Mathematics of Data: From Theory to Computation

Prof. Volkan Cevher
volkan.cevher@epfl.ch

Lecture 3: Some basics on optimization

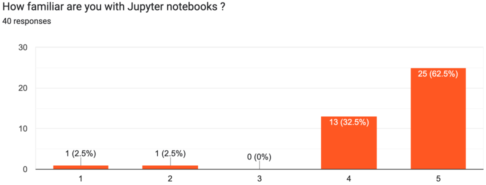
Laboratory for Information and Inference Systems (LIONS)
École Polytechnique Fédérale de Lausanne (EPFL)

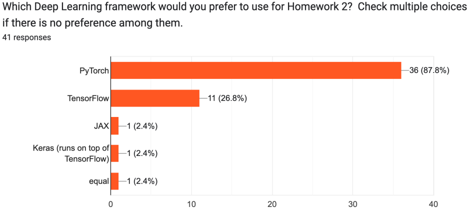
EE-556 (Fall 2023)

License Information for Mathematics of Data Slides

Survey responses

- A majority of respondents are familiar with `Python`.
  - Most are comfortable with `Jupyter` notebooks.
  - There is a clear preference for `PyTorch`.





Remark: ◦ Homeworks will be given as `Jupyter` notebooks.

Outline

- This lecture
    1. Linear algebra: Norms, matrix norms, dual norms
    2. Analysis: Continuity, Lipschitz continuity, differentiation
    3. Convexity: Convex sets, convex functions, subdifferentials, L-Lipschitz gradient functions, strong convexity
    4. Convergence rates and convergence plots
- Next lecture
    1. Gradient descent methods

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*

Observations:

- 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: $\|\mathbf{x}\|_\infty := \max_{i=1,\dots,p} |x_i|$ (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 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")

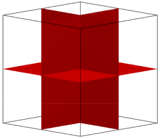
$\|\mathbf{x}\|_0 = \lim_{q\to 0}\|\mathbf{x}\|_q^q = |\{i : x_i \neq 0\}|$

Observations:
- ○ The ℓ_0-"norm" counts the non-zero components of $\mathbf{x}$. Hence, it is not a norm.
- ○ It does not satisfy the property (c) $\Rightarrow$ it is also neither a **quasi**- nor a **semi-norm**.

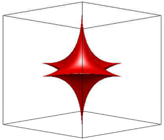
Vector norms contd.

Norm balls

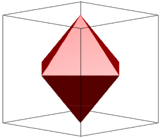
Radius r ball in ℓ_q-norm: $\mathcal{B}_q(r) = \{\mathbf{x} \in \mathbb{R}^p : \|\mathbf{x}\|_q \leq r\}$



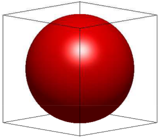
$\|\mathbf{x}\|_0 \leq 2$



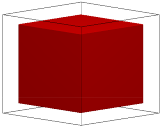
$\ell_{0.5}$-quasi norm ball



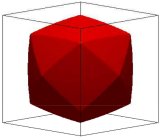
ℓ_1-norm ball



ℓ_2-norm ball



ℓ_∞-norm ball



TV-semi norm ball

Table: Some norm balls in $\mathbb{R}^3$

Vector norms contd.

Definition (Dual norm)

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

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

Observations:

- The **dual** of the *dual norm* is the original (primal) norm, i.e., $\|\mathbf{x}\|^{**} = \|\mathbf{x}\|$.
- The **dual** of $\|\cdot\|_q$ is $\|\cdot\|_p$ where p is such that $\frac{1}{q} + \frac{1}{p} = 1$.
- Hölder's inequality: $|\mathbf{x}^T\mathbf{y}| \leq \|\mathbf{x}\|_q \|\mathbf{y}\|_p$, where $p \in [1, +\infty)$ and $\frac{1}{q} + \frac{1}{p} = 1$.
- Cauchy-Schwarz is a special case of Hölder's inequality ($q = p = 2$).

Example

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*): $\sup\{\mathbf{z}^T\mathbf{x} \mid \|\mathbf{x}\|_\infty \leq 1\} = \|\mathbf{z}\|_1$.

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

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

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

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

Matrix norms contd.

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

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

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

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

Matrix norms contd.

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

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

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 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 $\mathbf{x}^T \mathbf{A} \mathbf{x} > 0$.

Observations:

- $\mathbf{A} \succeq 0$ 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 $\mathbf{A}^{-1} \succ 0$

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 $f \in \mathcal{C}(\mathcal{Q})$.

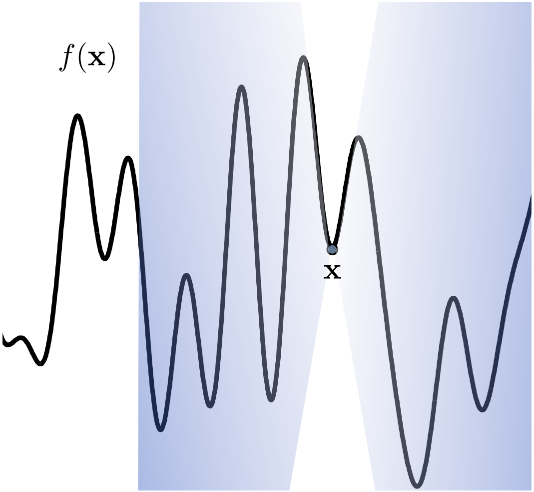
Definition (Lipschitz continuity)

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 the following holds

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

Observation: ○ “Small” changes in the input result into “small” changes in the function values.

Continuity in functions



Differentiability in functions

Definition (Differentiability)

Let $\mathcal{Q} \subseteq \mathbb{R}^p$. A function $f : \mathcal{Q} \to \mathbb{R}$ is said to be k-times continuously differentiable on $\mathcal{Q}$ if all its partial derivatives up to k-th order exist and are continuous over $\mathcal{Q}$. Notation: $f \in \mathcal{C}^k(\mathcal{Q})$.

◦ A key quantity is the gradient of the function $f : \mathcal{Q} \to \mathbb{R}$, which we denote as ∇f ($\mathbf{e}_i$ is the i-th unit vector):

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

◦ For $k = 2$, we dub $\nabla^2 f$ as the **Hessian** of f, i.e., $\left[\nabla^2 f\right]_{i,j} := \frac{\partial^2 f}{\partial x_i \partial x_j}$.

Gradients as linear approximations

A "Taylor" way of thinking about gradients:

Let $\mathcal{Q} \subseteq \mathbb{R}^p$. If $f \in \mathcal{C}^1(\mathcal{Q})$, then $\mathbf{u} \mapsto \langle \nabla f(\mathbf{x}), \mathbf{u} \rangle$ is the *unique* linear function from $\mathcal{Q}$ to $\mathbb{R}$ such that

$$\lim_{\mathbf{u} \rightrightarrows 0} \frac{|f(\mathbf{x}+\mathbf{u}) - f(\mathbf{x}) - \langle \nabla f(\mathbf{x}), \mathbf{u} \rangle|}{\|\mathbf{u}\|} \to 0$$

Example

The gradient of $f : \mathbf{x} \mapsto \|\mathbf{x}\|_2^2$ is

$$\nabla f(\mathbf{x}) = 2\mathbf{x}$$

Proof :

- To apply the Taylor way of thinking, we consider the following quantity:

$$\begin{aligned} f(\mathbf{x}+\mathbf{u}) - f(\mathbf{x}) = \|\mathbf{x}+\mathbf{u}\|_2^2 - \|\mathbf{x}\|_2^2 &= \|\mathbf{x}\|_2^2 + 2\langle \mathbf{x}, \mathbf{u} \rangle + \|\mathbf{u}\|_2^2 - \|\mathbf{x}\|_2^2 \\ &= 2\langle \mathbf{x}, \mathbf{u} \rangle + \|\mathbf{u}\|_2^2 \\ &= \langle 2\mathbf{x}, \mathbf{u} \rangle + o(\|\mathbf{u}\|_2). \end{aligned}$$

- Since the linear map is unique, we get that the gradient is $\nabla f(\mathbf{x}) = 2\mathbf{x}$.

To be or not to be differentiable

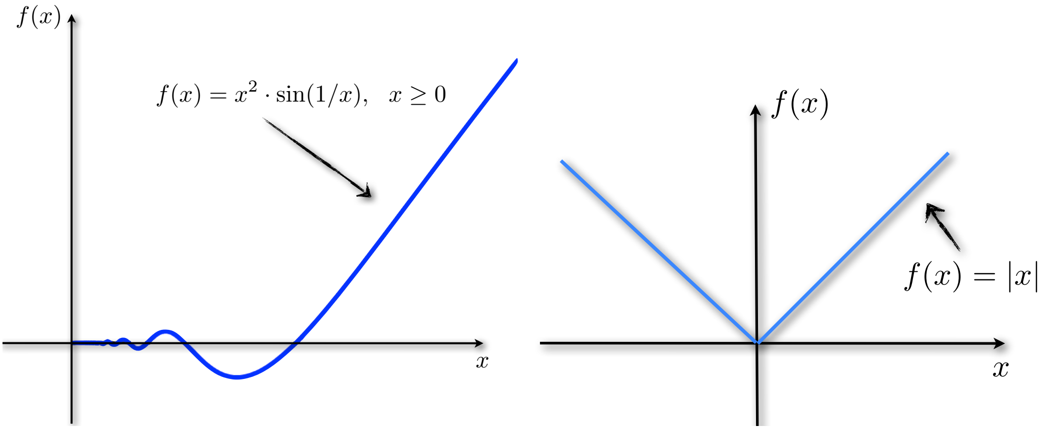


Figure: (**Left panel**) ∞-times continuously differentiable function in $\mathbb{R}$. (**Right panel**) Non-differentiable $f(x) = |x|$ in $\mathbb{R}$.

Gradients of vector valued functions

Jacobian

When $f : \mathbb{R}^n \rightrightarrows \mathbb{R}^d$ is a vector valued function, the following $d \times n$ matrix $\mathbf{J}$ of partial derivatives

$$\left[\mathbf{J}_f(\mathbf{x})\right]_{i,j} := \frac{\partial f_i}{\partial x_j}(\mathbf{x})$$

is called the Jacobian of f at $\mathbf{x}$.

Observations:

- The Jacobian is the transpose of the gradient, when f is real valued.
- Thinking in terms of Jacobians is really helpful when we need to use the chain rule.

Chain Rule via Jacobians

Let $\circ$ denote the functional composition: $g \circ f := g(f(\mathbf{x}))$. If $g \circ f$ is differentiable at $\mathbf{x}$, then the following holds

$$\mathbf{J}_{g \circ f}(\mathbf{x}) = \mathbf{J}_g(f(\mathbf{x}))\mathbf{J}_f(\mathbf{x}).$$

Hence, the chain rule, which is helpful in differentiating function compositions, can be related to a simple product of Jacobian matrices.

Example: Quadratic loss

Example

The gradient of the function $h : \mathbf{x} \mapsto \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2$ is given by the following expression:

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

Proof:

- We apply the chain rule:
  - The Jacobian of the affine function $f : \mathbf{x} \mapsto \mathbf{A}\mathbf{x} - \mathbf{b}$ is $\mathbf{J}_f(\mathbf{x}) = \mathbf{A}$.
  - The gradient of $g : \mathbf{x} \mapsto \|\mathbf{x}\|_2^2$ is $\nabla g(\mathbf{x}) = 2\mathbf{x} \Rightarrow \mathbf{J}_g(\mathbf{x}) = 2\mathbf{x}^T$.
  - Using the chain rule on the composition $h = g \circ f$:

$$\begin{aligned}\mathbf{J}_{g\circ f}(\mathbf{x}) &= \mathbf{J}_g(f(\mathbf{x}))\mathbf{J}_f(\mathbf{x})\\ &= \mathbf{J}_g(\mathbf{A}\mathbf{x} - \mathbf{b})\mathbf{J}_f(\mathbf{x})\\ &= 2(\mathbf{A}\mathbf{x} - \mathbf{b})^T\mathbf{A}.\end{aligned}$$

- Since h is real valued, the Jacobian is a row vector, we obtain the gradient by transposing.

Example: Logistic loss

Example

The gradient of the logistic loss $f(\mathbf{x}) = \log(1 + \exp(-b(\mathbf{a}^T\mathbf{x})))$ is given by the following expression:

$$\nabla f(\mathbf{x}) = -b\frac{\exp(-b(\mathbf{a}^T\mathbf{x}))}{1 + \exp(-b(\mathbf{a}^T\mathbf{x}))}\mathbf{a}.$$

Proof:

- f is a composition of the following functions:
  - $h(\mathbf{x}) = \mathbf{a}^T\mathbf{x}$, whose Jacobian is $\mathbf{J}_h(\mathbf{x}) = \mathbf{a}^T$
  - $g(u) = \log(1 + \exp(-bu))$, whose "$1 \times 1$ Jacobian" is $\mathbf{J}_g(u) = -b\frac{\exp(-bu)}{1+\exp(-bu)}$
  - By the chain rule:

$$\begin{aligned} \mathbf{J}_f(\mathbf{x}) &= \mathbf{J}_g(h(\mathbf{x})) \cdot \mathbf{J}_h(\mathbf{x}) \\ &= -b\frac{\exp(-b(\mathbf{a}^T\mathbf{x}))}{1 + \exp(-b(\mathbf{a}^T\mathbf{x}))}\mathbf{a}^T \end{aligned}$$

- The gradient is simply the transpose of $\mathbf{J}_f(\mathbf{x})$.

Use Jacobians !

With Jacobians, differentiating function compositions is a direct mechanical process.

A more complicated example here and another one at the advanced material!

Example

The gradient of $f : \mathbf{x} \mapsto \boldsymbol{w}_2^T \sigma(\mathbf{W}_1\mathbf{x} + \boldsymbol{\mu})$ is given by the following expression:

$$\nabla f(\mathbf{x}) = \mathbf{J}_f(\mathbf{x})^T = \mathbf{W}_1^T(\sigma'(\mathbf{W}_1\mathbf{x} + \boldsymbol{\mu}) \odot \boldsymbol{w}_2),$$

where σ is a non-linear function that applies to each coordinate, and $\odot$ denotes the component wise product.

Proof:

- We use the fact that f is a composition of the following functions:
  - $h(\mathbf{x}) = \mathbf{W}_1\mathbf{x} + \boldsymbol{\mu}$, whose Jacobian is $\mathbf{J}_h(\mathbf{x}) = \mathbf{W}_1$.
  - $g(\mathbf{x}) = \begin{bmatrix} \sigma(\mathbf{x}_1) \\ \vdots \\ \sigma(\mathbf{x}_n) \end{bmatrix}$, whose Jacobian is $\mathbf{J}_g(\mathbf{x}) = \mathrm{diag}(\sigma'(\mathbf{x}_1), \ldots, \sigma'(\mathbf{x}_n))$.
  - $k(\mathbf{x}) = \boldsymbol{w}_2^T\mathbf{x}$ whose Jacobian is $\mathbf{J}_k(\mathbf{x}) = \boldsymbol{w}_2^T$.
  - By the chain rule, we have that

$$\begin{aligned}\mathbf{J}_f(\mathbf{x}) &= \mathbf{J}_k(g(h(\mathbf{x}))) \cdot \mathbf{J}_g(h(\mathbf{x})) \cdot \mathbf{J}_h(\mathbf{x}) \\ &= \boldsymbol{w}_2^T \cdot \mathrm{diag}(\sigma'([\mathbf{W}_1\mathbf{x} + \boldsymbol{\mu}]_1), \ldots, \sigma'([\mathbf{W}_1\mathbf{x} + \boldsymbol{\mu}]_n)) \cdot \mathbf{W}_1.\end{aligned}$$

- Simply transpose the Jacobian to get the gradient and use $\odot$ to replace the diagonal matrix.

Some reminders on sets

Definition (Closed set)

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

Definition (Open set)

A set is *open* if its complement is closed.

Definition (Closure of a set)

Let $\mathcal{Q} \subseteq \mathbb{R}^p$ be a given open set, i.e., it contains a neighborhood of all its points. Then, the closure of $\mathcal{Q}$, denoted as $\mathrm{cl}(\mathcal{Q})$, is the smallest closed set in $\mathbb{R}^p$ that includes $\mathcal{Q}$.

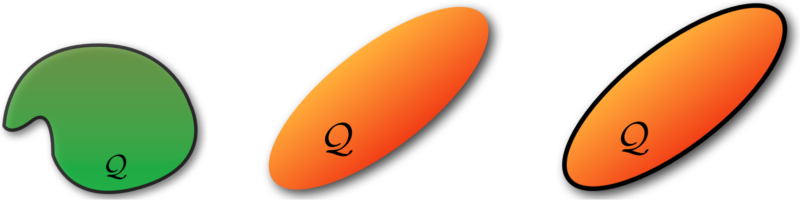


Figure: (**Left panel**) Closed set $\mathcal{Q}$. (**Middle panel**) Open set $\mathcal{Q}$ and its closure $\mathrm{cl}(\mathcal{Q})$ (**Right panel**).

Convexity of sets

Definition

- $\mathcal{Q} \subseteq \mathbb{R}^p$ is a convex set if

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

- $\mathcal{Q} \subseteq \mathbb{R}^p$ is a *strictly* convex set if

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

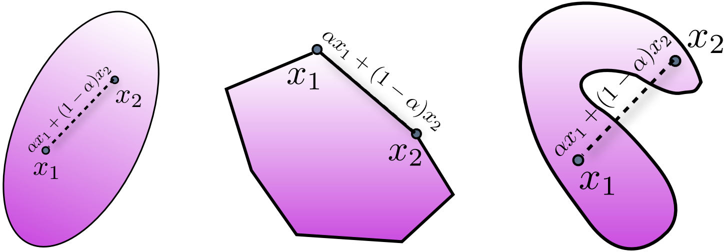


Figure: (**Left**) Strictly convex (**Middle**) Convex (**Right**) Non-convex

Convexity of functions

Definition

Let $\mathcal{Q}$ be a convex set in $\mathbb{R}^p$. A function $f \colon \mathcal{Q} \to \mathbb{R}$ is called *convex* if

$$f(\alpha\mathbf{x}_1 + (1-\alpha)\mathbf{x}_2) \leq \alpha f(\mathbf{x}_1) + (1-\alpha) f(\mathbf{x}_2), \qquad \forall \mathbf{x}_1, \mathbf{x}_2 \in \mathcal{Q}, \quad \forall \alpha \in [0,1].$$

- f is called concave, if $-f$ is convex.

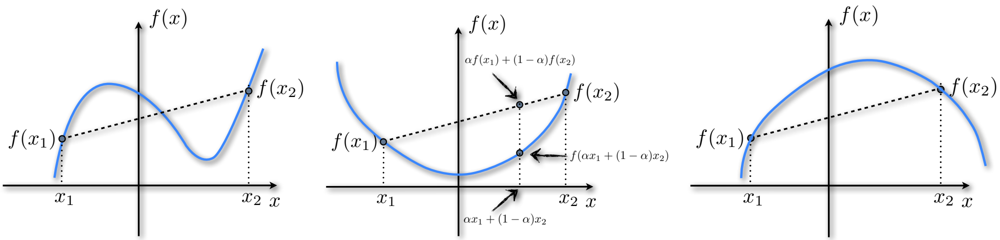


Figure: (**Left**) Non-convex (**Middle**) Convex (**Right**) Concave

Convexity of functions

Definition

Let $\mathcal{Q}$ be a convex set in $\mathbb{R}^p$. A function $f\colon \mathcal{Q} \to \mathbb{R}$ is called *convex* if

$$f(\alpha \mathbf{x}_1 + (1-\alpha)\mathbf{x}_2) \leq \alpha f(\mathbf{x}_1) + (1-\alpha) f(\mathbf{x}_2), \qquad \forall \mathbf{x}_1, \mathbf{x}_2 \in \mathcal{Q}, \quad \forall \alpha \in [0,1].$$

Question: ◦ Can we extend f from $\mathcal{Q}$ to $\mathbb{R}^p$ preserving convexity?

Convexity of functions

Definition

Let $\mathcal{Q}$ be a convex set in $\mathbb{R}^p$. A function $f\colon \mathcal{Q} \to \mathbb{R}$ is called *convex* if

$$f(\alpha\mathbf{x}_1 + (1-\alpha)\mathbf{x}_2) \leq \alpha f(\mathbf{x}_1) + (1-\alpha)f(\mathbf{x}_2), \qquad \forall \mathbf{x}_1, \mathbf{x}_2 \in \mathcal{Q}, \quad \forall \alpha \in [0,1].$$

Question: ◦ Can we extend f from $\mathcal{Q}$ to $\mathbb{R}^p$ preserving convexity?

Definition (Extended real-valued convex functions)

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

Recall, $\mathrm{dom}(f) = \mathcal{Q}$. If $\mathcal{Q} \neq \mathbb{R}^p$, extended f is never continuous, but it is l.s.c.

Convexity of functions

Definition

Let $\mathcal{Q}$ be a convex set in $\mathbb{R}^p$. A function $f\colon \mathcal{Q} \to \mathbb{R}$ is called *convex* if

$$f(\alpha\mathbf{x}_1 + (1-\alpha)\mathbf{x}_2) \leq \alpha f(\mathbf{x}_1) + (1-\alpha) f(\mathbf{x}_2), \qquad \forall \mathbf{x}_1, \mathbf{x}_2 \in \mathcal{Q}, \quad \forall \alpha \in [0,1].$$

Proposition

Every ℓ_q-norm $\|\cdot\|_q$ $(q \geq 1)$ in $\mathbb{R}^p$ is convex.

Proof :

Convexity of functions

Definition

Let $\mathcal{Q}$ be a convex set in $\mathbb{R}^p$. A function $f\colon \mathcal{Q} \to \mathbb{R}$ is called *convex* if

$$f(\alpha\mathbf{x}_1 + (1-\alpha)\mathbf{x}_2) \leq \alpha f(\mathbf{x}_1) + (1-\alpha) f(\mathbf{x}_2), \qquad \forall \mathbf{x}_1, \mathbf{x}_2 \in \mathcal{Q}, \quad \forall \alpha \in [0,1].$$

Proposition

Every ℓ_q-norm $\|\cdot\|_q$ $(q \geq 1)$ in $\mathbb{R}^p$ is convex.

Proof : ∘ Proof by intimidation.

Convexity of functions

Definition

Let $\mathcal{Q}$ be a convex set in $\mathbb{R}^p$. A function $f: \mathcal{Q} \to \mathbb{R}$ is called *convex* if

$$f(\alpha\mathbf{x}_1 + (1-\alpha)\mathbf{x}_2) \leq \alpha f(\mathbf{x}_1) + (1-\alpha)f(\mathbf{x}_2), \qquad \forall \mathbf{x}_1, \mathbf{x}_2 \in \mathcal{Q}, \quad \forall \alpha \in [0,1].$$

Proposition

Every ℓ_q-norm $\|\cdot\|_q$ $(q \geq 1)$ in $\mathbb{R}^p$ is convex.

Proof : ○ Kidding! By triangle inequality and homogeneity of the norm:

$$\|\alpha\mathbf{x}_1 + (1-\alpha)\mathbf{x}_2\|_q \leq \|\alpha\mathbf{x}_1\|_q + \|(1-\alpha)\mathbf{x}_2\|_q = \alpha\|\mathbf{x}_1\|_q + (1-\alpha)\|\mathbf{x}_2\|_q, \forall \mathbf{x}_1, \mathbf{x}_2 \in \mathcal{Q}, \forall \alpha \in [0,1].$$

Convexity of functions

Definition

Let $\mathcal{Q}$ be a convex set in $\mathbb{R}^p$. A function $f\colon \mathcal{Q} \to \mathbb{R}$ is called *convex* if

$$f(\alpha \mathbf{x}_1 + (1-\alpha)\mathbf{x}_2) \le \alpha f(\mathbf{x}_1) + (1-\alpha) f(\mathbf{x}_2), \qquad \forall \mathbf{x}_1, \mathbf{x}_2 \in \mathcal{Q}, \quad \forall \alpha \in [0,1].$$

Example

Function	Example	Attributes
ℓ_q vector norms, $q \ge 1$	$\|\mathbf{x}\|_2, \|\mathbf{x}\|_1, \|\mathbf{x}\|_\infty$	convex
ℓ_q matrix norms, $q \ge 1$	$\|\mathbf{X}\|_* = \sum_{i=1}^{\text{rank}(\mathbf{X})} \sigma_i$	convex
Square root function	$\sqrt{x}$	concave
Max of convex functions	$\max_i f_i(x)$, f_i convex	convex
Min of concave functions	$\min_i f_i(x)$, f_i concave	concave
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 $\mathbf{X} = \mathbf{X}^T$

Revisiting: Alternative definitions of function convexity II [2]

Recall, the epigraph of $f\colon \mathcal{Q} \to \mathbb{R} \cup \{+\infty\}$ is

$$\operatorname{epi}(f) = \{(\mathbf{x}, u) \in \mathcal{Q} \times \mathbb{R} \colon f(\mathbf{x}) \leq u\}.$$

Definition

A function $f : \mathcal{Q} \to \mathbb{R} \cup \{+\infty\}$ is convex if its epigraph is a convex set.

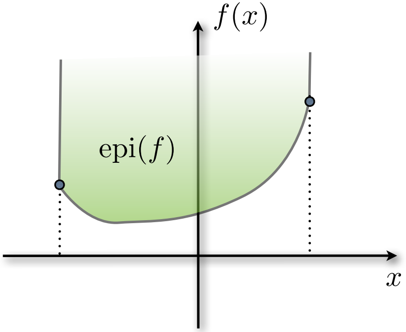
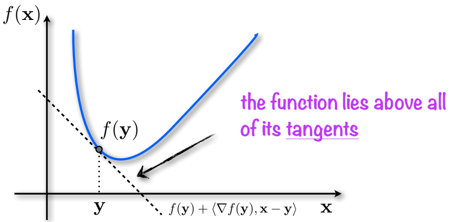


Figure: Epigraph — the region in green above graph f.

Revisiting: Alternative definition of function convexity III [2]



Definition

Let $\mathcal{Q}$ is a convex set in $\mathbb{R}^p$. A function $f \in \mathcal{C}^1(\mathcal{Q})$ is called convex on $\mathcal{Q}$ 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 $\mathcal{Q}$ 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 definition of function convexity IV [2]

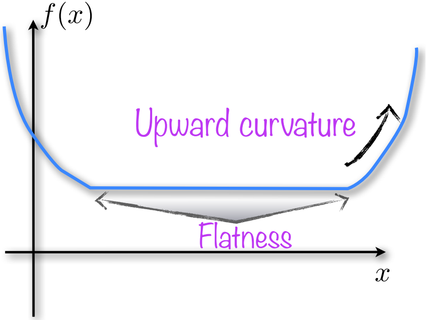
Definition

Let $\mathcal{Q}$ is a convex set in $\mathbb{R}^p$. A function $f \in \mathcal{C}^2(\mathcal{Q})$ is called convex on $\mathcal{Q}$ if for any $\mathbf{x} \in \mathcal{Q}$:

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

Remarks:

- Geometrical interpretation: the graph of f has zero or positive (upward) curvature.
- However, this does not exclude flatness of f.



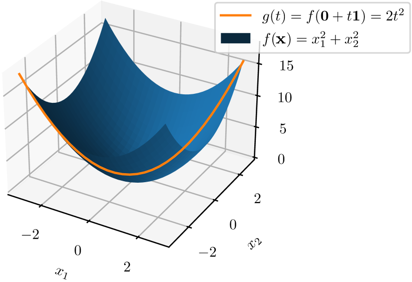
Revisiting: Alternative definition of function convexity V [2]

Definition

Let $\mathcal{Q}$ is a convex set in $\mathbb{R}^p$. A function $f \in \mathcal{C}^2(\mathcal{Q})$ is called convex on $\mathcal{Q}$ if for any $\mathbf{x} \in \mathcal{Q}$, $\mathbf{v} \in \mathbb{R}^p$, the function $g(t) = f(\mathbf{x} + t\mathbf{v})$ is convex on its domain $\{t | \mathbf{x} + t\mathbf{v} \in \mathcal{Q}\}$.

Remarks:

- This approach allows us to check the convexity long 1-dimensional lines.
- This concept generalizes to self-concordant functions (advanced material).



Strict convexity

Definition

A function $f\colon \mathcal{Q} \to \mathbb{R} \cup \{+\infty\}$ is called *strictly convex* on $\mathcal{Q}$ if

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

Theorem

If $\mathcal{Q} \subset \mathbb{R}^p$ is a convex set and $f\colon \mathbb{R}^p \to (-\infty, +\infty]$ is a proper and strictly convex function, then there exist at most one minimizer of f over $\mathcal{Q}$.

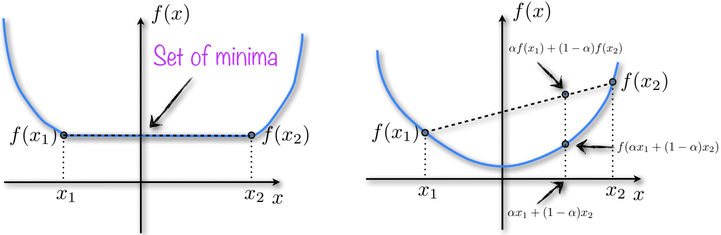


Figure: **(Left panel)** Convex function. **(Right panel)** Strictly convex function.

Subdifferentials and (sub)gradients in convex functions

Definition

Let $f : \mathcal{Q} \to \mathbb{R} \cup \{+\infty\}$ be a convex function. The subdifferential of f at a point $\mathbf{x} \in \mathcal{Q}$ is defined by the set:

$$\partial f(\mathbf{x}) = \{\mathbf{v} \in \mathbb{R}^p \ : \ f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \mathbf{v}, \ \mathbf{y} - \mathbf{x} \rangle \text{ for all } \mathbf{y} \in \mathcal{Q}\} \,.$$

Each element $\mathbf{v}$ of $\partial f(\mathbf{x})$ is called *subgradient* of f at $\mathbf{x}$.

Definition

Let $f : \mathcal{Q} \to \mathbb{R} \cup \{+\infty\}$ be a differentiable convex function. Then, the subdifferential of f at a point $\mathbf{x} \in \mathcal{Q}$ contains only the gradient, i.e., $\partial f(\mathbf{x}) = \{\nabla f(\mathbf{x})\}$.

Remark: ◦ Subdifferential generalizes ∇ to *nondifferentiable functions*

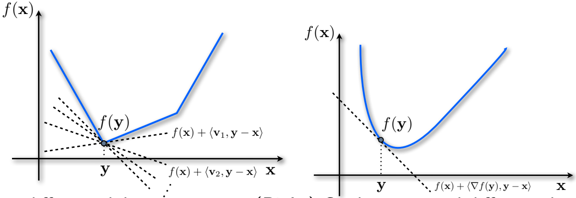


Figure: (**Left**) Non-differentiability at point $\mathbf{y}$. (**Right**) Gradient as a subdifferential with a singleton entry.

Generalized subdifferentials for nonconvex functions

Definition

Let $f : \mathcal{Q} \to \mathbb{R} \cup \{+\infty\}$ be a locally Lipschitz function. The Clarke subdifferential of f at a point $\mathbf{x} \in \mathcal{Q}$ is defined by the set:

$$\partial_C f(\mathbf{x}) = \mathrm{conv}\left(\left\{\mathbf{v} \in \mathbb{R}^p : \begin{array}{l} \exists \mathbf{x}^k \to \mathbf{x}, \nabla f\left(\mathbf{x}^k\right) \text{ exists,} \\ \nabla f\left(\mathbf{x}^k\right) \to \mathbf{v} \end{array} \right\}\right).$$

Remarks:

- For convex functions, the Clarke subdifferential reduces to subdifferential.
- If $\mathbf{x}^\star$ is a local minimum of f, then $\mathbf{0} \in \partial_C f(\mathbf{x}^\star)$.

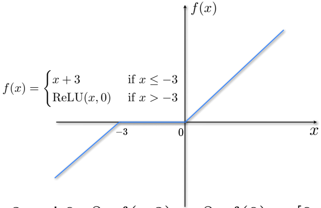


Figure: The Clarke subdifferential at -3 and 0: $\partial_C f(-3) = \partial_C f(0) = [0, 1]$. Non-subdifferentiability at -3 and 0.

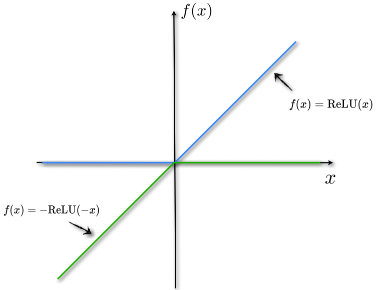
Heads up: Be careful with automatic differentiation!

Example (Simple)

The gradient of the function $f : x \mapsto \text{ReLU}(x) - \text{ReLU}(-x) = x$ at 0 is given by $g(0) = 1$.

Remark:

- Subdifferentials are tricky business!
- Automatic differentiation can be wrong [3]!
- We will revisit when we discuss the Moreau-Rockafellar's decomposition theorem.




```
import torch
x = torch.tensor([0.], requires_grad=True)
f = torch.nn.ReLU()(x) - torch.nn.ReLU()(-x)
f.backward()
print(x.grad)
```

```
tensor([0.])
```

Figure: (**Left panel**) ReLU function. (**Right panel**) Calculation of $g(0)$ in PyTorch.

L-Lipschitz gradient class of functions

Definition (L-Lipschitz gradient convex functions)

Let $f : \mathcal{Q} \to \mathbb{R}$ be differentiable and convex, i.e., $f \in \mathcal{F}^1(\mathcal{Q})$. Then, f has a Lipschitz gradient if there exists $L > 0$ (the Lipschitz constant) such that $\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_2 \leq L\|\mathbf{x} - \mathbf{y}\|_2, \ \ \forall \mathbf{x}, \mathbf{y} \in \mathcal{Q}$.

Proposition (L-Lipschitz gradient convex functions)

$f \in \mathcal{F}^1(\mathcal{Q})$ has L-Lipschitz gradient if and only if the following function is convex:

$$h(\mathbf{x}) = \frac{L}{2}\|\mathbf{x}\|_2^2 - f(\mathbf{x}) \quad \forall \mathbf{x} \in \mathcal{Q}.$$

Definition (Class of 2-nd order Lipschitz functions)

The class of twice continuously differentiable functions f on $\mathcal{Q}$ with Lipschitz continuous Hessian is denoted as $\mathcal{F}_L^{2,2}(\mathcal{Q})$ (with $2 \to 2$ denoting the spectral norm)

$$\|\nabla^2 f(\mathbf{x}) - \nabla^2 f(\mathbf{y})\|_{2\to 2} \leq L\|\mathbf{x} - \mathbf{y}\|_2, \ \ \forall \mathbf{x}, \mathbf{y} \in Q,$$

Remark: ◦ $\mathcal{F}_L^{l,m}$: functions that are l-times differentiable with m-th order Lipschitz property.

Example: Logistic regression

Problem (Logistic regression)

Given a sample vector $\mathbf{a}_i \in \mathbb{R}^p$ *and a binary class label* $b_i \in \{-1, +1\}$ $(i = 1, \ldots, n)$*, we define the conditional probability of* b_i *given* $\mathbf{a}_i$ *as:*

$$\mathbb{P}(b_i | \mathbf{a}_i, \mathbf{x}^\natural, \mu) \propto 1/(1 + e^{-b_i(\langle \mathbf{x}^\natural, \mathbf{a}_i \rangle + \mu)}),$$

where $\mathbf{x}^\natural \in \mathbb{R}^p$ *is some true weight vector,* $\mu \in \mathbb{R}$ *is called the intercept. How to estimate* $\mathbf{x}^\natural$ *given the sample vectors, the binary labels, and* μ*?*

Optimization formulation

$$\min_{\mathbf{x} \in \mathbb{R}^p} \underbrace{\frac{1}{n} \sum_{i=1}^{n} \log(1 + \exp(-b_i(\mathbf{a}_i^T \mathbf{x} + \mu)))}_{f(\mathbf{x})}$$

Structural properties

Let $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_n]^T$ (design matrix), then $f \in \mathcal{F}_L^{2,1}$, with $L = \frac{1}{4}\|\mathbf{A}^T\mathbf{A}\|$

μ-strongly convex functions

Definition

A function $f : \mathcal{Q} \to \mathbb{R} \cup \{+\infty\}$, $\mathcal{Q} \subseteq \mathbb{R}^p$ is called μ-*strongly* convex on its domain if and only if for any $\mathbf{x}, \ \mathbf{y} \in \mathcal{Q}$ and $\alpha \in [0,1]$ we have:

$$f(\alpha\mathbf{x} + (1-\alpha)\mathbf{y}) \leq \alpha f(\mathbf{x}) + (1-\alpha) f(\mathbf{y}) - \frac{\mu}{2}\alpha(1-\alpha)\|\mathbf{x}-\mathbf{y}\|_2^2.$$

The constant μ is called the convexity parameter of function f.

- The class of k-differentiable μ-strongly functions is denoted as $\mathcal{F}_\mu^k(\mathcal{Q})$.
- Strong convexity $\Rightarrow$ strict convexity, BUT strict convexity $\not\Rightarrow$ strong convexity

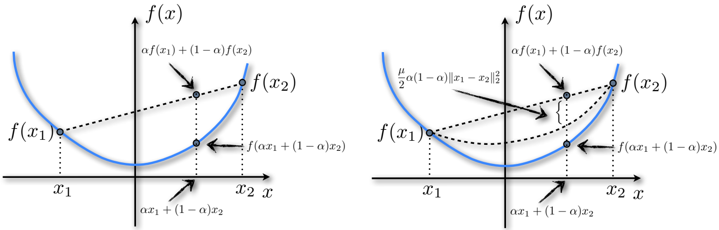


Figure: (**Left**) Convex (**Right**) Strongly convex

Alternative: μ-strongly convex functions

Definition

A convex function $f : \mathcal{Q} \to \mathbb{R}$ is said to be μ-strongly convex if

$$h(\mathbf{x}) = f(\mathbf{x}) - \frac{\mu}{2}\|\mathbf{x}\|_2^2$$

is convex, where μ is called the strong convexity parameter.

- The class of k-differentiable μ-strongly functions is denoted as $\mathcal{F}_{\mu}^{k}(\mathcal{Q})$.
- Non-smooth functions can be μ-strongly convex: e.g., $f(\mathbf{x}) = \|\mathbf{x}\|_1 + \frac{\mu}{2}\|\mathbf{x}\|_2^2$.

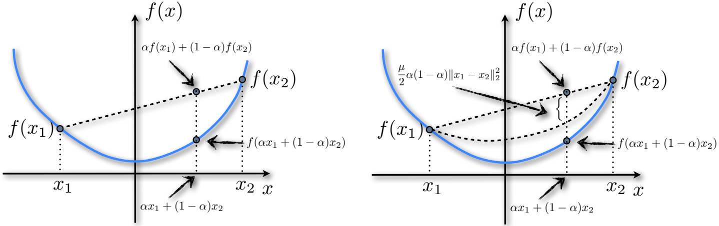


Figure: (**Left**) Convex (**Right**) Strongly convex

Properties of μ-strongly convex functions

Lemma

Let $f : \mathcal{Q} \to \mathbb{R}, \mathcal{Q} \subseteq \mathbb{R}^p$ be a twice differentiable convex function, i.e., $f \in \mathcal{F}^2(\mathcal{Q})$. Then, f is μ-strongly convex function if and only if

$$\nabla^2 f(\mathbf{x}) \succeq \mu \mathbf{I}, \quad \forall \mathbf{x} \in \mathbb{R}^p.$$

Properties of μ-strongly convex functions

Lemma

Let $f : \mathcal{Q} \to \mathbb{R}, \mathcal{Q} \subseteq \mathbb{R}^p$ be a twice differentiable convex function, i.e., $f \in \mathcal{F}^2(\mathcal{Q})$. Then, f is μ-strongly convex function if and only if

$$\nabla^2 f(\mathbf{x}) \succeq \mu \mathbf{I}, \quad \forall \mathbf{x} \in \mathbb{R}^p.$$

Example (Toy example)

Consider the quadratic function $f(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|_2^2$. Then, f is a μ-strongly convex since $\nabla^2 f(\mathbf{x}) = \mathbf{I} \implies \mu = 1$.

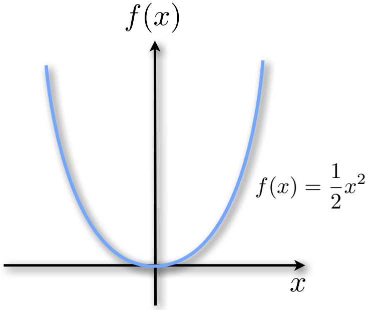


Figure: Toy example for μ-strongly convex functions.

Properties of μ-strongly convex functions

Lemma

Let $f : \mathcal{Q} \to \mathbb{R}, \mathcal{Q} \subseteq \mathbb{R}^p$ be a twice differentiable convex function, i.e., $f \in \mathcal{F}^2(\mathcal{Q})$. Then, f is μ-strongly convex function if and only if

$$\nabla^2 f(\mathbf{x}) \succeq \mu \mathbf{I}, \quad \forall \mathbf{x} \in \mathbb{R}^p.$$

Example (Overdetermined least squares)

Consider an *overdetermined* linear system of equations $\mathbf{b} = \mathbf{A}\mathbf{x}^\natural + \mathbf{w}$ where $\mathbf{A} \in \mathbb{R}^{n \times p}$ is a full column-rank matrix and $\mathbf{x}^\natural$ is unknown. Assume that $\mathbf{A}^T\mathbf{A} \succeq \rho\mathbf{I}, \rho > 0$ and let $f(\mathbf{x}) = \frac{1}{2}\|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2^2$. Then, f is a μ-strongly convex function, i.e., $f \in \mathcal{F}_\mu^2(\mathbb{R}^p)$ since:

$$\nabla^2 f(\mathbf{x}) = \mathbf{A}^T\mathbf{A} \quad \text{where} \quad \mathbf{A}^T\mathbf{A} \succeq \rho\mathbf{I} =: \mu\mathbf{I}.$$

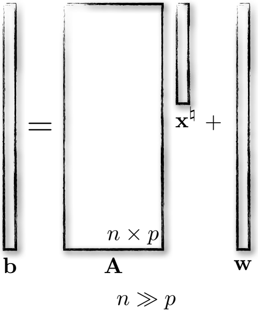


Figure: Overdetermined system of linear equations.

Properties of μ-strongly convex functions

Lemma

Let $f : \mathcal{Q} \to \mathbb{R}, \mathcal{Q} \subseteq \mathbb{R}^p$ be a twice differentiable convex function, i.e., $f \in \mathcal{F}^2(\mathcal{Q})$. Then, f is μ-strongly convex function if and only if

$$\nabla^2 f(\mathbf{x}) \succeq \mu \mathbf{I}, \quad \forall \mathbf{x} \in \mathbb{R}^p.$$

Example (Trivial)

Any linear function $f(\mathbf{x}) = \mathbf{c}^T \mathbf{x} + \beta \in \mathcal{F}^1_\mu(\mathbb{R}^p)$ for $\mu = 0$ since

$$\nabla f(\mathbf{x}) = \mathbf{c} \quad \text{and} \quad \nabla^2 f(\mathbf{x}) = \mathbf{0}.$$

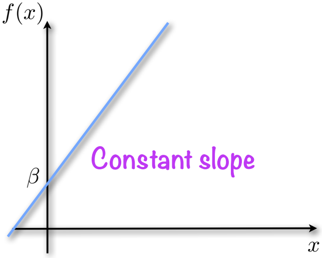


Figure: Counterexample for μ-strongly convex functions.

Properties of μ-strongly convex functions

Lemma

Let $f : \mathcal{Q} \to \mathbb{R}, \mathcal{Q} \subseteq \mathbb{R}^p$ be a twice differentiable convex function, i.e., $f \in \mathcal{F}^2(\mathcal{Q})$. Then, f is μ-strongly convex function if and only if

$$\nabla^2 f(\mathbf{x}) \succeq \mu \mathbf{I}, \quad \forall \mathbf{x} \in \mathbb{R}^p.$$

Lemma

A continuously differentiable function f belongs to $\mathcal{F}^1_\mu(\mathcal{Q})$ if there exists a constant $\mu > 0$ such that for any $\mathbf{x}, \mathbf{y} \in \mathcal{Q}$, we have:

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{\mu}{2} \|\mathbf{y} - \mathbf{x}\|_2^2$$

Lemma

Let f be continuously differentiable. The following condition, holding for all $\mathbf{x}, \mathbf{y} \in \mathcal{Q} \subseteq \mathbb{R}^p$, is equivalent to inclusion that f is μ-strongly convex function:

$$\langle \nabla f(\mathbf{x}) - \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \geq \mu \|\mathbf{x} - \mathbf{y}\|_2^2.$$

L-smooth, μ-strongly convex functions

Definition

Let $f : \mathcal{Q} \to \mathbb{R}, \mathcal{Q} \subseteq \mathbb{R}^p$ be a continuously differentiable function. Then, f is both μ-strongly and L-smooth convex function if for any $\mathbf{x}, \mathbf{y} \in \mathcal{Q}$, we have:

$$\frac{\mu}{2}\|\mathbf{y} - \mathbf{x}\|_2^2 \leq f(\mathbf{y}) - f(\mathbf{x}) - \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle \leq \frac{L}{2}\|\mathbf{y} - \mathbf{x}\|_2^2$$

for constants $0 < \mu \leq L$. We denote that $f \in \mathcal{F}_{\mu,L}^{1,1}(\mathcal{Q})$. If f is twice differentiable, an equivalent condition is

$$\mu \mathbf{I} \preceq \nabla^2 f(\mathbf{x}) \preceq L\mathbf{I}.$$

L-smooth, μ-strongly convex functions

Definition

Let $f : \mathcal{Q} \to \mathbb{R}, \mathcal{Q} \subseteq \mathbb{R}^p$ be a continuously differentiable function. Then, f is both μ-strongly and L-smooth convex function if for any $\mathbf{x}, \mathbf{y} \in \mathcal{Q}$, we have:

$$\frac{\mu}{2}\|\mathbf{y}-\mathbf{x}\|_2^2 \leq f(\mathbf{y}) - f(\mathbf{x}) - \langle \nabla f(\mathbf{x}), \mathbf{y}-\mathbf{x} \rangle \leq \frac{L}{2}\|\mathbf{y}-\mathbf{x}\|_2^2$$

for constants $0 < \mu \leq L$. We denote that $f \in \mathcal{F}^{1,1}_{\mu,L}(\mathcal{Q})$. If f is twice differentiable, an equivalent condition is

$$\mu\mathbf{I} \preceq \nabla^2 f(\mathbf{x}) \preceq L\mathbf{I}.$$

Example

Consider an linear system of equations $\mathbf{b} = \mathbf{A}\mathbf{x}^\natural$ where $\mu\mathbf{I} \preceq \mathbf{A}^T\mathbf{A} \preceq L\mathbf{I}$. Let $f(\mathbf{x}) = \frac{1}{2}\|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2^2$. Then, f is both μ-strongly convex and L-smooth function, i.e., $f \in \mathcal{F}^{2,1}_{\mu,L}(\mathbb{R}^p)$ since:

$$\nabla^2 f(\mathbf{x}) = \mathbf{A}^T\mathbf{A} \quad \text{where} \quad \mu\mathbf{I} \preceq \mathbf{A}^T\mathbf{A} \preceq L\mathbf{I}.$$

L-smooth, μ-strongly convex functions

Definition

Let $f : \mathcal{Q} \to \mathbb{R}, \mathcal{Q} \subseteq \mathbb{R}^p$ be a continuously differentiable function. Then, f is both μ-strongly and L-smooth convex function if for any $\mathbf{x}, \mathbf{y} \in \mathcal{Q}$, we have:

$$\frac{\mu}{2}\|\mathbf{y}-\mathbf{x}\|_2^2 \leq f(\mathbf{y}) - f(\mathbf{x}) - \langle \nabla f(\mathbf{x}), \mathbf{y}-\mathbf{x}\rangle \leq \frac{L}{2}\|\mathbf{y}-\mathbf{x}\|_2^2$$

for constants $0 < \mu \leq L$. We denote that $f \in \mathcal{F}^{1,1}_{\mu,L}(\mathcal{Q})$. If f is twice differentiable, an equivalent condition is

$$\mu\mathbf{I} \preceq \nabla^2 f(\mathbf{x}) \preceq L\mathbf{I}.$$

Observations:

- ◦ Both μ and L show up in convergence rate characterization of algorithms
- ◦ **Unfortunately, μ, L are usually not known a priori...**
- ◦ When they are known, they can help significantly (even in stopping algorithms)

Convergence rates

Definition (Convergence of a sequence)

The sequence $\mathbf{u}^1, \mathbf{u}^2, ..., \mathbf{u}^k, ...$ converges to $\mathbf{u}^\star$ (denoted $\lim_{k\to\infty} \mathbf{u}^k = \mathbf{u}^\star$), if

$$\forall\ \varepsilon > 0, \exists\ K \in \mathbb{N} : k \geq K \Rightarrow \|\mathbf{u}^k - \mathbf{u}^\star\| \leq \varepsilon$$

Convergence rates: the *"speed"* at which a sequence converges

- **sublinear:** if there exists $c > 0$ such that

$$\|\mathbf{u}^k - \mathbf{u}^\star\| = O(k^{-c})$$

- **linear:** if there exists $\alpha \in (0, 1)$ such that

$$\|\mathbf{u}^k - \mathbf{u}^\star\| = O(\alpha^k)$$

- **Q-linear:** if there exists a constant $r \in (0, 1)$ such that

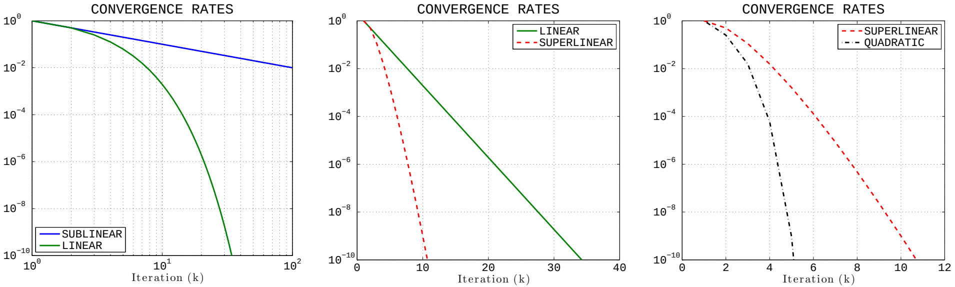
$$\lim_{k\to\infty} \frac{\|\mathbf{u}^{k+1} - \mathbf{u}^\star\|}{\|\mathbf{u}^k - \mathbf{u}^\star\|} = r$$

- **superlinear:** if $r = 0$, we say that the sequence converges *superlinearly*.
- **quadratic:** if there exists a constant $\mu > 0$ such that $\lim_{k\to\infty} \frac{\|\mathbf{u}^{k+1} - \mathbf{u}^\star\|}{\|\mathbf{u}^k - \mathbf{u}^\star\|^2} = \mu$

Example: Convergence rates

Examples of sequences that all converge to $u^\star = 0$:

- Sublinear: $u^k = 1/k$
- Linear: $u^k = 0.5^k$
- Superlinear: $u^k = k^{-k}$
- Quadratic: $u^k = 0.5^{2^k}$



Wrap up!

- Please take a look at the handout for rate examples!
- See advanced material for material beyond convexity!
  - Star-convexity
  - Invexity
- Lecture on Monday!

*Jacobian of the self-attention module [5]

Example

We consider the Jacobian of $f : \mathbf{X} \mapsto \sigma_s\left(\mathbf{X}\mathbf{W}_Q^\top \mathbf{W}_K \mathbf{X}^\top\right)\mathbf{X}\mathbf{W}_V^\top$, where σ_s is row-wise softmax, $\mathbf{X} \in \mathbb{R}^{d_s \times d}$, $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V \in \mathbb{R}^{d_m \times d}$, $f(\mathbf{X}) \in \mathbb{R}^{d_s \times d_m}$.

- Define $\boldsymbol{\beta}_i := \sigma_s\left(\mathbf{X}^{(i,:)}\mathbf{W}_Q^\top \mathbf{W}_K \mathbf{X}^\top\right)^\top \in \mathbb{R}^{d_s}$ We can reformulate the definition above as:

$$f(\mathbf{X}) = \begin{bmatrix} \boldsymbol{\beta}_1^\top \\ \vdots \\ \boldsymbol{\beta}_{d_s}^\top \end{bmatrix} \mathbf{X}\mathbf{W}_V^\top.$$

- By the product rule:

$$\frac{\partial f(\mathbf{X})}{\partial X^{(p,k)}} = \begin{bmatrix} \frac{\partial \boldsymbol{\beta}_1^\top}{\partial X^{(p,k)}} \\ \vdots \\ \frac{\partial \boldsymbol{\beta}_{d_s}^\top}{\partial X^{(p,k)}} \end{bmatrix} \mathbf{X}\mathbf{W}_V^\top + \begin{bmatrix} \boldsymbol{\beta}_1^\top \\ \vdots \\ \boldsymbol{\beta}_{d_s}^\top \end{bmatrix} \frac{\partial (\mathbf{X}\mathbf{W}_V^\top)}{\partial X^{(p,k)}}. \tag{1}$$

*Jacobian of self-attention module [5]

- Suppose $\boldsymbol{\beta} = \mathrm{Softmax}(\mathbf{u}) \in \mathbb{R}^{d_s}$, then $\frac{\partial \boldsymbol{\beta}}{\partial \mathbf{u}} = \mathrm{diag}(\boldsymbol{\beta}) - \boldsymbol{\beta}\boldsymbol{\beta}^\top$. This is because:
  - We can reformulate $\boldsymbol{\beta}$ as: $\boldsymbol{\beta} = \begin{bmatrix} \frac{\exp(u^{(1)})}{\sum_{i=1}^{d_s} \exp(u^{(i)})} \\ \vdots \\ \frac{\exp(u^{(d_s)})}{\sum_{i=1}^{d_s} \exp(u^{(i)})} \end{bmatrix}$.
  - Thus

$$
\frac{\partial \beta^{(j)}}{\partial u^{(k)}} = \frac{\partial \frac{\exp(u^{(j)})}{\sum_{i=1}^{d_s} \exp(u^{(i)})}}{\partial u^{(k)}} = \begin{cases} \frac{-\exp(u^{(j)}) - \exp(u^{(k)})}{(\sum_{i=1}^{d_s} \exp(u^{(i)}))^2} & \text{if } j \neq k \\ \frac{\exp(u^{(k)}) \sum_{i=1}^{d_s} \exp(u^{(i)}) - (\exp(u^{(k)}))^2}{\left(\sum_{i=1}^{d_s} \exp(u^{(i)})\right)^2} & \text{if } j = k \end{cases}
$$

$$
= \begin{cases} -\beta^{(j)}\beta^{(k)} & \text{if } j \neq k \\ \beta^{(k)} - \beta^{(j)}\beta^{(k)} & \text{if } j = k \end{cases}.
$$

  - Thus

$$
\frac{\partial \boldsymbol{\beta}}{\partial \mathbf{u}} = \mathrm{diag}(\boldsymbol{\beta}) - \boldsymbol{\beta}\boldsymbol{\beta}^\top. \quad (2)
$$

*Jacobian of self-attention module [5]

- Then we can calculate the term $\frac{\partial \boldsymbol{\beta}_i}{\partial X^{(p,k)}}$ for $i \in [d_s]$ in the first part of Eq. (1).

$$\begin{aligned}\frac{\partial \boldsymbol{\beta}_i}{\partial X^{(p,k)}} &= \left(\mathrm{diag}(\boldsymbol{\beta}_i) - \boldsymbol{\beta}_i \boldsymbol{\beta}_i^\top\right) \frac{\partial \left(\mathbf{X}\mathbf{W}_K^\top \mathbf{W}_Q {\mathbf{X}^{(i,:)}}^\top\right)}{\partial X^{(p,k)}} \\ &= \left(\mathrm{diag}(\boldsymbol{\beta}_i) - \boldsymbol{\beta}_i \boldsymbol{\beta}_i^\top\right) \left(\boldsymbol{e}_p \boldsymbol{e}_k^\top \mathbf{W}_K^\top \mathbf{W}_Q {\mathbf{X}^{(i,:)}}^\top + \mathbf{X}\mathbf{W}_K^\top \mathbf{W}_Q \boldsymbol{e}_k \delta_{ip}\right),\end{aligned} \tag{3}$$

where $\boldsymbol{e}_p$ is the p^{th} canonical basis vector of $\mathbb{R}^{d_s}$, $\boldsymbol{e}_k$ is the k^{th} canonical basis vector of $\mathbb{R}^{d}$.

- Next, let's consider the second term in Eq. (1):

$$\frac{\partial (\mathbf{X}\mathbf{W}_V^\top)}{\partial X^{(p,k)}} = \boldsymbol{e}_p \boldsymbol{e}_k^\top \mathbf{W}_V^\top . \tag{4}$$

- Lastly, substituting Eq. (3) and Eq. (4) into Eq. (1):

$$\frac{\partial f(\mathbf{X})}{\partial X^{(p,k)}} = \begin{bmatrix} \left(\mathrm{diag}(\boldsymbol{\beta}_1) - \boldsymbol{\beta}_1 \boldsymbol{\beta}_1^\top\right) \left(\boldsymbol{e}_p \boldsymbol{e}_k^\top \mathbf{W}_K^\top \mathbf{W}_Q {\mathbf{X}^{(1,:)}}^\top + \mathbf{X}\mathbf{W}_K^\top \mathbf{W}_Q \boldsymbol{e}_k \delta_{1p}\right) \\ \vdots \\ \left(\mathrm{diag}(\boldsymbol{\beta}_{d_s}) - \boldsymbol{\beta}_{d_s} \boldsymbol{\beta}_{d_s}^\top\right) \left(\boldsymbol{e}_p \boldsymbol{e}_k^\top \mathbf{W}_K^\top \mathbf{W}_Q {\mathbf{X}^{(d_s,:)}}^\top + \mathbf{X}\mathbf{W}_K^\top \mathbf{W}_Q \boldsymbol{e}_k \delta_{d_s p}\right) \end{bmatrix} \mathbf{X}\mathbf{W}_V^\top + \begin{bmatrix} \boldsymbol{\beta}_1^\top \\ \vdots \\ \boldsymbol{\beta}_{d_s}^\top \end{bmatrix} \boldsymbol{e}_p \boldsymbol{e}_k^\top \mathbf{W}_V^\top .$$

Convex hull

Definition (Convex hull)

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

Definition (Convex hull of points)

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

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

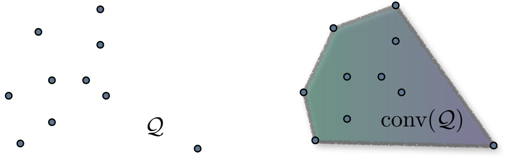


Figure: (**Left**) Discrete set of points $\mathcal{Q}$. (**Right**) Convex hull conv$(\mathcal{Q})$.

*Star convex sets

Definition

$\mathcal{Q} \subseteq \mathbb{R}^p$ is a *star-shaped* set if there exists a $\mathbf{x}_1 \in \mathcal{Q}$ such that

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

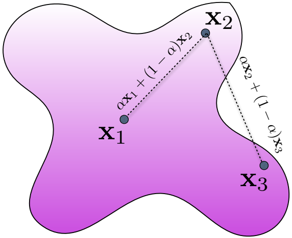


Figure: Example of a star-shaped but not convex set.

*Star convexity

Definition

A function $f\colon \mathcal{Q} \to \mathbb{R} \cup \{+\infty\}$ is called *star-convex* on $\mathcal{Q}$ if there exists a global minimum $\mathbf{x}^\star \in \mathcal{Q}$ such that

$$f(\alpha \mathbf{x}^\star + (1-\alpha)\mathbf{x}) \leq \alpha f(\mathbf{x}^\star) + (1-\alpha) f(\mathbf{x}) \quad \forall \mathbf{x} \in \mathcal{Q}, \quad \forall \alpha \in [0,1].$$

Remarks:

- Any convex function is star-convex.
- Star-convexity can be viewed as convexity between any point $\mathbf{x}$ and a global minimum $\mathbf{x}^\star$.
- Allows the negative gradient $-\nabla f(\mathbf{x})$ to the desired minimization direction.
- Consider the following objective function:

$$\min_{\mathbf{x}} f(\mathbf{x}) := \frac{1}{n} \left(\sum_{i=1}^{n} |b_i - \langle \mathbf{a}_i, \mathbf{x} \rangle|^q \right)^{1/q}.$$

  - Star-convex for any real number q when $n \leq p$.
  - Convex for $q \geq 1$.
  - ($q = 1$): the least-absolute deviation estimator. ($q = 2$): the least-squares estimator.

*Invex function

Definition

Let $\mathcal{Q}$ be an open set in $\mathbb{R}^p$. A differentiable function $f \colon \mathcal{Q} \to \mathbb{R}$ is called *invex* if there exists a function $\eta : \mathcal{Q} \times \mathcal{Q} \to \mathbb{R}^p$ such that

$$f(\mathbf{x}) \geq f(\mathbf{y}) + \langle \nabla f(\mathbf{y}),\ \eta(\mathbf{x}, \mathbf{y}) \rangle, \quad \forall \mathbf{x},\ \mathbf{y} \in \mathcal{Q}.$$

Remarks:

- Any convex function is invex function: $\eta(\mathbf{x}, \mathbf{y}) = \mathbf{x} - \mathbf{y}$.
- **Any local minima in an invex function is global minima!**

Proof :

- Suppose $\mathbf{x}^\star$ is a local minimum, then $\nabla f(\mathbf{x}^\star) = 0$. By the definition above, we have

$$f(\mathbf{x}) \geq f(\mathbf{x}^\star) + \langle 0,\ \eta(\mathbf{x}, \mathbf{y}) \rangle = f(\mathbf{x}^\star), \quad \forall \mathbf{x} \in \mathcal{Q}.$$

- $\Rightarrow \mathbf{x}^\star$ is also a global minimum.

Example (Causality via directed acyclic graph (DAG) learning [1])

For any $s > 0$, define $f^s : \{\boldsymbol{X} \in \mathbb{R}^{d \times d} \mid s > \rho(\boldsymbol{X} \circ \boldsymbol{X})\} \to \mathbb{R}$ as $f^s(\boldsymbol{X}) \stackrel{\text{def}}{=} -\log\det(s\boldsymbol{I} - \boldsymbol{X} \circ \boldsymbol{X}) + d \log s$, where $\circ$ is the Hadamard product, $\rho(\cdot)$ is the spectral radius, and $\boldsymbol{X}$ is the graph weighted adjacency matrix.

- Then, f^s is an invex function. $f^s(\boldsymbol{X}) \geq 0$ with $f^s(\boldsymbol{X}) = 0$ *if and only if* $\boldsymbol{X}$ is a DAG.

⋆Self-concordant functions [4]

Definition (Self-concordant functions in 1-dimension)

A convex function $\varphi : \mathbb{R} \to \mathbb{R}$ is self-concordant if

$$|\varphi'''(t)| \leq 2\varphi''(t)^{3/2}, \quad \forall t \in \mathbb{R}.$$

*Self-concordant functions [4]

Definition (Self-concordant functions in 1-dimension)

A convex function $\varphi : \mathbb{R} \to \mathbb{R}$ is self-concordant if

$$|\varphi'''(t)| \leq 2\varphi''(t)^{3/2}, \quad \forall t \in \mathbb{R}.$$

Affine Invariance of self-concordant functions

Let $\tilde{\varphi}(t) = \varphi(\alpha t + \beta)$ where $\alpha \neq 0$. Then, $\tilde{\varphi}$ is self-concordant iff φ is.

*Self-concordant functions [4]

Definition (Self-concordant functions in 1-dimension)

A convex function $\varphi : \mathbb{R} \to \mathbb{R}$ is self-concordant if

$$|\varphi'''(t)| \leq 2\varphi''(t)^{3/2}, \quad \forall t \in \mathbb{R}.$$

Affine Invariance of self-concordant functions

Let $\tilde{\varphi}(t) = \varphi(\alpha t + \beta)$ where $\alpha \neq 0$. Then, $\tilde{\varphi}$ is self-concordant iff φ is.

Important remarks of self-concordance

1. Generalize to higher dimension: A convex function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be (standard) self-concordant if $|\varphi'''(t)| \leq 2\varphi''(t)^{3/2}$, where $\varphi(t) := f(\mathbf{x} + t\mathbf{v})$ for all $t \in \mathbb{R}$, $\mathbf{x} \in \operatorname{dom} f$ and $\mathbf{v} \in \mathbb{R}^n$ such that $\mathbf{x} + t\mathbf{v} \in \operatorname{dom} f$.
2. Affine invariance still holds in high dimension.
3. Self-concordant functions are efficiently minimized by the Newton method and its variants.

References I

[1] Kevin Bello, Bryon Aragam, and Pradeep Ravikumar.
Dagma: Learning dags via m-matrices and a log-determinant acyclicity characterization.
In *Advances in Neural Information Processing Systems*, 2022.
(Cited on page 62.)

[2] S. Boyd, S.P. Boyd, L. Vandenberghe, and Cambridge University Press.
Convex Optimization.
Berichte uber verteilte messysteme. Cambridge University Press, 2004.
(Cited on pages 33, 34, 35, and 36.)

[3] Sham M Kakade and Jason D Lee.
Provably correct automatic sub-differentiation for qualified programs.
In *Advances in Neural Information Processing Systems*, volume 31, 2018.
(Cited on page 40.)

[4] Yurii Nesterov and Arkadii Nemirovskii.
Interior-point polynomial algorithms in convex programming.
SIAM, 1994.
(Cited on pages 63, 64, and 65.)

References II

[5] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin.
Attention is all you need.
In *Advances in Neural Information Processing Systems*, 2017.
(Cited on pages 56, 57, and 58.)