Mathematics of Data: From Theory to Computation

Prof. Volkan Cevher
volkan.cevher@epfl.ch

Lecture 3: Convex analysis and Linear Algebra

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

EE-556 (Fall 2018)

lions@epfl

MARIE CURIE ACTIONS

FNSNF
FONDS NATIONAL SUISSE
SCHWEIZERISCHER NATIONALFONDS
FONDO NAZIONALE SVIZZERO
SWISS NATIONAL SCIENCE FOUNDATION

erc

EPFL

License Information for Mathematics of Data Slides

Outline

- This lecture
  1. Basic concepts in convex analysis
  2. Basic review of linear algebra
- Next lecture
  1. Unconstrained convex optimization: the basics
  2. Gradient descent methods

Recommended reading

- Chapter 2 & 3 in S. Boyd, and L. Vandenberghe, *Convex Optimization*, Cambridge Univ. Press, 2009.
- Appendices A & B in D. Bertsekas, *Nonlinear Programming*, Athena Scientific, 1999.
- *Matrix computations*, G.H. Golub, C.F. Van Loan, JHU Press, 2012.
- *Linear algebra and its applications*, G. Strang, Thomson, Brooks/Cole, 2006.
- KC Border, *Quick Review of Matrix and Real Linear Algebra* `http://www.hss.caltech.edu/~kcb/Notes/LinearAlgebra.pdf`, 2013.

Motivation

Motivation

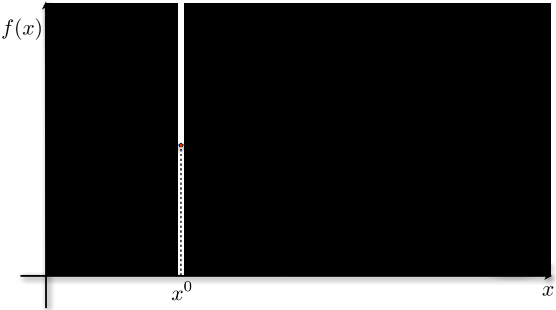
- The first part of this lecture introduces basic notions in convex analysis.
- The second part reviews some concepts in linear algebra.

Challenges for an iterative optimization algorithm

Problem

Find the minimum $x^\star$ of $f(x)$, given starting point x^0 based on only local information.

- **Fog of war**

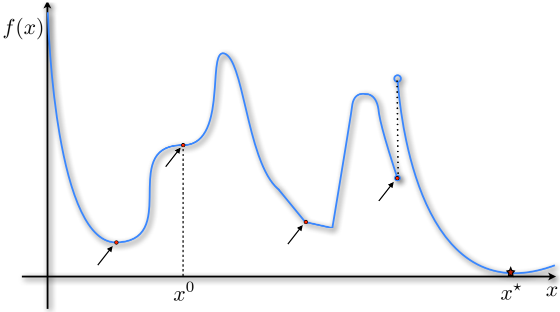


Challenges for an iterative optimization algorithm

Problem

Find the minimum $x^\star$ of $f(x)$, given starting point x^0 based on only local information.

- Fog of war, non-differentiability, discontinuities, local minima, stationary points...

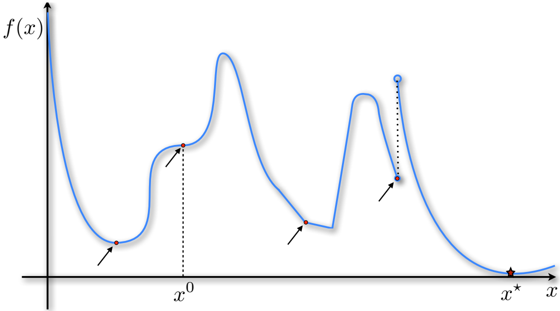


Challenges for an iterative optimization algorithm

Problem

Find the minimum $x^\star$ of $f(x)$, given starting point x^0 based on only local information.

- Fog of war, non-differentiability, discontinuities, local minima, stationary points...



We need a key structure on the function local minima: Convexity.

Basics of functions

Definition (Function)

A function f with domain $\mathcal{Q} \subseteq \mathbb{R}^p$ and codomain $\mathcal{U} \subseteq \mathbb{R}$ is denoted as:

$$f \ : \ \mathcal{Q} \to \mathcal{U}.$$

The domain $\mathcal{Q}$ represents the set of values in $\mathbb{R}^p$ on which f is defined and is denoted as $\mathrm{dom}(f) \equiv \mathcal{Q} = \{\mathbf{x} : -\infty < f(\mathbf{x}) < +\infty\}$. The codomain $\mathcal{U}$ is the set of function values of f for any input in $\mathcal{Q}$.

Continuity in functions

Definition (Continuity)

Let $f : \mathcal{Q} \to \mathbb{R}$ where $\mathcal{Q} \subseteq \mathbb{R}^p$. Then, f is a continuous function over its domain $\mathcal{Q}$ if and only if

$$\lim_{\mathbf{x} \to \mathbf{y}} f(\mathbf{x}) = f(\mathbf{y}), \quad \forall \mathbf{y} \in \mathcal{Q},$$

i.e., the limit of f—as $\mathbf{x}$ approaches $\mathbf{y}$—exists and is equal to $f(\mathbf{y})$.

Definition (Class of continuous functions)

We denote the class of continuous functions f over the domain $\mathcal{Q}$ as $f \in \mathcal{C}(\mathcal{Q})$.

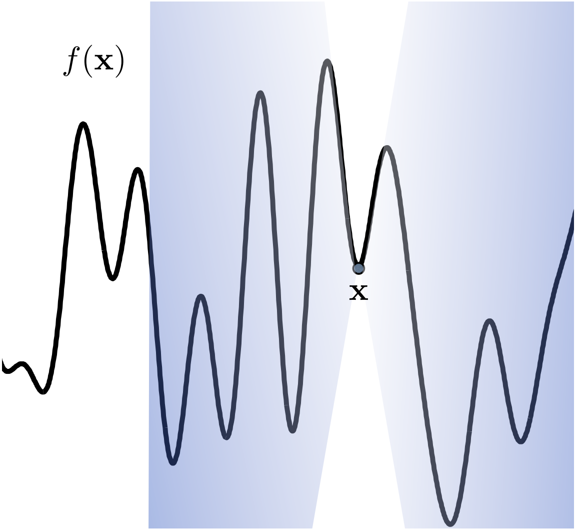
Definition (Lipschitz continuity)

Let $f : \mathcal{Q} \to \mathbb{R}$ where $\mathcal{Q} \subseteq \mathbb{R}^p$. Then, f is called Lipschitz continuous if there exists a constant value $K \geq 0$ such that:

$$|f(\mathbf{y}) - f(\mathbf{x})| \leq K \|\mathbf{y} - \mathbf{x}\|_2, \quad \forall \mathbf{x},\ \mathbf{y} \in \mathcal{Q}.$$

- "Small" changes in the input result into "small" changes in the function values.

Continuity in functions



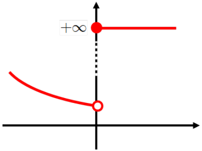
Lower semi-continuity

Definition

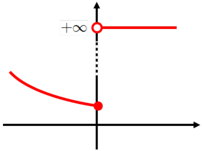
A function $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is lower semi-continuous (l.s.c.) if

$$\liminf_{\mathbf{x} \to \mathbf{y}} f(\mathbf{x}) \geq f(\mathbf{y}), \quad \text{for any } \mathbf{y} \in \mathrm{dom}(f).$$

$$f(x) = \begin{cases} e^{-x}, & \text{if } x < 0 \\ +\infty, & \text{if } x \geq 0 \end{cases}$$



$$f(x) = \begin{cases} e^{-x}, & \text{if } x \leq 0 \\ +\infty, & \text{if } x > 0 \end{cases}$$



Unless stated otherwise, we only consider l.s.c. functions.

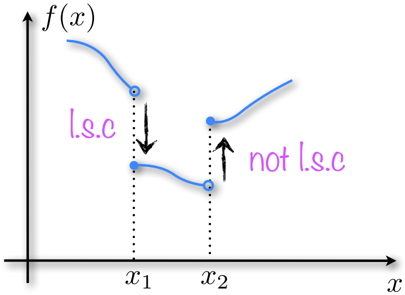
Lower semi-continuity

Definition

A function $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is lower semi-continuous (l.s.c.) if

$$\liminf_{\mathbf{x} \to \mathbf{y}} f(\mathbf{x}) \geq f(\mathbf{y}), \quad \text{for any } \mathbf{y} \in \mathrm{dom}(f).$$

- **Intuition:** A lower semi-continuous function *only jumps down.*



Differentiability in functions

- We use $\nabla f(\mathbf{x})$ to denote the *gradient* of f at $\mathbf{x} \in \mathbb{R}^p$ such that:

$$\nabla f(\mathbf{x}) = \sum_{i=1}^{p} \frac{\partial f}{\partial x_i} \mathbf{e}_i = \left[\frac{\partial f}{\partial x_1}, \cdots, \frac{\partial f}{\partial x_p} \right]^T$$

Example: $f(\mathbf{x}) = \|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2^2$

$\nabla f(\mathbf{x}) = -2\mathbf{A}^T(\mathbf{b} - \mathbf{A}\mathbf{x})$.

Definition (Differentiability)

Let $f \in \mathcal{C}(\mathcal{Q})$ where $\mathcal{Q} \subseteq \mathbb{R}^p$. Then, f is a k-times continuously differentiable on $\mathcal{Q}$ if its partial derivatives up to k-th order exist and are continuous $\forall \mathbf{x} \in \mathcal{Q}$.

Definition (Class of differentiable functions)

We denote the class of k-times continuously differentiable functions f on $\mathcal{Q}$ as $f \in \mathcal{C}^k(\mathcal{Q})$.

- In the special case of $k = 2$, we dub $\nabla^2 f(\mathbf{x})$ the **Hessian** of $f(\mathbf{x})$, where $[\nabla^2 f(\mathbf{x})]_{i,j} = \frac{\partial^2 f}{\partial x_i \partial x_j}$.
- We have $\mathcal{C}^q(\mathcal{Q}) \subseteq \mathcal{C}^k(\mathcal{Q})$ where $q \leq k$. For example, a twice differentiable function is also once differentiable.
- For the case of complex-valued matrices, we refer to the Matrix Cookbook online.

Differentiability in functions

- Some examples:

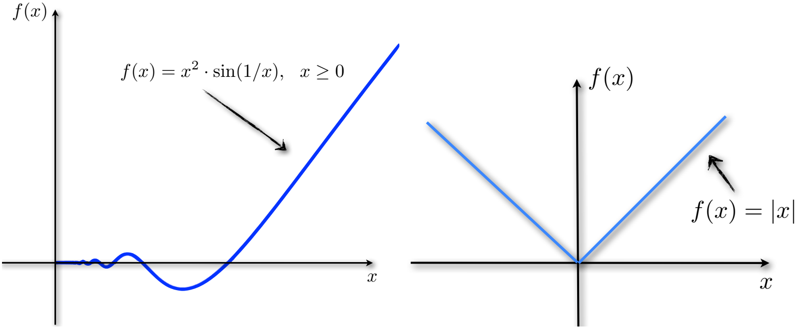


Figure: **(Left panel)** ∞-times continuously differentiable function in $\mathbb{R}$. **(Right panel)** Non-differentiable $f(x) = |x|$ in $\mathbb{R}$.

Stationary points of differentiable functions

Definition (Stationary point)

A point $\bar{\mathbf{x}}$ is called a stationary point of a twice differentiable function $f(\mathbf{x})$ if

$$\nabla f(\bar{\mathbf{x}}) = \mathbf{0}.$$

Definition (Local minima, maxima, and saddle points)

Let $\bar{\mathbf{x}}$ be a stationary point of a twice differentiable function $f(\mathbf{x})$.

- If $\nabla^2 f(\bar{\mathbf{x}}) \succ 0$, then the point $\bar{\mathbf{x}}$ is called a local minimum.
- If $\nabla^2 f(\bar{\mathbf{x}}) \prec 0$, then the point $\bar{\mathbf{x}}$ is called a local maximum.
- If $\nabla^2 f(\bar{\mathbf{x}}) = 0$, then the point $\bar{\mathbf{x}}$ can be a saddle point depending on the sign change.

Stationary points of smooth functions contd.

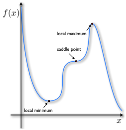
Intuition

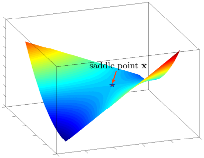
Recall Taylor's theorem for the function f around $\bar{\mathbf{x}}$ for all $\mathbf{y}$ that satisfy $\|\mathbf{y}-\bar{\mathbf{x}}\|_2 \leq r$ in a local region with radius r as follows

$$f(\mathbf{y}) = f(\bar{\mathbf{x}}) + \langle \nabla f(\bar{\mathbf{x}}), \mathbf{y}-\bar{\mathbf{x}}\rangle + \frac{1}{2}(\mathbf{y}-\bar{\mathbf{x}})^T \nabla^2 f(\mathbf{z})(\mathbf{y}-\bar{\mathbf{x}}),$$

where $\mathbf{z}$ is a point between $\bar{\mathbf{x}}$ and $\mathbf{y}$. When $r \to 0$, the second-order term becomes $\nabla^2 f(\mathbf{z}) \to \nabla^2 f(\bar{\mathbf{x}})$. Since $\nabla f(\bar{\mathbf{x}}) = 0$, Taylor's theorem leads to

- $f(\mathbf{y}) > f(\bar{\mathbf{x}})$ when $\nabla^2 f(\bar{\mathbf{x}}) \succ 0$. Hence, the point $\bar{\mathbf{x}}$ is a local minimum.
- $f(\mathbf{y}) < f(\bar{\mathbf{x}})$ when $\nabla^2 f(\bar{\mathbf{x}}) \prec 0$. Hence, the point $\bar{\mathbf{x}}$ is a local maximum.
- $f(\mathbf{y}) \gtrless f(\bar{\mathbf{x}})$ when $\nabla^2 f(\bar{\mathbf{x}}) = 0$. Hence, the point $\bar{\mathbf{x}}$ can be a saddle point (i.e., $f(x) = x^3$ at $\bar{x} = 0$), a local minima (i.e., $f(x) = x^4$ at $\bar{x} = 0$) or a local maxima (i.e., $f(x) = -x^4$ at $\bar{x} = 0$).





Convexity

Definition

A function $f : \mathcal{Q} \to \mathbb{R} \cup \{+\infty\}$ is called convex on its domain $\mathcal{Q}$ if, for any $\mathbf{x}_1,\ \mathbf{x}_2 \in \mathcal{Q}$ and $\alpha \in [0,1]$, we have:

$$f(\alpha \mathbf{x}_1 + (1-\alpha)\mathbf{x}_2) \leq \alpha f(\mathbf{x}_1) + (1-\alpha) f(\mathbf{x}_2).$$

- If $-f(\mathbf{x})$ is convex, then $f(\mathbf{x})$ is called concave.

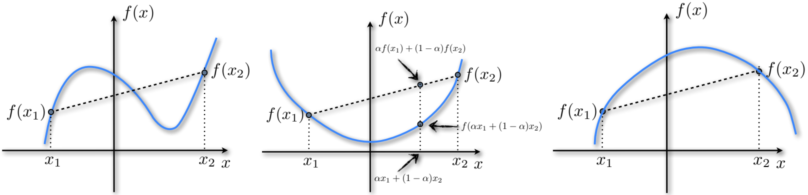


Figure: (**Left**) Non-convex (**Middle**) Convex (**Right**) Concave

Convexity

Definition

A function $f : \mathcal{Q} \to \mathbb{R} \cup \{+\infty\}$ is called convex on its domain $\mathcal{Q}$ if, for any $\mathbf{x}_1,\ \mathbf{x}_2 \in \mathcal{Q}$ and $\alpha \in [0,1]$, we have:

$$f(\alpha\mathbf{x}_1 + (1-\alpha)\mathbf{x}_2) \leq \alpha f(\mathbf{x}_1) + (1-\alpha) f(\mathbf{x}_2).$$

- Additional terms that you will encounter in the literature

Definition (Proper)

A convex function f is called proper if its domain satisfies $\mathrm{dom}(f) \neq \emptyset$ and, $f(\mathbf{x}) > -\infty,\ \forall x \in \mathrm{dom}(f)$.

Definition (Extended real-valued convex functions)

We define the extended real-valued convex functions f as

$$f(\mathbf{x}) = \begin{cases} f(\mathbf{x}) & \text{if } \mathbf{x} \in \mathrm{dom}(f) \\ +\infty & \text{if otherwise} \end{cases}$$

To denote this concept, we use $f : \mathrm{dom}(f) \to \mathbb{R} \cup \{+\infty\}$. (Note how l.s.c. might be useful)

Convexity

Definition

A function $f : \mathcal{Q} \to \mathbb{R} \cup \{+\infty\}$ is called convex on its domain $\mathcal{Q}$ if, for any $\mathbf{x}_1,\ \mathbf{x}_2 \in \mathcal{Q}$ and $\alpha \in [0, 1]$, we have:

$$f(\alpha\mathbf{x}_1 + (1-\alpha)\mathbf{x}_2) \leq \alpha f(\mathbf{x}_1) + (1-\alpha) f(\mathbf{x}_2).$$

Example

Function	Example	Attributes
ℓ_p vector norms, $p \geq 1$	$\\|\mathbf{x}\\|_2, \\|\mathbf{x}\\|_1, \\|\mathbf{x}\\|_\infty$	convex
ℓ_p matrix norms, $p \geq 1$	$\\|\mathbf{X}\\|_* = \sum_{i=1}^{\text{rank}(\mathbf{X})} \sigma_i$	convex
Square root function	$\sqrt{x}$	concave, nondecreasing
Maximum of functions	$\max\{x_1, \ldots, x_n\}$	convex, nondecreasing
Minimum of functions	$\min\{x_1, \ldots, x_n\}$	concave, nondecreasing
Sum of convex functions	$\sum_{i=1}^{n} f_i$, f_i convex	convex
Logarithmic functions	$\log(\det(\mathbf{X}))$	concave, assumes $\mathbf{X} \succ 0$
Affine/linear functions	$\sum_{i=1}^{n} X_{ii}$	both convex and concave
Eigenvalue functions	$\lambda_{max}(\mathbf{X})$	convex, assumes $\mathbf{X} = \mathbf{X}^T$

Strict convexity

Definition

A function $f : \mathcal{Q} \to \mathbb{R} \cup \{+\infty\}$ is called *strictly convex* on its domain $\mathcal{Q}$ if and only if for any $\mathbf{x}_1,\ \mathbf{x}_2 \in \mathcal{Q}$ and $\alpha \in [0, 1]$ we have:

$$f(\alpha\mathbf{x}_1 + (1-\alpha)\mathbf{x}_2) < \alpha f(\mathbf{x}_1) + (1-\alpha) f(\mathbf{x}_2).$$

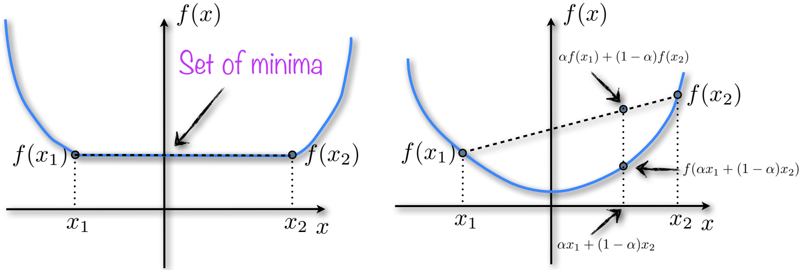


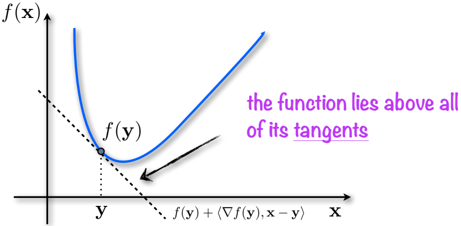
Figure: **(Left panel)** Convex function. **(Right panel)** Strictly convex function.

Revisiting: Alternative definitions of function convexity II

Definition

A function $f \in \mathcal{C}^1(\mathcal{Q})$ is called convex on its domain if for any $\mathbf{x},\ \mathbf{y} \in \mathcal{Q}$:

$$f(\mathbf{x}) \geq f(\mathbf{y}) + \langle \nabla f(\mathbf{y}),\ \mathbf{x} - \mathbf{y} \rangle.$$



Definition

A function $f \in \mathcal{C}^1(\mathcal{Q})$ is called convex on its domain if for any $\mathbf{x},\ \mathbf{y} \in \mathcal{Q}$:

$$\langle \nabla f(\mathbf{y}) - \nabla f(\mathbf{x}),\ \mathbf{y} - \mathbf{x} \rangle \geq 0.$$

*That is, if its gradient is a monotone operator.

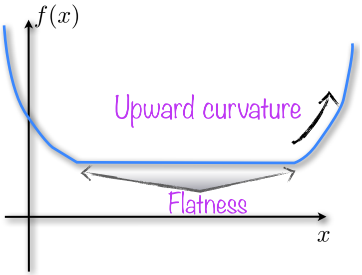
Revisiting: Alternative definitions of function convexity III

Definition

A function $f \in \mathcal{C}^2(\mathbb{R}^p)$ is called convex on its domain if for any $\mathbf{x},\ \mathbf{y} \in \mathbb{R}^p$:

$$\nabla^2 f(\mathbf{x}) \succeq 0.$$

- Geometrical interpretation: the graph of f has zero or positive (upward) curvature.
- However, this does not exclude flatness of f.
- $\nabla^2 f(\mathbf{x}) \succ 0$ is a sufficient condition for *strict* convexity.



Stationary points and convexity

Lemma

Let f be a **smooth convex** *function, i.e., $f \in \mathcal{F}^1$. Then, any stationary point of f is also a global minimum.*

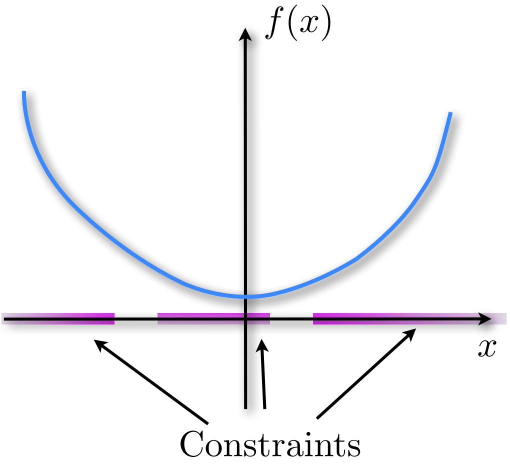
Proof.

Let $\mathbf{x}^\star$ be a stationary point, i.e., $\nabla f(\mathbf{x}^\star) = 0$. By convexity, we have:

$$f(\mathbf{x}) \geq f(\mathbf{x}^\star) + \langle \nabla f(\mathbf{x}^\star), \ \mathbf{x} - \mathbf{x}^\star \rangle \overset{\nabla f(\mathbf{x}^\star)=0}{=} f(\mathbf{x}^\star) \quad \text{for all } \mathbf{x} \in \mathbb{R}^p.$$

□

Is convexity of f enough for an iterative optimization algorithm?



Convexity over sets

Definition

- $\mathcal{Q} \subseteq \mathbb{R}^p$ is a convex set if $\mathbf{x}_1,\ \mathbf{x}_2 \in \mathcal{Q} \Rightarrow \forall \alpha \in [0,1],\quad \alpha \mathbf{x}_1 + (1-\alpha)\mathbf{x}_2 \in \mathcal{Q}$.
- $\mathcal{Q} \subseteq \mathbb{R}^p$ is a *strictly* convex set if

$$\mathbf{x}_1,\ \mathbf{x}_2 \in \mathcal{Q} \Longrightarrow \forall \alpha \in (0,1),\quad \alpha \mathbf{x}_1 + (1-\alpha)\mathbf{x}_2 \in \text{interior}(\mathcal{Q}).$$

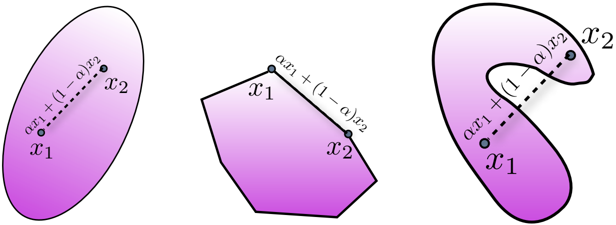


Figure: (**Left**) Strictly convex (**Middle**) Convex (**Right**) Non-convex

Convexity over sets

Definition

- $\mathcal{Q} \subseteq \mathbb{R}^p$ is a convex set if $\mathbf{x}_1,\ \mathbf{x}_2 \in \mathcal{Q} \Rightarrow \forall \alpha \in [0,1],\quad \alpha\mathbf{x}_1 + (1-\alpha)\mathbf{x}_2 \in \mathcal{Q}$.
- $\mathcal{Q} \subseteq \mathbb{R}^p$ is a *strictly* convex set if
$$\mathbf{x}_1,\ \mathbf{x}_2 \in \mathcal{Q} \Longrightarrow \forall \alpha \in (0,1),\quad \alpha\mathbf{x}_1 + (1-\alpha)\mathbf{x}_2 \in \text{interior}(\mathcal{Q}).$$

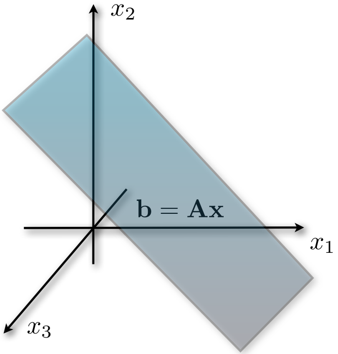


Figure: A linear set of equations $\mathbf{b} = \mathbf{A}\mathbf{x}$ defines an affine (thus convex) set.

Convexity over sets

Definition

- $\mathcal{Q} \subseteq \mathbb{R}^p$ is a convex set if $\mathbf{x}_1,\ \mathbf{x}_2 \in \mathcal{Q} \Rightarrow \forall \alpha \in [0,1],\quad \alpha\mathbf{x}_1 + (1-\alpha)\mathbf{x}_2 \in \mathcal{Q}$.
- $\mathcal{Q} \subseteq \mathbb{R}^p$ is a *strictly* convex set if
  $$\mathbf{x}_1,\ \mathbf{x}_2 \in \mathcal{Q} \Longrightarrow \forall \alpha \in (0,1),\quad \alpha\mathbf{x}_1 + (1-\alpha)\mathbf{x}_2 \in \text{interior}(\mathcal{Q}).$$

Why is this also important/useful?

- convex sets $<>$ convex optimization constraints

$$\begin{array}{ll} \underset{\mathbf{x}}{\text{minimize}} & f_0(\mathbf{x}) \\ \text{subject to} & \text{constraints} \end{array}$$

Some basic notions on sets

Definition (Closed set)

A set is called *closed* if it contains all its limit points.

Definition (Closure of a set)

Let $\mathcal{Q} \subseteq \mathbb{R}^p$ be a given open set, i.e., the limit points on the boundaries of $\mathcal{Q}$ do not belong into $\mathcal{Q}$. Then, the closure of $\mathcal{Q}$, denoted as $\mathrm{cl}(\mathcal{Q})$, is the smallest set in $\mathbb{R}^p$ that includes $\mathcal{Q}$ with its boundary points.

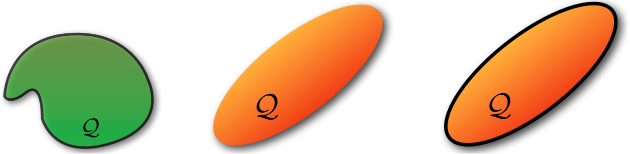


Figure: (**Left panel**) Closed set $\mathcal{Q}$. (**Middle panel**) Open set $\mathcal{Q}$ and its closure $\mathrm{cl}(\mathcal{Q})$ (**Right panel**).

Convex hull

Definition (Convex hull)

Let $\mathcal{V} \subseteq \mathbb{R}^p$ be a set. The convex hull of $\mathcal{V}$, i.e., conv$(\mathcal{V})$, is the *smallest* convex set that contains $\mathcal{V}$.

Definition (Convex hull of points)

Let $\mathcal{V} \subseteq \mathbb{R}^p$ be a finite set of points with cardinality $|\mathcal{V}|$. The convex hull of $\mathcal{V}$ is the set of all convex combinations of its points, i.e.,

$$\text{conv}(\mathcal{V}) = \left\{ \sum_{i=1}^{|\mathcal{V}|} \alpha_i \mathbf{x}_i \ : \ \sum_{i=1}^{|\mathcal{V}|} \alpha_i = 1,\ \alpha_i \geq 0, \forall i,\ \mathbf{x}_i \in \mathcal{V} \right\}.$$

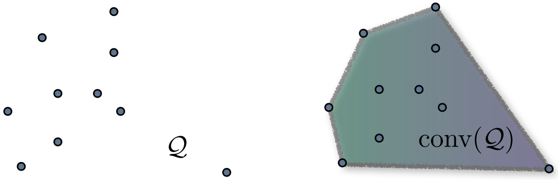


Figure: (**Left**) Discrete set of points $\mathcal{V}$. (**Right**) Convex hull conv$(\mathcal{V})$.

Revisiting: Alternative definitions of function convexity IV

Definition

The epigraph of a function $f : \mathcal{Q} \to \mathbb{R}, \mathcal{Q} \subseteq \mathbb{R}^p$ is the subset of $\mathbb{R}^{p+1}$ given by:

$$\text{epi}(f) = \{(\mathbf{x}, w) : \mathbf{x} \in \mathcal{Q}, w \in \mathbb{R}, f(\mathbf{x}) \leq w\}.$$

Lemma

A function $f : \mathcal{Q} \to \mathbb{R}$ is convex if and only if its epigraph, i.e, the region above its graph, is a convex set.

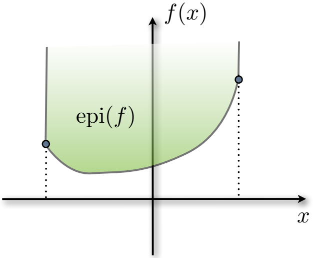


Figure: Epigraph — the region in green above graph $f(\cdot)$.

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 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

$$
\begin{array}{llll}
(1) & \ell_2\text{-norm:} & \|\mathbf{x}\|_2 := \sqrt{\sum_{i=1}^p x_i^2} & \text{(Euclidean norm)} \\
(2) & \ell_1\text{-norm:} & \|\mathbf{x}\|_1 := \sum_{i=1}^p |x_i| & \text{(Manhattan norm)} \\
(3) & \ell_\infty\text{-norm:} & \|\mathbf{x}\|_\infty := \max\limits_{i=1,\ldots,p} |x_i| & \text{(Chebyshev norm)}
\end{array}
$$

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 $\widehat{\mathcal{S}} = \mathrm{supp}\left(\widehat{\mathbf{y}}\right)$.

We now consider an optimization over sets

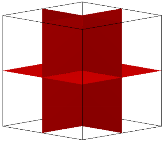
$$
\begin{aligned}
\widehat{\mathcal{S}} &\in \arg\min_{\mathcal{S}:|\mathcal{S}|\leq s} \|\mathbf{y}_{\mathcal{S}}-\mathbf{y}\|_2^2. \\
&\in \arg\max_{\mathcal{S}:|\mathcal{S}|\leq s} \left\{\|\mathbf{y}\|_2^2 - \|\mathbf{y}_{\mathcal{S}}-\mathbf{y}\|_2^2\right\} \\
&\in \arg\max_{\mathcal{S}:|\mathcal{S}|\leq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

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

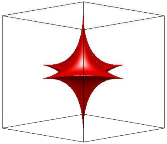
Vector norms contd.

Norm balls

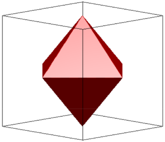
Radius r ball in ℓ_q-norm: $\mathcal{B}_q(r) = \{\mathbf{x} \in \mathbb{R}^p : \|\mathbf{x}\|_q \leq r\}$



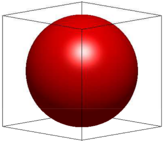
$\|\mathbf{x}\|_0 \leq 2$



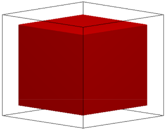
$\ell_{0.5}$-quasi norm ball



ℓ_1-norm ball



ℓ_2-norm ball



ℓ_∞-norm ball

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

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

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

Matrix definitions contd.

Definition (Eigenvalues & Eigenvectors)

The non-zero 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 $\mathbf{A}$.

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:

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

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

Matrix decompositions contd

Definition (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

Definition (floating-point operation)

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

Table: Complexity examples: vector are in $\mathbb{R}^p$, matrices in $\mathbb{R}^{n\times p}$, $\mathbb{R}^{p\times m}$ or $\mathbb{R}^{p\times p}$ [2]

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.

Computing eigenvalues and eigenvectors

- There are various algorithms to compute eigenpairs of matrices [3].
- One can choose an algorithm depending on the setting (computational complexity, number of eigenvalues/eigenvectors needed etc.)

Power Method

Starting with an initial vector $\mathbf{x}^0$, $\mathbf{x}^{k+1} = \frac{\mathbf{A}x^k}{\|\mathbf{A}x^k\|_2}$ converges to the leading eigenvector of the matrix $\mathbf{A}$ under certain conditions. Moreover, $\lambda^k = \frac{{\mathbf{x}^k}^* \mathbf{A}\mathbf{x}^k}{{\mathbf{x}^k}^* \mathbf{x}^k}$ converges to the leading eigenvalue.

- Useful when $\mathbf{A}$ is a large matrix with sparse entries as it does not require matrix decomposition, but only matrix-vector multiplications and normalizations.
- Used by PageRank algorithm of Google.

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 $(1 \le q, r \le \infty)$ of a matrix $\mathbf{A}$ is defined as

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

Problem

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

Solution

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

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**:

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

Matrix norms contd.

Useful relation for operator norms

The following **identity** holds

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

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 [1]; 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** [4]; these include
  1. $\|\mathbf{A}\|_{\infty\to 1}$ (via Gronthendieck factorization)
  2. $\|\mathbf{A}\|_{\infty\to 2}$ (via Pietzs factorization)

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

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

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 $\sigma_i = \sigma(\mathbf{A})_i$)

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

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

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

Matrix norms contd.

Problem (Rank-r approximation)

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

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)

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

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.

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.

References I

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

[2] Gene H Golub and Charles F Van Loan.
Matrix computations, volume 3.
JHU Press, 2012.

[3] Gilbert Strang.
Linear algebra and its applications.
Number 04; QA184, S8. 1976.

[4] 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.