Optimization for Data Science
Lecture 01: Preliminaries

Kimon Fountoulakis

School of Computer Science
University of Waterloo

10/09/2019

Vector space

- We will always work in vector spaces

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

Column vectors

- We will assume that a vector “x” is a column vector “n x 1”

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

Vector transpose

Transpose of "x"

$$x^T = [x_1, x_2, \ldots, x_n]$$

Transpose of the transpose

$$(x^T)^T = x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

Matrices

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \hdotsfor{4} \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}$$

$m < n$

More columns than rows

$m > n$

More rows than columns

$m = n$

columns = # rows

Matrix transpose

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}$$

Transpose (rows become columns and columns become rows)

$$A^T = \begin{bmatrix} a_{11} & a_{21} & \cdots & a_{m1} \\ a_{12} & a_{22} & \cdots & a_{m2} \\ \cdots & \cdots & \cdots & \cdots \\ a_{1n} & a_{2n} & \cdots & a_{mn} \end{bmatrix}$$

Transpose of the transpose

$$(A^T)^T = A$$

Symmetric matrices

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 4 & 5 \\ 3 & 5 & 6 \end{bmatrix}$$

lower triangular part = Upper triangular part

Inner product between vectors

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \quad y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

Column vectors

Inner product

$$x^T y = x_1 y_1 + x_2 y_2 + \cdots + x_n y_n$$

output is a scalar

Euclidean norm of a vector

$$\|x\|_2 = (x^T x)^{1/2} = \left(\sum_{i=1}^{n} x_i^2 \right)^{1/2}$$

- Shows the length of vector “x”

Euclidean norm of a matrix

$$\|A\|_2 = \max_{x \neq 0} \frac{\|Ax\|_2}{\|x\|_2}$$

- It shows how much vector “x” can be stretched by a matrix “A” relative to the initial length of “x”.
- Very important in implementing and analyzing optimization algorithms.

Cauchy-Schwartz inequality

$$z^T s \leq \|z\|_2 \|s\|_2$$

- Very useful when analyzing the running time of optimization algorithms.

Right matrix-vector product

$$Ax = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} a_1^T x \\ \vdots \\ a_n^T x \end{bmatrix}$$

Output is n x 1 column vector

- “alpha sub-i” is the i-th row of matrix “A”

Left matrix-vector product

$$y^T A = \begin{bmatrix} y_1 & y_2 & \dots & y_m \end{bmatrix} \begin{bmatrix} a_{11} & a_{21} & \cdots & a_{m1} \\ a_{12} & a_{22} & \cdots & a_{m2} \\ \cdots & \cdots & \cdots & \cdots \\ a_{1n} & a_{2n} & \cdots & a_{mn} \end{bmatrix} = [y^T a_1, y^T a_2, \dots, y^T a_m]$$

Output is
1 x n
row vector

- “alpha sub-i” is the i-th column of matrix “A”

Positive definite matrices

- A matrix is positive definite if

$$y^T A y > 0 \; \forall y \neq 0$$

Functions

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

Domain

Image

Gradient

- If “f” is differentiable, then the gradient “f” w.r.t “x” is

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

the derivative of “f” w.r.t to x_1,
by considering all other variables as constant

Gradient: example

$$f(x) = \frac{1}{2}a_{11}x_1^2 + a_{12}x_1x_2 + \frac{1}{2}a_{22}x_2^2 + b_1x_2 + c$$

$$\nabla f(x) \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \end{bmatrix} = \begin{bmatrix} a_{11}x_1 + a_{12}x_2 + b_1 \\ a_{12}x_1 + a_{22}x_2 + b_2 \end{bmatrix}$$

Gradient: example

$$f(x) = y^T A x$$

$$A = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \cdots & \cdots & \cdots \\ a_{n1} & \cdots & a_{nn} \end{bmatrix}$$

$$x = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \quad y = \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix}$$

$$\nabla f(x) = A^T y$$

Note the output is in column format
so this

$$\nabla f(x) = y^T A$$

is incorrect

Gradient: example of a quadratic function

$$f(x) = \frac{1}{2}x^T Ax + b^T x$$

$$\nabla f(x) = Ax + b$$

- Here we assume that matrix A is symmetric.

Second-order derivative

$$\nabla^2 f(x) = \begin{bmatrix} \frac{\partial}{\partial x_1}\left(\frac{\partial}{\partial x_1}\right) & \frac{\partial}{\partial x_1}\left(\frac{\partial f}{\partial x_2}\right) & \cdots & \frac{\partial}{\partial x_1}\left(\frac{\partial f}{\partial x_n}\right) \\ \frac{\partial}{\partial x_2}\left(\frac{\partial f}{\partial x_1}\right) & \frac{\partial}{\partial x_2}\left(\frac{\partial f}{\partial x_2}\right) & \cdots & \frac{\partial}{\partial x_2}\left(\frac{\partial f}{\partial x_n}\right) \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\partial}{\partial x_n}\left(\frac{\partial f}{\partial x_1}\right) & \frac{\partial}{\partial x_n}\left(\frac{\partial f}{\partial x_2}\right) & \cdots & \frac{\partial}{\partial x_n}\left(\frac{\partial f}{\partial x_n}\right) \end{bmatrix}$$

- This matrix is also called the Hessian matrix

Second derivative: example

$$f(x) = \frac{1}{2}a_{11}x_1^2 + a_{12}x_1x_2 + \frac{1}{2}a_{22}x_2^2 + b_1x_2 + c$$

$$\nabla f(x) \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \end{bmatrix} = \begin{bmatrix} a_{11}x_1 + a_{12}x_2 + b_1 \\ a_{12}x_1 + a_{22}x_2 + b_2 \end{bmatrix}$$

$$\nabla^2 f(x) = \begin{bmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{bmatrix}$$

Second derivative: example of a quadratic function

$$f(x) = \frac{1}{2} x^T A x + b^T x$$

$$\nabla f(x) = Ax + b$$

$$\nabla^2 f(x) = A$$

- Here we assume that matrix A is symmetric.

Summary

- We reviewed basic linear algebra and calculus concepts for vector spaces.

- We will use these for developing and analyzing numerical optimization algorithms