CSDS 440: Machine Learning

Soumya Ray (he/him, sray@case.edu)

Olin 516

Office hours T, Th 11:15-11:45 or by appointment

Announcements

- Test Thursday, 30 minutes, in class
  - Topics up to and including Decision Trees

Review of Calculus and Optimization

- Calculus classes/CSDS 477 / MATH 427/ MATH 433 for the less-crashy version

Functions

- A function *maps* an input set to an output set
- Usually denoted $f : X \rightarrow Y;\ f(x) = y$

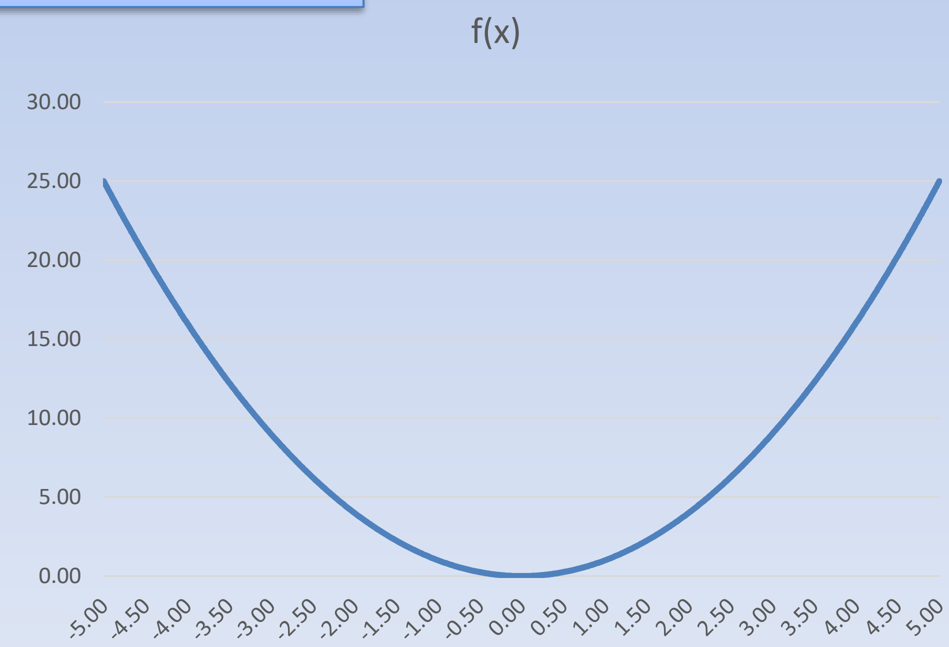
Domain

Range/Co-domain

"Scalar Function"

Example: $f : \mathbb{R} \rightarrow \mathbb{R};$
$f(x) = x^2$

Real numbers



Multivariate Functions

- A function can have *multiple* inputs, denoted by $f: A \times B \times \ldots \rightarrow Y$
- The "$\times$" is the "Cartesian product" or "cross product": all possible tuples
- Example: $\mathbb{R} \times \mathbb{R}$: all possible pairs of real numbers e.g. (1, 1.8), (5.98635435, -3.23456), (π, $\sqrt{2}$) etc.
  - Often abbreviated as $\mathbb{R}^2$ ($\mathbb{R}^D$ in general)

Example

- Suppose we have vectors of size 2
- The norm is a function $\|\cdot\| : \mathbb{R}^2 \rightarrow \mathbb{R}$

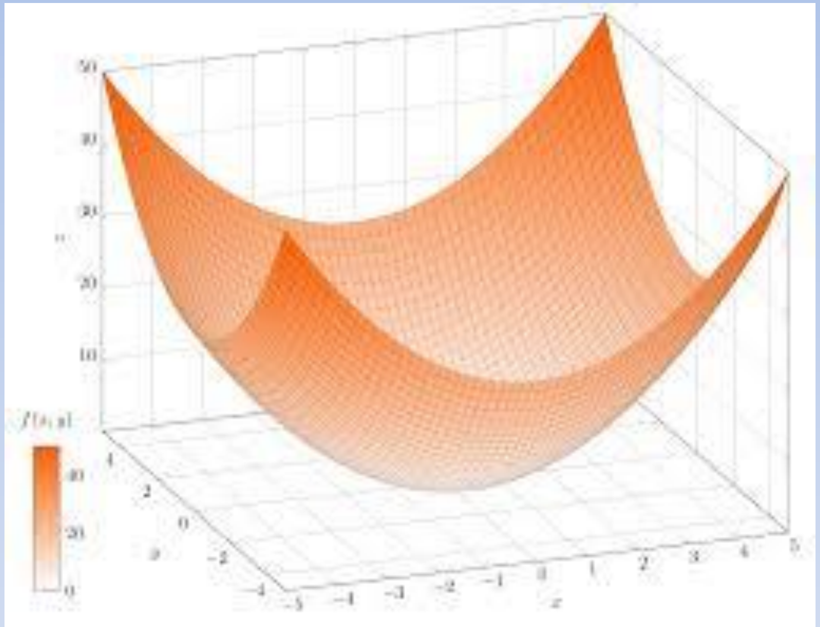
Vector functions

- A function can also have multiple *out*puts , denoted by $f: X \longrightarrow A \times B \times \ldots$
- Example: $f: \mathbb{R} \longrightarrow \mathbb{R}^2, f(x)=(x^2, x^3)$
- And a combination of both multiple inputs and outputs $f: A \times B \times \ldots \longrightarrow C \times D \times \ldots$
  - Multivariate vector functions

Gradient of multivariate functions

- When a function takes multiple inputs, we compute *partial derivatives* by varying each input at a time and holding the others fixed

- A function with m inputs will have m partial derivatives



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

Partial Derivatives

- Suppose $f : \mathbb{R}^2 \rightarrow \mathbb{R}$. The partial derivatives of f are:

$$\left.\frac{\partial f}{\partial x}\right|_{x,y} = \lim_{\Delta x \rightarrow 0} \frac{f(x+\Delta x, y) - f(x,y)}{\Delta x}$$

$$\left.\frac{\partial f}{\partial y}\right|_{x,y} = \lim_{\Delta y \rightarrow 0} \frac{f(x, y+\Delta y) - f(x,y)}{\Delta y}$$

- (Generalizes to functions of n inputs)

Gradient / Jacobian

- The *row* vector

$$\nabla_{x_1,\ldots,x_m} f = \begin{bmatrix} \dfrac{\partial f}{\partial x_1} & \dfrac{\partial f}{\partial x_2} \cdots \dfrac{\partial f}{\partial x_m} \end{bmatrix}$$

- Is called the gradient or Jacobian of *f*.

Example

Let $f(x, y) = \left(x^2 y + xy^3\right)$. Then

$$\frac{\partial f}{\partial x} = (2xy + y^3), \quad \frac{\partial f}{\partial y} = (x^2 + 3xy^2)$$

$$\nabla f = \begin{bmatrix} 2xy + y^3 & x^2 + 3xy^2 \end{bmatrix}$$

Vector functions

- A vector function $f : \mathbb{R} \rightarrow \mathbb{R}^n$ can be viewed as a *vector of scalar* functions

- Let $f(x)=(x^2, x^3)$; then $f(x)=(f_1(x), f_2(x))$; $f_1(x)= x^2$; $f_2(x)= x^3$

Jacobian of vector functions

- The Jacobian is then the *column* vector:

$$\nabla_x f = \begin{bmatrix} \dfrac{df_1}{dx} \\ \dfrac{df_2}{dx} \\ \vdots \\ \dfrac{df_n}{dx} \end{bmatrix}$$

$$f(x) = (x^2, x^3); \nabla_x f = \begin{bmatrix} 2x \\ 3x^2 \end{bmatrix}$$

Jacobian of Multivariate Vector functions

- Suppose $f:\mathbb{R}^m \rightarrow \mathbb{R}^n$, then

$$\nabla_x f = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_m} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_n}{\partial x_1} & \cdots & \frac{\partial f_n}{\partial x_m} \end{bmatrix}$$

Example

- Let $f(x, y)=(x^2y, xy^3)$.

$$\nabla_{x,y} f = \begin{bmatrix} 2xy & x^2 \\ y^3 & 3xy^2 \end{bmatrix}$$

Partial Derivative Shortcuts

$$\frac{\partial}{\partial x}(f(x)+g(x)) = \frac{\partial f(x)}{\partial x} + \frac{\partial g(x)}{\partial x}$$

"Partial Derivative Sum rule"

$$\frac{\partial}{\partial x}(f(x)g(x)) = f(x)\frac{\partial g(x)}{\partial x} + g(x)\frac{\partial f(x)}{\partial x}$$

"Partial Derivative Product rule"

$$\frac{\partial}{\partial x}\left(f\left(g(x)\right)\right) = \frac{\partial f}{\partial g}\frac{\partial g}{\partial x}$$

"Partial Derivative Chain rule"

$$\boxed{\frac{\partial f}{\partial g} = \left.\frac{\partial f}{\partial x}\right|_{g(x)}}$$

Note that, in the case of vector and multivariate functions, these are matrix products, so the order is important.

Example

- Let $f(x, y)=(e^x, x+y)$. Let $g(a, b)=(a^2, b^2)$. Then the Jacobian of $f(g(a,b))$ is:

$$\nabla_{x,y} f = \begin{bmatrix} e^x & 0 \\ 1 & 1 \end{bmatrix}; \nabla_{a,b} g = \begin{bmatrix} 2a & 0 \\ 0 & 2b \end{bmatrix}$$

$$\nabla_{a,b} f \circ g = \nabla_g f \nabla_{a,b} g$$

$$\nabla_g f = \nabla_{x,y} f \Big|_{x=a^2, y=b^2} = \begin{bmatrix} e^{a^2} & 0 \\ 1 & 1 \end{bmatrix}$$

$$\nabla_{a,b} f \circ g = \begin{bmatrix} e^{a^2} & 0 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 2a & 0 \\ 0 & 2b \end{bmatrix} = \begin{bmatrix} 2ae^{a^2} & 0 \\ 2a & 2b \end{bmatrix}$$

Higher Order Derivatives

- We can take the derivative of a function multiple times (if possible)

- The rate of change of the derivative is the *second derivative*:

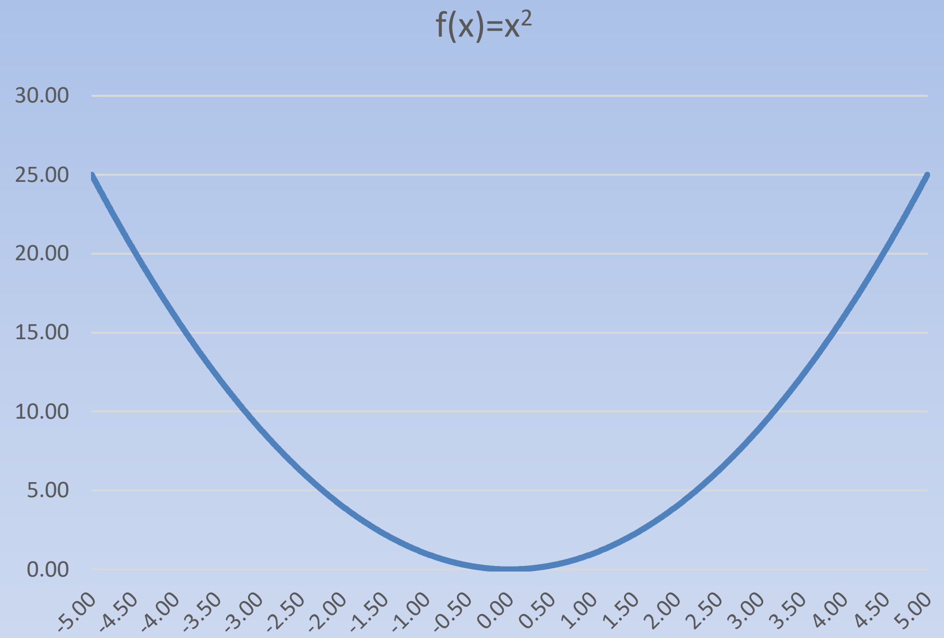
$$\left.\frac{d^2 f}{dx^2}\right|_{x_0} = f''(x_0) = \lim_{\Delta x \to 0} \frac{f'(x_0 + \Delta x) - f'(x_0)}{\Delta x}$$

Example and Geometry

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

$$\frac{df}{dx} = 2x$$

$$\frac{d^2 f}{dx^2} = 2.$$



Function with “positive curvature”

Example and Geometry

$$g'(x) = 4x^3 + 21x^2 + 10x - 17$$

$$g''(x) = 12x^2 + 42x + 10$$

$$= 64, -25.28, 45.28 \text{ (for different inputs } x\text{)}$$

Multivariate functions

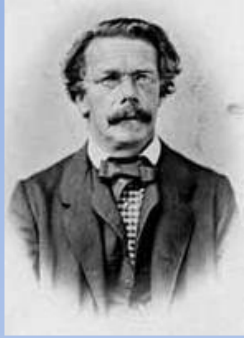
- Suppose we have a function $f : \mathbb{R}^m \to \mathbb{R}$

- The *derivative is itself a vector function* $\mathbb{R}^m \to \mathbb{R}^m$:

$$\nabla_{x_1,\ldots,x_m} f = \left[\frac{\partial f}{\partial x_1} \quad \frac{\partial f}{\partial x_2} \cdots \frac{\partial f}{\partial x_m} \right]$$

- The second derivative takes this as input

Hessian Matrix



- Taking the second derivative creates a $m \times m$ matrix:

$$\nabla_x^2 f = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \cdots & \frac{\partial^2 f}{\partial x_m \partial x_1} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_1 \partial x_m} & \cdots & \frac{\partial^2 f}{\partial x_m^2} \end{bmatrix}$$

Example

Let $f(x,y) = \left(x^2 y + xy^3\right)$. Then

$$\frac{\partial f}{\partial x} = (2xy + y^3), \quad \frac{\partial f}{\partial y} = (x^2 + 3xy^2)$$

$$\nabla_{x,y} f = g(x,y) = \begin{bmatrix} 2xy + y^3 & x^2 + 3xy^2 \end{bmatrix}$$

$$\nabla^2_{x,y} f = \nabla g_{x,y} = \begin{bmatrix} 2y & 2x + 3y^2 \\ 2x + 3y^2 & 6y \end{bmatrix}$$

Tensors

- What if $f : \mathbb{R}^m \rightarrow \mathbb{R}^n$?

$$\nabla_x f = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_m} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_n}{\partial x_1} & \cdots & \frac{\partial f_n}{\partial x_m} \end{bmatrix}$$

Then the second derivative is a three dimensional matrix H of size $(n \times m) \times m$ where H_{ijk}=$\frac{\partial f_i}{\partial x_k \partial x_j}$. Such matrices are called tensors.

Example

- Let $f(x, y)=(x^2y, xy^3)$.

$$\nabla_{x,y} f = \begin{bmatrix} 2xy & x^2 \\ y^3 & 3xy^2 \end{bmatrix}$$

$$\nabla^2_{x,y} f = \begin{bmatrix} 2y;2x & 2x;0 \\ 0;3y^2 & 3y^2;6xy \end{bmatrix}$$

Gradients of Matrices

- Suppose we have a function $f : \mathrm{B} \longrightarrow \mathrm{A}, \mathrm{B} \in \mathbb{R}^{m\times n} \longrightarrow \mathrm{A} \in \mathbb{R}^{p\times q}$
- The Jacobian of f will be a four dimensional tensor $\mathbb{R}^{(m\times n) \times (p\times q)}$
- Where $J(i, j, k, l) = \dfrac{\partial A_{ij}}{\partial B_{kl}}$

Gradient of a vector wrt a matrix

- Suppose we have function $f\,(A)$: $y=Ax$, $\mathbb{R}^{m\times n} \longrightarrow \mathbb{R}^m$
- What is $J(\,f\,)$?

$$\nabla_x f = \begin{bmatrix} \frac{df_1}{dx} \\ \frac{df_2}{dx} \\ \vdots \\ \frac{df_n}{dx} \end{bmatrix} \qquad \nabla_A f = \begin{bmatrix} \frac{\partial f_1}{\partial A} \\ \frac{\partial f_2}{\partial A} \\ \vdots \\ \frac{\partial f_m}{\partial A} \end{bmatrix} \qquad \frac{\partial f_i}{\partial A} \in \mathbf{R}^{1\times m\times n}$$

Reshaping Tensors

- In practice, it can often be useful to convert a tensor into a matrix or vector
  - Done by "stacking" tensors
- Idea: Every matrix in $\mathbb{R}^{m \times n}$ can be written as a vector in $\mathbb{R}^{mn}$
- Similarly, every tensor in $\mathbb{R}^{(m\times n) \times (p\times q)}$ can be written as a matrix in $\mathbb{R}^{mn \times pq}$

What is an optimization problem?

- Find the *extreme values* of a function (called an "objective function")
  - Sometimes we are interested in the extreme values themselves
  - Other times we are interested in the *arguments* to the function that produce those extreme values
    - argmax, argmin

Types of Optimization Problems

- Discrete vs continuous
  - Objective function is defined on discrete or continuous space
- Unconstrained vs constrained
  - Whether there are additional function constraints defining the “feasible region”
- In this class, we will mainly be interested in continuous problems, both unconstrained and constrained
  - Use tools from calculus and linear algebra

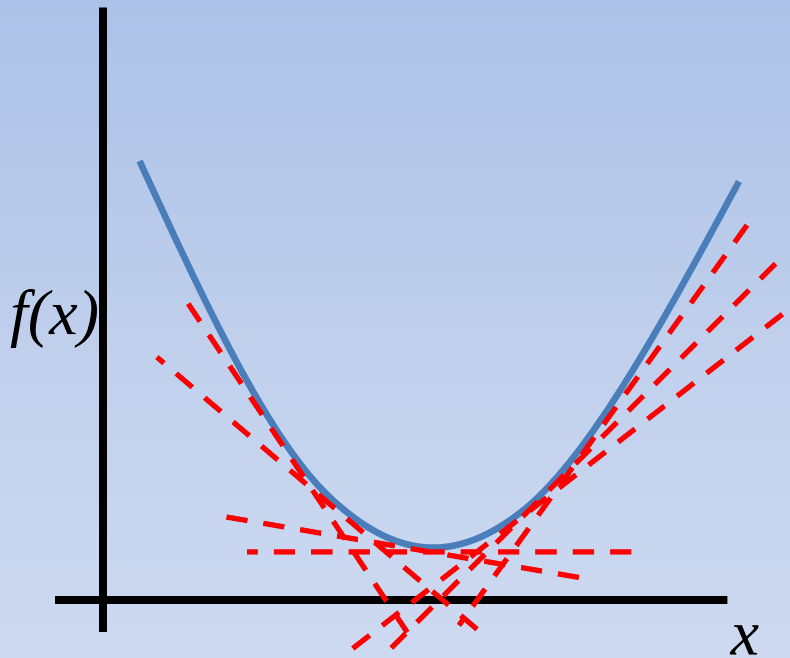
Unconstrained Continuous Optimization

Function of one variable:

$$\min_x f(x)$$

$$\frac{df}{dx} = 0$$

$$\frac{d^2 f}{dx^2} > 0$$



Example

$$g'(x) = 4x^3 + 21x^2 + 10x - 17$$

$$g'(x) = 0 \text{ for } x = -4.5, -1.4, 0.7$$

$$g''(x) = 12x^2 + 42x + 10$$

$$= 64, -25.28, 45.28$$

Multivariate functions

$$\min_{x_1,\ldots,x_m} f(x_1,\ldots,x_m)$$

$$J = \left(\frac{\partial f}{\partial x_i}\right) = 0$$

Jacobian is zero

$$H = \left(\frac{\partial^2 f}{\partial x_i \partial x_j}\right) > 0$$

Hessian is "positive definite"

Example

$$f(x,y,z) = x^5y^4 - z^6y^3 + x^4z^3$$

$$\nabla f = \begin{bmatrix} 5x^4y^4 + 4x^3z^3 & 4x^5y^3 - 3z^6y^2 & -6z^5y^3 + 3x^4z^2 \end{bmatrix} = 0$$

????

Observation

- In general, analytically solving for the zeros of the Jacobian is computationally (sometimes algebraically!) infeasible

- Alternative: switch to an *iterative* method