Lecture 1: Review of Linear Algebra

Ming Yan
Michigan State University, CMSE/Mathematics

CMSE 890: Optimization
Aug. 30th, 2017

Vectors

- $\mathbb{R}^n$: n-dimensional **Euclidean Space**
- A **vector** $\mathbf{x} \in \mathbb{R}^n/\mathbb{C}^n$ is an n-tuple $[x_1, x_2, \cdots, x_n]$, where $x_i \in \mathbb{R}/\mathbb{C}$.
- We also consider a vector $\mathbf{x} \in \mathbb{R}^n/\mathbb{C}^n$ as a column vector or a $n \times 1$ matrix.
- **Inner product in $\mathbb{R}^n$:** For $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$,
  $\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^\top \mathbf{y} = \sum_{i=1}^n x_i y_i = \mathbf{y}^\top \mathbf{x} = \langle \mathbf{y}, \mathbf{x} \rangle \in \mathbb{R}$.
- **Inner product in $\mathbb{C}^n$:** For $\mathbf{x}, \mathbf{y} \in \mathbb{C}^n$,
  $\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^H \mathbf{y} = \sum_{i=1}^n \overline{x_i} y_i = \overline{(\mathbf{y}^H \mathbf{x})} = \overline{\langle \mathbf{y}, \mathbf{x} \rangle} \in \mathbb{C}$
- **Euclidean norm in $\mathbb{R}^n$:** For $\mathbf{x} \in \mathbb{R}^n$, we have $\|\mathbf{x}\|_2 = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle} \geq 0$.
- **ℓ_2 norm in $\mathbb{C}^n$:** For $\mathbf{x} \in \mathbb{C}^n$, we have $\|\mathbf{x}\|_2 = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle} \geq 0$.

Cauchy-Schwarz inequality

Lemma (Cauchy-Schwarz inequality)

For any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ (or $\mathbb{C}^n$), we have

$$|\langle \mathbf{x}, \mathbf{y} \rangle| \leq \|\mathbf{x}\|_2 \|\mathbf{y}\|_2.$$

- When $\mathbf{y} = \mathbf{0}$, It is trivially true.
- When $\mathbf{y} \neq \mathbf{0}$, we have

$$\begin{aligned} 0 \leq & \|\mathbf{x} - \lambda \mathbf{y}\|^2 = \langle \mathbf{x}, \mathbf{x} \rangle - \langle \lambda \mathbf{y}, \mathbf{x} \rangle - \langle \mathbf{x}, \lambda \mathbf{y} \rangle + \langle \lambda \mathbf{y}, \lambda \mathbf{y} \rangle \\ = & \|\mathbf{x}\|^2 - \lambda \overline{\langle \mathbf{x}, \mathbf{y} \rangle} - \bar{\lambda} \langle \mathbf{x}, \mathbf{y} \rangle + |\lambda|^2 \|\mathbf{y}\|^2. \end{aligned}$$

  Letting $\lambda = \langle \mathbf{x}, \mathbf{y} \rangle / \|\mathbf{y}\|^2$, we have

$$0 \leq \|\mathbf{x}\|^2 - \frac{|\langle \mathbf{x}, \mathbf{y} \rangle|^2}{\|\mathbf{y}\|^2} - \frac{|\langle \mathbf{x}, \mathbf{y} \rangle|^2}{\|\mathbf{y}\|^2} + \frac{|\langle \mathbf{x}, \mathbf{y} \rangle|^2}{\|\mathbf{y}\|^2} = \|\mathbf{x}\|^2 - \frac{|\langle \mathbf{x}, \mathbf{y} \rangle|^2}{\|\mathbf{y}\|^2}.$$

- 11 Proofs for this lemma: Wu, Hui-Hua; Wu, Shanhe (April 2009). "Various proofs of the Cauchy-Schwarz inequality". OCTOGON MATHEMATICAL MAGAZINE. 17 (1): 221-229.

Cauchy-Schwarz inequality

Lemma (Cauchy-Schwarz inequality)

For any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ *(or* $\mathbb{C}^n$*), we have*

$$|\langle \mathbf{x}, \mathbf{y} \rangle| \leq \|\mathbf{x}\|_2 \|\mathbf{y}\|_2.$$

- It is one of the most important inequalities in all of mathematics.
- When does it hold with equality? $\mathbf{x} = (\langle \mathbf{x}, \mathbf{y} \rangle / \|\mathbf{y}\|^2)\mathbf{y}$.

Triangle inequality

Lemma (Triangle inequality)

For any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ *(or* $\mathbb{C}^n$*), we have*

$$\|\mathbf{x}+\mathbf{y}\|_2 \leq \|\mathbf{x}\|_2 + \|\mathbf{y}\|_2.$$

$$\begin{aligned}\|\mathbf{x}+\mathbf{y}\|_2^2 =& \langle \mathbf{x}+\mathbf{y}, \mathbf{x}+\mathbf{y}\rangle = \langle \mathbf{x}, \mathbf{x}\rangle + \langle \mathbf{x}, \mathbf{y}\rangle + \langle \mathbf{y}, \mathbf{x}\rangle + \langle \mathbf{y}, \mathbf{y}\rangle \\ =& \|\mathbf{x}\|_2^2 + \langle \mathbf{x}, \mathbf{y}\rangle + \overline{\langle \mathbf{y}, \mathbf{x}\rangle} + \|\mathbf{y}\|_2^2 \\ \leq& \|\mathbf{x}\|_2^2 + 2\|\mathbf{x}\|_2\|\mathbf{y}\|_2 + \|\mathbf{y}\|_2^2 = (\|\mathbf{x}\|_2 + \|\mathbf{y}\|_2)^2.\end{aligned}$$

- When does this inequality hold with equality?
- Other variants:

$$\begin{aligned}\|\mathbf{x}+\mathbf{y}\|_2 \geq& \|\mathbf{x}\|_2 - \|\mathbf{y}\|_2 \\ \|\mathbf{x}+\mathbf{y}\|_2 \geq& \|\mathbf{y}\|_2 - \|\mathbf{x}\|_2.\end{aligned}$$

More identities I

Lemma (Parallelogram identity)

$$\|\mathbf{x}+\mathbf{y}\|_2^2 + \|\mathbf{x}-\mathbf{y}\|_2^2 = 2\|\mathbf{x}\|_2^2 + 2\|\mathbf{y}\|_2^2$$

Lemma (Polarization identity)

$$\|\mathbf{x}+\mathbf{y}\|_2^2 - \|\mathbf{x}-\mathbf{y}\|_2^2 = 4\langle \mathbf{x}, \mathbf{y}\rangle$$

Lemma (Apollonius' identity)

$$\begin{aligned}\|\mathbf{x}-\mathbf{y}\|_2^2 =& 2\|\mathbf{x}\|_2^2 + 2\|\mathbf{y}\|_2^2 - \|\mathbf{x}+\mathbf{y}\|_2^2 \\ & - \|\mathbf{x}+\mathbf{y}\|_2^2 - 4\|\mathbf{z}\|_2^2 + 4\langle \mathbf{x}+\mathbf{y}, \mathbf{z}\rangle \\ =& 2\|\mathbf{x}-\mathbf{z}\|_2^2 + 2\|\mathbf{y}-\mathbf{z}\|_2^2 - 4\|\frac{1}{2}(\mathbf{x}+\mathbf{y}) - \mathbf{z}\|_2^2\end{aligned}$$

More identities II

Lemma (Cosine rule)

$$2\langle \mathbf{z}-\mathbf{x}, \mathbf{y}-\mathbf{z}\rangle = \|\mathbf{y}-\mathbf{x}\|_2^2 - \|\mathbf{z}-\mathbf{x}\|_2^2 - \|\mathbf{y}-\mathbf{z}\|_2^2$$

Lemma (Three-point identity)

$$\begin{aligned} 2\langle \mathbf{z}-\mathbf{x}, \mathbf{y}\rangle =& \|\mathbf{y}-\mathbf{x}\|_2^2 - \|\mathbf{z}-\mathbf{x}\|_2^2 - \|\mathbf{y}-\mathbf{z}\|_2^2 + 2\langle \mathbf{z}-\mathbf{x}, \mathbf{z}\rangle \\ =& \|\mathbf{y}-\mathbf{x}\|_2^2 - \|\mathbf{z}-\mathbf{x}\|_2^2 - \|\mathbf{y}-\mathbf{z}\|_2^2 + 2\langle \mathbf{z}-\mathbf{x}, \mathbf{z}\rangle \\ =& \|\mathbf{y}-\mathbf{x}\|_2^2 - \|\mathbf{y}-\mathbf{z}\|_2^2 + \|\mathbf{z}\|_2^2 - \|\mathbf{x}\|_2^2 \end{aligned}$$

Lemma (Four-point identity)

$$\begin{aligned} 2\langle \mathbf{z}-\mathbf{x}, \mathbf{y}-\mathbf{w}\rangle =& \|\mathbf{y}-\mathbf{x}\|_2^2 - \|\mathbf{y}-\mathbf{z}\|_2^2 + \|\mathbf{z}\|_2^2 - \|\mathbf{x}\|_2^2 \\ & - \|\mathbf{w}-\mathbf{x}\|_2^2 + \|\mathbf{w}-\mathbf{z}\|_2^2 - \|\mathbf{z}\|_2^2 + \|\mathbf{x}\|_2^2 \\ =& \|\mathbf{y}-\mathbf{x}\|_2^2 - \|\mathbf{w}-\mathbf{x}\|_2^2 - \|\mathbf{y}-\mathbf{z}\|_2^2 + \|\mathbf{w}-\mathbf{z}\|_2^2 \end{aligned}$$

What is a norm?

Norm properties

- **Absolute homogeneity/scalability:** $\|\alpha \mathbf{x}\| = |\alpha| \|\mathbf{x}\|$ for $\mathbf{x} \in \mathcal{V}$ and $\alpha \in \mathbb{R}/\mathbb{C}$
- **Triangle inequality or subadditivity:** $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$ for $\mathbf{x}, \mathbf{y} \in \mathcal{V}$
- **Separates points:** If $\|\mathbf{x}\| = 0$, then $\mathbf{x} = \mathbf{0}$

- From the absolute homogeneity, we have $\|\mathbf{0}\| = 0$ and $\|\mathbf{x}\| = \| - \mathbf{x}\|$, therefore $2\|\mathbf{x}\| = \|\mathbf{x}\| + \| - \mathbf{x}\| \geq \|\mathbf{x} + (-\mathbf{x})\| = 0$.
- A **seminorm** is a function that satisfies absolute homogeneity and triangle inequality.
- Two norms (or seminorms) $\| \cdot \|_p$ and $\| \cdot \|_q$ on a vector space $\mathcal{V}$ are equivalent if there exist two real constants c and C, with $c > 0$ such that

$$c\|\mathbf{x}\|_q \leq \|\mathbf{x}\|_p \leq C\|\mathbf{x}\|_q \quad \forall \mathbf{x} \in \mathcal{V}.$$

Norm in $\mathbb{R}^n$?

- $\|\mathbf{x}\|_2$
- $\|\mathbf{x}\|_0$: the number of non-zeros in $\mathbf{x}$
- $\|\mathbf{x}\|_1$: taxicab norm or Manhattan norm
- $\|\nabla \mathbf{x}\|_2 = \sqrt{\sum_{i=1}^{n-1} (x_{i+1} - x_i)^2}$
- $\sqrt{\mathbf{x}^\top \mathbf{A} \mathbf{x}}$ for some matrix $\mathbf{A}$
- ...

ℓ_p **norm**

Definition (ℓ_p norm in $\mathbb{R}^n/\mathbb{C}^n$)

When $p \geq 1$, $\|\mathbf{x}\|_p = \left(\sum_{i=1}^n |x_i|^p\right)^{1/p}$ for $\mathbf{x} \in \mathbb{R}^n/\mathbb{C}^n$.

- ℓ_2 norm: $\|\mathbf{x}\|_2 = \sqrt{\sum_{i=1}^n |x_i|^2}$
- ℓ_1 norm: $\|\mathbf{x}\|_1 = \sum_{i=1}^n |x_i|$
- ℓ_∞ norm: $\|\mathbf{x}\|_\infty = \max_{i=1}^n |x_i|$
- $\|\mathbf{x}\|_\infty \leq \|\mathbf{x}\|_2 \leq \sqrt{n}\|\mathbf{x}\|_\infty$
- $\|\mathbf{x}\|_\infty \leq \|\mathbf{x}\|_1 \leq n\|\mathbf{x}\|_\infty$
- $\|\mathbf{x}\|_2 \leq \|\mathbf{x}\|_1 \leq \sqrt{n}\|\mathbf{x}\|_2$
- $\|\mathbf{x}\|_q \leq \|\mathbf{x}\|_p \leq n^{1/p-1/q}\|\mathbf{x}\|_q$ for $1 \leq p < q$

Holder's inequality

Lemma (Holder's inequality)

$|\langle \mathbf{x}, \mathbf{y} \rangle| \leq \|\mathbf{x}\|_p \|\mathbf{y}\|_q$ *with* $1/p + 1/q = 1$ *and* $p, q \in [1, \infty]$

- $\| \cdot \|_p$ and $\| \cdot \|_q$ are dual norms.

Collection of vectors, subspaces

A set of m vectors, $\mathbf{V} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_m\}$

- **Linear combination**: $\sum_{j=1}^{m} \alpha_j \mathbf{x}_j$, $\alpha_j \in \mathbb{R}/\mathbb{C}$.
- **Linearly independent**: No vector in $\mathbf{V}$ can be written as linear combination of other. If $\mathbf{0} = \sum_{j=1}^{m} \alpha_j \mathbf{x}_j$, then $\alpha_j = 0$ for all $j = 1, \cdots, m$.
- **Span**: $\mathrm{Span}(\mathbf{V}) = \{\mathbf{x} | \mathbf{x} = \sum_{j=1}^{m} \alpha_j \mathbf{x}_j, \alpha_j \in \mathbb{R}/\mathbb{C}\}$.

Definition (Subspace)

A collection of vectors $\mathbf{V} \subset \mathbb{R}^n/\mathbb{C}^n$ is a subspace iff it is closed under linear combination, i.e.,

$$\mathbf{x}, \mathbf{y} \in \mathbf{V} \Rightarrow \alpha\mathbf{x} + \beta\mathbf{y} \in \mathbf{V}, \forall \alpha, \beta \in \mathbb{R}/\mathbb{C}$$

- **Basis of a subspace**: A linearly independent spanning set
- **Dimensionality of a subspace**: the number of elements in a basis

Matrix

- $\mathbf{A} \in \mathbb{R}^{m\times n}$: A matrix of dimension $m \times n$.
- $\mathbf{A} = [a_{ij}] = [\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_b]$, here $\mathbf{a}_i \in \mathbb{R}^m/\mathbb{C}^m$
- $\mathrm{Rank}(\mathbf{A})$ is the largest number of linearly independent columns, which is equivalent to the largest number of linear independent rows.
- $\mathrm{Rank}(\mathbf{A}) = \mathrm{Rank}(\mathbf{A}^H) \leq \min(m, n)$
- $\mathbf{A}$ is full-rank if $\mathrm{Rank}(\mathbf{A}) = \min(m, n)$.
- $\mathbf{A}$ is full-row-rank if $\mathrm{Rank}(\mathbf{A}) = m$ and full-column-rank if $\mathrm{Rank}(\mathbf{A}) = n$.

Matrices are representations of linear operators.

$$\mathbf{A} : \mathbb{R}^n/\mathbb{C}^n \to \mathbb{R}^m/\mathbb{C}^m \qquad \mathbf{x} \in \mathbb{R}^n/\mathbb{C}^m \mapsto \mathbf{A}\mathbf{x} \in \mathbb{R}^m/\mathbb{C}^m.$$

Examples of linear operators that aren't matrices?

Matrix structure and algorithm complexity

cost (execution time) of solving $\mathbf{Ax} = \mathbf{b}$ with $\mathbf{A} \in \mathbf{R}^{n \times n}$

- n^3 for general methods
- less if $\mathbf{A}$ is structured (banded, sparse, Toeplitz, ...)

Flop counts

- flop (floating-point operation): one addition, subtraction, multiplication, or division of two floating-point numbers
- to estimate complexity of an algorithm: express number of flops as a (polynomial) function of the problem dimensions, and simplify by keeping only the leading terms
- not an accurate predictor of computation time on modern computers
- useful as a rough estimate of complexity

vector-vector operations ($\mathbf{x}, \mathbf{y} \in \mathbf{R}^n$)

- inner product $\mathbf{x}^\top \mathbf{y}$: $2n - 1$ flops (or $2n$ if n is large)
- sum $\mathbf{x} + \mathbf{y}$, scalar multiplication $\alpha \mathbf{x}$: n flops

matrix-vector product $\mathbf{y} = \mathbf{A}\mathbf{x}$ with $\mathbf{A} \in \mathbf{R}^{m \times n}$

- $m(2n - 1)$ flops (or $2mn$ if n large)
- $2N$ if $\mathbf{A}$ is sparse with N nonzero elements
- $2p(n + m)$ if $\mathbf{A}$ is given as $\mathbf{A} = \mathbf{U}\mathbf{V}^\top$, $U \in \mathbf{R}^{m \times p}$, $\mathbf{V} \in \mathbf{R}^{n \times p}$

matrix-matrix product $\mathbf{C} = \mathbf{A}\mathbf{B}$ with $\mathbf{A} \in \mathbf{R}^{m \times n}$, $\mathbf{B} \in \mathbf{R}^{n \times p}$

- $mp(2n - 1)$ flops (or $2mnp$ if n large)
- less if $\mathbf{A}$ and/or $\mathbf{B}$ are sparse
- $(1/2)m(m + 1)(2n - 1) \approx m^2 n$ if $m = p$ and $\mathbf{C}$ symmetric

Linear equations that are easy to solve $\mathbf{Ax} = \mathbf{b}$

- **diagonal matrices** ($a_{ij} = 0$ if $i \neq j$ and $a_{ii} \neq 0$ for all i): n flops

$$\mathbf{x} = \mathbf{A}^{-1}\mathbf{b} = [b_1/a_{11}; b_2/a_{22}; \cdots; b_n/a_{nn}]$$

- **lower triangular** ($a_{ij} = 0$ if $j > i$ and $a_{ii} \neq 0$ for all i): n^2 flops

$$\begin{aligned} x_1 &= b_1/a_{11} \\ x_2 &= (b_2 - a_{21}x_1)/a_{22} \\ &\cdots \\ x_n &= (b_n - a_{n1}x_1 - \cdots - a_{n,n-1}x_{n-1})/a_{nn} \end{aligned}$$

  called forward substitution

- **upper triangular** ($a_{ij} = 0$ if $j < i$ and $a_{ii} \neq 0$ for all i): n^2 flops via backward substitution

Special matrix for $\mathbf{Ax} = \mathbf{b}$

- **orthogonal matrices** $(\mathbf{A}^{-1} = \mathbf{A}^H)$
  - $2n^2$ flops to compute $\mathbf{x} = \mathbf{A}^H\mathbf{b}$ for general $\mathbf{A}$
  - less with structure, e.g., if $\mathbf{A} = \mathbf{I} - 2\mathbf{u}\mathbf{u}^\top$ with $\|\mathbf{u}\|_2 = 1$, we can compute $\mathbf{x} = \mathbf{A}^\top\mathbf{b} = \mathbf{b} - 2(\mathbf{u}^\top\mathbf{b})\mathbf{u}$ in $4n$ flops
  - permutation matrices:

$$a_{ij} = \begin{cases} 1 & j = \pi_i \\ 0 & \text{otherwise} \end{cases}$$

    where $\pi = (\pi_1, \pi_2, \cdots, \pi_n)$ is a permutation of $(1, 2, \cdots, n)$.

    - interpretation: $\mathbf{Ax} = (x_{\pi_1}, \ldots, x_{\pi_n})$
    - cost of solving $\mathbf{Ax} = \mathbf{b}$ is 0 flops

    example:

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}, \quad \mathbf{A}^{-1} = \mathbf{A}^\top = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$$

- **orthogonal matrices** ($\mathbf{A}^{-1} = \mathbf{A}^H$)
  - Discrete Fourier transform: $n\log(n)$ for FFT

$$W = \frac{1}{\sqrt{n}} \begin{bmatrix} 1 & 1 & 1 & 1 & \cdots & 1 \\ 1 & \omega & \omega^2 & \omega^3 & \cdots & \omega^{n-1} \\ 1 & \omega^2 & \omega^4 & \omega^6 & \cdots & \omega^{2(n-1)} \\ 1 & \omega^3 & \omega^6 & \omega^9 & \cdots & \omega^{3(n-1)} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \omega^{n-1} & \omega^{2(n-1)} & \omega^{3(n-1)} & \cdots & \omega^{(n-1)(n-1)} \end{bmatrix},$$

    where $\omega = e^{-2\pi i/n}$
  - Discrete Wavelet transform: only $O(n)$ in certain cases

The factor-solve method for solving $\mathbf{Ax} = \mathbf{b}$

- factor $\mathbf{A}$ as a product of simple matrices (usually 2 or 3):

$$\mathbf{A} = \mathbf{A}_1 \mathbf{A}_2 \cdots \mathbf{A}_k$$

  (A_i diagonal, upper or lower triangular, etc)

- compute $\mathbf{x} = \mathbf{A}^{-1}\mathbf{b} = \mathbf{A}_k^{-1} \cdots \mathbf{A}_2^{-1}\mathbf{A}_1^{-1}\mathbf{b}$ by solving k 'easy' equations

$$\mathbf{A}_1\mathbf{x}_1 = b, \quad \mathbf{A}_2\mathbf{x}_2 = \mathbf{x}_1, \quad \ldots, \quad \mathbf{A}_k\mathbf{x} = \mathbf{x}_{k-1}$$

  cost of factorization step usually dominates cost of solve step

equations with multiple righthand sides

$$\mathbf{Ax}_1 = b_1, \quad \mathbf{Ax}_2 = \mathbf{b}_2, \quad \ldots, \quad \mathbf{Ax}_m = \mathbf{b}_m$$

cost: one factorization plus m solves

LU factorization

every nonsingular matrix $\mathbf{A}$ can be factored as

$$\mathbf{A} = \mathbf{PLU}$$

with $\mathbf{P}$ a permutation matrix, $\mathbf{L}$ lower triangular, $\mathbf{U}$ upper triangular
cost: $(2/3)n^3$ flops

$$\mathbf{x} = \mathbf{A}^{-1}\mathbf{b} = \mathbf{U}^{-1}\mathbf{L}^{-1}\mathbf{P}^{-1}\mathbf{b}$$

1. LU factorization: $(2/3)n^3$
2. Permutation: 0
3. Lower triangular: n^2
4. Upper triangular: n^2

Total costs: $(2/3)n^3 + 0 + n^2 + n^2 \approx (2/3)n^3$ flops for large n.

sparse LU factorization

$$\mathbf{A} = \mathbf{P}_1\mathbf{L}\mathbf{U}\mathbf{P}_2$$

- adding permutation matrix $\mathbf{P}_2$ offers possibility of sparser $\mathbf{L}$, $\mathbf{U}$ (hence, cheaper factor and solve steps)
- $\mathbf{P}_1$ and $\mathbf{P}_2$ chosen (heuristically) to yield sparse $\mathbf{L}$, $\mathbf{U}$
- choice of $\mathbf{P}_1$ and $\mathbf{P}_2$ depends on sparsity pattern and values of $\mathbf{A}$
- cost is usually much less than $(2/3)n^3$; exact value depends in a complicated way on n, number of zeros in $\mathbf{A}$, sparsity pattern

Cholesky factorization

every symmetric/Hermitian positive definite matrix $\mathbf{A}$ can be factored as

$$\mathbf{A} = \mathbf{L}\mathbf{L}^H$$

with $\mathbf{L}$ lower triangular.
cost: $(1/3)n^3$ flops

$$\mathbf{x} = \mathbf{A}^{-1}\mathbf{b} = (\mathbf{L}\mathbf{L}^H)^{-1}\mathbf{b} = \mathbf{L}^{-H}\mathbf{L}^{-1}\mathbf{b}$$

1. Cholesky factorization: $(1/3)n^3$
2. Lower triangular: n^2
3. Upper triangular: n^2

Total costs: $(1/3)n^3 + 0 + n^2 + n^2 \approx (1/3)n^3$ flops for large n.

sparse Cholesky factorization

$$\mathbf{A} = \mathbf{P}\mathbf{L}\mathbf{L}^H\mathbf{P}^\top$$

- adding permutation matrix $\mathbf{P}$ offers possibility of sparser $\mathbf{L}$
- $\mathbf{P}$ chosen (heuristically) to yield sparse $\mathbf{L}$
- choice of $\mathbf{P}$ only depends on sparsity pattern of $\mathbf{A}$ (unlike sparse $\mathbf{LU}$)

$\mathbf{LDL}^\top$ factorization

every nonsingular symmetric/Hermitian matrix $\mathbf{A}$ can be factored as

$$\mathbf{A} = \mathbf{PLDL}^H\mathbf{P}^\top$$

with $\mathbf{P}$ a permutation matrix, $\mathbf{L}$ lower triangular, $\mathbf{D}$ block diagonal with 1×1 or 2×2 diagonal blocks.
cost: $(1/3)n^3$

$$\mathbf{x} = \mathbf{A}^{-1}\mathbf{b} = (\mathbf{PLDL}^H\mathbf{P}^\top)^{-1}\mathbf{b} = \mathbf{P}^{-\top}\mathbf{L}^{-H}\mathbf{D}^{-1}\mathbf{L}^{-1}\mathbf{P}^{-1}\mathbf{b}$$

- total costs: $(1/3)n^3 + 0 + n^2 + n + n^2 + 0 \approx (1/3)n^3$ flops for large n
- for sparse $\mathbf{A}$, can choose $\mathbf{P}$ to yield sparse $\mathbf{L}$; cost $\ll (1/3)n^3$

Equations with structured sub-blocks

$$\begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \end{bmatrix}$$

- variables $\mathbf{x}_1 \in \mathbf{R}^{n_1}$, $\mathbf{x}_2 \in \mathbf{R}^{n_2}$; blocks $\mathbf{A}_{ij} \in \mathbf{R}^{n_i \times n_j}$
- if $\mathbf{A}_{11}$ is nonsingular, we can eliminate $\mathbf{x}_1$:

$$\Rightarrow \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ & \mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12} \end{bmatrix} \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{b}_1 \end{bmatrix}$$

1. Form $\mathbf{A}_{11}^{-1}\mathbf{A}_{12}$ and $\mathbf{A}_{11}^{-1}\mathbf{b}_1$
2. Form $\mathbf{S} = \mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}$ and $\tilde{\mathbf{b}} = \mathbf{b}_2 - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{b}_1$
3. Determine $\mathbf{x}_2$ by solving $\mathbf{S}\mathbf{x}_2 = \tilde{\mathbf{b}}$
4. Determine $\mathbf{x}_1$ by solving $\mathbf{A}_{11}\mathbf{x}_1 = \mathbf{b}_1 - \mathbf{A}_{12}\mathbf{x}_2$

dominant terms in flop count

- step 1+4: $f + n_2 s$ (f is cost of factoring A11; s is cost of solve step)
- step 2: $2n_2^2 n_1$ (cost dominated by product of $\mathbf{A}_{21}$ and $\mathbf{A}_{11}^{-1}\mathbf{A}_{12}$)
- step 3: $(2/3)n_2^3$

total: $f + n_2 s + 2n_2^2 n_1 + (2/3)n_2^3$

examples

- general $\mathbf{A}_{11}$ ($f = (2/3)n_1^3$, $s = 2n_1^2$): no gain over standard method

$$\#\text{flops} = (2/3)n_1^3 + 2n_1^2 n_2 + 2n_2^2 n_1 + (2/3)n_2^3 = (2/3)(n_1 + n_2)^3$$

- block elimination is useful for structured $\mathbf{A}_{11}$ ($f \ll n_1^3$)
  for example, diagonal ($f = 0$, $s = n_1$): # flops$\approx 2n_2^2 n_1 + (2/3)n_2^3$

Structured matrix plus low rank term

$$(\mathbf{A} + \mathbf{BC})\mathbf{x} = \mathbf{b}$$

where $\mathbf{A} \in \mathbb{R}^{n \times n}$, $\mathbf{B} \in \mathbb{R}^{n \times p}$, and $\mathbf{C} \in \mathbb{R}^{p \times n}$. Assume that $\mathbf{A}$ has structure such that $\mathbf{Ax} = \mathbf{b}$ is easy to solve. We can rewrite is as

$$\begin{bmatrix} \mathbf{A} & \mathbf{B} \\ -\mathbf{C} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} = \begin{bmatrix} \mathbf{b} \\ \mathbf{0} \end{bmatrix}$$

$$\Rightarrow \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ & \mathbf{I} + \mathbf{CA}^{-1}\mathbf{B} \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} = \begin{bmatrix} \mathbf{b} \\ \mathbf{0} + \mathbf{CA}^{-1}\mathbf{b} \end{bmatrix}$$

$$(\mathbf{A} + \mathbf{BC})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{B}(\mathbf{I} + \mathbf{CA}^{-1}\mathbf{B})^{-1}\mathbf{CA}^{-1}$$

example: $\mathbf{A}$ diagonal, $\mathbf{B}$, $\mathbf{C}$ dense

- method 1: form $\mathbf{D} = \mathbf{A} + \mathbf{BC}$, then solve $\mathbf{Dx} = \mathbf{b}$
  cost: $(2/3)n^3 + 2pn^2$
- method 2 (via matrix inversion lemma): solve

$$(\mathbf{I} + \mathbf{CA}^{-1}\mathbf{B})\mathbf{y} = \mathbf{CA}^{-1}\mathbf{b},$$

  then compute $\mathbf{x} = \mathbf{A}^{-1}\mathbf{b} - \mathbf{A}^{-1}\mathbf{By}$
  cost: $2p^2n + (2/3)p^3$ (i.e., linear in n)

Underdetermined linear equations

if $\mathbf{A} \in \mathbf{R}^{p \times n}$ with $p < n$, $rank \mathbf{A} = p$,

$$\{\mathbf{x} | \mathbf{A}\mathbf{x} = \mathbf{b}\} = \{\mathbf{F}\mathbf{z} + \hat{\mathbf{x}} | \mathbf{z} \in \mathbf{R}^{n-p}\}$$

- $\hat{\mathbf{x}}$ is (any) particular solution
- columns of $\mathbf{F} \in \mathbf{R}^{n \times (n-p)}$ span nullspace of $\mathbf{A}$
- there exist several numerical methods for computing $\mathbf{F}$ (QR factorization, rectangular LU factorization, . . .)

Eigenvalues and eigenvectors

Definition (Eigenvalues and eigenvectors)

Let $\mathbf{A}$ be a $n \times n$ square matrix, λ is an eigenvalue of $\mathbf{A}$ if

$$\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$$

for some nonzero $\mathbf{x}$, and $\mathbf{x}$ is the corresponding eigenvectors.

- **Intuition:** eigenvectors are vectors in $\mathbb{R}^n/\mathbb{C}^n$ whose direction is preserved under action of $\mathbf{A}$; however, length may change.
- $\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{U}^{-1}$ where $\mathbf{D}$ is diagonal, then the diagonal entries of $\mathbf{D}$ are eigenvalues and the columns of $\mathbf{U}$ are eigenvectors.

Spectral theorem

Theorem (Spectral Theorem)

If $\mathbf{A} = \mathbf{A}^H$*, then*

- *The matrix is symmetric/Hermitian ,*
- *all eigenvalues are real,*
- *eigenvectors with different eigenvalues are perpendicular*
- *there exists a complete orthogonal basis of eigenvectors*

Singular value decomposition

Definition (SVD)

Any matrix $\mathbf{A} \in \mathbb{R}^{m\times n}/\mathbb{C}^{m\times n}$ can be written as

$$\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^H,$$

where $\mathbf{U} \in \mathbb{R}^{m\times m}/\mathbb{C}^{m\times m}$ and $\mathbf{V} \in \mathbb{R}^{n\times n}/\mathbb{C}^{n\times n}$ are unitary and $\Sigma \in \mathbb{R}^{m\times n}$ is diagonal.

- $\mathbf{U}\mathbf{U}^H = \mathbf{U}^H\mathbf{U} = \mathbf{I}$ and $\mathbf{V}\mathbf{V}^H = \mathbf{V}^H\mathbf{V} = \mathbf{I}$ (unitary)
- Diagonal entries of Σ are called the singular values; they are positive and real. Typically, $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$, where r is the rank of $\mathbf{A}$.
- $\mathbf{A}^H\mathbf{A} = \mathbf{V}\Sigma^\top\mathbf{U}^H\mathbf{U}\Sigma\mathbf{V}^H = \mathbf{V}\Sigma^\top\Sigma\mathbf{V}^H$. Therefore, $\sqrt{\mathbf{A}^H\mathbf{A}} = \mathbf{V}\sqrt{\Sigma^\top\Sigma}\mathbf{V}^H$
- Singular values are the eigenvalues of $\sqrt{\mathbf{A}^H\mathbf{A}}$ and $\sqrt{\mathbf{A}\mathbf{A}^H}$.
- The columns of $\mathbf{V}$ are eigenvectors of $\mathbf{A}^H\mathbf{A}$, and the columns of $\mathbf{U}$ are eigenvectors of $\mathbf{A}\mathbf{A}^H$.
- If $\mathbf{A} = \mathbf{A}^H$, singular values are the same as the eigenvalues
- Geometric picture and other properties, read Wikipedia

Singular value decomposition

Definition (SVD2)

Any matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ can be written as

$$\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^\top,$$

where $\mathbf{U} \in \mathbb{R}^{m \times r}$ and $\mathbf{V} \in \mathbb{R}^{n \times r}$ are unitary and $\Sigma \in \mathbb{R}^{r \times r}$ is diagonal.

- If $\mathbf{A}^{-1}$ exists, then $\mathbf{A}^{-1} = \mathbf{V}\Sigma^{-1}\mathbf{U}^\top$
- Even if $\mathbf{A}$ is singular, we can deïňĄne a pseudo-inverse $\mathbf{A}^*$ as follows:
- The ratio of the largest to smallest singular value is the so-called condition number of $\mathbf{A}$

Matrix norm

Definition (Spectral norm)

$$\|\mathbf{A}\|_{2,2} = \max_{\mathbf{x}\neq\mathbf{0}} \frac{\|\mathbf{A}\mathbf{x}\|_2}{\|\mathbf{x}\|_2} = \max_{\|\mathbf{x}\|_2=1} \|\mathbf{A}\mathbf{x}\|_2$$

- The norm is call the induced norm or the ℓ_2-norm
- Quantifies the maximum increase in length of unit-norm vectors due to the operation of the matrix $\mathbf{A}$
- $\|\mathbf{A}\|_{2,2}$ is equal to the largest singular value of $\mathbf{A}$
- $\|\mathbf{A}\mathbf{x}\|_2 \leq \|\mathbf{A}\|_{2,2}\|\mathbf{x}\|_2$ (Q: When is it equal?)

Lemma

$$\|\mathbf{A}\mathbf{B}\|_{2,2} \leq \|\mathbf{A}\|_{2,2}\|\mathbf{B}\|_{2,2}$$

Induced matrix norm

Definition

$$\|\mathbf{A}\|_{p,q} = \max_{\mathbf{z}\neq\mathbf{0}} \frac{\|\mathbf{A}\mathbf{x}\|_q}{\|\mathbf{x}\|_p} = \max_{\|\mathbf{x}\|_p=1} \|\mathbf{A}\mathbf{x}\|_q$$

- $\|\mathbf{A}\|_{2,2}$: the maximum singular value of $\mathbf{A}$.
- $\|\mathbf{A}\|_{1,1}$: the maximum of the absolute column sums.
- $\|\mathbf{A}\|_{\infty,\infty}$: the maximum of the absolute row sums.

- $\|\mathbf{A}\mathbf{x}\|_q \leq \|\mathbf{A}\|_{p,q}\|\mathbf{x}\|_p$
- $\|\mathbf{A}\|_{2,2}^2 \leq \|\mathbf{A}\|_{1,1}\|\mathbf{A}\|_{\infty,\infty}$

Other frequently-used matrix norms

Definition (Frobenius norm)

$$\|\mathbf{A}\|_F = \sqrt{\sum_{i,j} |a_{ij}|^2} = \sqrt{\mathrm{tr}(\mathbf{A}^H\mathbf{A})} = \sqrt{\mathrm{tr}(\mathbf{A}\mathbf{A}^H)}$$

Definition (Nuclear norm)

$$\|\mathbf{A}\|_* = \mathrm{tr}(\sqrt{\mathbf{A}^H\mathbf{A}}) = \sum_{i=1}^{\min(m,n)} \sigma_i$$

where σ_i are the singular values of the matrix $\mathbf{A}$.

- All matrix norms are equivalent.

Derivatives with vectors I

Vector-by-scalar $\mathbf{y} \in \mathbf{R}^n$

$$\frac{\partial \mathbf{y}}{\partial x} = \left[\frac{\partial y_1}{\partial x}, \frac{\partial y_2}{\partial x}, \cdots, \frac{\partial y_n}{\partial x}\right]$$

- $\frac{\partial x\mathbf{a}}{\partial x} = \mathbf{a}^\top$

Scalar-by-vector $\mathbf{x} \in \mathbf{R}^n$

$$\frac{\partial y}{\partial \mathbf{x}} = \left[\frac{\partial y}{\partial x_1}; \frac{\partial y}{\partial x_2}; \cdots; \frac{\partial y}{\partial x_n}\right]$$

- $\nabla_{\mathbf{u}} f(\mathbf{x}) = \nabla f(\mathbf{x})^\top \mathbf{u}$

Derivatives with vectors II

Vector-by-vector $\mathbf{x} \in \mathbf{R}^n$ and $\mathbf{y} \in \mathbf{R}^m$

$$\frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_2}{\partial x_1} & \cdots & \frac{\partial y_m}{\partial x_1} \\ \frac{\partial y_1}{\partial x_2} & \frac{\partial y_2}{\partial x_2} & \cdots & \frac{\partial y_m}{\partial x_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial y_1}{\partial x_n} & \frac{\partial y_2}{\partial x_n} & \cdots & \frac{\partial y_m}{\partial x_n} \end{bmatrix}$$

- $\dfrac{\partial \mathbf{A}\mathbf{x}}{\partial \mathbf{x}} = \mathbf{A}^\top$
- $\dfrac{\partial (\mathbf{u}(\mathbf{x}) \cdot \mathbf{a})}{\partial \mathbf{x}} = \dfrac{\partial (\mathbf{u}(\mathbf{x})^\top \mathbf{a})}{\partial \mathbf{x}} = \frac{\partial \mathbf{u}(\mathbf{x})}{\partial \mathbf{x}} \mathbf{a}$

Vector-by-vector identities

- $\mathbf{a}$ is not a function of $\mathbf{x}$: $\dfrac{\partial \mathbf{a}}{\partial \mathbf{x}} = \mathbf{0}$
- $\dfrac{\partial \mathbf{x}}{\partial \mathbf{x}} = \mathbf{I}$
- $\mathbf{A}$ is not a function of $\mathbf{x}$: $\dfrac{\partial \mathbf{A}\mathbf{x}}{\mathbf{x}} = \mathbf{A}^\top$
- $\mathbf{A}$ is not a function of $\mathbf{x}$: $\dfrac{\partial \mathbf{x}^\top \mathbf{A}}{\partial \mathbf{x}} = \dfrac{\partial \mathbf{A}^\top \mathbf{x}}{\mathbf{x}} = \mathbf{A}$
- a is not a function of $\mathbf{x}$, $\mathbf{u} = \mathbf{u}(\mathbf{x})$: $\dfrac{\partial a\mathbf{u}}{\partial \mathbf{x}} = a\dfrac{\partial \mathbf{u}}{\partial \mathbf{x}}$
- $\dfrac{\partial (\mathbf{u} + \mathbf{v})}{\partial \mathbf{x}} = \dfrac{\partial \mathbf{u}}{\partial \mathbf{x}} + \dfrac{\partial \mathbf{v}}{\partial \mathbf{x}}$
- $a = a(\mathbf{x})$, $\mathbf{u} = \mathbf{u}(\mathbf{x})$: $\dfrac{\partial a\mathbf{u}}{\partial \mathbf{x}} = a\dfrac{\partial \mathbf{u}}{\partial \mathbf{x}} + \dfrac{\partial a}{\partial \mathbf{x}}\mathbf{u}^\top$
- $\mathbf{u} = \mathbf{u}(\mathbf{x})$: $\dfrac{\partial \mathbf{g}(\mathbf{u})}{\partial \mathbf{x}} = \dfrac{\partial \mathbf{u}}{\partial \mathbf{x}}\dfrac{\partial \mathbf{g}(\mathbf{u})}{\partial \mathbf{u}}$
- $\mathbf{A}$ is not a function of $\mathbf{x}$, $\mathbf{u} = \mathbf{u}(\mathbf{x})$: $\dfrac{\partial \mathbf{A}\mathbf{u}}{\partial \mathbf{x}} = \dfrac{\partial \mathbf{u}}{\partial \mathbf{x}}\mathbf{A}^\top$
- $\mathbf{u} = \mathbf{u}(\mathbf{x})$: $\dfrac{\partial \mathbf{f}(\mathbf{g}(\mathbf{u}))}{\partial \mathbf{x}} = \dfrac{\partial \mathbf{u}}{\partial \mathbf{x}}\dfrac{\partial \mathbf{g}(\mathbf{u})}{\partial \mathbf{u}}\dfrac{\partial \mathbf{f}(\mathbf{g})}{\partial \mathbf{g}}$

Scalar-by-vector identities

- a is not a function of $\mathbf{x}$: $\dfrac{\partial a}{\partial \mathbf{x}} = \mathbf{0}$
- a is not a function of $\mathbf{x}$, $u = u(\mathbf{x})$: $\dfrac{\partial au}{\partial \mathbf{x}} = a\dfrac{\partial u}{\partial \mathbf{x}}$
- $u = u(\mathbf{x})$, $v = v(\mathbf{x})$: $\dfrac{\partial(u+v)}{\partial \mathbf{x}} = \dfrac{\partial u}{\partial \mathbf{x}} + \dfrac{\partial v}{\partial \mathbf{x}}$
- $u = u(\mathbf{x})$, $v = v(\mathbf{x})$: $\dfrac{\partial uv}{\partial \mathbf{x}} = \dfrac{\partial u}{\partial \mathbf{x}}v + u\dfrac{\partial v}{\partial \mathbf{x}}$
- $u = u(\mathbf{x})$: $\dfrac{\partial g(u)}{\partial \mathbf{x}} = \dfrac{\partial u}{\partial \mathbf{x}}\dfrac{\partial g(u)}{\partial u}$
- $u = u(\mathbf{x})$: $\dfrac{\partial f(g(u))}{\partial \mathbf{x}} = \dfrac{\partial u}{\partial \mathbf{x}}\dfrac{\partial g(u)}{\partial u}\dfrac{\partial f(g)}{\partial g}$
- $\mathbf{u} = \mathbf{u}(\mathbf{x})$, $\mathbf{v} = \mathbf{v}(\mathbf{x})$: $\dfrac{\partial(\mathbf{u}^\top\mathbf{v})}{\partial \mathbf{x}} = \dfrac{\partial \mathbf{u}}{\partial \mathbf{x}}\mathbf{v} + \dfrac{\partial \mathbf{v}}{\partial \mathbf{x}}\mathbf{u}$
- $\mathbf{u} = \mathbf{u}(\mathbf{x})$, $\mathbf{v} = \mathbf{v}(\mathbf{x})$, $\mathbf{A}$ is not a function of $\mathbf{x}$:
  $\dfrac{\partial(\mathbf{u}^\top\mathbf{A}\mathbf{v})}{\partial \mathbf{x}} = \dfrac{\partial \mathbf{u}}{\partial \mathbf{x}}\mathbf{A}\mathbf{v} + \dfrac{\partial \mathbf{v}}{\partial \mathbf{x}}\mathbf{A}^\top\mathbf{u}$

Scalar-by-vector identities

- $\mathbf{a}$ is not a function of $\mathbf{x}$: $\dfrac{\partial \mathbf{a}^\top \mathbf{x}}{\partial \mathbf{x}} = \dfrac{\partial \mathbf{x}^\top \mathbf{a}}{\partial \mathbf{x}} = \mathbf{a}$
- $\mathbf{A}$ is not a function of $\mathbf{x}$, $\mathbf{b}$ is not a function of $\mathbf{x}$: $\dfrac{\partial \mathbf{b}^\top \mathbf{A}\mathbf{x}}{\partial \mathbf{x}} = \mathbf{A}^\top \mathbf{b}$
- $\dfrac{\partial \mathbf{x}^\top \mathbf{x}}{\partial \mathbf{x}} = 2\mathbf{x}$
- $\mathbf{A}$ is not a function of $\mathbf{x}$: $\dfrac{\partial \mathbf{x}^\top \mathbf{A}\mathbf{x}}{\partial \mathbf{x}} = (\mathbf{A} + \mathbf{A}^\top)\mathbf{x}$
- $\mathbf{A}$ is not a function of $\mathbf{x}$: $\dfrac{\partial^2 \mathbf{x}^\top \mathbf{A}\mathbf{x}}{\partial \mathbf{x}^2} = \mathbf{A} + \mathbf{A}^\top$
- $\mathbf{a}$ is not a function of $\mathbf{x}$, $u = u(\mathbf{x})$: $\dfrac{\partial \mathbf{a}^\top \mathbf{u}}{\partial \mathbf{x}} = \dfrac{\partial \mathbf{u}^\top \mathbf{a}}{\partial \mathbf{x}} = \dfrac{\partial \mathbf{u}}{\partial \mathbf{x}}\mathbf{a}$
- $\mathbf{a}$, $\mathbf{b}$ are not functions of $\mathbf{x}$: $\dfrac{\partial \mathbf{a}^\top \mathbf{x}\mathbf{x}^\top \mathbf{b}}{\partial \mathbf{x}} = (\mathbf{a}\mathbf{b}^\top + \mathbf{b}\mathbf{a}^\top)\mathbf{x}$
- $\mathbf{A}$, $\mathbf{b}$, $\mathbf{C}$, $\mathbf{D}$, $\mathbf{e}$ are not functions of $\mathbf{x}$:
  $\dfrac{\partial (\mathbf{A}\mathbf{x} + \mathbf{b})^\top \mathbf{C}(\mathbf{D}\mathbf{x} + \mathbf{e})}{\partial \mathbf{x}} = \mathbf{D}^\top \mathbf{C}^\top (\mathbf{A}\mathbf{x} + \mathbf{b}) + \mathbf{A}^\top \mathbf{C}(\mathbf{D}\mathbf{x} + \mathbf{e})$
- $\mathbf{a}$ is not a function of $\mathbf{x}$: $\dfrac{\partial \|\mathbf{x} - \mathbf{a}\|}{\partial \mathbf{x}} = \dfrac{\mathbf{x} - \mathbf{a}}{\|\mathbf{x} - \mathbf{a}\|}$

Derivatives with matrices

Scalar-by-matrix $\mathbf{X} \in \mathbf{R}^{p \times q}$

$$\frac{\partial y}{\partial \mathbf{X}} = \begin{bmatrix} \frac{\partial y}{\partial x_{11}} & \frac{\partial y}{\partial x_{12}} & \cdots & \frac{\partial y}{\partial x_{1q}} \\ \frac{\partial y}{\partial x_{21}} & \frac{\partial y}{\partial x_{22}} & \cdots & \frac{\partial y}{\partial x_{2q}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial y}{\partial x_{p1}} & \frac{\partial y}{\partial x_{p2}} & \cdots & \frac{\partial y}{\partial x_{pq}} \end{bmatrix}$$

Scalar-by-matrix identities

- $\mathrm{tr}(\mathbf{A}) = \mathrm{tr}(\mathbf{A}^\top)$
- $\mathrm{tr}(\mathbf{ABCD}) = \mathrm{tr}(\mathbf{BCDA}) = \mathrm{tr}(\mathbf{CDAB}) = \mathrm{tr}(\mathbf{DABC})$
- $\dfrac{\partial\,\mathrm{tr}(\mathbf{AX})}{\partial\mathbf{X}} = \mathbf{A}^\top$
- $\dfrac{\partial\,\mathrm{tr}(\mathbf{X}^\top\mathbf{AX})}{\partial\mathbf{X}} = (\mathbf{A} + \mathbf{A}^\top)\mathbf{X}$
- $\dfrac{\partial\,\mathrm{tr}(\mathbf{X}^{-1}\mathbf{A})}{\partial\mathbf{X}} = -(\mathbf{X}^{-1})^\top\mathbf{A}^\top(\mathbf{X}^{-1})^\top$
- $\dfrac{\partial\,\mathrm{tr}(\mathbf{AXBX}^\top\mathbf{C})}{\partial\mathbf{X}} = \mathbf{A}^\top\mathbf{C}^\top\mathbf{XB}^\top + \mathbf{CAXB}$
- $\dfrac{\partial\,\mathrm{tr}(\mathbf{X}^n)}{\partial\mathbf{X}} = n(\mathbf{X}^{n-1})^\top$
- $\dfrac{\partial\,\mathrm{tr}(\mathbf{AX}^n)}{\partial\mathbf{X}} = \displaystyle\sum_{i=0}^{n-1}(\mathbf{X}^i\mathbf{AX}^{n-i-1})^\top$
- $\dfrac{\partial\,\mathrm{tr}(e^{\mathbf{X}})}{\partial\mathbf{X}} = (e^{\mathbf{X}})^\top$
- $\dfrac{\partial\,\mathrm{tr}(\sin(\mathbf{X}))}{\partial\mathbf{X}} = (\cos(\mathbf{X}))^\top$