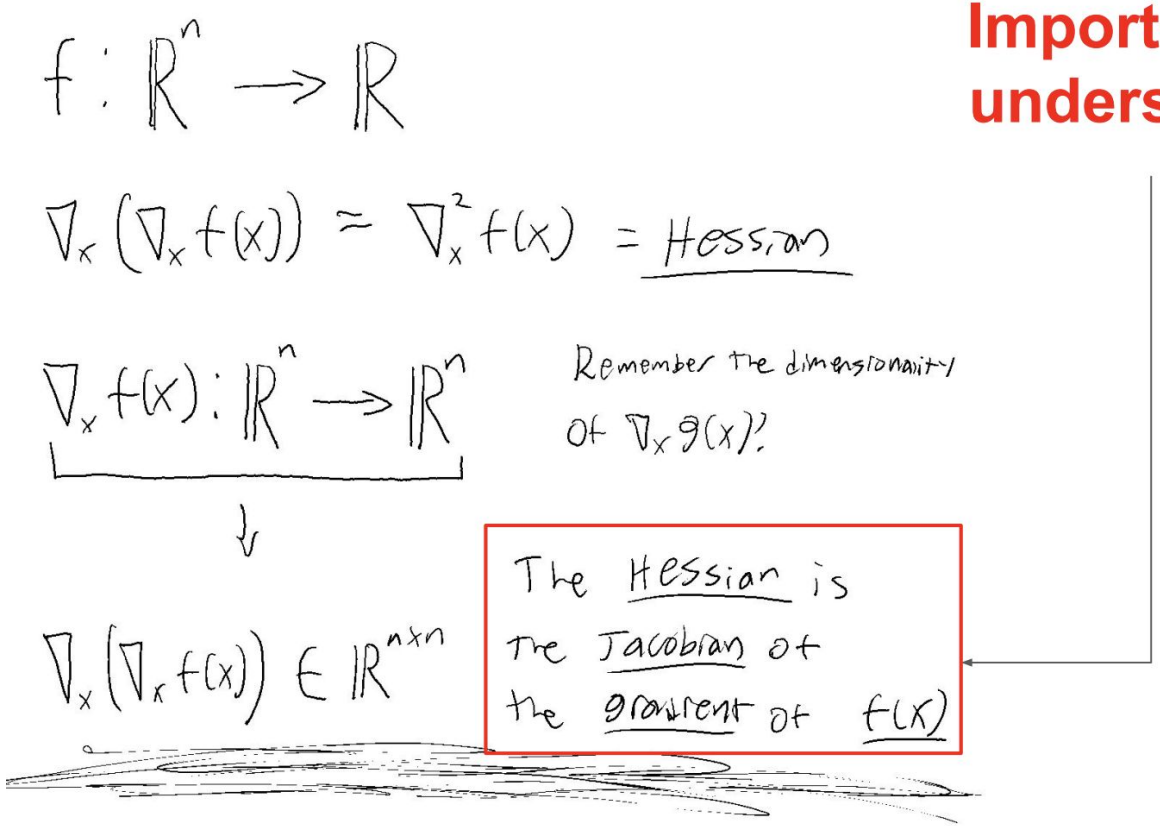
$$\nabla_x g(x) = \begin{bmatrix} \frac{\partial g_1(x)}{\partial x_1} & \cdots & \frac{\partial g_1(x)}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial g_m(x)}{\partial x_1} & \cdots & \frac{\partial g_m(x)}{\partial x_n} \end{bmatrix} = \begin{bmatrix} \nabla_x^T g_1(x) \\ \vdots \\ \nabla_x^T g_m(x) \end{bmatrix} \in \mathbb{R}^{m \times n}$$

This matrix gives us an idea of how the output will change if we slightly change the value of x.

For example, if we increase x_1, how is g(x) affected?

Slides by Marco D.

Hessian



Important to understand!

$$\nabla_x^2 f(x) = \begin{bmatrix} \frac{\partial^2 f(x)}{\partial x_1^2} & \frac{\partial^2 f(x)}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f(x)}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f(x)}{\partial x_2 \partial x_1} & \frac{\partial^2 f(x)}{\partial x_2^2} & \cdots & \frac{\partial^2 f(x)}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f(x)}{\partial x_n \partial x_1} & \frac{\partial^2 f(x)}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f(x)}{\partial x_n^2} \end{bmatrix} \in \mathbb{R}^{n \times n}$$

Notice the combinations of variables:

- Derive by the same variable twice for the diagonal
- Derive by every combination of x_i,x_j where $i \neq j$ for the off-diagonals

$$\nabla_x^2 f(x) = \begin{bmatrix} \frac{\partial^2 f(x)}{\partial x_1^2} & \frac{\partial^2 f(x)}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f(x)}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f(x)}{\partial x_2 \partial x_1} & \frac{\partial^2 f(x)}{\partial x_2^2} & \cdots & \frac{\partial^2 f(x)}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f(x)}{\partial x_n \partial x_1} & \frac{\partial^2 f(x)}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f(x)}{\partial x_n^2} \end{bmatrix} \in \mathbb{R}^{n \times n}$$

$$f(X) = X_1^2 + X_2^2$$

$$\text{Gradient: } \nabla_x f(X) = \begin{bmatrix} 2X_1 \\ 2X_2 \end{bmatrix} \rightarrow \begin{array}{c|c} \frac{\partial}{\partial X_1}(2X_1) = 2 & \frac{\partial}{\partial X_2}(2X_1) = 0 \\ \hline \frac{\partial}{\partial X_1}(2X_2) = 0 & \frac{\partial}{\partial X_2}(2X_2) = 2 \end{array}$$

$$\text{Hessian: } \nabla_x^2 f(x) = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$$

Problems 1.2 a-b

Solve them! Ask for help if you are stuck. Look at section 1.1 for help remembering how these gradients, Jacobians, and Hessians compute.

(a) Let $f(x_1, x_2) = x_1^2 + e^{x_1 x_2} + 2\log(x_2)$. What are the gradient and the Hessian of f?

(b) Note that $\nabla_x f : \mathbb{R}^n \to \mathbb{R}^n$. What is the Jacobian of $\nabla_x f$?

(a) Let $f(x_1, x_2) = x_1^2 + e^{x_1 x_2} + 2\log(x_2)$. What are the gradient and the Hessian of f?

Solution:

$$\nabla_x f(x) = \begin{bmatrix} \frac{\partial f(x)}{\partial x_1} \\ \frac{\partial f(x)}{\partial x_2} \end{bmatrix} = \begin{bmatrix} 2x_1 + x_2 e^{x_1 x_2} \\ x_1 e^{x_1 x_2} + \frac{2}{x_2} \end{bmatrix} \text{ and } \nabla_x^2 f(x) = \begin{bmatrix} \frac{\partial^2 f(x)}{\partial x_1^2} & \frac{\partial^2 f(x)}{\partial x_1 \partial x_2} \\ \frac{\partial^2 f(x)}{\partial x_2 \partial x_1} & \frac{\partial^2 f(x)}{\partial x_2^2} \end{bmatrix} = \begin{bmatrix} 2 + x_2^2 e^{x_1 x_2} & e^{x_1 x_2} + x_1 x_2 e^{x_1 x_2} \\ e^{x_1 x_2} + x_1 x_2 e^{x_1 x_2} & x_1^2 e^{x_1 x_2} - \frac{2}{x_2^2} \end{bmatrix}$$

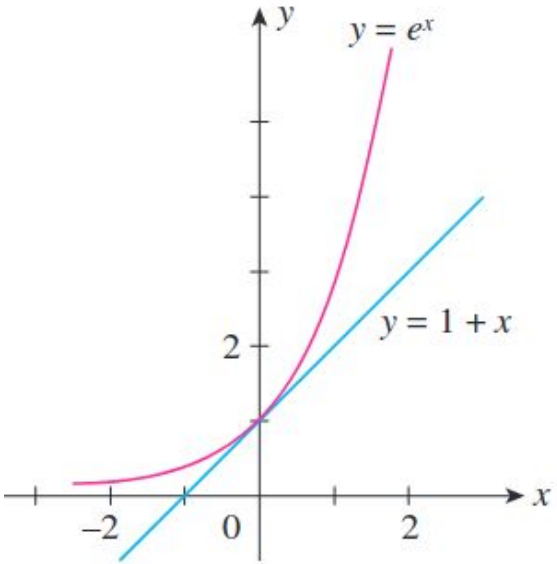
(b) Note that $\nabla_x f : \mathbb{R}^n \to \mathbb{R}^n$. What is the Jacobian of $\nabla_x f$?

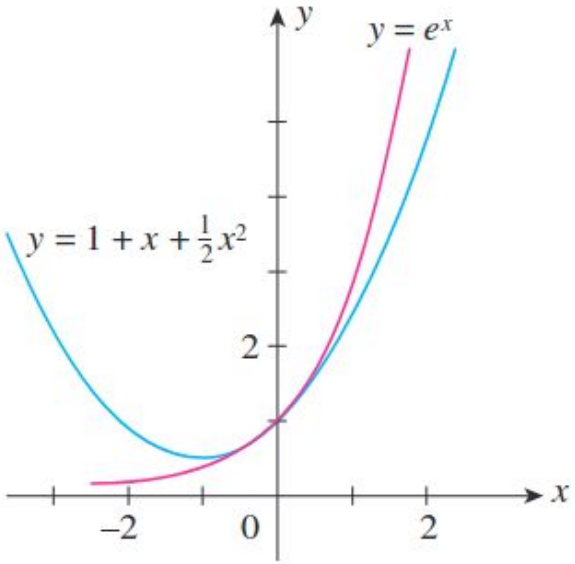
Equivalent

Solution:

$$\nabla_x (\nabla_x f)(x) = \begin{bmatrix} \frac{\partial (\nabla_x f)_1(x)}{\partial x_1} & \frac{\partial (\nabla_x f)_1(x)}{\partial x_2} \\ \frac{\partial (\nabla_x f)_2(x)}{\partial x_1} & \frac{\partial (\nabla_x f)_2(x)}{\partial x_2} \end{bmatrix} = \begin{bmatrix} 2 + x_2^2 e^{x_1 x_2} & e^{x_1 x_2} + x_1 x_2 e^{x_1 x_2} \\ e^{x_1 x_2} + x_1 x_2 e^{x_1 x_2} & x_1^2 e^{x_1 x_2} - \frac{2}{x_2^2} \end{bmatrix} = \nabla_x^2 f(x)$$

Approximations





$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \dots$$

Taylor Series

$$f(x) = f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \frac{f^{(3)}(a)}{3!}(x-a)^3 + \dots + \frac{f^{(n)}(a)}{n!}(x-a)^n + \dots$$

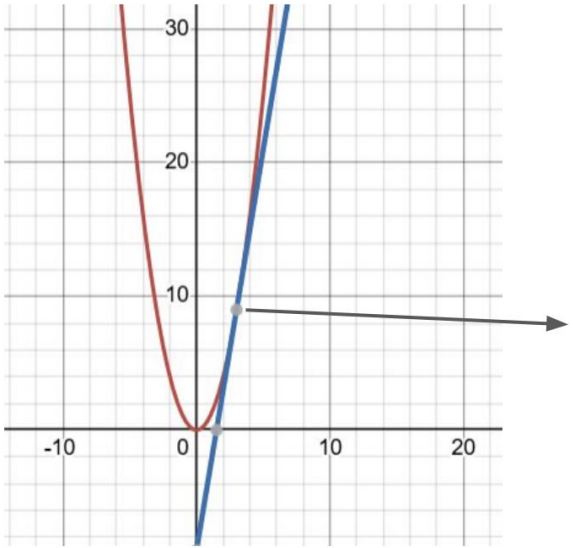
Approximations

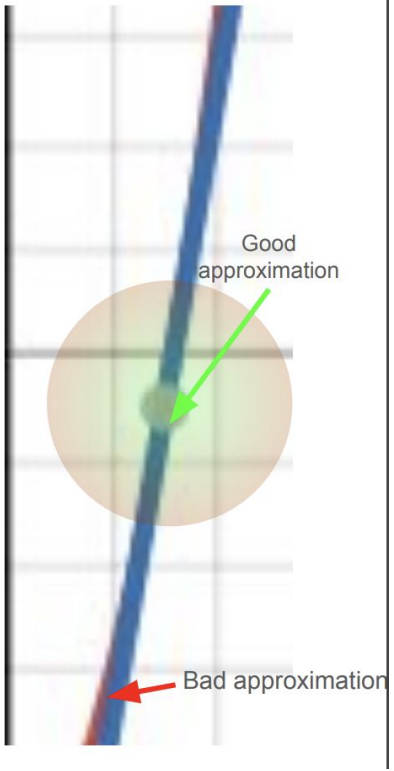
Linear Approximation

The derivative of **f(x)** at some **(x,y)** can be used to linearly approximate **f(x ± ε)**

Where ε is very tiny!

This extends to multivariate functions… proof in your notes





Approximation

$$f : \mathbb{R} \to \mathbb{R}$$

$$\frac{df}{dx}(x) = \lim_{\epsilon \to 0} \frac{f(x+\epsilon) - f(x)}{\epsilon} \Leftrightarrow \frac{df}{dx}(x) \approx \frac{f(x+\epsilon) - f(x)}{\epsilon} \Leftrightarrow f(x+\epsilon) \approx f(x) + \epsilon \frac{df}{dx}(x)$$

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

$$f(x_1 + \epsilon_1, \ldots, x_n) \approx f(x_1, \ldots, x_n) + \epsilon_1 \frac{\partial f}{\partial x_1}(x_1, \ldots, x_n)$$

$$f(x_1 + \epsilon_1, x_2 + \epsilon_2, \ldots, x_n) \approx f(x_1, x_2 + \epsilon_2, \ldots, x_n) + \epsilon_1 \frac{\partial f}{\partial x_1}(x_1, x_2 + \epsilon_2, \ldots, x_n)$$

$$\approx f(x_1, x_2, \ldots, x_n) + \epsilon_2 \frac{\partial f}{\partial x_2}(x_1, x_2, \ldots, x_n) +$$

$$+ \epsilon_1 \frac{\partial f}{\partial x_1}(x_1, x_2, \ldots, x_n) + \epsilon_1 \epsilon_2 \frac{\partial f}{\partial x_2} \frac{\partial f}{\partial x_1}(x_1, x_2, \ldots, x_n)$$

$$\approx f(x_1, x_2, \ldots, x_n) + \epsilon_1 \frac{\partial f}{\partial x_1}(x_1, x_2, \ldots, x_n) + \epsilon_2 \frac{\partial f}{\partial x_2}(x_1, x_2, \ldots, x_n)$$

$$\approx f(x_1, x_2, \ldots, x_n) + \epsilon_1 \frac{\partial f}{\partial x_1}(x_1, x_2, \ldots, x_n) + \epsilon_2 \frac{\partial f}{\partial x_2}(x_1, x_2, \ldots, x_n)$$

Linear Approximation

For a “many-to-one” function, the gradient gives us a vector we can use to linearly approximate a small area around some **x**

$$f : \mathbb{R}^n \rightarrow \mathbb{R}$$

$$\text{Let } \epsilon = [\epsilon_1, \ldots, \epsilon_n]^T \text{ and } x = [x_1, \ldots, x_n]^T$$

$$f(x + \epsilon) \approx f(x) + \nabla_x f(x)^T \epsilon$$

What about a “many-to-many” function?

$$g : \mathbb{R}^n \rightarrow \mathbb{R}^m$$

Problem 1.2 c

Remember that the Jacobian is just the gradient of a "many-to-many" function.

$$\begin{bmatrix} \nabla_x^T g_1(x) \\ \vdots \\ \nabla_x^T g_m(x) \end{bmatrix} \in \mathbb{R}^{m \times n}$$

Also remember: *For a "many-to-one" function, the gradient gives us a vector we can use to linearly approximate a small area around some* ***x***

(c) The gradient $\nabla_x f(x)$ offers the best linear approximation of f around the point x. What does the Jacobian of a function $g : \mathbb{R}^n \to \mathbb{R}^m$ offer?

(c) The gradient $\nabla_x f(x)$ offers the best linear approximation of f around the point x. What does the Jacobian of a function $g : \mathbb{R}^n \to \mathbb{R}^m$ offer?

Solution:

The Jacobian also offers the best linear approximation of g around a point x, but now it approximates a vector, instead of a scalar,

$$g(x + \epsilon) \approx g(x) + \nabla_x g(x)\epsilon$$

where $\nabla_x g(x)\epsilon$ is a matrix multiplication instead of a dot product.

Problem 1.2 d (walkthrough)

(d) If we use the gradient and the Hessian of $f : \mathbb{R}^n \to \mathbb{R}$, what type of an approximation for the function f around a point x can we create.

Remember Taylor expansion?

↳ To approximate a function around a point a

$$f(x) \approx f(a) + \frac{f'(a)}{1!}(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \frac{f'''(a)}{3!}(x-a)^3 \cdots$$

Exact at a, close around a

Better and better approximations

Remember Taylor expansion?

↳ To approximate a function around a point a

$$f(x) \simeq f(a) + \frac{f'(a)}{1!}(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \frac{f'''(a)}{3!}(x-a)^3 \cdots$$

↑ Exact at a, close around a

Better and better approximations

Set $a = x$, we want to estimate $x + \epsilon$

$$f(x+\epsilon) \simeq f(x) + \frac{f'(x)}{1!}(x+\epsilon-x) + \frac{f''(x)}{2!}(x+\epsilon-x)^2 + \cdots$$

↓

$$f(x+\epsilon) \simeq f(x) + f'(x)\epsilon + \frac{1}{2}f''(x)\epsilon^2 + \cdots$$

Generalizing to vectors: $f: \mathbb{R}^n \rightarrow \mathbb{R}$, $x \in \mathbb{R}^n$, $\epsilon \in \mathbb{R}^n$

$$f(x+\epsilon) \approx f(x) + (\nabla_x f(x))^T \epsilon$$

gradient = first order derivative of $f(x)$

So what is the second order derivative?

Second order derivative = $\nabla_x (\nabla_x f(x))$ = Hessian

↳ gives us a Quadratic Approximation

2nd order Taylor expansion around **x** generalized to vectors

$$f(x+\epsilon) \approx f(x) + (\nabla_x f(x))^T \epsilon + \frac{1}{2} \epsilon^T (\nabla_x^2 f(x))^T \epsilon$$

Answer!

Problem 1.2 g (IMPORTANT!)

(g) Draw the gradient on the picture. Describe what happens to the values of the approximation of f if we move from x in directions d_1, d_2, d_3 for which $\nabla_x f(x)^T d_1 > 0, \nabla_x f(x)^T d_2 < 0, \nabla_x f(x)^T d_3 = 0$? Can the same conclusions be drawn about the function of f?

$(\nabla_x f(x))^T d_1 > 0$

↳ Direction d_1 points generally toward the gradient

$(\nabla_x f(x))^T d_2 < 0$

↳ Direction d_2 points generally away from the gradient

$(\nabla_x f(x))^T d_3 = 0$

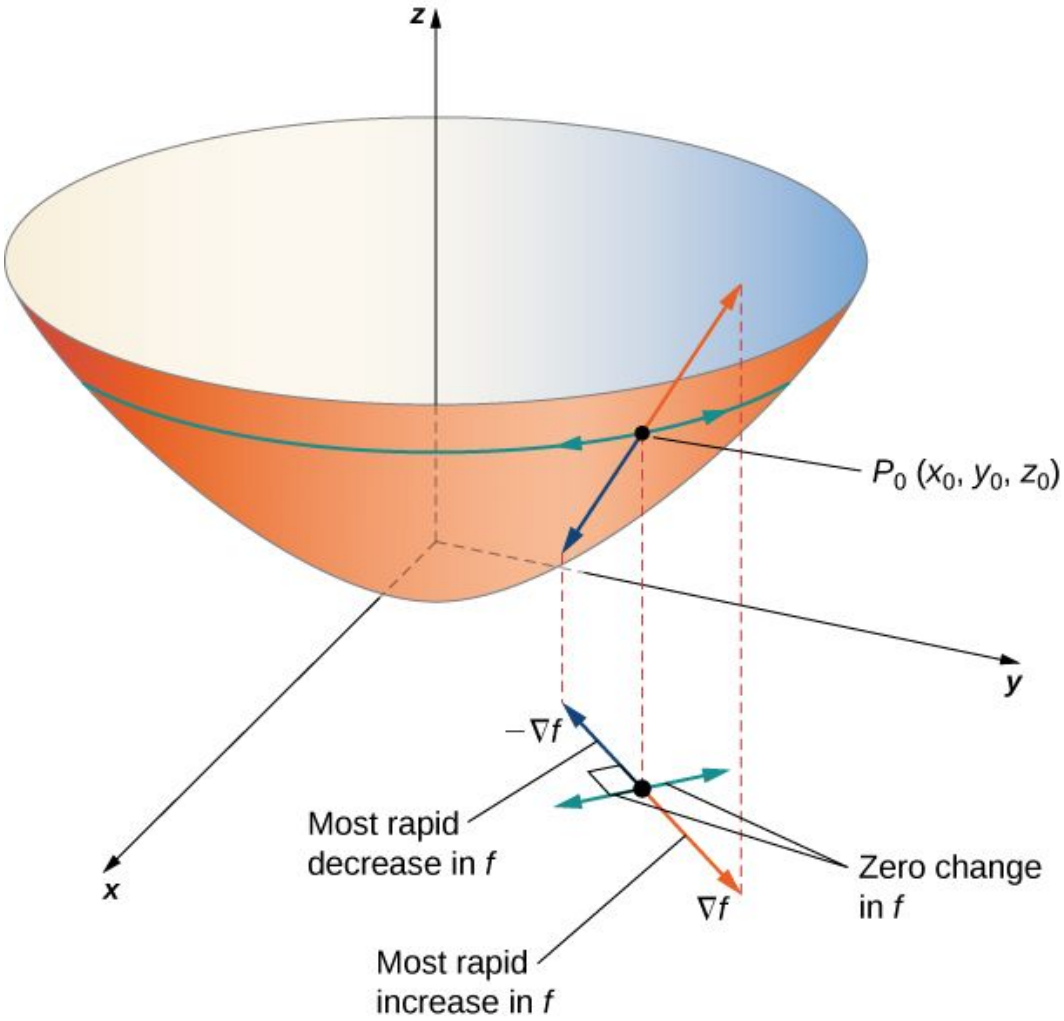
↳ Direction d_3 points orthogonal to the gradient

(g) Draw the gradient on the picture. Describe what happens to the values of the approximation of f if we move from x in directions d_1, d_2, d_3 for which $\nabla_x f(x)^T d_1 > 0, \nabla_x f(x)^T d_2 < 0, \nabla_x f(x)^T d_3 = 0$? Can the same conclusions be drawn about the function of f?

Solution:

- d_1: Value of approximation goes up.
- d_2: Value of approximation goes down.
- d_3: Value of approximation stays the same.

The same can be said for f, but only in the immediate vicinity of the point x.


z
$P_0\ (x_0, y_0, z_0)$
y
x
$-\nabla f$
Most rapid
decrease in f
∇f
Zero change
in f
Most rapid
increase in f

1.3. Algebra

Useful rules!

Let $f : \mathbb{R}^n \to \mathbb{R}$, $g : \mathbb{R}^n \to \mathbb{R}$, . Below is a list of important gradient properties:

- **Gradient of constant:** $\nabla_x c = 0 \in \mathbb{R}^n$ for a constant $c \in \mathbb{R}^n$.
- **Linearity:** $\nabla_x(\alpha f + \beta g)(x) = \alpha \nabla_x f(x) + \beta \nabla_x g(x)$ for a scalars $\alpha, \beta \in \mathbb{R}$.
- **Product rule:** $\nabla_x(fg)(x) = \nabla_x f(x) \cdot g(x) + \nabla_x g(x) \cdot f(x)$.

Let $f : \mathbb{R}^n \to \mathbb{R}^m$, $g : \mathbb{R}^n \to \mathbb{R}^m$, $h : \mathbb{R}^m \to \mathbb{R}^k$, $l : \mathbb{R}^m \to \mathbb{R}$. Below is a list of important Jacobian properties:

- **Jacobian of constant:** $\nabla_x c = 0 \in \mathbb{R}^{n \times m}$ for a constant $c \in \mathbb{R}^n$.
- **Linearity:** $\nabla_x(\alpha f + \beta g)(x) = \alpha \nabla_x f(x) + \beta \nabla_x g(x)$ for a scalars $\alpha, \beta \in \mathbb{R}$.
- **Product rule:** $\nabla_x(f^T g)(x) = [\nabla_x f(x)]^T g(x) + [\nabla_x g(x)]^T f(x)$.
- **Chain rule:** $\nabla_x(h \circ g)(x) = \nabla_{g(x)} h(g(x)) \nabla_x g(x)$ and $\nabla_x(l \circ g)(x) = \left[[\nabla_{g(x)} l(g(x))]^T \nabla_x g(x)\right]^T$.

(a) Let $f : \mathbb{R}^n \to \mathbb{R}$ be $f(x) = v^T x$ for $v \in \mathbb{R}^n$. Using the definition of the gradient, write out $\nabla_x f(x)$ and specify its dimensions.

(b) Let $f : \mathbb{R}^n \to \mathbb{R}^n$ be $f(x) = x$. Using the definition of the Jacobian, write out $\nabla_x f(x)$ and specify its dimensions.

(c) Let $f : \mathbb{R}^n \to \mathbb{R}^m$ be $f(x) = Ax$ for $A \in \mathbb{R}^{m \times n}$. Using the definition of the Jacobian, write out $\nabla_x f(x)$ and specify its dimensions.

Solution:

$$\nabla_x f(x) = \begin{bmatrix} \frac{\partial (Ax)_1}{\partial x_1} & \cdots & \frac{\partial (Ax)_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial (Ax)_m}{\partial x_1} & \cdots & \frac{\partial (Ax)_m}{\partial x_n} \end{bmatrix} = \begin{bmatrix} \frac{\partial}{\partial x_1} \sum_k A_{1k} x_k & \cdots & \frac{\partial}{\partial x_n} \sum_k A_{1k} x_k \\ \vdots & \ddots & \vdots \\ \frac{\partial}{\partial x_1} \sum_k A_{mk} x_k & \cdots & \frac{\partial}{\partial x_n} \sum_k A_{mk} x_k \end{bmatrix}$$

$$= \begin{bmatrix} A_{11} & \cdots & A_{1n} \\ \vdots & \ddots & \vdots \\ A_{m1} & \cdots & A_{mn} \end{bmatrix} = A \in \mathbb{R}^{m \times n}$$

(d) Let $f : \mathbb{R}^n \to \mathbb{R}$ be $f(x) = \alpha v^T x + \beta w^T x$ where $\alpha, \beta \in \mathbb{R}$ and $v, w \in \mathbb{R}^n$. Using the properties at the beginning of the section and previous results, write out $\nabla_x f(x)$.

Solution:

$$\nabla_x f(x) = \nabla_x \left(\alpha v^T x + \beta w^T x\right) = \alpha \nabla_x v^T x + \beta \nabla_x w^T x = \alpha v + \beta w$$

(e) Let $f : \mathbb{R}^n \to \mathbb{R}$ be $f(x) = x^T A x$ and $A \in \mathbb{R}^{n \times n}$. Using the properties at the beginning of the section and previous results, write out $\nabla_x f(x)$.

Solution:

$$\nabla_x f(x) = \nabla_x \left(x^T A x\right) = (\nabla_x x)^T (Ax) + (\nabla_x Ax)^T x = IAx + A^T x = (A + A^T)x$$

where we used the product rule and split $x^T Ax$ into $g(x)^T h(x)$ where $g, h : \mathbb{R}^n \to \mathbb{R}^n$ and $g(x) = x$ and $h(x) = Ax$

(f) With f defined as in the previous part, what is the Hessian of f. Only use previously proven facts and recall that the Hessian is the Jacobian of the gradient.

Solution:

$$\nabla_x^2 f(x) = \nabla_x(\nabla_x f)(x) = \nabla_x(A + A^T)x = A + A^T$$

where we used part 3 in the last step.

(g) Let $f : \mathbb{R}^m \to \mathbb{R}$ be $f(x) = (Ax - y)^T W(Ax - y)$ and $A \in \mathbb{R}^{m \times n}, W \in \mathbb{R}^{n \times n}, y \in \mathbb{R}^n$. Using the properties at the beginning of the section and previous results, write out $\nabla_x f(x)$.

Solution:

Let $f = h \circ g$ where $g(x) = Ax - y$ and $h(z) = z^T W z$. Using the chain rule and parts 3 and 5, we can derive:

$$\begin{aligned}
\nabla_x f(x) = \nabla_x (h \circ g)(x) &= \left[[\nabla_{g(x)} h(g(x))]^T \nabla_x g(x) \right]^T \\
&= \left[[(W + W^T)(Ax - y)]^T A \right]^T \\
&= A^T (W + W^T)(Ax - y)
\end{aligned}$$

Two types of layout for matrix differentiation: Numerator Layout and Denominator Layout

$$\frac{\partial \mathbf{y}}{\partial x} = \begin{bmatrix} \frac{\partial y_1}{\partial x} \\ \frac{\partial y_2}{\partial x} \\ \vdots \\ \frac{\partial y_m}{\partial x} \end{bmatrix}_{m\times 1} \equiv \begin{bmatrix} \frac{\partial y_1}{\partial x} & \frac{\partial y_2}{\partial x} & \cdots & \frac{\partial y_m}{\partial x} \end{bmatrix}_{1\times m}$$

$$\frac{\partial y}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial y}{\partial x_1} & \frac{\partial y}{\partial x_2} & \cdots & \frac{\partial y}{\partial x_n} \end{bmatrix}_{1\times n} \equiv \begin{bmatrix} \frac{\partial y}{\partial x_1} \\ \frac{\partial y}{\partial x_2} \\ \vdots \\ \frac{\partial y}{\partial x_n} \end{bmatrix}_{n\times 1}$$

Numerator layout ~ Denominator Layout