Mathematics of Data: From Theory to Computation

Prof. Volkan Cevher
volkan.cevher@epfl.ch

Lecture 1: Objects in space

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

EE-556 (Fall 2016)

License Information for Mathematics of Data Slides

Outline

- This class:
  1. Linear algebra review
     - Notation
     - Vectors
     - Matrices
     - Tensors
- Next class
  1. Review of probability theory

Recommended reading material

- Zico Kolter and Chuong 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.
- Simon Foucart and Holger Rauhut, *A mathematical introduction to compressive sensing* (Appendix A: Matrix Analysis), Springer, 2013.
- Joel A 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 lecture is intended to help you follow mathematical discussions in data sciences, which rely heavily on basic linear algebra concepts.

Notation

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

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 ($\mathbb{C}$).

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. $(\mathbf{x} + \mathbf{y}) + \mathbf{z} = \mathbf{x} + (\mathbf{y} + \mathbf{z}), \forall \mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathcal{V}$ — *associative under addition*
3. $\mathbf{0} + \mathbf{x} = \mathbf{x}, \forall \mathbf{x} \in \mathcal{V}$ — $\mathbf{0}$ *being additive identity*
4. $\forall \mathbf{x} \in \mathcal{V} \ \ \exists\, (-\mathbf{x}) \in \mathcal{V}$ 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. $1\mathbf{x} = \mathbf{x}, \forall \mathbf{x} \in \mathcal{V}$ — $\mathbf{1}$ *being multiplicative identity*

Vector spaces contd.

Example (Vector space)

1. $\mathcal{V}_1 = \{\mathbf{0}\}$ for $\mathbf{0} \in \mathbb{R}^p$
2. $\mathcal{V}_2 = \mathbb{R}^p$
3. $\mathcal{V}_3 = \sum_{i=1}^{k} \alpha_i \mathbf{x}_i$ for $\alpha_i \in \mathbb{R}$ 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)

$\mathcal{V}_1$, $\mathcal{V}_2$, and $\mathcal{V}_3$ in the example above are subspaces 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.,

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

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) $\mathcal{V} = \operatorname{span}\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k\}$, (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)

A 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) $\|\lambda\mathbf{x}\| = |\lambda|\|\mathbf{x}\|$ *homogeniety*

(d) $\|\mathbf{x}+\mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$ *triangle inequality*

- There is a 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: $\|\mathbf{x}\|_\infty := \max_{i=1,\ldots,p} |x_i|$ (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 is in fact 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. any $\mathbf{x} = c(1,1,\ldots,1)^T$ for $c \neq 0$ will have $\|\mathbf{x}\|_{\mathrm{TV}} = 0$ 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 the property (c) $\Rightarrow$ it is also neither a **quasi-** nor a **semi-norm**.

Vector norms contd.

Problem (s-sparse approximation)

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

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

Define $\quad \widehat{\mathbf{y}} \in \arg\min_{\mathbf{x} \in \mathbb{R}^p : \|\mathbf{x}\|_0 \leq s} \|\mathbf{x} - \mathbf{y}\|_2^2 \quad$ and let $\widehat{\mathcal{S}} = \text{supp}\left(\widehat{\mathbf{y}}\right)$.

We now consider an optimization over sets

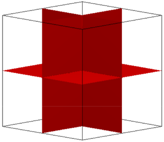
$$
\begin{aligned}
\widehat{\mathcal{S}} &\in \arg\min_{\mathcal{S}:|\mathcal{S}| \leq s} \|\mathbf{y}_{\mathcal{S}} - \mathbf{y}\|_2^2. \\
&\in \arg\max_{\mathcal{S}:|\mathcal{S}| \leq s} \left\{ \|\mathbf{y}\|_2^2 - \|\mathbf{y}_{\mathcal{S}} - \mathbf{y}\|_2^2 \right\} \\
&\in \arg\max_{\mathcal{S}:|\mathcal{S}|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

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

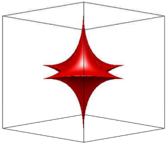
Vector norms contd.

Norm balls

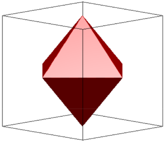
Radius r ball in ℓ_q-norm: $\mathcal{B}_q(r) = \{\mathbf{x} \in \mathbb{R}^p : \|\mathbf{x}\|_q \leq r\}$



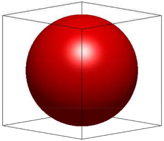
$\|\mathbf{x}\|_0 \leq 2$



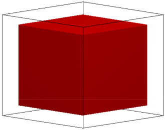
$\ell_{0.5}$-quasi norm ball



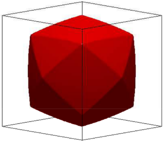
ℓ_1-norm ball



ℓ_2-norm ball



ℓ_∞-norm ball



TV-semi norm ball

Table: Example norm balls 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. $\langle \mathbf{x}, \mathbf{y} \rangle = \langle \mathbf{y}, \mathbf{x} \rangle, \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^p$ *symmetry*
2. $\langle (\alpha \mathbf{x} + \beta \mathbf{y}), \mathbf{z} \rangle = \langle \alpha \mathbf{x}, \mathbf{z} \rangle + \langle \beta \mathbf{y}, \mathbf{z} \rangle, \forall \alpha, \beta \in \mathbb{R}, \forall \mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbb{R}^p$ *linearity*
3. $\langle \mathbf{x}, \mathbf{x} \rangle \geq 0, \forall \mathbf{x} \in \mathbb{R}^p$ *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 $(q = r = 2)$

Definition (Inner product space)

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

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 $q = 1 + 1/\log(p)$?

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 a set, then a function $d(\cdot,\cdot) : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a metric 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: $d_E(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_2$
- Bregman distance: $d_B(\cdot,\cdot)$ ...more on this later!

Basic 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 $\mathbf{A}$.

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

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

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

Matrix definitions contd.

Definition (Eigenvalues & Eigenvectors)

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

- $\mathbf{A}$ scales its eigenvectors by it eigenvalues.

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

Lemma

The eigenvalues of a **symmetric** $\mathbf{A}$ *are real.*

Proof.

Assume $\mathbf{A}\mathbf{x} = \lambda\mathbf{x},\ \mathbf{x} \in \mathbb{C}^p, \mathbf{x} \neq \mathbf{0}$, then $\overline{\mathbf{x}}^T\mathbf{A}\mathbf{x} = \overline{\mathbf{x}}^T(\mathbf{A}\mathbf{x}) = \overline{\mathbf{x}}^T(\lambda\mathbf{x}) = \lambda\sum_{i=1}^n |x_i|^2$
but $\overline{\mathbf{x}}^T\mathbf{A}\mathbf{x} = \overline{(\mathbf{A}\mathbf{x})}^T\mathbf{x} = \overline{(\lambda\mathbf{x})}^T\mathbf{x} = \overline{\lambda}\sum_{i=1}^n |x_i|^2 \ \Rightarrow\ \lambda = \overline{\lambda}$ i.e. $\lambda \in \mathbb{R}$ □

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 $\mathbf{x}^T \mathbf{A} \mathbf{x} \geq 0$ for all $\mathbf{x} \neq \mathbf{0}$; 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 $\mathbf{A}^{-1} \succ 0$

Matrix decompositions

Definition (Eigenvalue decomposition)

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

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

- the columns of $\mathbf{X} \in \mathbb{R}^{n\times n}$, i.e. $\mathbf{x}_i$, are eigenvectors of $\mathbf{A}$
- $\mathbf{\Lambda} = \mathbf{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$ where λ_i (also denoted $\lambda_i(\mathbf{A})$) are eigenvalues of $\mathbf{A}$
- A matrix that admits this decomposition is therefore called **diagonalizable** matrix

Eigendecomposition of symmetric matrices

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

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

- $\lambda_n(\mathbf{A}) = \lambda_{\min}(\mathbf{A})$ 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 (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 $\mathbf{A}$.

Matrix decompositions contd

Definition (Singular value decomposition)

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

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

- $\mathrm{rank}(\mathbf{A}) = r \leq \min(n, p)$ 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. $\mathbf{U}^T\mathbf{U} = \mathbf{I}$)
- $\mathbf{\Sigma} = \mathbf{diag}\,(\sigma_1, \sigma_2, \ldots, \sigma_r)$ where $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r \geq 0$

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

Matrix decompositions contd

Definition (LU)

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

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

where $\mathbf{P}$ is a **permutation matrix**[1], $\mathbf{L}$ is **lower triangular** and $\mathbf{U}$ is **upper triangular**.

Definition (QR)

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

$$\mathbf{A} = \mathbf{QR}$$

where $\mathbf{Q} \in \mathbb{R}^{n \times n}$ is an **orthonormal** 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 and symmetric matrix, $\mathbf{A} \in \mathbb{R}^{p \times p}$, is given by:

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

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

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

Complexity of matrix operations

Complexity of matrix operations

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.

Complexity of matrix operations contd

Table: Complexity examples: vector are in $\mathbb{R}^p$, matrices in $\mathbb{R}^{n\times p}$ or $\mathbb{R}^{p\times m}$ for square matrices

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 much less if $\mathbf{A}$ is sparse[1]
LU decomposition	$\frac{2}{3}p^3 + 2p^2$ flops	or $\frac{2}{3}p^3$ for p large much less if $\mathbf{A}$ is sparse[1]
Cholesky decomposition	$\frac{1}{3}p^3 + 2p^2$ flops	or $\frac{1}{3}p^3$ for p large much less if $\mathbf{A}$ is sparse[1]
SVD	$C_1 n^2 p + C_2 p^3$ flops	$C_1 = 4$, $C_2 = 22$ for R-SVD algo.
Determinant	complexity of SVD	

[1] Complexity depends on p, no. of nonzeros in $\mathbf{A}$ and the sparsity pattern.

Matrix norms

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

Definition (Matrix norm)

A 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

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

Linear operators

- Matrices are often given in an **implicit** form.
- 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* $\mathcal{M} : \mathbb{R}^{n \times p} \to \mathbb{R}^{np}$, a linear operator can define the following implicit mapping

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

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* $\left(\mathbf{B}^T \otimes \mathbf{A}\right) \mathrm{vec}(\mathbf{X})$.

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

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

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**:

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

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

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

Matrix norms contd.

Useful relation for operator norms

The following **identity** holds

$$\|\mathbf{A}\|_{q\to r} := \max_{\|\mathbf{z}\|_r\le 1, \|\mathbf{x}\|_q=1} \langle \mathbf{z}, \mathbf{A}\mathbf{x}\rangle = \max_{\|\mathbf{x}\|_{q'}\le 1, \|\mathbf{z}\|_{r'}=1} \langle \mathbf{A}^T\mathbf{z}, \mathbf{x}\rangle =: \|\mathbf{A}^T\|_{q'\to r'}$$

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

Example

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

*Matrix norms contd.

Computation of operator norms

- The computation of some **operator norms** is NP-hard* [3]; 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** [5]; these include
  1. $\|\mathbf{A}\|_{\infty\to 1}$ (via Gronthendieck factorization)
  2. $\|\mathbf{A}\|_{\infty\to 2}$ (via Pietzs factorization)

*: See Lecture 3.

Matrix norms contd.

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

Definition (Schatten q-norms)

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

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

$$\|\mathbf{A}\|_1 = \|\mathbf{A}\|_* := \sum_{i=1}^{r} \sigma_i \equiv \mathrm{trace}\left(\sqrt{\mathbf{A}^T\mathbf{A}}\right) \quad \text{(Nuclear/trace)}$$

$$\|\mathbf{A}\|_2 = \|\mathbf{A}\|_F := \sqrt{\sum_{i=1}^{r} (\sigma_i)^2} \equiv \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{p} |a_{ij}|^2} \quad \text{(Frobenius)}$$

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

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: $\quad \operatorname{rank}(\mathbf{X}) \leq r.$

Solution (Eckart–Young–Mirsky Theorem)

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

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 of a matrix)

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

$$\|\mathbf{A}\|^* = \sup\left\{\operatorname{trace}\left(\mathbf{A}^T\mathbf{X}\right) \mid \|\mathbf{X}\| \leq 1\right\}.$$

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	$\|\mathbf{X}\|$	$\|\mathbf{X}\|_F$	$\|\mathbf{X}\|_*$

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. $\|\mathbf{A}\|_{2\to 2}^2 \leq \|\mathbf{A}\|_{1\to 1}$ 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 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)

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

Theorem (Interlacing property)

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

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

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

*Tensors

- Tensors provide a natural and concise mathematical represention of **data**.

Definition (Tensor)

An m^{th}-**rank** 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, order is used interchangeably with rank, i.e., k^{th}-**rank** tensor is also referred to as an **order**-k tensor.

- Tensors are **multidimensional arrays** and are a generalization of:
  1. scalars - **tensors** with no *indices*; i.e., zeroth-rank tensor.
  2. vectors - **tensors** with exactly one *index*; i.e., first-rank tensor.
  3. matrices - **tensors** with exactly two *indices*; i.e., second-rank tensor.

- Think of the third-order Taylor series expansion

*Tensors contd.

Caveat!

Not much is known about tensors and the generalizability of matrix notions to tensors:

- 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 [2].
  2. Removal of the best rank-1 approximation of a general tensor may increase the tensor rank of the residual [4].
- It is NP-hard to compute the rank of a tensor in general; only approximations of **(super) symmetric** tensors possible [1].

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] 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.

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

[4] 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.

[5] 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.