15-884/15-484 – Linear Algebra and Matrix Calculus Review

J. Zico Kolter

September 5, 2013

- Linear algebra notation is used extensively in machine learning, optimization, power systems
- This lecture reviews some basic notation and methods, and introduces matrix calculus used in class
- For more advanced material (mainly matrix calculus), we will present the methods again when used, but these slides serve as a single reference for all the methods we will use

Linear equations

- Set of linear equations (two equations, two unknowns)

$$\begin{aligned} 4x_1 - 5x_2 &= -13 \\ -2x_1 + 3x_2 &= 9 \end{aligned}$$

- Set of linear equations (two equations, two unknowns)

$$\begin{array}{rrcr} 4x_1 & -\,5x_2 & = & -13 \\ -2x_1 & +\,3x_2 & = & 9 \end{array}$$

- Can represent compactly using matrix notation

$$Ax = b$$

with

$$A = \begin{bmatrix} 4 & -5 \\ -2 & 3 \end{bmatrix}, \quad x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \quad b = \begin{bmatrix} -13 \\ 9 \end{bmatrix}$$

Basic notation

- A matrix with real-valued entries, m rows, and n columns

$$A \in \mathbb{R}^{m \times n}$$

A_{ij} denotes the entry in the ith row and jth column

- A (column) vector with n real-valued entries

$$x \in \mathbb{R}^n$$

x_i denotes the ith entry

The transpose

- The transpose operator A^T switches rows and columns of a matrix

$$A_{ij} = (A^T)_{ji}$$

- For a vector $x \in \mathbb{R}^n$, $x^T \in \mathbb{R}^{1 \times n}$ would represent a row vector

Elements of a matrix

- Can write a matrix in terms of its columns

$$A = \begin{bmatrix} | & | & & | \\ a_1 & a_2 & \cdots & a_n \\ | & | & & | \end{bmatrix}$$

- Careful, a_i here corresponds to an entire *vector* $a_i \in \mathbb{R}^m$, not an element of a vector

- Similarly, can write a matrix in terms of rows

$$A = \begin{bmatrix} - & a_1^T & - \\ - & a_2^T & - \\ & \vdots & \\ - & a_m^T & - \end{bmatrix}$$

- $a_1 \in \mathbb{R}^n$ here and $a_1 \in \mathbb{R}^m$ from previous slide are *not* the same vector

Matrix addition

- For two matrices *of the same size and type*, $A, B \in \mathbb{R}^{m \times n}$ addition is just sum of corresponding elements

$$A + B = C \in \mathbb{R}^{m \times n} \iff C_{ij} = A_{ij} + B_{ij}$$

- Addition is undefined for matrices of different sizes $A \in \mathbb{R}^{m \times n}$, $B \in \mathbb{R}^{p \times q}$

Matrix multiplication

- For two matrices $A \in \mathbb{R}^{m \times n}$, $B \in \mathbb{R}^{n \times p}$, their product is

$$AB = C \in \mathbb{R}^{m \times p} \iff C_{ij} = \sum_{k=1}^{n} A_{ik} B_{kj}$$

- Multiplication is undefined when number of columns in A doesn't equal number or rows in B (one exception: cA for $c \in \mathbb{R}$ taken to mean scaling A by c)

- Some special cases:
  - Inner product, $x, y \in \mathbb{R}^n$

$$x^T y \in \mathbb{R} = \sum_{i=1}^{n} x_i y_i$$

  - Matrix-vector product, $A \in \mathbb{R}^{m \times n}$, $x \in \mathbb{R}^n \iff Ax \in \mathbb{R}^m$

$$A = \left[\begin{array}{cccc} | & | & & | \\ a_1 & a_2 & \cdots & a_n \\ | & | & & | \end{array} \right], \;\; Ax \in \mathbb{R}^m = \sum_{i=1}^{n} a_i x_i$$

- Some imporant properties
  - Associative: ($A \in \mathbb{R}^{n \times n}$, $B \in \mathbb{R}^{n \times p}$, $C \in \mathbb{R}^{p \times q}$)

$$A(BC) = (AB)C$$

  - Distributive: ($A \in \mathbb{R}^{m \times n}$, $B, C \in \mathbb{R}^{n \times p}$)

$$A(B + C) = AB + AC$$

  - *NOT* commutative: (the dimensions might not even make sense, but this doesn't hold even when the dimensions are correct)

$$AB \neq BA$$

  - Transpose of matrix product: $(AB)^T = B^T A^T$

Special matrices

- The identity:

$$I \in \mathbb{R}^{n \times n} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}$$

(ones on the diagonal, zeros everywhere else)

- Has the property that for any $A \in \mathbb{R}^{m \times n}$

$$AI = A = IA$$

(note that the identity matrices on the left and right are *different sizes*, $n \times n$ or $m \times m$, to make the multiplication work)

- The zero matrix

$$0 \in \mathbb{R}^{m \times n} = \begin{bmatrix} 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix}$$

- Useful in defining block forms for matrices; e.g. $A \in \mathbb{R}^{m \times n}$, $B \in \mathbb{R}^{p \times q}$

$$C \in \mathbb{R}^{(m+p)\times(n+q)} = \begin{bmatrix} A & 0 \\ 0 & B \end{bmatrix}$$

- The all-ones vector

$$1 \in \mathbb{R}^n = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$$

- Useful, for example, in compactly representing sums

$$a \in \mathbb{R}^n, \;\; 1^T a = \sum_{i=1}^{n} a_i$$

- Symmetric matrix: $A \in \mathbb{R}^{n \times n}$ with $A = A^T$
- Arise naturally in many settings
  - For $A \in \mathbb{R}^{m \times n}$, $A^T A \in \mathbb{R}^{m \times m}$ is symmetric
  - Many matrices in power systems will be symmetric

- Diagonal matrix: for $d \in \mathbb{R}^n$

$$\operatorname{diag}(d) \in \mathbb{R}^{n \times n} = \begin{bmatrix} d_1 & 0 & \cdots & 0 \\ 0 & d_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & d_n \end{bmatrix}$$

- For example, the identity is given by $I = \operatorname{diag}(1)$

- Inverse of a square matrix $A \in \mathbb{R}^{n \times n}$ denoted A^{-1}

$$AA^{-1} = I = A^{-1}A$$

- May not exist (*non-singular* matrix has inverse, *singular* matrix does not)

$$A^{-1} \text{ exists } \iff Ax \neq 0 \text{ for all } x \neq 0$$

Notation for matrix functions

- $f(x) = x^2$, $f : \mathbb{R} \to \mathbb{R}$
- Function with matrix inputs/outputs

$$f : \mathbb{R}^{m \times n} \to \mathbb{R}^{p \times q}$$

- Transpose: $f(A) = A^T$

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

- Inverse: $f(A) = A^{-1}$

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

- Multiplication: $f(x) = Ax$ for $A \in \mathbb{R}^{m \times n}$

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

- A vector norm is any function $f : \mathbb{R}^n \to \mathbb{R}$ with

    1. $f(x) \geq 0$ and $f(x) = 0 \Leftrightarrow x = 0$

    2. $f(ax) = |a| f(x)$ for $a \in \mathbb{R}$

    3. $f(x + y) \leq f(x) + f(y)$

- ℓ_2 norm: $\|x\|_2 = \sqrt{x^T x} = \sqrt{\sum_{i=1}^n x_i^2}$

- ℓ_1 norm: $\|x\|_1 = \sum_{i=1}^n |x_i|$

Putting equations in matrix form

- Given $a_i \in \mathbb{R}^n$, $b_i \in \mathbb{R}$ for $i = 1, \ldots, m$, $f : \mathbb{R}^n \to \mathbb{R}$

$$f(x) = \sum_{i=1}^{m} (a_i^T x - b_i)^2$$

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

$$f(x) = \begin{bmatrix} x_1^2 \\ x_2^2 \\ \vdots \\ x_n^2 \end{bmatrix}$$

Eigenvalues and eigenvectors

- For $A \in \mathbb{R}^{n \times n}$, $\lambda \in \mathbb{C}$ is an *eigenvalue* and $x \in \mathbb{C}^n \neq 0$ an eigenvector if

$$Ax = \lambda x$$

- Write equations for all n eigenvalues as

$$A \begin{bmatrix} | & & | \\ x_1 & \cdots & x_n \\ | & & | \end{bmatrix} = \begin{bmatrix} | & & | \\ x_1 & \cdots & x_n \\ | & & | \end{bmatrix} \begin{bmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_n \end{bmatrix}$$

- Write as $AX = X\Lambda \iff A = X\Lambda X^{-1}$ (if X invertible)

- An example: Given $A \in \mathbb{R}^{n \times n}$, what can we say about A^k as $k \to \infty$?

Matrix Calculus

- The *Jacobian*: for vector-input, vector-output function $f : \mathbb{R}^n \to \mathbb{R}^m$

$$D_x f(x) \in \mathbb{R}^{m \times n} = \begin{bmatrix} \frac{\partial f_1(x)}{\partial x_1} & \cdots & \frac{\partial f_1(x)}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m(x)}{\partial x_1} & \cdots & \frac{\partial f_m(x)}{\partial x_n} \end{bmatrix}$$

- Example: $f : \mathbb{R}^3 \to \mathbb{R}^2$

$$f(x) = \begin{bmatrix} x_1^2 x_2 + x_3 \\ x_2 / x_3 \end{bmatrix}$$

what is $D_x f(x)$?

- The *gradient*: for vector-input, scalar-output function $f : \mathbb{R}^n \to \mathbb{R}$

$$\nabla_x f(x) \in \mathbb{R}^n = \begin{bmatrix} \frac{\partial f(x)}{\partial x_1} \\ \frac{\partial f(x)}{\partial x_2} \\ \vdots \\ \frac{\partial f(x)}{\partial x_n} \end{bmatrix} = (D_x f(x))^T$$

- Important rules and common gradient

$$\nabla_x(af(x) + bg(x)) = a\nabla_x f(x) + b\nabla_x g(x), \ \ (a, b \in \mathbb{R})$$
$$\nabla_x(x^T Ax) = (A + A^T)x, \ \ (A \in \mathbb{R}^{k\times k})$$
$$\nabla_x(b^T x) = b, \ \ (b \in \mathbb{R}^k)$$

- The *Hessian*: for vector-input, scalar-output function $f : \mathbb{R}^n \to \mathbb{R}$

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

$$= D_x(\nabla_x f(x)) \text{ (Jacobian of the gradient)}$$

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

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

what is $\nabla_x^2 f(x)$?