Mathematics of Data: From Theory to Computation

Prof. Volkan Cevher
volkan.cevher@epfl.ch

Laboratory for Information and Inference Systems (LIONS)
École Polytechnique Fédérale de Lausanne (EPFL)

EE-556 (Fall 2014)

License Information for Mathematics of Data Slides

Outline

- This class: Linear algebra review
  1. Notation
  2. Vectors
  3. Matrices
  4. Tensors
- Next class
  1. Learning and convexity

Recommended reading material

- Z Kolter and C Do, *Linear Algebra Review and Reference* `http://cs229.stanford.edu/section/cs229-linalg.pdf`, 2012.
- KC Border, *Quick Review of Matrix and Real Linear Algebra* `http://www.hss.caltech.edu/~kcb/Notes/LinearAlgebra.pdf`, 2013.
- KB Petersen and MS Pedersen, *The matrix cookbook* `http://orion.uwaterloo.ca/~hwolkowi/matrixcookbook.pdf`, 2012.
- S Foucart and H Rauhut, *A mathematical introduction to compressive sensing* (Appendix A: Matrix Analysis), Springer, 2013.
- JA Tropp, *Column subset selection, matrix factorization, and eigenvalue optimization*, In Proc. of the 20th Annual ACM-SIAM Symposium on Discrete Algorithms, pp 978–986, SIAM, 2009.

Motivation

Motivation

This review is intended to help you follow mathematical discussions in data sciences, which rely heavily on basic linear algebra concepts:

- **Data and unknown parameters** are usually represented in the form of finite dimensional **linear algebra** objects like *vectors*, *matrices*, or *tensors*.
- **Computation** revolving around these objects invariably requires **numerical linear algebra** routines.

Notation

- Scalars are denoted by lowercase letters (e.g. k)
- Vectors by lowercase boldface letter (e.g., $\mathbf{x}$)
- Matrices and tensors by uppercase boldface letter (e.g. $\mathbf{A}$)
- Component of a *vector* $\mathbf{x}$, *matrix* $\mathbf{A}$ & *tensor* $\mathbf{A}$ as x_i, a_{ij} & $A_{i,j,k,\ldots}$ respectively.
- Sets by uppercase calligraphic letters (e.g. $\mathcal{S}$)

Vectors

1. Vector spaces
2. Vector norms
3. Inner products
4. Dual norms
5. *Extensions to Banach spaces

*: advanced

Vector spaces

Note:

We focus on the **field of real** numbers ($\mathbb{R}$) but most of the results can be generalized to the **field of complex** numbers ($\mathbb{C}$) in a straightforward fashion.

A vector space or *linear space* (over the field $\mathbb{R}$) consists of

(a) a set of vectors $\mathcal{V}$

(b) an addition operation: $\mathcal{V} \times \mathcal{V} \to \mathcal{V}$

(c) a scalar multiplication operation: $\mathbb{R} \times \mathcal{V} \to \mathcal{V}$

(d) a distinguished element $\mathbf{0} \in \mathcal{V}$

and satisfies the following properties:

1. $\mathbf{x} + \mathbf{y} = \mathbf{y} + \mathbf{x}, \quad \forall \mathbf{x}, \mathbf{y} \in \mathcal{V}$ (commutative under addition)
2. $(\mathbf{x} + \mathbf{y}) + \mathbf{z} = \mathbf{x} + (\mathbf{y} + \mathbf{z}), \quad \forall \mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathcal{V}$ (associative under addition)
3. $\mathbf{0} + \mathbf{x} = \mathbf{x}, \quad \forall \mathbf{x} \in \mathcal{V}$ ($\mathbf{0}$ being additive identity)
4. $\forall \mathbf{x} \in \mathcal{V} \;\; \exists\, (-\mathbf{x})$ such that $\mathbf{x} + (-\mathbf{x}) = \mathbf{0}$ ($-\mathbf{x}$ being additive inverse)
5. $(\alpha\beta)\mathbf{x} = \alpha(\beta\mathbf{x}), \quad \forall \alpha, \beta \in \mathbb{R} \quad \forall \mathbf{x} \in \mathcal{V}$ (associative under scalar multiplication)
6. $\alpha(\mathbf{x} + \mathbf{y}) = \alpha\mathbf{x} + \alpha\mathbf{y}, \quad \forall \alpha \in \mathbb{R} \quad \forall \mathbf{x}, \mathbf{y} \in \mathcal{V}$ (distributive)
7. $1\mathbf{x} = \mathbf{x}, \quad \forall \mathbf{x} \in \mathcal{V}$ ($\mathbf{1}$ being multiplicative identity)

Vector spaces contd.

Example (Vector space)

- $\mathcal{V}_1 = \{\mathbf{0}\}$ for $\mathbf{0} \in \mathbb{R}^p$
- $\mathcal{V}_2 = \mathbb{R}^p$
- $\mathcal{V}_3 = \sum_{i=1}^{k} \alpha_i \mathbf{x}_i$ for $\alpha_i \in \mathbb{R}$, $k < p$, and $\mathbf{x}_i \in \mathbb{R}^p$

It is straight forward to show that $\mathcal{V}_1$, $\mathcal{V}_2$, and $\mathcal{V}_3$ satisfy properties 1–7 above.

Definition (Subspace)

A **subspace** is a vector space that is a *subset* of another vector space.

Example (Subspace)

$\mathcal{V}_3$ (and actually $\mathcal{V}_1$ as well as $\mathcal{V}_2$) in the example above is subspace of $\mathbb{R}^p$.

Vector spaces contd.

Definition (Span)

The **span** of a set of vectors, $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k\}$, is the set of all possible linear combinations of these vectors; i.e.,

$$\mathrm{span}\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k\} = \{\alpha_1\mathbf{x}_1 + \alpha_2\mathbf{x}_2 + \cdots + \alpha_k\mathbf{x}_k \mid \alpha_1, \alpha_2, \ldots, \alpha_k \in \mathbb{R}\}.$$

Definition (Linear independence)

A set of vectors, $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k\}$, is **linearly independent** if

$$\alpha_1\mathbf{x}_1 + \alpha_2\mathbf{x}_2 + \cdots + \alpha_k\mathbf{x}_k = \mathbf{0} \;\Rightarrow\; \alpha_1 = \alpha_2 = \ldots = \alpha_k = 0.$$

Definition (Basis)

The **basis** of a vector space, $\mathcal{V}$, is a set of vectors $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k\}$ that satisfy
(a) $\mathcal{V} = \mathrm{span}\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k\}$, (b) $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k\}$ are linearly independent.

Definition (Dimension*)

The **dimension** of a vector space, $\mathcal{V}$, (denoted $\mathbf{dim}(\mathcal{V})$) is the number of vectors in the basis of $\mathcal{V}$.

*We will generalize the concept of affine dimension to the ***statistical dimension*** of convex objects.

Vector Norms

Definition (Vector norm)

The norm of a vector in $\mathbb{R}^p$ is a function $\|\cdot\| : \mathbb{R}^p \to \mathbb{R}$ such that for all vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^p$ and scalar $\lambda \in \mathbb{R}$

(a)	$\|\mathbf{x}\| \geq 0$ for all $\mathbf{x} \in \mathbb{R}^p$	(*nonnegativity*)
(b)	$\|\mathbf{x}\| = 0$ if and only if $\mathbf{x} = \mathbf{0}$	(*definitiveness*)
(c)	$\|\lambda\mathbf{x}\| = \lvert\lambda\rvert\|\mathbf{x}\|$	(*homogeniety*)
(d)	$\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$	(*triangle inequality*)

- There are an important family of ℓ_q-norms parameterized by $q \in [1, \infty]$.
- For $\mathbf{x} \in \mathbb{R}^p$, the ℓ_q-norm is defined as $\|\mathbf{x}\|_q := \left(\sum_{i=1}^p |x_i|^q\right)^{1/q}$.

Example

(1)	ℓ_2-norm:	$\|\mathbf{x}\|_2 := \sqrt{\sum_{i=1}^p x_i^2}$	(Euclidean norm)
(2)	ℓ_1-norm:	$\|\mathbf{x}\|_1 := \sum_{i=1}^p |x_i|$	(Manhattan norm)
(3)	ℓ_∞-norm:	$\|\mathbf{x}\|_\infty := \max_{i=1,\dots,p} |x_i|$	(Chebyshev norm)

Vector norms contd.

Definition (Quasi-norm)

A quasi-norm satisfies all the norm properties except (d) triangle inequality, which is replaced by $\|\mathbf{x}+\mathbf{y}\| \leq c\,(\|\mathbf{x}\| + \|\mathbf{y}\|)$ for a constant $c \geq 1$.

Definition (Semi(pseudo)-norm)

A semi(pseudo)-norm satisfies all the norm properties except (b) definiteness.

Example

- The ℓ_q-norm becomes a quasi-norm when $q \in (0,1)$ with $c = 2^{1/q} - 1$.
- The **total variation norm** (TV-norm) defined (in 1D): $\|\mathbf{x}\|_{\mathrm{TV}} := \sum_{i=1}^{p-1} |x_{i+1} - x_i|$ is a **semi-norm** since it fails to satisfy (b); e.g., $\mathbf{x} = (1, 1, \ldots, 1)^T$ has $\|\mathbf{x}\|_{\mathrm{TV}} = 0$ even though $\mathbf{x} \neq \mathbf{0}$.

Definition (ℓ_0-"norm")

$\|\mathbf{x}\|_0 = \lim_{q\to 0}\|\mathbf{x}\|_q^q = |\{i : x_i \neq 0\}|$

The ℓ_0-"norm" counts the non-zero components of $\mathbf{x}$. It is not a norm – it does not satisfy norm properties (c) and (d) $\Rightarrow$ it is also neither a **quasi**- nor a **semi-norm**.

Vector norms contd.

Problem (s-sparse approximation)

Find $\displaystyle \arg\min_{\mathbf{x}\in\mathbb{R}^p} \|\mathbf{x}-\mathbf{y}\|_2$ subject to: $\|\mathbf{x}\|_0 \leq s$.

Vector norms contd.

Problem (s-sparse approximation)

Find $\arg\min_{\mathbf{x}\in\mathbb{R}^p} \|\mathbf{x}-\mathbf{y}\|_2$ subject to: $\|\mathbf{x}\|_0 \leq s$.

Notation for the solution

- Ground set is denoted by $\mathcal{N} := \{1, \ldots, p\}$
- Base set $\mathcal{S}$ defined as $\mathcal{S} \subseteq 2^{\mathcal{N}}$ (a subset of the power set of $\mathcal{N}$)
- $\mathcal{S}^c$ denotes the complement of $\mathcal{S}$, i.e., $\mathcal{S}^c \equiv \mathcal{N} \setminus \mathcal{S}$
- $|\mathcal{S}|$ denotes the cardinality of a set $\mathcal{S}$
- $\mathbf{x}_{\mathcal{S}}$ for the restriction of $\mathbf{x}$ onto $\mathcal{S}$, i.e. $(\mathbf{x}_{\mathcal{S}})_i = \begin{cases} x_i & \text{if } i \in \mathcal{S} \\ 0 & \text{otherwise} \end{cases}$
- $\mathbf{x}_{|\mathcal{S}}$ maps the indices $\mathcal{S}$ of $\mathbf{x}$ into another vector in $\mathbb{R}^{|\mathcal{S}|}$ for the restriction of $\mathbf{x}$ onto $\mathcal{S}$, i.e. $(\mathbf{x}_{\mathcal{S}})_i$ is the entry of $\mathbf{x}$ corresponding to the i-th index in $\mathcal{S}$
- Support supp of a vector $\mathbf{x}$ is index set of its non-zero coefficients, i.e., $\text{supp}(\mathbf{x}) := \{\mathcal{S} | \mathbf{x}_{\mathcal{S}} \neq 0\}$

Vector norms contd.

Problem (s-sparse approximation)

Find $\quad \arg\min_{\mathbf{x} \in \mathbb{R}^p} \|\mathbf{x} - \mathbf{y}\|_2 \quad$ subject to: $\quad \|\mathbf{x}\|_0 \leq s$.

Solution

Let $\widehat{\mathbf{y}} \in \arg\min_{\mathbf{x} \in \mathbb{R}^p} \{\|\mathbf{x} - \mathbf{y}\|_2^2 : \|\mathbf{x}\|_0 \leq s\}$ and $\widehat{\mathcal{S}} = \text{supp}\left(\widehat{\mathbf{y}}\right)$. Assume we know $\widehat{\mathcal{S}}$ *a priori*. Then $\widehat{\mathbf{y}}_{\widehat{\mathcal{S}}^c} = \mathbf{0}$ and $\widehat{\mathbf{y}}_{|\widehat{\mathcal{S}}} = \arg\min_{\mathbf{x} \in \mathbb{R}^s} \|\mathbf{x} - \mathbf{y}_{|\widehat{\mathcal{S}}}\|_2 = \mathbf{y}_{|\widehat{\mathcal{S}}}$.

Therefore, the underlying difficulty in the s-sparse approximation problem boils down to finding $\widehat{\mathcal{S}}$:

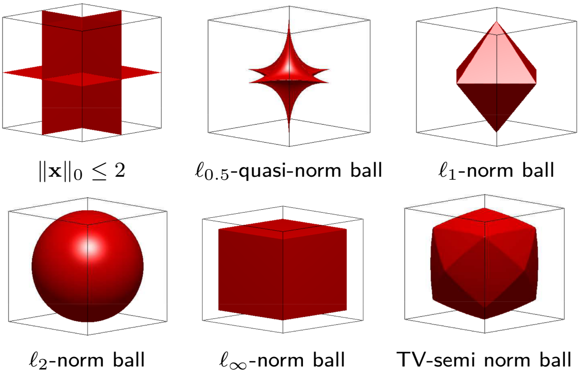
$$
\begin{aligned}
\widehat{\mathcal{S}} &\in \arg\min_{\mathcal{S}:|\mathcal{S}| \leq s} \|\mathbf{y}_{\mathcal{S}} - \mathbf{y}\|_2^2. \\
&\in \arg\max_{\mathcal{S}:|\mathcal{S}| \leq s} \left\{\|\mathbf{y}\|_2^2 - \|\mathbf{y}_{\mathcal{S}} - \mathbf{y}\|_2^2\right\} \\
&\in \arg\max_{\mathcal{S}:|\mathcal{S}| \leq s} \left\{\|\mathbf{y}_{\mathcal{S}}\|_2^2\right\} = \arg\max_{\mathcal{S}:|\mathcal{S}| \leq s} \sum_{i \in \mathcal{S}} \|y_i\|^2 \quad (\equiv \text{modular approximation problem}).
\end{aligned}
$$

Thus, the best s-sparse approximation of a vector is a vector with the s **largest components** of the vector in *magnitude*.

Vector norms contd.

Norm and "Norm" balls

Radius r ball in ℓ_q-norm: $\mathcal{B}_q(r) = \{\mathbf{x} \in \mathbb{R}^p : \|\mathbf{x}\|_q \leq r\}$



Example ℓ_q-(quasi) and TV-(semi) norm balls along with the set of 2-sparse vectors in $\mathbb{R}^3$

Inner products

Definition (Inner product)

The **inner product** of any two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^p$ (denoted by $\langle \cdot, \cdot \rangle$) is defined as $\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^T \mathbf{y} = \sum_i^p x_i y_i$.

The inner product satisfies the following properties:

1. $\langle \mathbf{x}, \mathbf{y} \rangle = \langle \mathbf{y}, \mathbf{x} \rangle, \quad \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^p$ (symmetry)
2. $\langle (\alpha \mathbf{x} + \beta \mathbf{y}), \mathbf{z} \rangle = \alpha \langle \mathbf{x}, \mathbf{z} \rangle + \beta \langle \mathbf{y}, \mathbf{z} \rangle, \quad \forall \alpha, \beta \in \mathbb{R} \quad \forall \mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbb{R}^p$ (linearity)
3. $\langle \mathbf{x}, \mathbf{x} \rangle \geq 0 \quad \forall \mathbf{x} \in \mathbb{R}^p$ (positive definiteness)

Important relations involving the inner product:

- Hölder's inequality: $|\langle \mathbf{x}, \mathbf{y} \rangle| \leq \|\mathbf{x}\|_q \|\mathbf{y}\|_r$, where $r > 1$ and $\frac{1}{q} + \frac{1}{r} = 1$
- Cauchy-Schwarz is a special case of Hölder's inequality ($q = r = 2$)

Inner products contd.

Definition (Inner product space)

An **inner product space** is a vector space endowed with an inner product.

Example

A **Hilbert space** (denoted $\mathcal{H}$) is an inner product space.

A vector space endowed with a norm is known as a **normed vector space**. For example, $\mathcal{H}$ is a normed vector space equipped with the ℓ_2-norm.

Vector norms contd.

Definition (Dual norm)

Let $\|\cdot\|$ be a norm in $\mathbb{R}^p$, then the **dual norm** denoted by $\|\cdot\|^*$ is defined:

$$\|\mathbf{x}\|^* = \sup_{\|\mathbf{y}\|\leq 1} \mathbf{x}^T\mathbf{y}, \quad \text{for all } \mathbf{x}, \mathbf{y} \in \mathbb{R}^p$$

Example 1

i) $\|\cdot\|_2$ is **dual** of $\|\cdot\|_2$ (i.e., $\|\cdot\|_2$ is *self-dual*): $\sup\{\mathbf{z}^T\mathbf{x} \mid \|\mathbf{x}\|_2 \leq 1\} = \|\mathbf{z}\|_2$.

ii) $\|\cdot\|_1$ is **dual** of $\|\cdot\|_\infty$, (and *vice versa*): $\sup\{\mathbf{z}^T\mathbf{x} \mid \|\mathbf{x}\|_\infty \leq 1\} = \|\mathbf{z}\|_1$.

Example 2

What is the **dual norm** of $\|\cdot\|_q$ for $q = 1 + 1/\log(p)$?

Vector norms contd.

Definition (Dual norm)

Let $\|\cdot\|$ be a norm in $\mathbb{R}^p$, then the **dual norm** denoted by $\|\cdot\|^*$ is defined:

$$\|\mathbf{x}\|^* = \sup_{\|\mathbf{y}\|\leq 1} \mathbf{x}^T\mathbf{y}, \quad \text{for all } \mathbf{x},\mathbf{y} \in \mathbb{R}^p$$

- The **dual** of the *dual norm* is the original (primal) norm, i.e., $\|\mathbf{x}\|^{**} = \|\mathbf{x}\|$.
- Hölder's inequality $\Rightarrow$ $\|\cdot\|_q$ is a dual norm of $\|\cdot\|_r$ when $\frac{1}{q} + \frac{1}{r} = 1$.

Example 1

i) $\|\cdot\|_2$ is **dual** of $\|\cdot\|_2$ (i.e., $\|\cdot\|_2$ is *self-dual*): $\sup\{\mathbf{z}^T\mathbf{x} \mid \|\mathbf{x}\|_2 \leq 1\} = \|\mathbf{z}\|_2$.

ii) $\|\cdot\|_1$ is **dual** of $\|\cdot\|_\infty$, (and *vice versa*): $\sup\{\mathbf{z}^T\mathbf{x} \mid \|\mathbf{x}\|_\infty \leq 1\} = \|\mathbf{z}\|_1$.

Example 2

What is the **dual norm** of $\|\cdot\|_q$ for $q = 1 + 1/\log(p)$?

Solution

By Hölder's inequality, $\|\cdot\|_r$ is the **dual norm** of $\|\cdot\|_q$ if $\frac{1}{q} + \frac{1}{r} = 1$. Therefore, $r = 1 + \log(p)$ for $q = 1 + 1/\log(p)$.

Metrics

- A metric on a set is a function that satisfies the minimal properties of a distance.

Definition (Metric)

Let $\mathcal{X}$ be some Hilbert space, then a metric $d(\cdot,\cdot) : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ if $\forall \mathbf{x}, \mathbf{y} \in \mathcal{X}$:

(a)	$d(\mathbf{x},\mathbf{y}) \geq 0$ for all $\mathbf{x}$ and $\mathbf{y}$	(*nonnegativity*)
(b)	$d(\mathbf{x},\mathbf{y}) = 0$ if and only if $\mathbf{x} = \mathbf{y}$	(*definiteness*)
(c)	$d(\mathbf{x},\mathbf{y}) = d(\mathbf{y},\mathbf{x})$	(*symmetry*)
(d)	$d(\mathbf{x},\mathbf{y}) \leq d(\mathbf{x},\mathbf{z}) + d(\mathbf{z},\mathbf{y})$	(*triangle inequality*)

- A pseudo-metric satisfies (a), (c) and (d) but not necessarily (b)
- A metric space $(\mathcal{X}, d)$ is a set $\mathcal{X}$ with a metric d defined on $\mathcal{X}$
- **Norms** induce metrics while **pseudo-norms** induce pseudo-metrics

Example

- Euclidean distance: $d_E(\mathbf{x},\mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_2^2$
- q-distances: $d_E(\mathbf{x},\mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_q^q$ for $q \in (0,1)$
- *Bregman distances $d_B(\cdot,\cdot)$ (more on this in Lecture 3)

*Banach spaces on $\mathbb{R}^p$

We only work with Banach spaces on $\mathbb{R}^p$ in this course. In general, a Banach space can be infinite-dimensional.

Proposition

The space $\mathbb{R}^p$ with any norm is a Banach space.

Example

Any Hilbert space on $\mathbb{R}^p$ is a Banach space.

A Banach space is not necessarily an inner product space.

Example

The space $\mathbb{R}^p$ with the ℓ_q-norm, $q \in [1, \infty)$, is a Banach space. But it is an inner product space only when $q = 2$.

*Banach spaces on $\mathbb{R}^p$

Theorem (Representer)

For every linear function $f : \mathbb{R}^p \to \mathbb{R}$*, we can always find a vector* $\mathbf{x}_f \in \mathbb{R}^p$ *such that* $\left\langle \mathbf{x}_f, \mathbf{x} \right\rangle := \sum_{i=1}^{p} x_i (x_f)_i = f(\mathbf{x})$ *for all* $\mathbf{x} \in \mathbb{R}^p$*.*

Definition (Dual space)

The *dual space* of a Banach space E on $\mathbb{R}^p$ with a norm $\|\cdot\|$ is the space E^* of all linear functions $f : \mathbb{R}^p \to \mathbb{R}$ with the dual norm $\|\cdot\|^*$.

Thus E^* is equivalent to $\mathbb{R}^p$ with the dual norm $\|\cdot\|^*$, since for each $f \in E^*$, we can always find the corresponding $\mathbf{x}_f \in \mathbb{R}^n$, and vice versa.

Definition (Dual pairing)

Let E be a Banach space and E^* be the dual space. For each $\mathbf{x} \in E$ and $f \in E^*$, we denote by $\langle f, \mathbf{x} \rangle$ the value of the linear function f at $\mathbf{x}$.

Thus for each $f \in E^*$ and its corresponding $\mathbf{x}_f \in \mathbb{R}^p$, we have $\langle f, \mathbf{x} \rangle = \left\langle \mathbf{x}_f, \mathbf{x} \right\rangle$.

Note that $\langle f, \mathbf{x} \rangle$ denotes a dual pairing, and $\left\langle \mathbf{x}_f, \mathbf{x} \right\rangle$ corresponds to the inner product with respect to the ℓ_2-norm.

Matrices

1. Special matrix types
2. Basic matrix definitions
3. Matrix decompositions
4. Complexity of matrix operations
5. Matrix norms

Matrices

- A matrix is a rectangular array of numbers arranged by rows and columns.
- We first describe a set of special matrices to get started.

Definition (Identity matrix)

The *identity* matrix (denoted $\mathbf{I} \in \mathbb{R}^{p\times p}$) is a square matrix of zero entries except on the *main diagonal*, which has ones on it. For compatible matrices $\mathbf{A}$ and $\mathbf{B}$, it satisfies:

$$\mathbf{IA} = \mathbf{A} \text{ and } \mathbf{BI} = \mathbf{B}.$$

Definition (Orthogonal (or Unitary) matrix)

A matrix $\mathbf{A} \in \mathbb{R}^{p\times p}$ is **orthogonal** or **unitary** if $\mathbf{A}^T\mathbf{A} = \mathbf{A}\mathbf{A}^T = \mathbf{I}$.

Definition (Triangular matrix)

A matrix $\mathbf{A} \in \mathbb{R}^{p\times p}$ is **lower triangular** if all its entries above the *main diagonal* are zero, i.e., $a_{ij} = 0$ for $j > i$; while it is **upper triangular** if $\mathbf{A}^T$ is lower triangular.

Definition (Permutation matrix)

A matrix $\mathbf{P} \in \mathbb{R}^{n\times p}$ is **permutation** if it has only one 1 in each row and each column and satisfies $\mathbf{P}\mathbf{P}^T = \mathbf{I}$.

Special matrices

Definition (Incidence matrix)

An incidence matrix shows the relationship between two sets $\mathcal{X}$ and $\mathcal{Y}$. The i-th row corresponding to entry $x_i \in \mathcal{X}$ and the j-th column corresponding to entry $y_j \in \mathcal{Y}$ of an incidence matrix is 1 if x_i and x_j are related and 0 if they are not.

Definition (Adjacency matrix)

An adjacency matrix is a symmetric square matrix with $\{0,1\}$ entries where 1 or 0 at the (i,j)-th location indicates the i-th and the j-th vertices of a graph are adjacent (i.e., share an edge) or not.

- The diagonal entries of adjacency matrices take different values depending on different conventions.

Definition (Stochastic matrix)

A matrix $\mathbf{P} \in \mathbb{R}^{n\times p}$ is **stochastic** (also know as transition or probability) matrix if $\sum_j p_{ij} = 1 \quad \text{for} \quad 0 \leq p_{ij} \leq 1$; while $\mathbf{A}$ is **doubly stochastic** if $\sum_i p_{ij} = \sum_j p_{ij} = 1$.

Definition (Gaussian matrix)

A matrix $\mathbf{A} \in \mathbb{R}^{p\times p}$ is **Gaussian** if its entries $a_{lk} \sim \mathcal{N}\left(\mu, \sigma^2\right) \quad \text{for} \quad l,k \in [p]$. That is, its entries are independent and identically distributed (*i.i.d.*) with mean μ & variance σ^2 according to the Gaussian distribution.

Special matrices contd.

Definition (Fourier matrix)

A matrix $\mathbf{F} \in \mathbb{C}^{p\times p}$ is **Fourier matrix** if its entries

$$f_{lk} = \frac{1}{\sqrt{p}} e^{i2\pi lk/p}, \quad \text{for} \quad l,k \in [p],\; i = \sqrt{-1}.$$

Definition (Discrete Cosine Transform matrix)

A matrix $\mathbf{A} \in \mathbb{R}^{p\times p}$ is **Discrete Cosine Transform** (DCT) matrix if its entries

$$a_{lk} = \sqrt{\frac{2}{p}} \cos\left(\frac{\pi}{p}(l-1)\left(k-\frac{1}{2}\right)\right); 1 \leq l \leq p,\; 1 \leq k \leq p.$$

- The **Fourier and DCT matrices** are both orthogonal, i.e., $\mathbf{F}^H\mathbf{F} = \mathbf{F}\mathbf{F}^H = \mathbf{I}$, where $\mathbf{F}^H =$ complex-conjugate$(\mathbf{F}^T)$.
- Both matrices are rarely stored since they have an implicit fast matrix-vector multiplication algorithm.

Special matrices contd.

Definition (Hadamard matrix [4])

Let the indices $l, k \in [2^n]$ be defined as $\quad l = \sum_{j=1}^{n} l_j 2^{j-1} + 1, \quad k = \sum_{j=1}^{n} k_j 2^{j-1} + 1.$

A matrix $\mathbf{H} = \mathbf{H}_n \in \mathbb{R}^{2^n \times 2^n}$ is a **Hadamard matrix** (or **Hadamard transform**) if

$$h_{lk} = \frac{1}{2^{n/2}} (-1)^{\sum_{j=1}^{n} k_j l_j}.$$

- The **Hadamard matrix** is orthogonal and self-adjoint, i.e., $\mathbf{H}_n = \mathbf{H}_n^T$.
- The **Hadamard matrix** is rarely stored since it has a fast matrix-vector multiplication algorithm that uses the recursive identity:

$$\mathbf{H}_n = \frac{1}{\sqrt{2}} \begin{pmatrix} \mathbf{H}_{n-1} & \mathbf{H}_{n-1} \\ \mathbf{H}_{n-1} & -\mathbf{H}_{n-1} \end{pmatrix}, \quad \mathbf{H}_0 = 1.$$

Special matrices contd.

Definition (Toeplitz matrix [2])

Let a $\mathbf{t} = (t_1, t_2, \ldots, t_{2p-1})$ be fixed or drawn from a probability distribution $\mathcal{P}(\mathbf{t})$. Then $\mathbf{T} \in \mathbb{R}^{p \times p}$ is **Toeplitz matrix** if

$$\mathbf{T} = \begin{pmatrix} t_1 & t_2 & t_3 & \cdots & t_{p-1} & t_p \\ t_{p+1} & t_1 & t_2 & \cdots & t_{p-2} & t_{p-1} \\ t_{p+2} & t_{p+1} & t_1 & \cdots & t_{p-3} & t_{p-2} \\ \vdots & \vdots & \vdots & \ddots & \ddots & \vdots \\ t_{2p-2} & t_{2p-3} & \cdots & \cdots & t_1 & t_2 \\ t_{2p-1} & t_{2p-2} & t_{2p-3} & \cdots & t_{p+1} & t_1 \end{pmatrix}.$$

Definition (Circulant matrix [7])

Let a $\mathbf{c} = (c_1, c_2, \ldots, c_p)$ be fixed or drawn from a probability distribution $\mathcal{P}(\mathbf{c})$, then $\mathbf{C} \in \mathbb{R}^{p \times p}$ is **Circulant matrix** if

$$\mathbf{C} = \begin{pmatrix} c_1 & c_p & \cdots & c_3 & c_2 \\ c_2 & c_1 & \cdots & c_4 & c_3 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ c_p & c_{p-1} & \cdots & c_2 & c_1 \end{pmatrix}.$$

Special matrices contd.

Partial Fourier, Partial Toeplitz, Partial Circulant, ...

A **partial** Fourier, Toeplitz or Circulant matrix refers to a matrix consisting of a subset of the rows of a Fourier, Toeplitz or Circulant matrix, respectively.

- Fourier, Hadamard, Toeplitz and Circulant matrices are structured matrices. In addition, Toeplitz and Circulant matrices are **banded**.
- These matrices also have lower degrees-of-freedom as compared to a general matrix in $\mathbb{R}^{p \times p}$. Hence, computations revolving around these matrices are typically cheaper than the computation we need for a general matrix.
- Incident and adjacency matrices are often used in graph theory. They have important decompositional and computational properties, which we will revisit in Lecture 11.

Basics of matrix definitions

Definition (Nullspace of a matrix)

The **nullspace** of a matrix, $\mathbf{A} \in \mathbb{R}^{n\times p}$, (denoted by $\mathrm{null}(\mathbf{A})$) is defined as

$$\mathrm{null}(\mathbf{A}) = \{\mathbf{x} \in \mathbb{R}^p \mid \mathbf{A}\mathbf{x} = \mathbf{0}\}$$

- $\mathrm{null}(\mathbf{A})$ is the set of vectors mapped to zero by $\mathbf{A}$.
- $\mathrm{null}(\mathbf{A})$ is the set of vectors orthogonal to the rows of $\mathbf{A}$.

Definition (Range of a matrix)

The **range** of a matrix, $\mathbf{A} \in \mathbb{R}^{n\times p}$, (denoted by $\mathrm{range}(\mathbf{A})$) is defined as

$$\mathrm{range}(\mathbf{A}) = \{\mathbf{A}\mathbf{x} \mid \mathbf{x} \in \mathbb{R}^p\} \subseteq \mathbb{R}^n$$

- $\mathrm{range}(\mathbf{A})$ is the span of the columns (or the **column space**) of $\mathbf{A}$.
- $\mathrm{range}(\mathbf{A})$ is the set of vectors $\mathbf{y} = \mathbf{A}\mathbf{x}$ for which the system has a solution.

Definition (Rank of a matrix)

The **rank** of a matrix, $\mathbf{A} \in \mathbb{R}^{n\times p}$, (denoted by $\mathrm{rank}(\mathbf{A})$) is defined as

$$\mathrm{rank}(\mathbf{A}) = \mathbf{dim}\left(\mathrm{range}(\mathbf{A})\right)$$

- $\mathrm{rank}(\mathbf{A})$ is the maximum number of independent columns (or rows) of $\mathbf{A}$, $\Rightarrow\ \mathrm{rank}(\mathbf{A}) \leq \min(n, p)$. We also have $\mathrm{rank}(\mathbf{A}) = \mathrm{rank}(\mathbf{A}^T)$; and $\mathrm{rank}(\mathbf{A}) + \mathbf{dim}\left(\mathrm{null}(\mathbf{A})\right) = p$.

Matrix definitions contd.

Definition (Eigenvalues & Eigenvectors)

The vector $\mathbf{x}$ is an **eigenvector** of a *square* matrix $\mathbf{A} \in \mathbb{R}^{p \times p}$ if $\mathbf{A}\mathbf{x} = \lambda \mathbf{x}$ where $\lambda \in \mathbb{R}$ is called an **eigenvalue** of $\mathbf{A}$.

Definition (Singular values & singular vectors)

For $\mathbf{A} \in \mathbb{R}^{n \times p}$ and *unit* vectors $\mathbf{u} \in \mathbb{R}^n$ and $\mathbf{v} \in \mathbb{R}^p$ if

$$\mathbf{A}\mathbf{v} = \sigma \mathbf{u} \quad \text{and} \quad \mathbf{A}^T \mathbf{u} = \sigma \mathbf{v}$$

then $\sigma \in \mathbb{R}$ $(\sigma \geq 0)$ is a **singular value** of $\mathbf{A}$; $\mathbf{v}$ and $\mathbf{u}$ are the **right singular vector** and the **left singular vector** respectively of $\mathbf{A}$.

Definition (Symmetric matrix)

A matrix $\mathbf{A} \in \mathbb{R}^{p \times p}$ is **symmetric** if $\mathbf{A} = \mathbf{A}^T$.

Definition (Matrix inverse)

The **inverse** of a square matrix $\mathbf{A} \in \mathbb{R}^{p \times p}$ (denoted by $\mathbf{A}^{-1}$), if it exists, satisfies:

$$\mathbf{A}^{-1}\mathbf{A} = \mathbf{A}\mathbf{A}^{-1} = \mathbf{I}, \quad \text{where } \mathbf{I} \text{ is the identity matrix.}$$

- If $\mathbf{A}^{-1}$ exists we say $\mathbf{A}$ is *invertible*. We also refer to it as nonsingular or nondegenerate.
- If $\mathbf{A}$ is unitary, then $\mathbf{A}^{-1} = \mathbf{A}^T$.

Matrix decompositions

Definition (Singular value decomposition)

The **singular value decomposition** (SVD) of a matrix, $\mathbf{A} \in \mathbb{R}^{n \times p}$, is given by:

$$\mathbf{A} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^T$$

- $\mathrm{rank}(\mathbf{A}) = r \leq \min(n, p)$ and σ_i is the i^{th} **singular value** of $\mathbf{A}$
- $\mathbf{u}_i$ and $\mathbf{v}_i$ are the i^{th} **left** and **right singular vectors** of $\mathbf{A}$ respectively
- $\mathbf{U} \in \mathbb{R}^{n \times r}$ and $\mathbf{V} \in \mathbb{R}^{p \times r}$ are **unitary** matrices (i.e., $\mathbf{U}^T\mathbf{U} = \mathbf{I}$)
- $\boldsymbol{\Sigma} = \mathbf{diag}\,(\sigma_1, \sigma_2, \ldots, \sigma_r)$ where $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r \geq 0$

- $\mathbf{v}_i$ are eigenvectors of $\mathbf{A}^T\mathbf{A}$; $\sigma_i = \sqrt{\lambda_i\,(\mathbf{A}^T\mathbf{A})}$ (and $\lambda_i\left(\mathbf{A}^T\mathbf{A}\right) = 0$ for $i > r$)

  since $\quad \mathbf{A}^T\mathbf{A} = \left(\mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T\right)^T\left(\mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T\right) = \left(\mathbf{V}\boldsymbol{\Sigma}^2\mathbf{V}^T\right)$
- $\mathbf{u}_i$ are eigenvectors of $\mathbf{A}\mathbf{A}^T$; $\sigma_i = \sqrt{\lambda_i\,(\mathbf{A}\mathbf{A}^T)}$ (and $\lambda_i\left(\mathbf{A}\mathbf{A}^T\right) = 0$ for $i > r$)

  since $\quad \mathbf{A}\mathbf{A}^T = \left(\mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T\right)\left(\mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T\right)^T = \left(\mathbf{U}\boldsymbol{\Sigma}^2\mathbf{U}^T\right)$

Matrix decompositions contd

Definition (Eigenvalue decomposition)

The **eigenvalue decomposition** of a square matrix, $\mathbf{A} \in \mathbb{R}^{p\times p}$, is given by:

$$\mathbf{A} = \mathbf{X}\mathbf{\Lambda}\mathbf{X}^{-1}$$

- the columns of $\mathbf{X} \in \mathbb{R}^{p\times p}$, i.e., $\mathbf{x}_i$, are eigenvectors of $\mathbf{A}$
- $\mathbf{\Lambda} = \mathbf{diag}\,(\lambda_1, \lambda_2, \ldots, \lambda_p)$ where λ_i (also denoted $\lambda_i(\mathbf{A})$) are eigenvalues of $\mathbf{A}$
- Note that not all matrices are diagonalizable. This happens if at least one eigenvalue has multiplicity $m > 1$ and if there are less than m linearly independent eigenvectors associated with that eigenvalue.

Eigendecomposition of symmetric matrices

If $\mathbf{A} \in \mathbb{R}^{p\times p}$ is symmetric, the decomposition becomes $\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$ where $\mathbf{U} \in \mathbb{R}^{p\times p}$ is unitary (or **orthonormal**), i.e., $\mathbf{U}^T\mathbf{U} = \mathbf{U}\mathbf{U}^T = \mathbf{I}$ and λ_i are **real**.

If we order $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$, $\lambda_i(\mathbf{A})$ becomes the i^{th} largest eigenvalue of $\mathbf{A}$:

- $\lambda_p(\mathbf{A}) = \lambda_{\min}(\mathbf{A})$ is the **minimum** eigenvalue of $\mathbf{A}$
- $\lambda_1(\mathbf{A}) = \lambda_{\max}(\mathbf{A})$ is the **maximum** eigenvalue of $\mathbf{A}$

Matrix decompositions contd

Definition (LU)

The **LU factorization** of a nonsingular square matrix, $\mathbf{A} \in \mathbb{R}^{p \times p}$, is given by:

$$\mathbf{A} = \mathbf{L}\mathbf{U}$$

where the matrix $\mathbf{L}$ is **lower triangular** and the matrix $\mathbf{U}$ is **upper triangular**.

Definition (QR)

The **QR factorization** of any matrix, $\mathbf{A} \in \mathbb{R}^{n \times p}$, is given by:

$$\mathbf{A} = \mathbf{Q}\mathbf{R}$$

where $\mathbf{Q} \in \mathbb{R}^{n \times n}$ is an **orthogonal** matrix, i.e., $\mathbf{Q}^T\mathbf{Q} = \mathbf{I}$, and $\mathbf{R} \in \mathbb{R}^{n \times p}$ is **upper triangular**.

Definition (Cholesky)

The **Cholesky factorization** of a positive definite matrix, $\mathbf{A} \in \mathbb{R}^{p \times p}$, is given by:

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

where $\mathbf{L}$ is a **lower triangular** matrix with positive entries on the *diagonal*.

Matrix definitions contd.

Definition (Moore–Penrose pseudoinverse)

The **Moore–Penrose pseudoinverse** of a matrix $\mathbf{A} \in \mathbb{R}^{n \times p}$ (denoted by $\mathbf{A}^{\dagger}$) can be constructed using its singular value decomposition $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ as follows:

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

where the operation † preserves the zero entries of the diagonal matrix $\mathbf{\Sigma}$, reciprocates the non-zero entries, and then transposes the matrix.

Definition (Determinant of a matrix)

The **determinant** of a square matrix $\mathbf{A} \in \mathbb{R}^{p \times p}$, denoted by $\det(\mathbf{A})$, is given by:

$$\det(\mathbf{A}) = \Pi_{i=1}^{p} \lambda_i$$

where λ_i are *eigenvalues* of $\mathbf{A}$.

Definition (Trace of a matrix)

The **trace** of a square matrix $\mathbf{A} \in \mathbb{R}^{p \times p}$, denoted by $\mathrm{trace}(\mathbf{A})$, is given by:

$$\mathrm{trace}(\mathbf{A}) = \sum_{i=1}^{p} a_{ii} = \sum_{i=1}^{p} \lambda_i$$

where a_{ii} are the elements of the *main diagonal* of $\mathbf{A}$ and λ_i are *eigenvalues* of $\mathbf{A}$.

Matrix definitions contd.

Definition (Positive semidefinite & positive definite matrices)

A symmetric matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ is **positive semidefinite** (denoted $\mathbf{A} \succeq 0$) if $\mathbf{x}^T\mathbf{A}\mathbf{x} \geq 0$; while it is **positive definite** (denoted $\mathbf{A} \succ 0$) if $\mathbf{x}^T\mathbf{A}\mathbf{x} > 0$

- $\mathbf{A} \succeq 0$ iff all its eigenvalues are **nonnegative**, i.e., $\lambda_{\min}(\mathbf{A}) \geq 0$.
- Similarly, $\mathbf{A} \succ 0$ iff all its eigenvalues are **positive**, i.e., $\lambda_{\min}(\mathbf{A}) > 0$.
- $\mathbf{A}$ is negative semidefinite if $-\mathbf{A} \succeq 0$; while $\mathbf{A}$ is negative definite if $-\mathbf{A} \succ 0$.
- Semidefinite ordering of two *symmetric* matrices, $\mathbf{A}$ and $\mathbf{B}$: $\mathbf{A} \succeq \mathbf{B}$ if $\mathbf{A} - \mathbf{B} \succeq 0$.

Example (Matrix inequalities)

1. If $\mathbf{A} \succeq 0$ and $\mathbf{B} \succeq 0$, then $\mathbf{A} + \mathbf{B} \succeq 0$
2. If $\mathbf{A} \succeq \mathbf{B}$ and $\mathbf{C} \succeq \mathbf{D}$, then $\mathbf{A} + \mathbf{C} \succeq \mathbf{B} + \mathbf{D}$
3. If $\mathbf{B} \preceq 0$ then $\mathbf{A} + \mathbf{B} \preceq \mathbf{A}$
4. If $\mathbf{A} \succeq 0$ and $\alpha \geq 0$, then $\alpha\mathbf{A} \succeq 0$
5. If $\mathbf{A} \succ 0$, then $\mathbf{A}^2 \succ 0$
6. If $\mathbf{A} \succ 0$, then $\mathbf{A}^{-1} \succ 0$

Complexity of matrix operations

Complexity of an algorithm

The complexity or *cost* of an algorithm is expressed in terms of **floating-point operations** (flops) as a function of the *problem dimension*.

Definition (floating-point operation)

A **floating-point operation** (flop) is one addition, subtraction, multiplication, or division of two floating-point numbers.

- In computing, flops, i.e., the plural form of flop, also stands for FLoating-point Operations Per Second, which measures the rate. We can disambiguate depending on the context.

Complexity of matrix operations

Table: Complexity illustrations. Vector are in $\mathbb{R}^p$. Matrices are in $\mathbb{R}^{m\times n}$ or $\mathbb{R}^{n\times p}$ or $\mathbb{R}^{p\times p}$.

Operation	Complexity	Remarks
vector addition	p flops	
vector inner product	$2p-1$ flops	or $\approx 2p$ for p large
matrix-vector product	$n(2p-1)$ flops	or $\approx 2np$ for p large $2m$ if $\mathbf{A}$ is sparse with m nonzeros
matrix-matrix product	$mn(2p-1)$ flops	or $\approx 2mnp$ for p large (naïve method) much less if the matrices are sparse[1,2]
LU decomposition	$\frac{2}{3}p^3 + 2p^2$ flops	or $\approx \frac{2}{3}p^3$ for p large much less if the matrix is sparse[1]
Cholesky decomposition	$\frac{1}{3}p^3 + 2p^2$ flops	or $\approx \frac{1}{3}p^3$ for p large much less if the matrix is sparse[1]
Matrix SVD	$C_1 n^2 p + C_2 p^3$ flops	$C_1 = 4$, $C_2 = 22$ for R-SVD algo.
Matrix determinant	complexity of SVD$+p$ flops	much less for sparse $\mathbf{A}$ using Cholesky
Matrix inverse	$Cp^{\log_2 7}$ flops,	$4 < C < 5$ using Strassen algorithm

[1] Computational complexity depends on the number of nonzeros in the matrices.
[2] For multiplying $p \times p$ matrices, the best computational complexity result is currently $O(p^{2.373})$.

Matrix norms

Similar to vector norms, **matrix norms** are a metric over matrices:

Definition (Matrix norm)

The norm of an $n \times p$ matrix is a map $\|\cdot\| : \mathbb{R}^{n \times p} \to \mathbb{R}$ such that for all matrices $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n \times p}$ and scalar $\lambda \in \mathbb{R}$

(a)	$\|\mathbf{A}\| \geq 0$ for all $\mathbf{A} \in \mathbb{R}^{n \times p}$	(*nonnegativity*)
(b)	$\|\mathbf{A}\| = 0$ if and only if $\mathbf{A} = \mathbf{0}$	(*definitiveness*)
(c)	$\|\lambda \mathbf{A}\| = \|\lambda\| \|\mathbf{A}\|$	(*homogeniety*)
(d)	$\|\mathbf{A} + \mathbf{B}\| \leq \|\mathbf{A}\| + \|\mathbf{B}\|$	(*triangle inequality*)

Definition (Matrix inner product)

Matrix inner product is defined as follows

$$\langle \mathbf{A}, \mathbf{B} \rangle = \operatorname{trace}\left(\mathbf{A}\mathbf{B}^T\right).$$

For complex matrices, we replace the transpose operation with the conjugate transpose (i.e., Hermitian).

Matrix norms contd.

- Similar to vector ℓ_p-norms we have Schatten q-norms for matrices.

Definition (Schatten q-norms)

$\|\mathbf{A}\|_{S_q} := \left(\sum_{i=1}^{p} (\sigma(\mathbf{A})_i)^q\right)^{1/q}$, where $\sigma(\mathbf{A})_i$ is the i^{th} singular value of $\mathbf{A}$.

Example (with $r = \min\{n, p\}$ and $\sigma_i = \sigma(\mathbf{A})_i$)

$$\|\mathbf{A}\|_{S_1} = \|\mathbf{A}\|_* := \sum_{i=1}^{r} \sigma_i \equiv \operatorname{trace}\left(\sqrt{\mathbf{A}^T\mathbf{A}}\right) \quad \text{(Nuclear/trace)}$$

$$\|\mathbf{A}\|_{S_2} = \|\mathbf{A}\|_F := \sqrt{\sum_{i=1}^{r} (\sigma_i)^2} \equiv \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{p} |a_{ij}|^2} \quad \text{(Frobenius)}$$

$$\|\mathbf{A}\|_{S_\infty} = \|\mathbf{A}\| := \max_{i=1,\ldots,r} \{\sigma_i\} \equiv \max_{\mathbf{x}\neq\mathbf{0}} \frac{\|\mathbf{A}\mathbf{x}\|}{\|\mathbf{x}\|} \quad \text{(Spectral/matrix)}$$

Matrix norms contd.

Problem (Rank-r approximation)

Find $\quad \arg\min_{\mathbf{X}} \ \|\mathbf{X} - \mathbf{Y}\|_F \quad$ subject to: $\quad \mathrm{rank}(\mathbf{X}) \leq r.$

Matrix norms contd.

Problem (Rank-r approximation)

Find $\quad \arg\min_{\mathbf{X}} \|\mathbf{X}-\mathbf{Y}\|_F \quad$ subject to: $\quad \mathrm{rank}(\mathbf{X}) \leq r.$

Solution (**Eckart–Young–Mirsky Theorem**)

$$
\begin{aligned}
\arg\min_{\mathbf{X}:\mathrm{rank}(\mathbf{X})\leq r} \|\mathbf{X}-\mathbf{Y}\|_F &= \arg\min_{\mathbf{X}:\mathrm{rank}(\mathbf{X})\leq r} \|\mathbf{X}-\mathbf{U}\boldsymbol{\Sigma}_{\mathbf{Y}}\mathbf{V}^T\|_F, \quad \text{(SVD)}\\
&= \arg\min_{\mathbf{X}:\mathrm{rank}(\mathbf{X})\leq r} \|\mathbf{U}^T\mathbf{X}\mathbf{V}-\boldsymbol{\Sigma}_{\mathbf{Y}}\|_F, \quad \text{(unitary invariance of } \|\cdot\|_F)\\
&= \mathbf{U}\left(\arg\min_{\mathbf{M}:\mathrm{rank}(\mathbf{M})\leq r} \|\mathbf{M}-\boldsymbol{\Sigma}_{\mathbf{Y}}\|_F\right)\mathbf{V}^T, \quad \text{(sparse approx.)}\\
&= \mathbf{U}H_r\left(\boldsymbol{\Sigma}_{\mathbf{Y}}\right)\mathbf{V}^T, \quad (r\text{-sparse approx. of the diagonal entries})
\end{aligned}
$$

Singular value hard thresholding operator H_r performs the best rank-r approximation of a matrix via sparse approximation: We keep the r **largest singular values** of the matrix and set the rest to zero.

Matrix norms contd.

- The last step of the above solution makes use of the Mirsky inequality.

Theorem (Mirsky inequality)

If $\mathbf{A}, \mathbf{B}$ *are* $p \times p$ *matrices with singular values*

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_p \geq 0, \quad \tau_1 \geq \tau_2 \geq \cdots \geq \tau_p \geq 0$$

respectively. Let $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_p)^T$ *and* $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_p)^T$, *then*

$$\|\mathbf{A} - \mathbf{B}\|_F \geq \|\boldsymbol{\sigma} - \boldsymbol{\tau}\|_2.$$

- **Mirsky theorem** is proved using the following simplified version of von Neumann trace inequality.

Theorem (von Neumann trace inequality)

If $\mathbf{A}, \mathbf{B}$ *are* $p \times p$ *matrices with singular values*

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_p \geq 0, \quad \tau_1 \geq \tau_2 \geq \cdots \geq \tau_p \geq 0$$

respectively. Let $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_p)^T$ *and* $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_p)^T$, *then*

$$\langle \mathbf{A}, \mathbf{B} \rangle \leq \langle \boldsymbol{\sigma}, \boldsymbol{\tau} \rangle$$

Matrix norms contd.

Matrix & vector norm analogy

Vectors	$\|\mathbf{x}\|_1$	$\|\mathbf{x}\|_2$	$\|\mathbf{x}\|_\infty$
Matrices	$\|\mathbf{X}\|_*$	$\|\mathbf{X}\|_F$	$\|\mathbf{X}\|$

Definition (Dual norm for matrices)

The **dual norm** of $\mathbf{A} \in \mathbb{R}^{n \times p}$ is defined as

$$\|\mathbf{A}\|^* = \sup_{\mathbf{X}} \left\{ \langle \mathbf{X}, \mathbf{A} \rangle \;|\; \|\mathbf{X}\| \leq 1 \right\}.$$

Matrix & vector dual norm analogy

Vector primal norm	$\|\mathbf{x}\|_1$	$\|\mathbf{x}\|_2$	$\|\mathbf{x}\|_\infty$
Vector dual norm	$\|\mathbf{x}\|_\infty$	$\|\mathbf{x}\|_2$	$\|\mathbf{x}\|_1$
Matrix primal norm	$\|\mathbf{X}\|_*$	$\|\mathbf{X}\|_F$	$\|\mathbf{X}\|$
Matrix dual norm	$\|\mathbf{X}\|$	$\|\mathbf{X}\|_F$	$\|\mathbf{X}\|_*$

Linear operators

- Matrices are often given in an **implicit** form (e.g., partial Fourier, DCT, and Hadamard matrices). It is convenient to think of them as *linear operators*.

Proposition (Linear operators & matrices)

Any **linear operator** *in finite dimensional spaces can be represented as a* **matrix***.*

Example

Given matrices $\mathbf{A}, \mathbf{B}$ and $\mathbf{X}$ with compatible dimensions and the *linear operator* $\mathcal{M} : \mathbb{R}^{n \times p} \to \mathbb{R}^{np}$, we can define an implicit mapping through the linear operator

$$\mathcal{M}(\mathbf{X}) := \left(\mathbf{B}^T \otimes \mathbf{A}\right) \mathrm{vec}(\mathbf{X}) = \mathrm{vec}(\mathbf{AXB}),$$

where $\otimes$ is the Kronecker product and $\mathrm{vec} : \mathbb{R}^{n \times p} \to \mathbb{R}^{np}$ is yet another linear operator that vectorizes its entries.
Note: Clearly, it is more efficient to compute $\mathrm{vec}(\mathbf{AXB})$ than to perform the *matrix multiplication* $\left(\mathbf{B}^T \otimes \mathbf{A}\right) \mathrm{vec}(\mathbf{X})$.

Example

Define a partial Hadamard matrix $\bar{\mathbf{H}}_n$ as $\bar{\mathbf{H}}_n = \bar{\mathbf{I}}\mathbf{H}_n$ where $\bar{\mathbf{I}}$ be a partial identity matrix. While we can store $\bar{\mathbf{H}}_n$ and use standard matrix multiplication techniques, it is often more efficient (both space and computation-wise) to apply the fast Hadamard transform algorithm and then apply $\bar{\mathbf{I}}$.

Matrix norms contd.

Definition (Operator norm)

The **operator norm** between ℓ_q and ℓ_r $(1 \leq q, r \leq \infty)$ of a matrix $\mathbf{A}$ is defined as

$$\|\mathbf{A}\|_{q \to r} = \sup_{\|\mathbf{x}\|_q \leq 1} \|\mathbf{A}\mathbf{x}\|_r$$

Problem

Show that $\|\mathbf{A}\|_{2\to 2} = \|\mathbf{A}\|$, i.e., ℓ_2-to-ℓ_2 *operator* norm is the *spectral* norm.

Matrix norms contd.

Definition (Operator norm)

The **operator norm** between ℓ_q and ℓ_r $(1 \leq q, r \leq \infty)$ of a matrix $\mathbf{A}$ is defined as

$$\|\mathbf{A}\|_{q\to r} = \sup_{\|\mathbf{x}\|_q \leq 1} \|\mathbf{A}\mathbf{x}\|_r$$

Problem

Show that $\|\mathbf{A}\|_{2\to 2} = \|\mathbf{A}\|$, i.e., ℓ_2-to-ℓ_2 *operator* norm is the *spectral* norm.

Solution

$$\begin{aligned}
\|\mathbf{A}\|_{2\to 2} = \sup_{\|\mathbf{x}\|_2 \leq 1} \|\mathbf{A}\mathbf{x}\|_2 &= \sup_{\|\mathbf{x}\|_2 \leq 1} \|\mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T\mathbf{x}\|_2 \quad \text{(using SVD of } \mathbf{A}\text{)} \\
&= \sup_{\|\mathbf{x}\|_2 \leq 1} \|\boldsymbol{\Sigma}\mathbf{V}^T\mathbf{x}\|_2 \quad \text{(unitary invariance of } \|\cdot\|_2\text{)} \\
&= \sup_{\|\mathbf{z}\|_2 \leq 1} \|\boldsymbol{\Sigma}\mathbf{z}\|_2 \quad \text{(letting } \mathbf{V}^T\mathbf{x} = \mathbf{z}\text{)} \\
&= \sup_{\|\mathbf{z}\|_2 \leq 1} \sqrt{\sum_{i=1}^{\min(n,p)} \sigma_i^2 z_i^2} = \sigma_{\max} = \|\mathbf{A}\| \qquad \square
\end{aligned}$$

Matrix norms contd.

Other examples

- The $\|\mathbf{A}\|_{\infty\to\infty}$ (norm induced by ℓ_∞-norm) also denoted $\|\mathbf{A}\|_\infty$, is the **max-row-sum norm**:

$$\|\mathbf{A}\|_{\infty\to\infty} := \sup_{\mathbf{x}}\{\|\mathbf{A}\mathbf{x}\|_\infty \mid \|\mathbf{x}\|_\infty \leq 1\} = \max_{i=1,\ldots,n} \sum_{j=1}^{p} |a_{ij}|.$$

- The $\|\mathbf{A}\|_{1\to1}$ (norm induced by ℓ_1-norm) also denoted $\|\mathbf{A}\|_1$, is the **max-column-sum norm**:

$$\|\mathbf{A}\|_{1\to1} := \sup_{\mathbf{x}}\{\|\mathbf{A}\mathbf{x}\|_1 \mid \|\mathbf{x}\|_1 \leq 1\} = \max_{j=1,\ldots,p} \sum_{i=1}^{n} |a_{ij}|.$$

Matrix norms contd.

Useful relation for operator norms

The following **identity** holds

$$\|\mathbf{A}\|_{q\to r} = \|\mathbf{A}^T\|_{r'\to q'}$$

whenever $1/q + 1/q' = 1 = 1/r + 1/r'$.

Example

1. $\|\mathbf{A}\|_{\infty\to 1} = \|\mathbf{A}^T\|_{\infty\to 1}$.
2. $\|\mathbf{A}\|_{2\to 1} = \|\mathbf{A}^T\|_{\infty\to 2}$.
3. $\|\mathbf{A}\|_{1\to 1} = \|\mathbf{A}^T\|_{\infty\to\infty}$.

Matrix norms contd.

Computation of operator norms

- The computation of some **operator norms** is NP-hard [4]; these include:
  1. $\|\mathbf{A}\|_{\infty\to 1}$
  2. $\|\mathbf{A}\|_{2\to 1}$
  3. $\|\mathbf{A}\|_{\infty\to 2}$
- But some of them are **approximable** [9]; these include:
  1. $\|\mathbf{A}\|_{\infty\to 1}$ (using Gronthendieck factorization)
  2. $\|\mathbf{A}\|_{\infty\to 2}$ (using Pietzs factorization)

Matrix norms contd.

Definition (Nuclear norm computation)

$$\|\mathbf{A}\|_* := \|\boldsymbol{\sigma}(\mathbf{A})\|_1 \quad \text{where } \boldsymbol{\sigma}(\mathbf{A}) \text{ is a vector of singular values of } \mathbf{A}$$

$$= \min_{\mathbf{U},\mathbf{V}:\mathbf{A}=\mathbf{U}\mathbf{V}^H} \|\mathbf{U}\|_F\|\mathbf{V}\|_F \;= \min_{\mathbf{U},\mathbf{V}:\mathbf{A}=\mathbf{U}\mathbf{V}^H} \frac{1}{2}\left(\|\mathbf{U}\|_F^2 + \|\mathbf{V}\|_F^2\right)$$

Additional useful properties are below:

- Nuclear vs. Frobenius: $\|\mathbf{A}\|_F \leq \|\mathbf{A}\|_* \leq \sqrt{\text{rank}(\mathbf{A})} \cdot \|\mathbf{A}\|_F$
- Hölder for matrices: $|\langle \mathbf{A}, \mathbf{B}\rangle| \leq \|\mathbf{A}\|_p\|\mathbf{B}\|_q$, when $\frac{1}{p} + \frac{1}{q} = 1$
- We have
  1. $\|\mathbf{A}\|_{2\to 2} \leq \|\mathbf{A}\|_F$
  2. $\|\mathbf{A}\|_{2\to 2}^2 \leq \|\mathbf{A}\|_{1\to 1}\|\mathbf{A}\|_{\infty\to\infty}$
  3. $\|\mathbf{A}\|_{2\to 2} \leq \|\mathbf{A}\|_{1\to 1}$ when $\mathbf{A}$ is self-adjoint.

Matrix perturbation inequalities

- In the theorems below $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{p\times p}$ are symmetric positive semi-definite matrices with spectra $\{\lambda_i(\mathbf{A})\}_{i=1}^p$ and $\{\lambda_i(\mathbf{B})\}_{i=1}^p$ where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$.

Theorem (Lidskii inequality)

$$\lambda_{i_1}(\mathbf{A}+\mathbf{B}) + \cdots + \lambda_{i_n}(\mathbf{A}+\mathbf{B}) \leq \lambda_{i_1}(\mathbf{A}) + \cdots \lambda_{i_n}(\mathbf{A}) + \lambda_{i_1}(\mathbf{B}) + \cdots + \lambda_{i_n}(\mathbf{B}),$$

for any $1 \leq i_1 \leq \cdots \leq i_n \leq p$.

Theorem (Weyl inequality)

$$\lambda_{i+j-1}(\mathbf{A}+\mathbf{B}) \leq \lambda_i(\mathbf{A}) + \lambda_j(\mathbf{B}), \quad \text{for any} \;\; i,j \geq 1 \;\; \text{and} \;\; i+j-1 \leq p.$$

Theorem (Interlacing property)

Let $\mathbf{A}_n = \mathbf{A}(1:n, 1:n)$*, then*

$$\lambda_{n+1}(\mathbf{A}_{n+1}) \leq \lambda_n(\mathbf{A}_n) + \lambda_n(\mathbf{A}_{n+1}) \qquad \text{for} \;\; n = 1, \ldots, p.$$

- These inequalities **hold** in the more general setting when λ_i are replaced by σ_i.
- The list goes on to include Wedin bounds, Wielandt-Hoffman bounds and so on.
- More on such inequalities can be found in Terry Tao's blog (**254A, Notes 3a**).

Tensors

1. Basic tensor definitions
2. Notation and preliminaries
3. Tensors decompositions
4. Tensor rank
5. Advanced material

Basic definitions

- Tensors provide natural and concise mathematical representations of **data**.

Definition (Tensor)

An **order** m tensor in p-**dimensional** space is a mathematical object that has p *indices* and p^m *components* and obeys certain transformation rules.

- In the literature, rank is used interchangeably with order, i.e., an **order**-k tensor is also referred to as kth-**rank** tensor.
- In this course, we will use order instead of rank so that it is not confused with the **rank of a tensor**.
- Furthermore, mode or way is also used to refer to the **order** of a tensor.
- Tensors are **multidimensional arrays** and are a generalization of:
  1. scalars - **tensors** with *no indices*; i.e., order zero tensor.
  2. vectors - **tensors** with exactly *one index*; i.e., order one tensor.
  3. matrices - **tensors** with exactly *two indices*; i.e., order two tensor.
- A third-order tensor has exactly *three indices*.
- A higher-order tensor has *greater than two indices*; i.e., a tensor of order ≥ 2.

Notation & preliminaries

Notation & preliminaries

- The notation conforms to [6] which is the main reference for this material.
- Higher-order tensors are denoted by boldface Euler script letters, e.g. $\boldsymbol{\mathcal{A}}$.
- Element $(i, j, k, \ldots)$ of a tensor $\boldsymbol{\mathcal{A}}$ are denoted by $a_{ijk\ldots}$
- The mth element in a sequence is denoted by a superscript in parentheses, e.g. $\mathbf{A}^{(m)}$ denotes the mth matrix in a sequence.
- Subarrays of a tensor are formed when a **subset of the indices** of the elements of a tensor are fixed.
- Fibers are the higher-order analogue of matrix rows and columns, defined by *fixing every index but one*.
- Slices are 2-dimensional sections of a tensor, defined by **fixing all but** 2 **indices**. For instance, the horizontal, lateral, and frontal slices of a third-order tensor $\boldsymbol{\mathcal{A}}$ are denoted by $\mathbf{A}_{i::}$, $\mathbf{A}_{:j:}$, & $\mathbf{A}_{::k}$ (or more compactly $\mathbf{A}_i$, $\mathbf{A}_j$, & $\mathbf{A}_k$) respectively.

Notation & preliminaries

Notation & preliminaries

- The notation conforms to [6] which is the main reference for this material.
- Higher-order tensors are denoted by boldface Euler script letters, e.g. $\boldsymbol{\mathcal{A}}$.
- Element $(i, j, k, \ldots)$ of a tensor $\boldsymbol{\mathcal{A}}$ are denoted by $a_{ijk\ldots}$
- The mth element in a sequence is denoted by a superscript in parentheses, e.g. $\mathbf{A}^{(m)}$ denotes the mth matrix in a sequence.
- Subarrays of a tensor are formed when a **subset of the indices** of the elements of a tensor are fixed.
- Fibers are the higher-order analogue of matrix rows and columns, defined by *fixing every index but one*.
- Slices are 2-dimensional sections of a tensor, defined by **fixing all but 2 indices**. For instance, the horizontal, lateral, and frontal slices of a third-order tensor $\boldsymbol{\mathcal{A}}$ are denoted by $\mathbf{A}_{i::}$, $\mathbf{A}_{:j:}$, & $\mathbf{A}_{::k}$ (or more compactly $\mathbf{A}_i$, $\mathbf{A}_j$, & $\mathbf{A}_k$) respectively.

Curse of dimensionality

Storage of an order-m tensor with mode sizes p requires p^m elements.

Notation & preliminaries contd.

- Tensors are linear vector spaces.

Definition (Norm)

The **norm** of a tensor $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{p_1 \times p_2 \times \cdots \times p_k}$ is given by

$$\|\mathbf{A}\| = \sqrt{\sum_{i_1=1}^{p_1} \sum_{i_2=1}^{p_2} \cdots \sum_{i_k=1}^{p_k} a^2_{i_1 i_2 \ldots i_k}}$$

- This is the analogue to the matrix Frobenius norm.

Definition (Inner product)

The **inner product** of two same-sized tensors $\boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{Y}} \in \mathbb{R}^{p_1 \times p_2 \times \cdots \times p_k}$ is given by

$$\langle \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{Y}} \rangle = \sum_{i_1=1}^{p_1} \sum_{i_2=1}^{p_2} \cdots \sum_{i_k=1}^{p_k} x_{i_1 i_2 \ldots i_k} y_{i_1 i_2 \ldots i_k}$$

- It follows immediately that $\langle \boldsymbol{\mathcal{A}}, \boldsymbol{\mathcal{A}} \rangle = \|\boldsymbol{\mathcal{A}}\|$.

Notation & preliminaries contd.

Rank-one tensors

A k-way tensor $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{p_1 \times p_2 \times \cdots \times p_k}$ is **rank-one** if it can be written as the *outer product* of k vectors, i.e.

$$\boldsymbol{\mathcal{A}} = \mathbf{v}^{(1)} \circ \mathbf{v}^{(2)} \circ \cdots \circ \mathbf{v}^{(k)}$$

where "$\circ$" represents the *vector outer product*.

- Each element of the tensor is the product of the corresponding vector elements:

$$x_{i_1 i_2 \cdots i_k} = v^{(1)}_{i_1} v^{(2)}_{i_2} \cdots v^{(k)}_{i_k} \quad \forall 1 \leq i_n \leq p_n.$$

Notation & preliminaries contd.

Rank-one tensors

A k-way tensor $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{p_1 \times p_2 \times \cdots \times p_k}$ is **rank-one** if it can be written as the *outer product* of k vectors, i.e.

$$\boldsymbol{\mathcal{A}} = \mathbf{v}^{(1)} \circ \mathbf{v}^{(2)} \circ \cdots \circ \mathbf{v}^{(k)}$$

where "$\circ$" represents the *vector outer product*.

- Each element of the tensor is the product of the corresponding vector elements:

$$x_{i_1 i_2 \cdots i_k} = v^{(1)}_{i_1} v^{(2)}_{i_2} \cdots v^{(k)}_{i_k} \quad \forall 1 \leq i_n \leq p_n.$$

Definition (Cubical tensors)

A tensor $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{p_1 \times \cdots \times p_k}$ is **cubical** if every mode is same size, i.e. $p_1 = \cdots = p_k = p$; as a shorthand an order-k cubical tensors is denoted as $\mathbf{A} \in \otimes^k \mathbb{R}^p$.

Definition (Symmetric tensors)

A cubical tensor $\mathbf{A} \in \otimes^k \mathbb{R}^p$ is **symmetric** (also referred to as **super-symmetric**) if its k-way representations are invariant to permutations of the array indices: i.e. for all indices $i_i, i_2, \ldots, i_k \in [p]$ and any permutation π on k:

$$a_{i_1 i_2 \ldots i_k} = a_{i_{\pi(1)} i_{\pi(2)} \cdots i_{\pi(k)}}.$$

Notation & preliminaries contd.

Why tensors are important?

Multivariate functions are related to **multidimensional arrays** or *tensors*:

Take a function $f(\mathbf{x}_1, \ldots, \mathbf{x}_p)$; take a tensor-product grid and get a **tensor**, i.e.

$$a_{i_1 i_2 \ldots i_p} = f(\mathbf{x}_1(i_1), \ldots, \mathbf{x}_p(i_p))$$

Notation & preliminaries contd.

Why tensors are important?

Multivariate functions are related to **multidimensional arrays** or *tensors*:

Take a function $f(\mathbf{x}_1, \ldots, \mathbf{x}_p)$; take a tensor-product grid and get a **tensor**, i.e.

$$a_{i_1 i_2 \ldots i_p} = f(\mathbf{x}_1(i_1), \ldots, \mathbf{x}_p(i_p))$$

Where does tensors come from?

- n-th derivative of a multivariate function $f(x_1, \ldots, x_p)$, i.e. $\nabla^n f(x_1, \ldots, x_p)$
- p-dimensional PDE: $\Delta u = f$, $u = u(\mathbf{x}_1, \ldots, \mathbf{x}_p)$
- Data (images, video, hyperspectral images, etc)
- Latent variable models, joint probability distributions
- Many others

Tensor decomposition

Definition (Tensor decomposition [6])

Tensor decomposition refers to the factorization of a tensor into a finite sum of component rank-one tensors.

- This is the analogue of the SVD for matrices and is also known as **parallel factors** and **canonical decompositions**.

Tensor decomposition

Definition (Tensor decomposition [6])

Tensor decomposition refers to the factorization of a tensor into a finite sum of component rank-one tensors.

- This is the analogue of the SVD for matrices and is also known as **parallel factors** and **canonical decompositions**.

Example

Given a order-3 tensor $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{p_1 \times p_2 \times p_3}$, it's decomposition attempts to express it as

$$\boldsymbol{\mathcal{A}} \approx \sum_{r=1}^{R} \mathbf{x}_r \circ \mathbf{y}_r \circ \mathbf{z}_r,$$

where $R > 0$ is integer and for $r = 1, \ldots, R$, $\mathbf{x}_r \in \mathbb{R}^{p_1}$, $\mathbf{y}_r \in \mathbb{R}^{p_2}$, and $\mathbf{z}_r \in \mathbb{R}^{p_3}$. Elementwise, this decomposition can be written as

$$a_{ijk} \approx \sum_{r=1}^{R} x_{ir} y_{jr} z_{kr} \quad \text{for} \quad i = 1, \ldots, p_1, \; j = 1, \ldots, p_2, \; k = 1, \ldots, p_3.$$

Tensor decomposition contd.

Definition (Factor matrices)

Given a decomposition $\boldsymbol{\mathcal{A}} \approx \sum_{r=1}^{R} \mathbf{x}_r \circ \mathbf{y}_r \circ \mathbf{z}_r$, the **factor matrices** refers to the combination of the vectors from the rank-one components, i.e. $\mathbf{X} = [\mathbf{x}_1 \ \mathbf{x}_2 \ \cdots \ \mathbf{x}_R]$ and similarly for $\mathbf{Y}$ and $\mathbf{Z}$.

- Thus tensor decomposition can be concisely written as

$$\boldsymbol{\mathcal{A}} \approx [[\mathbf{X}, \mathbf{Y}, \mathbf{Z}]] \equiv \sum_{r=1}^{R} \mathbf{x}_r \circ \mathbf{y}_r \circ \mathbf{z}_r.$$

- If we assume that the columns of $\mathbf{X}, \mathbf{Y}$, and $\mathbf{Z}$ are **normalized** with the weights absorbed in a vector $\boldsymbol{\lambda}$, then the tensor decomposition can further be expressed as

$$\boldsymbol{\mathcal{A}} = [[\boldsymbol{\lambda}; \mathbf{X}, \mathbf{Y}, \mathbf{Z}]] \equiv \sum_{r=1}^{R} \lambda_r \mathbf{x}_r \circ \mathbf{y}_r \circ \mathbf{z}_r.$$

Tensor rank

Definition (Tensor rank)

The **rank** of a tensor $\boldsymbol{\mathcal{A}}$ denoted $\text{rank}(\boldsymbol{\mathcal{A}})$ is the smallest number of rank-one tensors that generate $\boldsymbol{\mathcal{A}}$ as their sum.

- This is the smallest number of components in an exact tensor decomposition where "exact" means the decomposition holds with *equality*:

$$\boldsymbol{\mathcal{A}} = [[\mathbf{X}, \mathbf{Y}, \mathbf{Z}]] \equiv \sum_{r=1}^{R} \mathbf{x}_r \circ \mathbf{y}_r \circ \mathbf{z}_r.$$

- An **exact** tensor decomposition with $R = \mathrm{rank}(\boldsymbol{\mathcal{A}})$ is called rank decomposition.
- This is the exact analogue of the definition of a matrix rank but the properties of a matrix and a tensor ranks are quite different.

Tensors rank contd.

Tensor rank approximation: caveat!

Not much is known about the generalizability of matrix notions to tensors particularly *rank approximation*.

- The equivalence of the Eckart-Young-Mirsky theorem for rank-k approximation of matrices does not exist for tensors.
  1. For instance, summing k of the factors of a third-order tensor of rank R does not necessarily yield a best rank-k approximation.
  2. Kolda [5] gave an example where the best rank-k approximation of a tensor is not a factor in the best rank-2 approximation.
- The notion of tensor (symmetric) rank is considerably more delicate than matrix (symmetric) rank. For instance:
  1. Not clear *a priori* that the symmetric rank should even be finite [3].
  2. Removal of the best rank-1 approximation of a general tensor may increase the tensor rank of the residual [8].
- It is NP-hard to compute the rank of a tensor in general; only approximations of **(super) symmetric** tensors possible [1].

⋆ Tensors as multilinear maps

- Just as a matrix can be **pre-** & **post-multiplied** by a pair of matrices, an order-k tensor can be *multiplied on k-sides* by k-matrices.

Definition (Multilinear maps with tensors)

For a set of matrices $\left\{\mathbf{X}_i \in \mathbb{R}^{p\times m_i} \mid i \in [k]\right\}$, the $(i_1, i_2, \ldots, i_k)$-th entry of a k-way array representation of $\boldsymbol{\mathcal{A}}(\mathbf{X}_1, \ldots, \mathbf{X}_k) \in \mathbb{R}^{m_1\times\cdots\times m_k}$ is

$$[\boldsymbol{\mathcal{A}}(\mathbf{X}_1, \ldots, \mathbf{X}_k)]_{i_1\ldots i_k} := \sum_{j_1,\ldots,j_k \in [p]} a_{j_1 j_2 \ldots j_k} [X_1]_{j_1 i_1} [X_2]_{j_2 i_2} \ldots [X_k]_{j_k i_k},$$

where $[\mathbf{X}_i]_{jk}$ is the (j, k) entry of a matrix $\mathbf{X}_i$.

Example

1. If $\mathbf{A}$ is a matrix $(k = 2)$, then
$$\mathbf{A}(\mathbf{X}_1, \mathbf{X}_2) = \mathbf{X}_1^T \mathbf{A} \mathbf{X}_2$$
2. For a matrix $\mathbf{A}$ and a vector $\mathbf{x} \in \mathbb{R}^p$, we can express $\mathbf{A}\mathbf{x}$ as
$$\mathbf{A}(\mathbf{I}, \mathbf{x}) = \mathbf{A}\mathbf{x}$$
3. With the canonical basis $\left\{\mathbf{e}_{i_1}, \mathbf{e}_{i_2}, \ldots, \mathbf{e}_{i_k}\right\}$ we have
$$\mathbf{A}\left(\mathbf{e}_{i_1}, \mathbf{e}_{i_2}, \ldots, \mathbf{e}_{i_k}\right) = A_{i_1, i_2, \ldots, i_k}$$

⋆ Tensor compression and Tucker decomposition

- The Tucker decomposition is a form of **higher-order PCA**.
- It also goes by many other names, see [6].

Definition (Tucker decomposition [6])

The **Tucker decomposition** decomposes a tensor into a core tensor multiplied (or transformed) by a matrix along each mode.

Example

- In the case of a third-order tensor $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{p_1 \times p_2 \times p_3}$, we have

$$\boldsymbol{\mathcal{A}} = \sum_{r_1=1}^{R_1} \sum_{r_2=1}^{R_2} \sum_{r_1=3}^{R_3} g_{r_1 r_2 r_3} \mathbf{x}_{r_1} \circ \mathbf{y}_{r_2} \circ \mathbf{z}_{r_3} = [[\boldsymbol{\mathcal{G}}; \mathbf{X}, \mathbf{Y}, \mathbf{Z}]].$$

- The matrices $\mathbf{X} \in \mathbb{R}^{p_1 \times R_1}$, $\mathbf{Y} \in \mathbb{R}^{p_2 \times R_2}$, and $\mathbf{Z} \in \mathbb{R}^{p_3 \times R_3}$ are the factor matrices and are the principal components in each mode.
- The tensor $\boldsymbol{\mathcal{G}} \in \mathbb{R}^{R_1 \times R_2 \times R_3}$ is the core tensor and its entries show the level of interaction between different components.

⋆ Banach's results for tensors

- Banach proved that the maximal overlap between a symmetric tensor and a rank-1 tensor is attained at a symmetric rank-1 tensor.
- Unfortunately, this–seemingly trivial result—is not obvious. That is, if $\mathbf{U} \in \mathrm{Sym}^k(\mathbb{C}^p)$ is a k-index totally symmetric vector with d dimensions per index, then
$$\max \arg_{\mathbf{X}=\mathbf{x}_1 \circ \ldots \circ \mathbf{x}_k, \|\mathbf{x}_i\|_2=1} |\langle \mathbf{X}, \mathbf{U} \rangle|^2$$
fulfills $\mathbf{x}_1 = \ldots = \mathbf{x}_n$.

References

[1] Anima Anandkumar, Rong Ge, Daniel Hsu, Sham M Kakade, and Matus Telgarsky.
Tensor decompositions for learning latent variable models.
arXiv preprint arXiv:1210.7559, 2012.

[2] Waheed U Bajwa, Jarvis D Haupt, Gil M Raz, Stephen J Wright, and Robert D Nowak.
Toeplitz-structured compressed sensing matrices.
In *Statistical Signal Processing, 2007. SSP'07. IEEE/SP 14th Workshop on*, pages 294–298. IEEE, 2007.

[3] Pierre Comon, Gene Golub, Lek-Heng Lim, and Bernard Mourrain.
Symmetric tensors and symmetric tensor rank.
SIAM Journal on Matrix Analysis and Applications, 30(3):1254–1279, 2008.

[4] Simon Foucart and Holger Rauhut.
A mathematical introduction to compressive sensing.
Springer, 2013.

[5] Tamara G Kolda.
Orthogonal tensor decompositions.
SIAM Journal on Matrix Analysis and Applications, 23(1):243–255, 2001.

[6] Tamara G Kolda and Brett W Bader.
Tensor decompositions and applications.
SIAM review, 51(3):455–500, 2009.

References

[7] Holger Rauhut, Justin Romberg, and Joel A Tropp.
Restricted isometries for partial random circulant matrices.
Applied and Computational Harmonic Analysis, 32(2):242–254, 2012.

[8] Alwin Stegeman and Pierre Comon.
Subtracting a best rank-1 approximation may increase tensor rank.
Linear Algebra and its Applications, 433(7):1276–1300, 2010.

[9] Joel A Tropp.
Column subset selection, matrix factorization, and eigenvalue optimization.
In *Proceedings of the Twentieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 978–986, 2009.