446 Section 3.00000001

TA: Varun Ananth

Plans for today!

1. This
2. Reminders
3. Review Vector Calculus/Approximations
   a. Will be lecture/follow along style

Jacobians and Hessians

And how it is just a fancy way to describe gradients

Slides by Marco D.

1.1. Definitions

Let $f : \mathbb{R}^n \to \mathbb{R}$ and let $g : \mathbb{R}^n \to \mathbb{R}^m$. The **gradient** of f(with respect to x) evaluated at x is the vector of partial derivatives:

$$\nabla_x f(x) = \begin{bmatrix} \frac{\partial f(x)}{\partial x_1} \\ \vdots \\ \frac{\partial f(x)}{\partial x_n} \end{bmatrix} \in \mathbb{R}^n$$

The **Jacobian** of g(with respect to x) evaluated at x is the matrix of partial derivatives:

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

WHAT IS THIS?!?!?!

Sometimes the Jacobian is denoted by $J_g(x)$, but we use $\nabla_x g(x)$ to highlight that the Jacobian is nothing more than the generalization of the gradient to functions which have a vector output.

The **Hessian** of f(with respect to x) evaluated at x is the matrix of partial derivatives:

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

Sometimes the Hessian is denoted by $H_f(x)$, but we use $\nabla_x^2 f(x)$ to highlight that the Hessian is the Jacobian of the gradient of f.

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

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

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

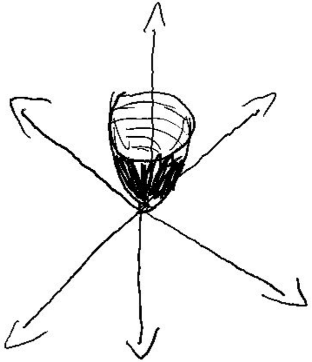
Slides by Marco D.

Visualizing the Gradient

i.e: scalar value

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

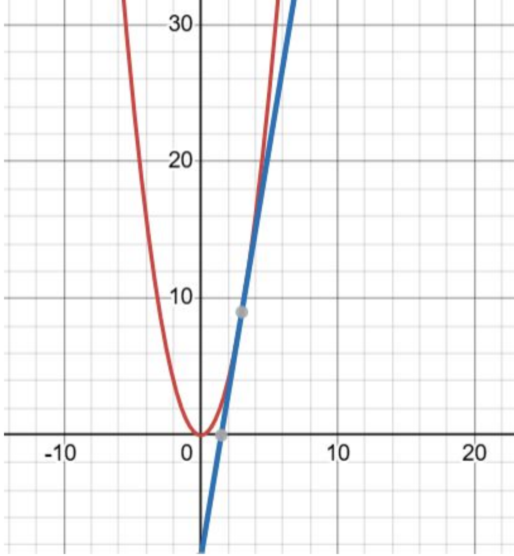
Example: $z = \underbrace{x_1^2 + x_2^2}_{\text{can also be thought of as x and y}}$:

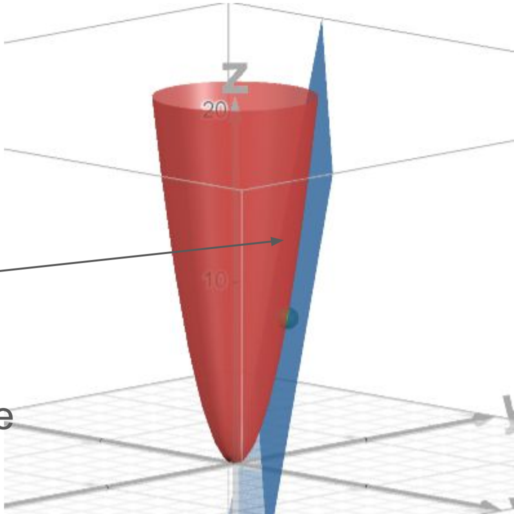


$$\nabla_x f(x) = \begin{bmatrix} \frac{\partial f(x)}{\partial x_1} \\ \frac{\partial f(x)}{\partial x_2} \end{bmatrix} = \begin{bmatrix} 2x_1 \\ 2x_2 \end{bmatrix}$$

all x's

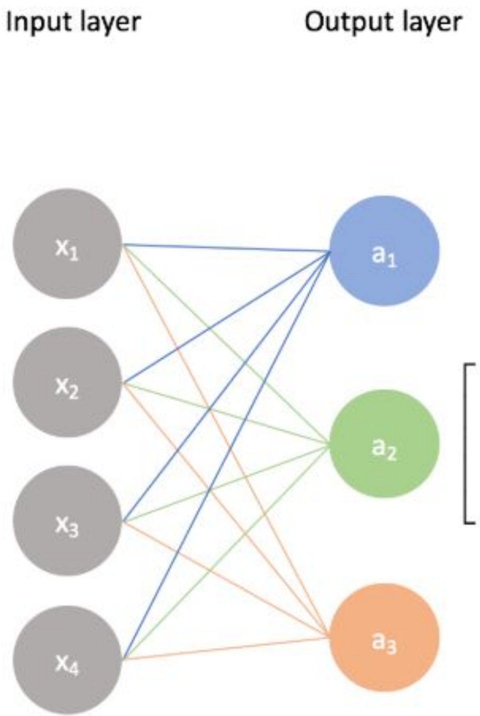
Tells you the slope of the tangent plane in the x_1 and x_2 directions





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

$$\begin{bmatrix} w_1 & w_2 & w_3 & w_4 \\ w_1 & w_2 & w_3 & w_4 \\ w_1 & w_2 & w_3 & w_4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} + \begin{bmatrix} b \\ b \\ b \end{bmatrix} = \begin{bmatrix} w_1x_1 + w_2x_2 + w_3x_3 + w_4x_4 + b \\ w_1x_1 + w_2x_2 + w_3x_3 + w_4x_4 + b \\ w_1x_1 + w_2x_2 + w_3x_3 + w_4x_4 + b \end{bmatrix} \xrightarrow[activation]{} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix}$$

What do we do now???

Slides by Marco D.

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

Example: $g(x) = Wx$

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

$$\begin{array}{c} m \end{array}\underbrace{\begin{bmatrix} \\ W \\ \\ \end{bmatrix}}_{n} \; \begin{array}{c} n \end{array}\underbrace{\begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}}_{1} = \begin{array}{c} m \end{array}\underbrace{\begin{bmatrix} \\ \\ \\ \end{bmatrix}}_{1}$$

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

$$\begin{array}{c} m \end{array}\underbrace{\begin{bmatrix} w_1 \\ \\ W \\ \\ \end{bmatrix}}_{n} \; \begin{array}{c} n \end{array}\begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = \begin{array}{c} m \end{array}\underbrace{\begin{bmatrix} w_1 \cdot x \\ w_2 \cdot x \\ \vdots \\ w_m \cdot x \end{bmatrix}}_{1}$$

dot product

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

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

Gradient!

$$\nabla_x g(x) = \begin{array}{c|c} m & \begin{bmatrix} \nabla_x w_1 \cdot x \\ \nabla_x w_2 \cdot x \\ \vdots \\ \nabla_x w_m \cdot x \end{bmatrix} \\ & 1 \end{array}$$

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

Gradient!

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

Each row becomes an *n* element gradient

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

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

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

$\vdots$

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

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

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

$$\nabla_x (\nabla_x f(x)) = \nabla_x^2 f(x) = \underline{\text{Hessian}}$$

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

Remember the dimensionality of $\nabla_x g(x)$?

$$\downarrow$$

$$\nabla_x (\nabla_x f(x)) \in \mathbb{R}^{n \times n}$$

Important to understand!

The Hessian is the Jacobian of the gradient of $f(x)$

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

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

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

$$f(x) = x_1^2 + x_2^2$$

Gradient: $\nabla_x f(x) = \begin{bmatrix} 2x_1 \\ 2x_2 \end{bmatrix} \rightarrow \begin{array}{c|c} \frac{\partial}{\partial x_1}(2x_1) = 2 & \frac{\partial}{\partial x_2}(2x_1) = 0 \\ \hline \frac{\partial}{\partial x_1}(2x_2) = 0 & \frac{\partial}{\partial x_2}(2x_2) = 2 \end{array}$

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 $\nabla_x f$?

Answers

(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:

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

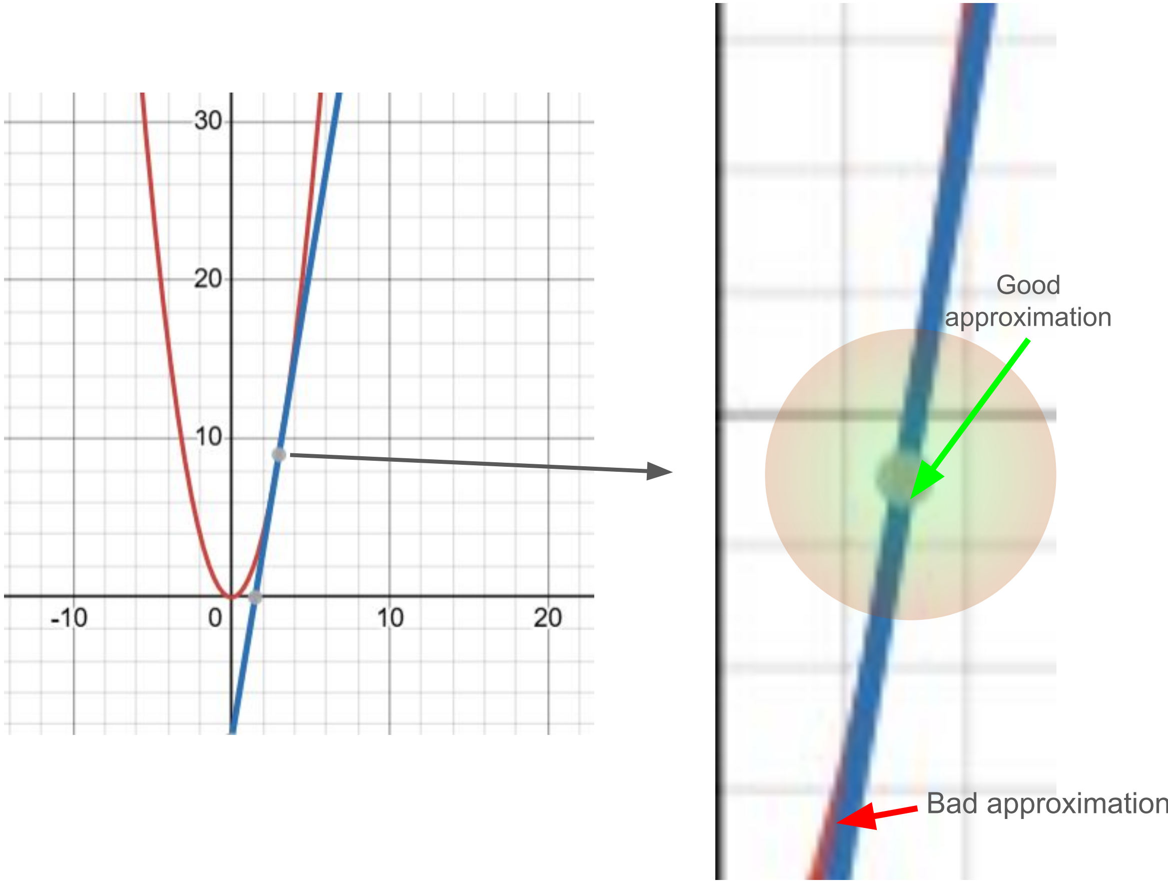
Approximations

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



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**

What about a "many-to-many" function?

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

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

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

? ? ? ? ?

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

Problem 1.2 c

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

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?

Answer

(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

$$f(x) \approx f(a) + \frac{f'(a)}{1!}(x-a) + \underbrace{\frac{f''(a)}{2!}(x-a)^2 + \frac{f'''(a)}{3!}(x-a)^3 \ldots}_{\text{Better and better approximations}}$$

Exact at a, close around a

Remember Taylor expansion?

↳ To approximate a function around a point a

$$f(x) \approx f(a) + \frac{f'(a)}{1!}(x-a) + \underbrace{\frac{f''(a)}{2!}(x-a)^2 + \frac{f'''(a)}{3!}(x-a)^3 \cdots}_{\text{Better and better approximations}}$$

Exact at a, close around a

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

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

$$\Downarrow$$

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

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

So what is the second order derivative?

$$f(x+\epsilon) \approx f(x) + \underbrace{(\nabla_x f(x))}_{}^T \epsilon$$

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

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

Answer

(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.

Intuition used here will be useful on the exam

Properties

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 $\nabla_x(l \circ g)(x) = \left[[\nabla_{g(x)} l(g(x))]^T \nabla_x g(x)\right]^T$.

Ludwig Otto Hesse

Hessian

vs

Carl Gustav Jacob Jacobi

Jacobian

vs

William Grady Hamilton

~~William Rowan Hamilton~~

Gradient

Questions/Chat Time!