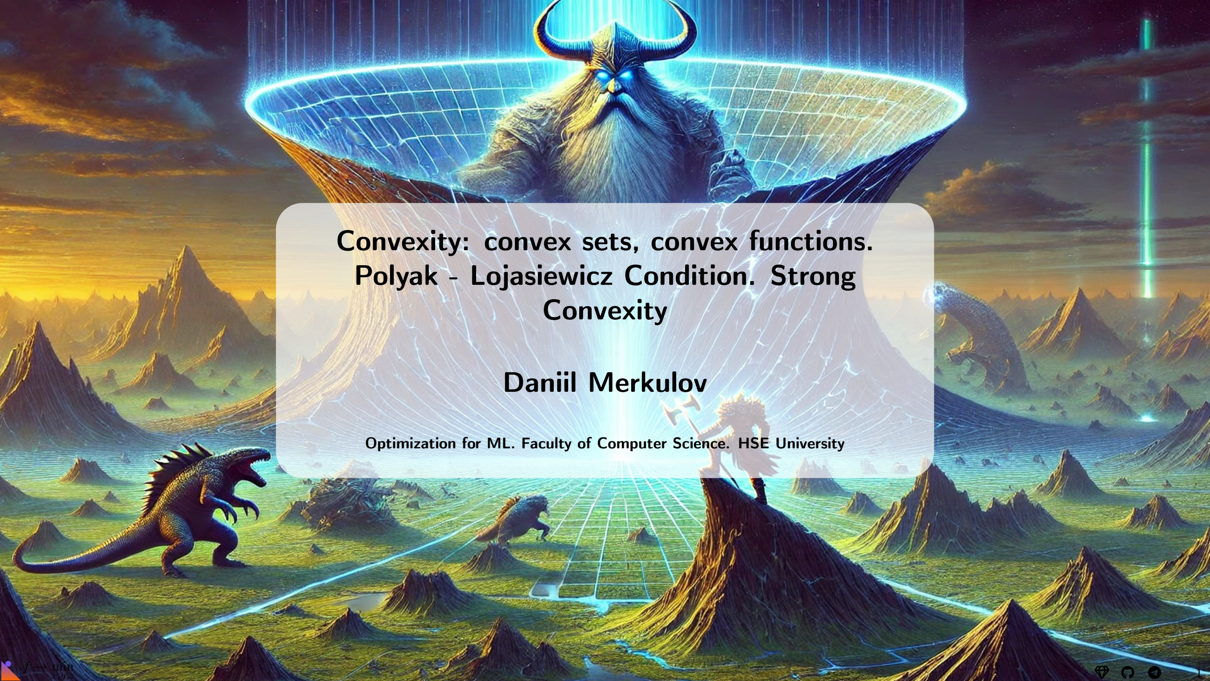


Convexity: convex sets, convex functions. Polyak - Lojasiewicz Condition. Strong Convexity

Daniil Merkulov

Optimization for ML. Faculty of Computer Science. HSE University

Convex sets

Affine set

Suppose x_1, x_2 are two points in $\mathbb{R}^{\ltimes}$. Then the line passing through them is defined as follows:

$$x = \theta x_1 + (1-\theta)x_2, \theta \in \mathbb{R}$$

The set A is called **affine** if for any x_1, x_2 from A the line passing through them also lies in A, i.e.

$$\forall \theta \in \mathbb{R}, \forall x_1, x_2 \in A : \theta x_1 + (1-\theta)x_2 \in A$$

Example

- $\mathbb{R}^n$ is an affine set.

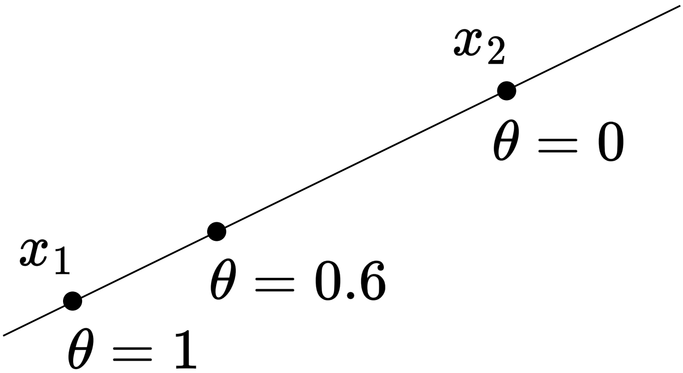


Figure 1: Illustration of a line between two vectors x_1 and x_2

Affine set

Suppose x_1, x_2 are two points in $\mathbb{R}^{\ltimes}$. Then the line passing through them is defined as follows:

$$x = \theta x_1 + (1-\theta)x_2, \theta \in \mathbb{R}$$

The set A is called **affine** if for any x_1, x_2 from A the line passing through them also lies in A, i.e.

$$\forall \theta \in \mathbb{R}, \forall x_1, x_2 \in A : \theta x_1 + (1-\theta)x_2 \in A$$

ℹ Example

- $\mathbb{R}^n$ is an affine set.
- The set of solutions $\{x \mid \mathbf{A}x = \mathbf{b}\}$ is also an affine set.

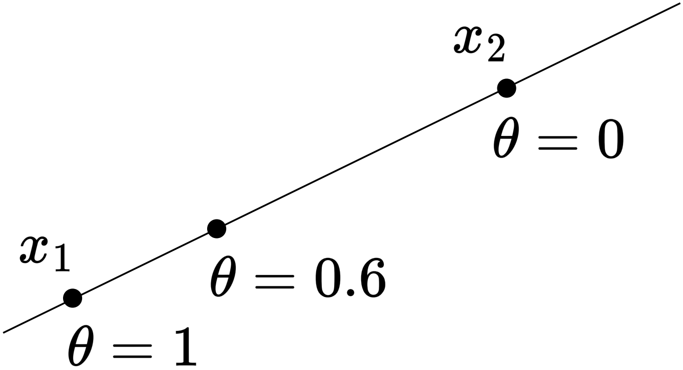


Figure 1: Illustration of a line between two vectors x_1 and x_2

Cone

A non-empty set S is called a cone, if:

$$\forall x \in S,\ \theta \geq 0 \ \ \rightarrow \ \ \theta x \in S$$

For any point in the cone, it also contains a beam through this point.

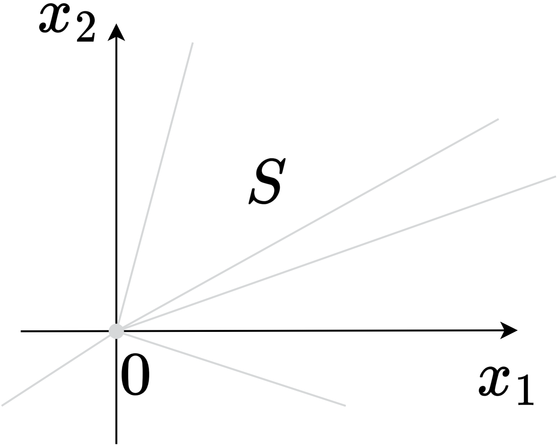


Figure 2: Illustration of a cone

Convex cone

The set S is called a convex cone, if:

$$\forall x_1, x_2 \in S,\ \theta_1, \theta_2 \geq 0 \ \ \rightarrow \ \ \theta_1 x_1 + \theta_2 x_2 \in S$$

A Convex cone is just like a cone, but it is also convex.

Example

- $\mathbb{R}^n$

Convex cone: set that contains all conic combinations of points in the set

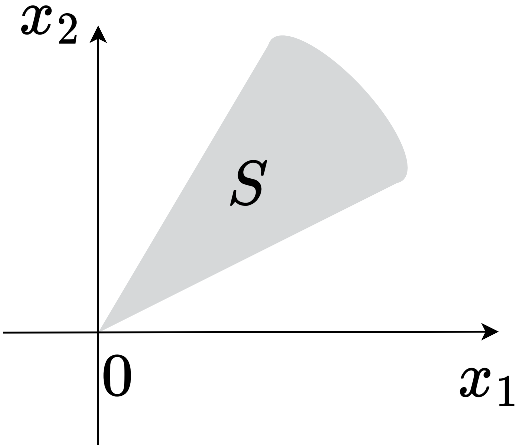


Figure 3: Illustration of a convex cone

Convex cone

The set S is called a convex cone, if:

$$\forall x_1, x_2 \in S,\ \theta_1, \theta_2 \geq 0 \ \ \rightarrow \ \ \theta_1 x_1 + \theta_2 x_2 \in S$$

A Convex cone is just like a cone, but it is also convex.

ℹ Example

- $\mathbb{R}^n$
- Affine sets, containing 0

Convex cone: set that contains all conic combinations of points in the set

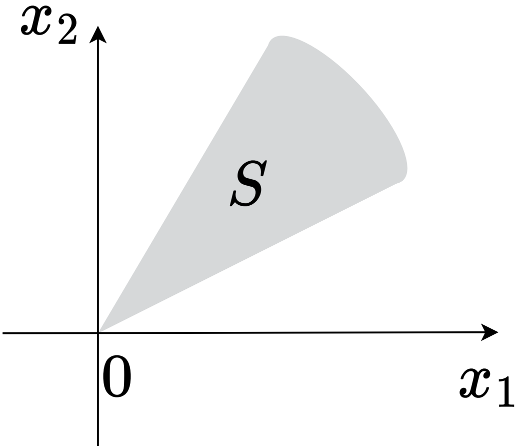


Figure 3: Illustration of a convex cone

Convex cone

The set S is called a convex cone, if:

$$\forall x_1, x_2 \in S,\ \theta_1, \theta_2 \geq 0 \ \ \rightarrow \ \ \theta_1 x_1 + \theta_2 x_2 \in S$$

A Convex cone is just like a cone, but it is also convex.

Example

- $\mathbb{R}^n$
- Affine sets, containing 0
- Ray

Convex cone: set that contains all conic combinations of points in the set

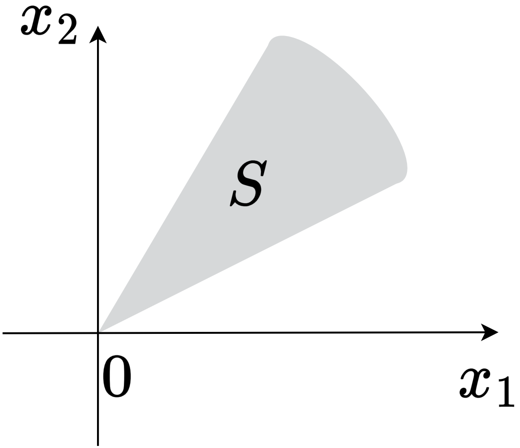


Figure 3: Illustration of a convex cone

Convex cone

The set S is called a convex cone, if:

$$\forall x_1, x_2 \in S,\ \theta_1, \theta_2 \geq 0 \quad \to \quad \theta_1 x_1 + \theta_2 x_2 \in S$$

A Convex cone is just like a cone, but it is also convex.

ℹ Example

- $\mathbb{R}^n$
- Affine sets, containing 0
- Ray
- $\mathbf{S}^n_+$ - the set of symmetric positive semi-definite matrices

Convex cone: set that contains all conic combinations of points in the set

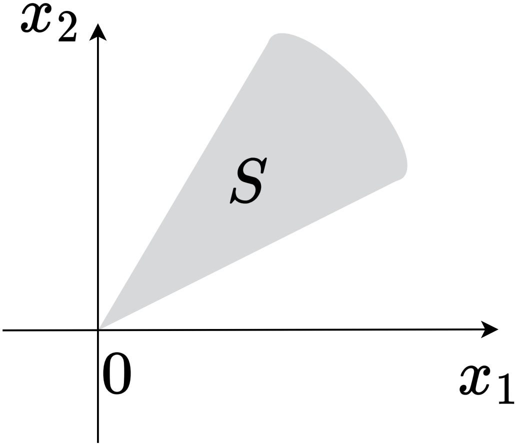


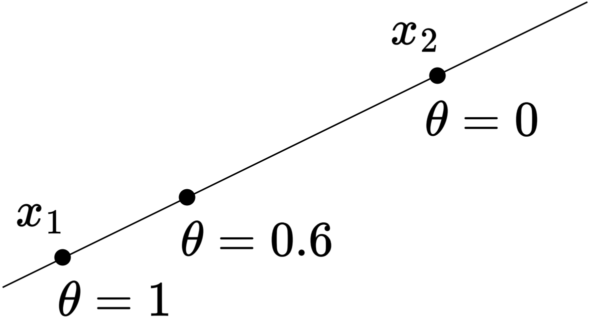
Figure 3: Illustration of a convex cone

Line segment

Suppose x_1, x_2 are two points in $\mathbb{R}^n$.
Then the line segment between them is defined as follows:

$$
x = \theta x_1 + (1-\theta)x_2, \ \theta \in [0,1]
$$

A Convex set contains a line segment between any two points in the set.



Convex set

The set S is called **convex** if for any x_1, x_2 from S the line segment between them also lies in S, i.e.

$$\forall \theta \in [0,1], \; \forall x_1, x_2 \in S : \theta x_1 + (1-\theta) x_2 \in S$$

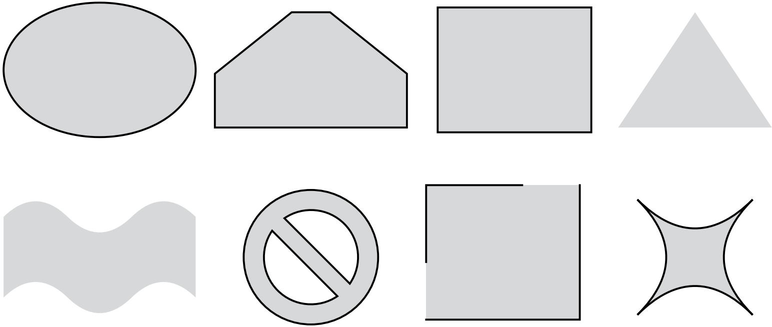


Figure 5: Top: examples of convex sets. Bottom: examples of non-convex sets.

ℹ Example

An empty set and a set from a single vector are convex by definition.

ℹ Example

Any affine set, a ray, or a line segment are all convex sets.

Convex combination

Let $x_1, x_2, \dots, x_k \in S$, then the point $\theta_1 x_1 + \theta_2 x_2 + \dots + \theta_k x_k$ is called the convex combination of points $x_1, x_2, \dots, x_k$ if $\sum\limits_{i=1}^{k} \theta_i = 1, \ \theta_i \geq 0$.

Convex hull

The set of all convex combinations of points from S is called the convex hull of the set S.

$$\mathbf{conv}(S) = \left\{ \sum_{i=1}^{k} \theta_i x_i \mid x_i \in S, \sum_{i=1}^{k} \theta_i = 1, \; \theta_i \geq 0 \right\}$$

- The set $\mathbf{conv}(S)$ is the smallest convex set containing S.

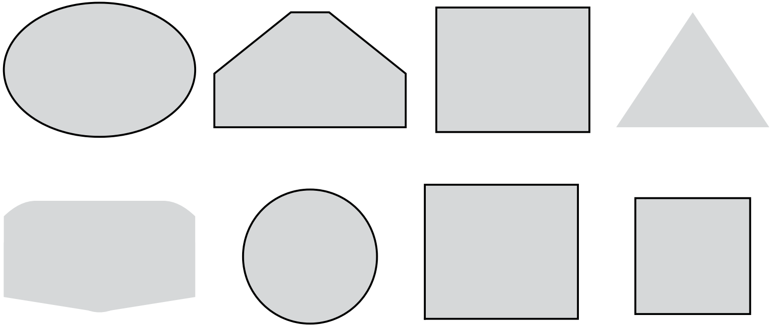


Figure 6: Top: convex hulls of the convex sets. Bottom: the convex hull of the non-convex sets.

Convex hull

The set of all convex combinations of points from S is called the convex hull of the set S.

$$\mathbf{conv}(S) = \left\{ \sum_{i=1}^{k} \theta_i x_i \mid x_i \in S, \sum_{i=1}^{k} \theta_i = 1, \ \theta_i \geq 0 \right\}$$

- The set $\mathbf{conv}(S)$ is the smallest convex set containing S.
- The set S is convex if and only if $S = \mathbf{conv}(S)$.

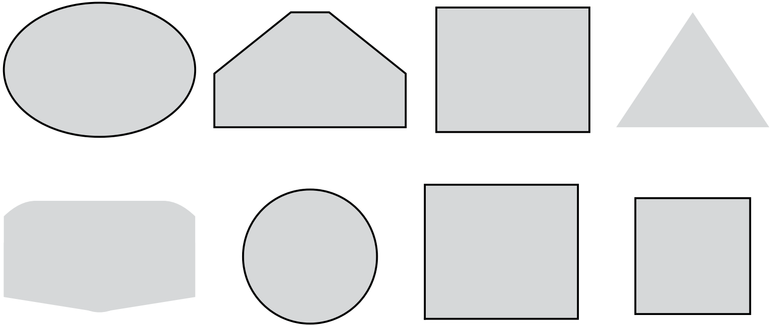


Figure 6: Top: convex hulls of the convex sets. Bottom: the convex hull of the non-convex sets.

Minkowski addition

The Minkowski sum of two sets of vectors S_1 and S_2 in Euclidean space is formed by adding each vector in S_1 to each vector in S_2.

$$S_1 + S_2 = \{\mathbf{s_1} + \mathbf{s_2} \,|\, \mathbf{s_1} \in S_1,\; \mathbf{s_2} \in S_2\}$$

Similarly, one can define a linear combination of the sets.

ℹ Example

We will work in the $\mathbb{R}^2$ space. Let's define:

$$S_1 := \{x \in \mathbb{R}^2 : x_1^2 + x_2^2 \leq 1\}$$

This is a unit circle centered at the origin. And:

$$S_2 := \{x \in \mathbb{R}^2 : -4 \leq x_1 \leq -1, -3 \leq x_2 \leq -1\}$$

This represents a rectangle. The sum of the sets S_1 and S_2 will form an enlarged rectangle S_2 with rounded corners. The resulting set will be convex.

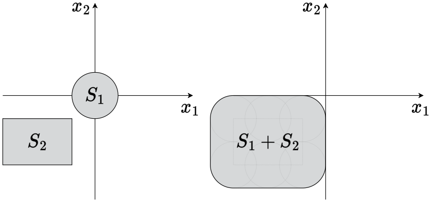


Figure 7: $S = S_1 + S_2$

Finding convexity

In practice, it is very important to understand whether a specific set is convex or not. Two approaches are used for this depending on the context.

- By definition.

Finding convexity

In practice, it is very important to understand whether a specific set is convex or not. Two approaches are used for this depending on the context.

- By definition.
- Show that S is derived from simple convex sets using operations that preserve convexity.

Finding convexity by definition

$$x_1, x_2 \in S,\ 0 \leq \theta \leq 1 \ \ \rightarrow \ \ \theta x_1 + (1-\theta)x_2 \in S$$

Example

Prove, that the set of symmetric positive definite matrices $\mathbf{S}^n_{++} = \{\mathbf{X} \in \mathbb{R}^{n \times n} \mid \mathbf{X} = \mathbf{X}^\top,\ \mathbf{X} \succ 0\}$ is convex.

Operations, that preserve convexity

The linear combination of convex sets is convex Let there be 2 convex sets S_x, S_y, let the set

$$S = \left\{ s \mid s = c_1 x + c_2 y, \ x \in S_x, \ y \in S_y, \ c_1, c_2 \in \mathbb{R} \right\}$$

Take two points from S: $s_1 = c_1 x_1 + c_2 y_1, s_2 = c_1 x_2 + c_2 y_2$ and prove that the segment between them $\theta s_1 + (1 - \theta) s_2, \theta \in [0, 1]$ also belongs to S

$$\theta s_1 + (1 - \theta) s_2$$

$$\theta (c_1 x_1 + c_2 y_1) + (1 - \theta)(c_1 x_2 + c_2 y_2)$$

$$c_1 (\theta x_1 + (1 - \theta) x_2) + c_2 (\theta y_1 + (1 - \theta) y_2)$$

$$c_1 x + c_2 y \in S$$

The intersection of any (!) number of convex sets is convex

If the desired intersection is empty or contains one point, the property is proved by definition. Otherwise, take 2 points and a segment between them. These points must lie in all intersecting sets, and since they are all convex, the segment between them lies in all sets and, therefore, in their intersection.

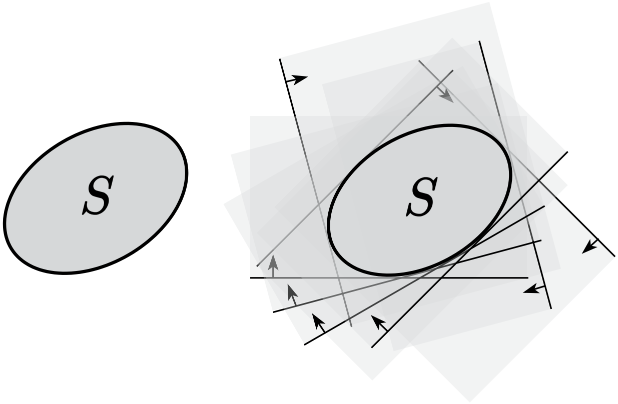


Figure 8: Intersection of halfplanes

The image of the convex set under affine mapping is convex

$$S \subseteq \mathbb{R}^n \text{ convex} \quad \rightarrow \quad f(S) = \{f(x) \mid x \in S\} \text{ convex} \qquad (f(x) = \mathbf{A}x + \mathbf{b})$$

Examples of affine functions: extension, projection, transposition, set of solutions of linear matrix inequality $\{x \mid x_1 A_1 + ... + x_m A_m \preceq B\}$. Here $A_i, B \in \mathbf{S}^p$ are symmetric matrices $p \times p$.

Note also that the prototype of the convex set under affine mapping is also convex.

$$S \subseteq \mathbb{R}^m \text{ convex} \ \rightarrow \ f^{-1}(S) = \{x \in \mathbb{R}^n \mid f(x) \in S\} \text{ convex} \ \ (f(x) = \mathbf{A}x + \mathbf{b})$$

Example

Let $x \in \mathbb{R}$ is a random variable with a given probability distribution of $\mathbb{P}(x = a_i) = p_i$, where $i = 1, \dots, n$, and $a_1 < \dots < a_n$. It is said that the probability vector of outcomes of $p \in \mathbb{R}^n$ belongs to the probabilistic simplex, i.e.

$$P = \{p \mid \mathbf{1}^T p = 1, p \succeq 0\} = \{p \mid p_1 + \dots + p_n = 1, p_i \ge 0\}.$$

Determine if the following sets of p are convex:

- $\mathbb{P}(x > \alpha) \le \beta$

Example

Let $x \in \mathbb{R}$ is a random variable with a given probability distribution of $\mathbb{P}(x = a_i) = p_i$, where $i = 1, \ldots, n$, and $a_1 < \ldots < a_n$. It is said that the probability vector of outcomes of $p \in \mathbb{R}^n$ belongs to the probabilistic simplex, i.e.

$$P = \{p \mid \mathbf{1}^T p = 1, p \succeq 0\} = \{p \mid p_1 + \ldots + p_n = 1, p_i \geq 0\}.$$

Determine if the following sets of p are convex:

- $\mathbb{P}(x > \alpha) \leq \beta$
- $\mathbb{E}|x^{201}| \leq \alpha \mathbb{E}|x|$

Example

Let $x \in \mathbb{R}$ is a random variable with a given probability distribution of $\mathbb{P}(x = a_i) = p_i$, where $i = 1, \ldots, n$, and $a_1 < \ldots < a_n$. It is said that the probability vector of outcomes of $p \in \mathbb{R}^n$ belongs to the probabilistic simplex, i.e.

$$P = \{p \mid \mathbf{1}^T p = 1, p \succeq 0\} = \{p \mid p_1 + \ldots + p_n = 1, p_i \geq 0\}.$$

Determine if the following sets of p are convex:

- $\mathbb{P}(x > \alpha) \leq \beta$
- $\mathbb{E}|x^{201}| \leq \alpha \mathbb{E}|x|$
- $\mathbb{E}|x^2| \geq \alpha \mathbb{V}x \geq \alpha$

Convex functions

Jensen's inequality

The function $f(x)$, **which is defined on the convex set** $S \subseteq \mathbb{R}^n$, is called **convex** on S, if:

$$f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda) f(x_2)$$

for any $x_1, x_2 \in S$ and $0 \leq \lambda \leq 1$.
If the above inequality holds as strict inequality $x_1 \neq x_2$ and $0 < \lambda < 1$, then the function is called strictly convex on S.

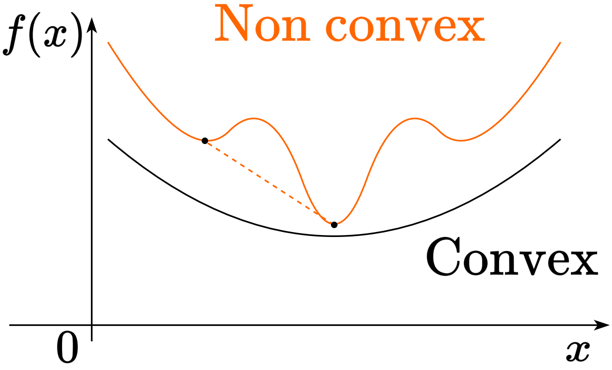


Figure 9: Difference between convex and non-convex function

Jensen's inequality

ℹ Theorem

Let $f(x)$ be a convex function on a convex set $X \subseteq \mathbb{R}^n$ and let $x_i \in X, 1 \leq i \leq m$, be arbitrary points from X. Then

$$f\left(\sum_{i=1}^{m} \lambda_i x_i\right) \leq \sum_{i=1}^{m} \lambda_i f(x_i)$$

for any $\lambda = [\lambda_1, \ldots, \lambda_m] \in \Delta_m$ - probability simplex.

Proof

1. First, note that the point $\sum_{i=1}^{m} \lambda_i x_i$ as a convex combination of points from the convex set X belongs to X.

Jensen's inequality

Theorem

Let $f(x)$ be a convex function on a convex set $X \subseteq \mathbb{R}^n$ and let $x_i \in X, 1 \leq i \leq m$, be arbitrary points from X. Then

$$f\left(\sum_{i=1}^{m} \lambda_i x_i\right) \leq \sum_{i=1}^{m} \lambda_i f(x_i)$$

for any $\lambda = [\lambda_1, \ldots, \lambda_m] \in \Delta_m$ - probability simplex.

Proof

1. First, note that the point $\sum_{i=1}^{m} \lambda_i x_i$ as a convex combination of points from the convex set X belongs to X.
2. We will prove this by induction. For $m = 1$, the statement is obviously true, and for $m = 2$, it follows from the definition of a convex function.

Jensen's inequality

3. Assume it is true for all m up to $m = k$, and we will prove it for $m = k + 1$. Let $\lambda \in \Delta k + 1$ and

$$x = \sum_{i=1}^{k+1} \lambda_i x_i = \sum_{i=1}^{k} \lambda_i x_i + \lambda_{k+1} x_{k+1}.$$

Assuming $0 < \lambda_{k+1} < 1$, as otherwise, it reduces to previously considered cases, we have

$$x = \lambda_{k+1} x_{k+1} + (1 - \lambda_{k+1}) \bar{x},$$

where $\bar{x} = \sum_{i=1}^{k} \gamma_i x_i$ and $\gamma_i = \frac{\lambda_i}{1 - \lambda_{k+1}} \geq 0, 1 \leq i \leq k$.

$$f\left(\sum_{i=1}^{k+1} \lambda_i x_i\right) = f\left(\lambda_{k+1} x_{k+1} + (1 - \lambda_{k+1}) \bar{x}\right) \leq \lambda_{k+1} f(x_{k+1}) + (1 - \lambda_{k+1}) f(\bar{x}) \leq \sum_{i=1}^{k+1} \lambda_i f(x_i)$$

Thus, initial inequality is satisfied for $m = k + 1$ as well.

Jensen's inequality

3. Assume it is true for all m up to $m = k$, and we will prove it for $m = k+1$. Let $\lambda \in \Delta k + 1$ and

$$x = \sum_{i=1}^{k+1} \lambda_i x_i = \sum_{i=1}^{k} \lambda_i x_i + \lambda_{k+1} x_{k+1}.$$

Assuming $0 < \lambda_{k+1} < 1$, as otherwise, it reduces to previously considered cases, we have

$$x = \lambda_{k+1} x_{k+1} + (1 - \lambda_{k+1})\bar{x},$$

where $\bar{x} = \sum_{i=1}^{k} \gamma_i x_i$ and $\gamma_i = \frac{\lambda_i}{1-\lambda_{k+1}} \geq 0, 1 \leq i \leq k$.

4. Since $\lambda \in \Delta_{k+1}$, then $\gamma = [\gamma_1, \ldots, \gamma_k] \in \Delta_k$. Therefore $\bar{x} \in X$ and by the convexity of $f(x)$ and the induction hypothesis:

$$f\left(\sum_{i=1}^{k+1} \lambda_i x_i\right) = f\left(\lambda_{k+1} x_{k+1} + (1 - \lambda_{k+1})\bar{x}\right) \leq \lambda_{k+1} f(x_{k+1}) + (1 - \lambda_{k+1}) f(\bar{x}) \leq \sum_{i=1}^{k+1} \lambda_i f(x_i)$$

Thus, initial inequality is satisfied for $m = k+1$ as well.

Examples of convex functions

- $f(x) = x^p, \; p > 1, \; x \in \mathbb{R}_+$
- $f(x) = \|x\|^p, \; p > 1, x \in \mathbb{R}^n$
- $f(x) = e^{cx}, \; c \in \mathbb{R}, x \in \mathbb{R}$
- $f(x) = -\ln x, \; x \in \mathbb{R}_{++}$
- $f(x) = x \ln x, \; x \in \mathbb{R}_{++}$
- The sum of the largest k coordinates $f(x) = x_{(1)} + ... + x_{(k)}, \; x \in \mathbb{R}^n$
- $f(X) = \lambda_{max}(X), \; X = X^T$
- $f(X) = -\log \det X, \; X \in S^n_{++}$

Epigraph

For the function $f(x)$, defined on $S \subseteq \mathbb{R}^n$, the following set:

$$\text{epi } f = \{[x, \mu] \in S \times \mathbb{R} : f(x) \leq \mu\}$$

is called **epigraph** of the function $f(x)$.

ℹ Convexity of the epigraph is the convexity of the function

For a function $f(x)$, defined on a convex set X, to be convex on X, it is necessary and sufficient that the epigraph of f is a convex set.

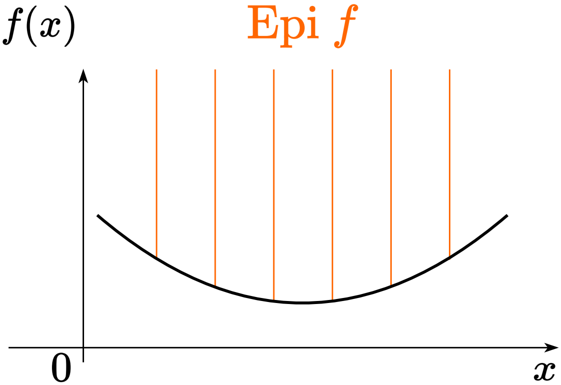


Figure 10: Epigraph of a function

Convexity of the epigraph is the convexity of the function

1. **Necessity**: Assume $f(x)$ is convex on X. Take any two arbitrary points $[x_1, \mu_1] \in \text{epi} f$ and $[x_2, \mu_2] \in \text{epi} f$. Also take $0 \leq \lambda \leq 1$ and denote $x_\lambda = \lambda x_1 + (1-\lambda)x_2, \mu_\lambda = \lambda \mu_1 + (1-\lambda)\mu_2$. Then,

$$\lambda \begin{bmatrix} x_1 \\ \mu_1 \end{bmatrix} + (1-\lambda) \begin{bmatrix} x_2 \\ \mu_2 \end{bmatrix} = \begin{bmatrix} x_\lambda \\ \mu_\lambda \end{bmatrix}.$$

From the convexity of the set X, it follows that $x_\lambda \in X$. Moreover, since $f(x)$ is a convex function,

$$f(x_\lambda) \leq \lambda f(x_1) + (1-\lambda) f(x_2) \leq \lambda \mu_1 + (1-\lambda)\mu_2 = \mu_\lambda$$

Inequality above indicates that $\begin{bmatrix} x_\lambda \\ \mu_\lambda \end{bmatrix} \in \text{epi} f$. Thus, the epigraph of f is a convex set.

Convexity of the epigraph is the convexity of the function

1. **Necessity**: Assume $f(x)$ is convex on X. Take any two arbitrary points $[x_1, \mu_1] \in$ epi$f$ and $[x_2, \mu_2] \in$ epi$f$. Also take $0 \leq \lambda \leq 1$ and denote $x_\lambda = \lambda x_1 + (1-\lambda)x_2, \mu_\lambda = \lambda \mu_1 + (1-\lambda)\mu_2$. Then,

$$\lambda \begin{bmatrix} x_1 \\ \mu_1 \end{bmatrix} + (1-\lambda) \begin{bmatrix} x_2 \\ \mu_2 \end{bmatrix} = \begin{bmatrix} x_\lambda \\ \mu_\lambda \end{bmatrix}.$$

From the convexity of the set X, it follows that $x_\lambda \in X$. Moreover, since $f(x)$ is a convex function,

$$f(x_\lambda) \leq \lambda f(x_1) + (1-\lambda) f(x_2) \leq \lambda \mu_1 + (1-\lambda)\mu_2 = \mu_\lambda$$

Inequality above indicates that $\begin{bmatrix} x_\lambda \\ \mu_\lambda \end{bmatrix} \in$ epi$f$. Thus, the epigraph of f is a convex set.

2. **Sufficiency**: Assume the epigraph of f, epi$f$, is a convex set. Then, from the membership of the points $[x_1, \mu_1]$ and $[x_2, \mu_2]$ in the epigraph of f, it follows that

$$\begin{bmatrix} x_\lambda \\ \mu_\lambda \end{bmatrix} = \lambda \begin{bmatrix} x_1 \\ \mu_1 \end{bmatrix} + (1-\lambda) \begin{bmatrix} x_2 \\ \mu_2 \end{bmatrix} \in \text{epi} f$$

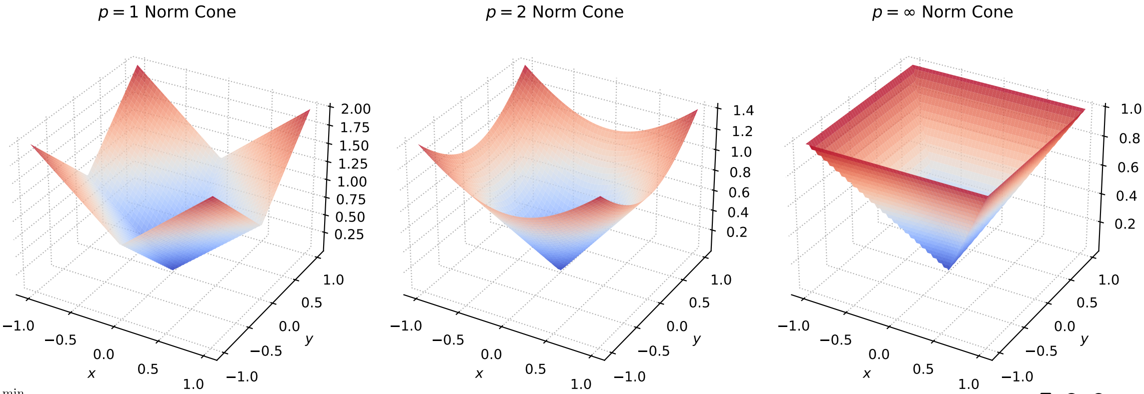
for any $0 \leq \lambda \leq 1$, i.e., $f(x_\lambda) \leq \mu_\lambda = \lambda \mu_1 + (1-\lambda)\mu_2$. But this is true for all $\mu_1 \geq f(x_1)$ and $\mu_2 \geq f(x_2)$, particularly when $\mu_1 = f(x_1)$ and $\mu_2 = f(x_2)$. Hence we arrive at the inequality

Example: norm cone

Let a norm $\|\cdot\|$ be defined in the space U. Consider the set:

$$K := \{(x, t) \in U \times \mathbb{R}^+ : \|x\| \leq t\}$$

which represents the epigraph of the function $x \mapsto \|x\|$. This set is called the cone norm. According to the statement above, the set K is convex. Code for the figures


$p = 1$ Norm Cone

$p = 2$ Norm Cone

$p = \infty$ Norm Cone

Sublevel set

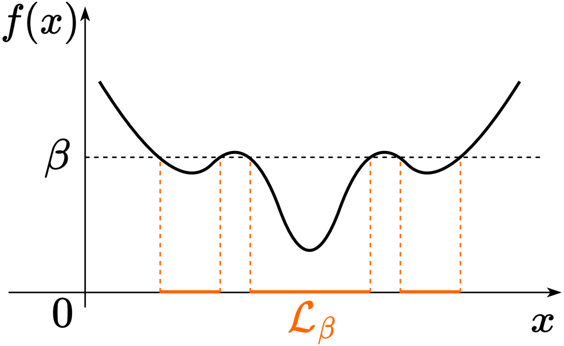


Figure 12: Sublevel set of a function with respect to level β

For the function $f(x)$, defined on $S \subseteq \mathbb{R}^n$, the following set:

$$\mathcal{L}_\beta = \{x \in S : f(x) \leq \beta\}$$

is called **sublevel set** or Lebesgue set of the function $f(x)$.

Sublevel set

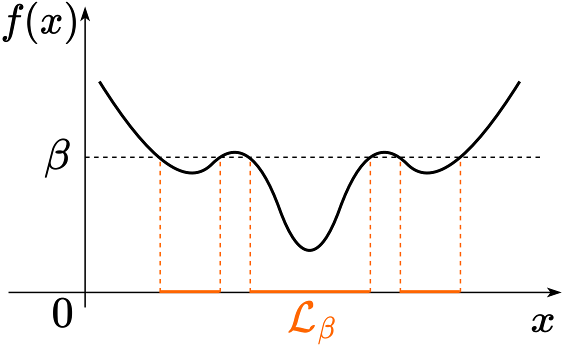


Figure 12: Sublevel set of a function with respect to level β

For the function $f(x)$, defined on $S \subseteq \mathbb{R}^n$, the following set:

$$\mathcal{L}_\beta = \{x \in S : f(x) \leq \beta\}$$

is called **sublevel set** or Lebesgue set of the function $f(x)$. Note, that if the function $f(x)$ is convex, then its sublevel sets are convex for any $\beta \in \mathbb{R}$.

While the **converse is not true**. (For example, consider the function $f(x) = \sqrt{|x|}$)

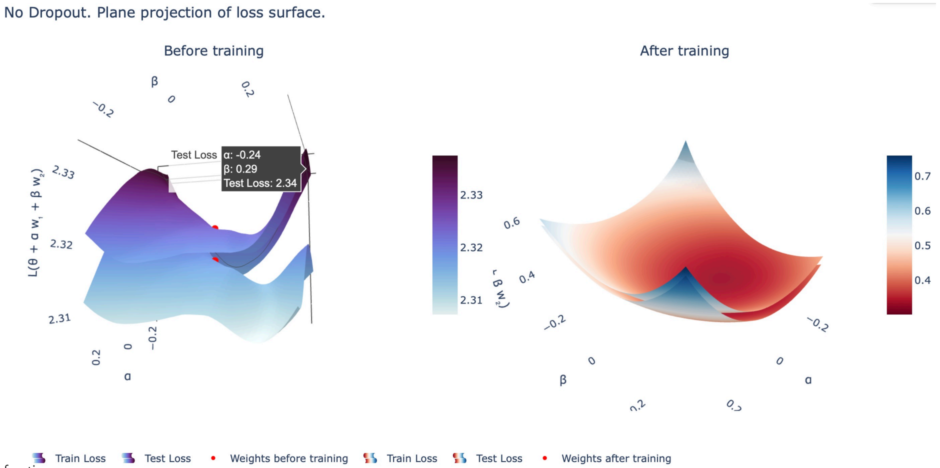
Reduction to a line

$f : S \to \mathbb{R}$ is convex if and only if S is a convex set and the function $g(t) = f(x + tv)$ defined on $\{t \mid x + tv \in S\}$ is convex for any $x \in S, v \in \mathbb{R}^n$, which allows checking convexity of the scalar function to establish convexity of the vector function.

Reduction to a line

$f : S \to \mathbb{R}$ is convex if and only if S is a convex set and the function $g(t) = f(x + tv)$ defined on $\{t \mid x + tv \in S\}$ is convex for any $x \in S, v \in \mathbb{R}^n$, which allows checking convexity of the scalar function to establish convexity of the vector function.

If you find a direction v for which $g(t)$ is not convex, then f is not convex.



Operations that preserve convexity

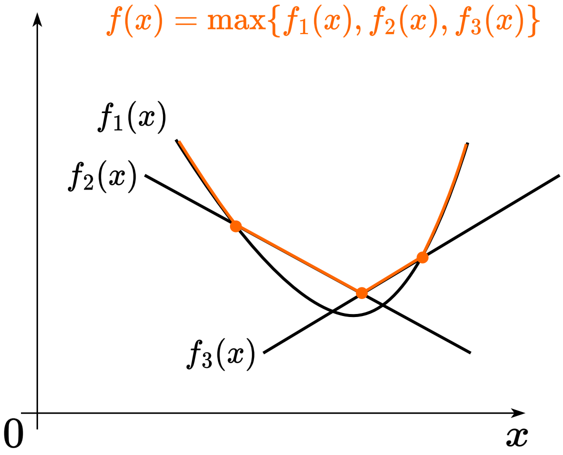


Figure 13: Pointwise maximum (supremum) of convex functions is convex

- Pointwise maximum (supremum) of any number of functions: If $f_1(x), \dots, f_m(x)$ are convex, then $f(x) = \max\{f_1(x), \dots, f_m(x)\}$ is convex.

Operations that preserve convexity

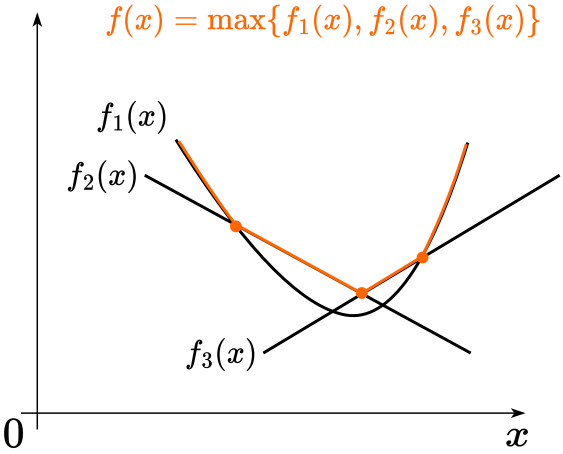


Figure 13: Pointwise maximum (supremum) of convex functions is convex

- Pointwise maximum (supremum) of any number of functions: If $f_1(x), \dots, f_m(x)$ are convex, then $f(x) = \max\{f_1(x), \dots, f_m(x)\}$ is convex.
- Non-negative sum of the convex functions: $\alpha f(x) + \beta g(x), (\alpha \geq 0, \beta \geq 0)$.

Operations that preserve convexity

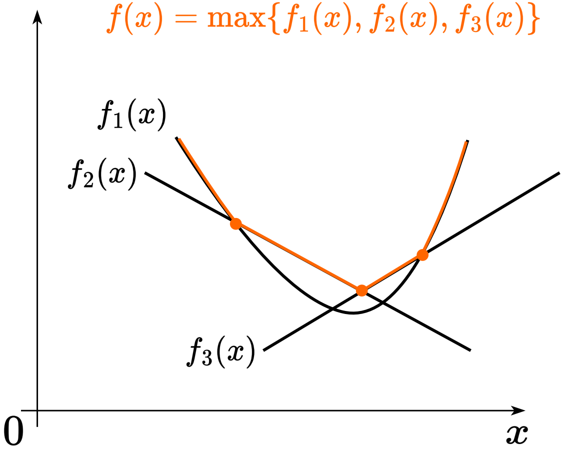


Figure 13: Pointwise maximum (supremum) of convex functions is convex

- Pointwise maximum (supremum) of any number of functions: If $f_1(x), \dots, f_m(x)$ are convex, then $f(x) = \max\{f_1(x), \dots, f_m(x)\}$ is convex.
- Non-negative sum of the convex functions: $\alpha f(x) + \beta g(x), (\alpha \geq 0, \beta \geq 0)$.
- Composition with affine function $f(Ax + b)$ is convex, if $f(x)$ is convex.

Operations that preserve convexity

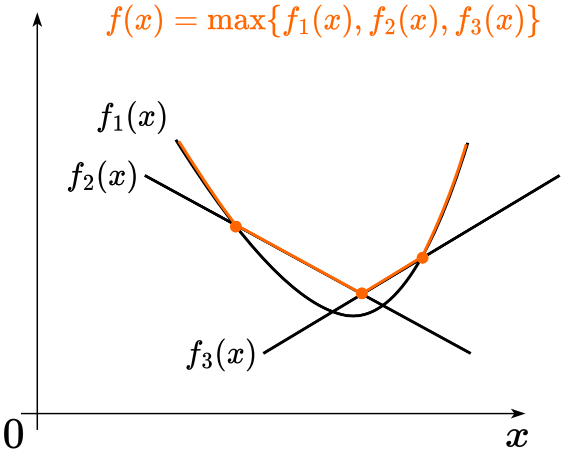


Figure 13: Pointwise maximum (supremum) of convex functions is convex

- Pointwise maximum (supremum) of any number of functions: If $f_1(x), \ldots, f_m(x)$ are convex, then $f(x) = \max\{f_1(x), \ldots, f_m(x)\}$ is convex.
- Non-negative sum of the convex functions: $\alpha f(x) + \beta g(x), (\alpha \geq 0, \beta \geq 0)$.
- Composition with affine function $f(Ax + b)$ is convex, if $f(x)$ is convex.
- If $f(x, y)$ is convex on x for any $y \in Y$: $g(x) = \sup\limits_{y \in Y} f(x, y)$ is convex.

Operations that preserve convexity

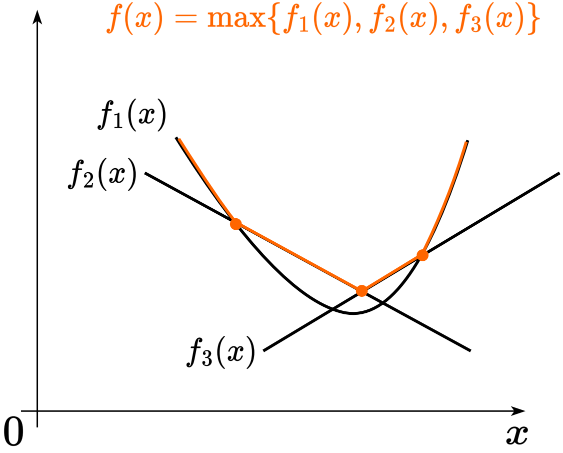


Figure 13: Pointwise maximum (supremum) of convex functions is convex

- Pointwise maximum (supremum) of any number of functions: If $f_1(x), \ldots, f_m(x)$ are convex, then $f(x) = \max\{f_1(x), \ldots, f_m(x)\}$ is convex.
- Non-negative sum of the convex functions: $\alpha f(x) + \beta g(x), (\alpha \geq 0, \beta \geq 0)$.
- Composition with affine function $f(Ax+b)$ is convex, if $f(x)$ is convex.
- If $f(x,y)$ is convex on x for any $y \in Y$: $g(x) = \sup\limits_{y \in Y} f(x,y)$ is convex.
- If $f(x)$ is convex on S, then $g(x,t) = tf(x/t)$ - is convex with $x/t \in S, t > 0$.

Operations that preserve convexity

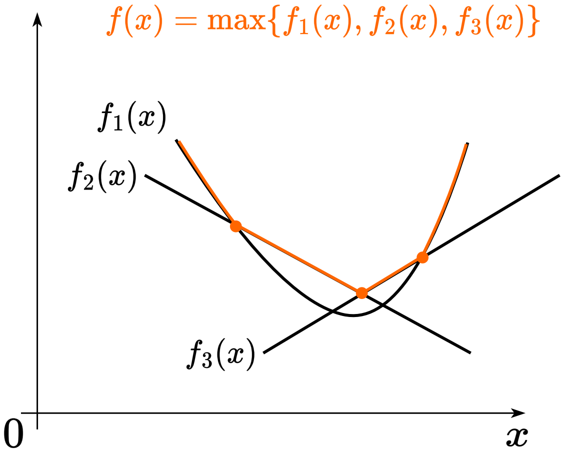


Figure 13: Pointwise maximum (supremum) of convex functions is convex

- Pointwise maximum (supremum) of any number of functions: If $f_1(x), \ldots, f_m(x)$ are convex, then $f(x) = \max\{f_1(x), \ldots, f_m(x)\}$ is convex.
- Non-negative sum of the convex functions: $\alpha f(x) + \beta g(x), (\alpha \geq 0, \beta \geq 0)$.
- Composition with affine function $f(Ax + b)$ is convex, if $f(x)$ is convex.
- If $f(x, y)$ is convex on x for any $y \in Y$: $g(x) = \sup\limits_{y \in Y} f(x, y)$ is convex.
- If $f(x)$ is convex on S, then $g(x, t) = tf(x/t)$ - is convex with $x/t \in S, t > 0$.
- Let $f_1 : S_1 \to \mathbb{R}$ and $f_2 : S_2 \to \mathbb{R}$, where range$(f_1) \subseteq S_2$. If f_1 and f_2 are convex, and f_2 is increasing, then $f_2 \circ f_1$ is convex on S_1.

Maximum eigenvalue of a matrix is a convex function

Example

Show, that $f(A) = \lambda_{max}(A)$ - is convex, if $A \in S^n$.

Strong convexity criteria

First-order differential criterion of convexity

The differentiable function $f(x)$ defined on the convex set $S \subseteq \mathbb{R}^n$ is convex if and only if $\forall x, y \in S$:

$$f(y) \geq f(x) + \nabla f^T(x)(y - x)$$

Let $y = x + \Delta x$, then the criterion will become more tractable:

$$f(x + \Delta x) \geq f(x) + \nabla f^T(x) \Delta x$$

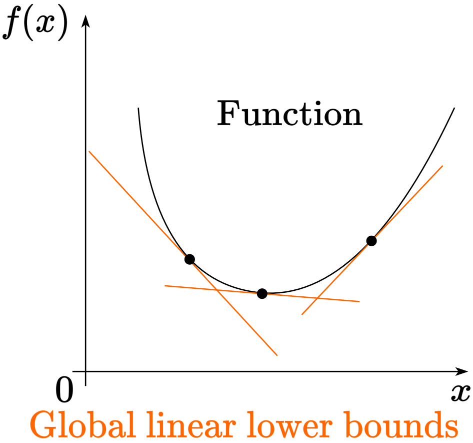


Figure 14: Convex function is greater or equal than Taylor linear approximation at any point

Second-order differential criterion of convexity

Twice differentiable function $f(x)$ defined on the convex set $S \subseteq \mathbb{R}^n$ is convex if and only if $\forall x \in \mathbf{int}(S) \neq \emptyset$:

$$\nabla^2 f(x) \succeq 0$$

In other words, $\forall y \in \mathbb{R}^n$:

$$\langle y, \nabla^2 f(x) y \rangle \geq 0$$

Strong convexity

$f(x)$, **defined on the convex set** $S \subseteq \mathbb{R}^n$, is called μ-strongly convex (strongly convex) on S, if:

$$f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda)f(x_2) - \frac{\mu}{2}\lambda(1-\lambda)\|x_1 - x_2\|^2$$

for any $x_1, x_2 \in S$ and $0 \leq \lambda \leq 1$ for some $\mu > 0$.

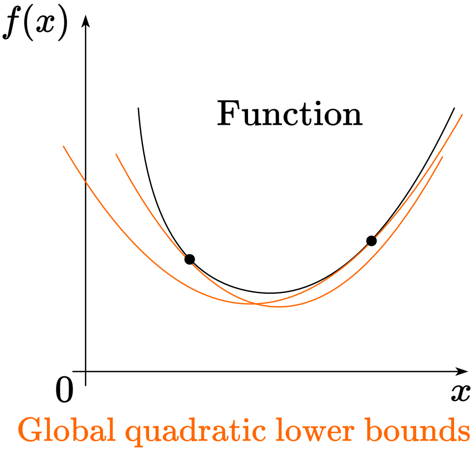


Figure 15: Strongly convex function is greater or equal than Taylor quadratic approximation at any point

First-order differential criterion of strong convexity

Differentiable $f(x)$ defined on the convex set $S \subseteq \mathbb{R}^n$ is μ-strongly convex if and only if $\forall x, y \in S$:

$$f(y) \geq f(x) + \nabla f^T(x)(y-x) + \frac{\mu}{2}\|y-x\|^2$$

First-order differential criterion of strong convexity

Differentiable $f(x)$ defined on the convex set $S \subseteq \mathbb{R}^n$ is μ-strongly convex if and only if $\forall x, y \in S$:

$$f(y) \geq f(x) + \nabla f^T(x)(y - x) + \frac{\mu}{2}\|y - x\|^2$$

Let $y = x + \Delta x$, then the criterion will become more tractable:

$$f(x + \Delta x) \geq f(x) + \nabla f^T(x)\Delta x + \frac{\mu}{2}\|\Delta x\|^2$$

First-order differential criterion of strong convexity

Differentiable $f(x)$ defined on the convex set $S \subseteq \mathbb{R}^n$ is μ-strongly convex if and only if $\forall x, y \in S$:

$$f(y) \geq f(x) + \nabla f^T(x)(y - x) + \frac{\mu}{2}\|y - x\|^2$$

Let $y = x + \Delta x$, then the criterion will become more tractable:

$$f(x + \Delta x) \geq f(x) + \nabla f^T(x)\Delta x + \frac{\mu}{2}\|\Delta x\|^2$$

Theorem

Let $f(x)$ be a differentiable function on a convex set $X \subseteq \mathbb{R}^n$. Then $f(x)$ is strongly convex on X with a constant $\mu > 0$ if and only if

$$f(x) - f(x_0) \geq \langle \nabla f(x_0), x - x_0 \rangle + \frac{\mu}{2}\|x - x_0\|^2$$

for all $x, x_0 \in X$.

Proof of first-order differential criterion of strong convexity

Necessity: Let $0 < \lambda \leq 1$. According to the definition of a strongly convex function,

$$
f(\lambda x + (1-\lambda)x_0) \leq \lambda f(x) + (1-\lambda)f(x_0) - \frac{\mu}{2}\lambda(1-\lambda)\|x - x_0\|^2
$$

Proof of first-order differential criterion of strong convexity

Necessity: Let $0 < \lambda \leq 1$. According to the definition of a strongly convex function,

$$f(\lambda x + (1-\lambda)x_0) \leq \lambda f(x) + (1-\lambda) f(x_0) - \frac{\mu}{2}\lambda(1-\lambda)\|x - x_0\|^2$$

or equivalently,

$$f(x) - f(x_0) - \frac{\mu}{2}(1-\lambda)\|x - x_0\|^2 \geq \frac{1}{\lambda}[f(\lambda x + (1-\lambda)x_0) - f(x_0)] =$$

Proof of first-order differential criterion of strong convexity

Necessity: Let $0 < \lambda \leq 1$. According to the definition of a strongly convex function,

$$f(\lambda x + (1-\lambda)x_0) \leq \lambda f(x) + (1-\lambda)f(x_0) - \frac{\mu}{2}\lambda(1-\lambda)\|x - x_0\|^2$$

or equivalently,

$$f(x) - f(x_0) - \frac{\mu}{2}(1-\lambda)\|x - x_0\|^2 \geq \frac{1}{\lambda}[f(\lambda x + (1-\lambda)x_0) - f(x_0)] =$$

$$= \frac{1}{\lambda}[f(x_0 + \lambda(x - x_0)) - f(x_0)] = \frac{1}{\lambda}[\lambda\langle\nabla f(x_0), x - x_0\rangle + o(\lambda)] =$$

Proof of first-order differential criterion of strong convexity

Necessity: Let $0 < \lambda \leq 1$. According to the definition of a strongly convex function,

$$f(\lambda x + (1-\lambda)x_0) \leq \lambda f(x) + (1-\lambda) f(x_0) - \frac{\mu}{2}\lambda(1-\lambda)\|x - x_0\|^2$$

or equivalently,

$$f(x) - f(x_0) - \frac{\mu}{2}(1-\lambda)\|x - x_0\|^2 \geq \frac{1}{\lambda}[f(\lambda x + (1-\lambda)x_0) - f(x_0)] =$$

$$= \frac{1}{\lambda}[f(x_0 + \lambda(x - x_0)) - f(x_0)] = \frac{1}{\lambda}[\lambda\langle\nabla f(x_0), x - x_0\rangle + o(\lambda)] =$$

$$= \langle\nabla f(x_0), x - x_0\rangle + \frac{o(\lambda)}{\lambda}.$$

Thus, taking the limit as $\lambda \downarrow 0$, we arrive at the initial statement.

Proof of first-order differential criterion of strong convexity

Sufficiency: Assume the inequality in the theorem is satisfied for all $x, x_0 \in X$. Take $x_0 = \lambda x_1 + (1-\lambda)x_2$, where $x_1, x_2 \in X$, $0 \leq \lambda \leq 1$. According to the inequality, the following inequalities hold:

Proof of first-order differential criterion of strong convexity

Sufficiency: Assume the inequality in the theorem is satisfied for all $x, x_0 \in X$. Take $x_0 = \lambda x_1 + (1-\lambda)x_2$, where $x_1, x_2 \in X$, $0 \leq \lambda \leq 1$. According to the inequality, the following inequalities hold:

$$
\begin{aligned}
f(x_1) - f(x_0) &\geq \langle \nabla f(x_0), x_1 - x_0 \rangle + \frac{\mu}{2}\|x_1 - x_0\|^2, \\
f(x_2) - f(x_0) &\geq \langle \nabla f(x_0), x_2 - x_0 \rangle + \frac{\mu}{2}\|x_2 - x_0\|^2.
\end{aligned}
$$

Multiplying the first inequality by λ and the second by $1-\lambda$ and adding them, considering that

Proof of first-order differential criterion of strong convexity

Sufficiency: Assume the inequality in the theorem is satisfied for all $x, x_0 \in X$. Take $x_0 = \lambda x_1 + (1-\lambda)x_2$, where $x_1, x_2 \in X$, $0 \leq \lambda \leq 1$. According to the inequality, the following inequalities hold:

$$f(x_1) - f(x_0) \geq \langle \nabla f(x_0), x_1 - x_0 \rangle + \frac{\mu}{2}\|x_1 - x_0\|^2,$$

$$f(x_2) - f(x_0) \geq \langle \nabla f(x_0), x_2 - x_0 \rangle + \frac{\mu}{2}\|x_2 - x_0\|^2.$$

Multiplying the first inequality by λ and the second by $1-\lambda$ and adding them, considering that

$$x_1 - x_0 = (1-\lambda)(x_1 - x_2), \quad x_2 - x_0 = \lambda(x_2 - x_1),$$

and $\lambda(1-\lambda)^2 + \lambda^2(1-\lambda) = \lambda(1-\lambda)$, we get

Proof of first-order differential criterion of strong convexity

Sufficiency: Assume the inequality in the theorem is satisfied for all $x, x_0 \in X$. Take $x_0 = \lambda x_1 + (1-\lambda)x_2$, where $x_1, x_2 \in X$, $0 \leq \lambda \leq 1$. According to the inequality, the following inequalities hold:

$$
\begin{aligned}
f(x_1) - f(x_0) &\geq \langle \nabla f(x_0), x_1 - x_0 \rangle + \frac{\mu}{2}\|x_1 - x_0\|^2, \\
f(x_2) - f(x_0) &\geq \langle \nabla f(x_0), x_2 - x_0 \rangle + \frac{\mu}{2}\|x_2 - x_0\|^2.
\end{aligned}
$$

Multiplying the first inequality by λ and the second by $1 - \lambda$ and adding them, considering that

$$
x_1 - x_0 = (1-\lambda)(x_1 - x_2), \quad x_2 - x_0 = \lambda(x_2 - x_1),
$$

and $\lambda(1-\lambda)^2 + \lambda^2(1-\lambda) = \lambda(1-\lambda)$, we get

$$
\begin{aligned}
\lambda f(x_1) + (1-\lambda)f(x_2) - f(x_0) - \frac{\mu}{2}\lambda(1-\lambda)\|x_1 - x_2\|^2 \geq \\
\langle \nabla f(x_0), \lambda x_1 + (1-\lambda)x_2 - x_0 \rangle = 0.
\end{aligned}
$$

Thus, inequality from the definition of a strongly convex function is satisfied. It is important to mention, that $\mu = 0$ stands for the convex case and corresponding differential criterion.

Second-order differential criterion of strong convexity

Twice differentiable function $f(x)$ defined on the convex set $S \subseteq \mathbb{R}^n$ is called μ-strongly convex if and only if $\forall x \in \mathbf{int}(S) \neq \emptyset$:

$$\nabla^2 f(x) \succeq \mu I$$

In other words:

$$\langle y, \nabla^2 f(x) y \rangle \geq \mu \|y\|^2$$

Second-order differential criterion of strong convexity

Twice differentiable function $f(x)$ defined on the convex set $S \subseteq \mathbb{R}^n$ is called μ-strongly convex if and only if $\forall x \in \mathbf{int}(S) \neq \emptyset$:

$$\nabla^2 f(x) \succeq \mu I$$

In other words:

$$\langle y, \nabla^2 f(x) y \rangle \geq \mu \|y\|^2$$

Theorem

Let $X \subseteq \mathbb{R}^n$ be a convex set, with int$X \neq \emptyset$. Furthermore, let $f(x)$ be a twice continuously differentiable function on X. Then $f(x)$ is strongly convex on X with a constant $\mu > 0$ if and only if

$$\langle y, \nabla^2 f(x) y \rangle \geq \mu \|y\|^2$$

for all $x \in X$ and $y \in \mathbb{R}^n$.

Proof of second-order differential criterion of strong convexity

The target inequality is trivial when $y = \mathbf{0}_n$, hence we assume $y \neq \mathbf{0}_n$.

Necessity: Assume initially that x is an interior point of X. Then $x + \alpha y \in X$ for all $y \in \mathbb{R}^n$ and sufficiently small α. Since $f(x)$ is twice differentiable,

$$f(x + \alpha y) = f(x) + \alpha \langle \nabla f(x), y \rangle + \frac{\alpha^2}{2} \langle y, \nabla^2 f(x) y \rangle + o(\alpha^2).$$

Proof of second-order differential criterion of strong convexity

The target inequality is trivial when $y = \mathbf{0}_n$, hence we assume $y \neq \mathbf{0}_n$.

Necessity: Assume initially that x is an interior point of X. Then $x + \alpha y \in X$ for all $y \in \mathbb{R}^n$ and sufficiently small α. Since $f(x)$ is twice differentiable,

$$f(x + \alpha y) = f(x) + \alpha \langle \nabla f(x), y \rangle + \frac{\alpha^2}{2} \langle y, \nabla^2 f(x) y \rangle + o(\alpha^2).$$

Based on the first-order criterion of strong convexity, we have

$$\frac{\alpha^2}{2} \langle y, \nabla^2 f(x) y \rangle + o(\alpha^2) = f(x + \alpha y) - f(x) - \alpha \langle \nabla f(x), y \rangle \geq \frac{\mu}{2} \alpha^2 \|y\|^2.$$

This inequality reduces to the target inequality after dividing both sides by α^2 and taking the limit as $\alpha \downarrow 0$.

If $x \in X$ but $x \notin \text{int}X$, consider a sequence $\{x_k\}$ such that $x_k \in \text{int}X$ and $x_k \to x$ as $k \to \infty$. Then, we arrive at the target inequality after taking the limit.

Proof of second-order differential criterion of strong convexity

Sufficiency: Using Taylor's formula with the Lagrange remainder and the target inequality, we obtain for $x + y \in X$:

$$f(x+y) - f(x) - \langle \nabla f(x), y \rangle = \frac{1}{2} \langle y, \nabla^2 f(x + \alpha y) y \rangle \geq \frac{\mu}{2} \|y\|^2,$$

where $0 \leq \alpha \leq 1$. Therefore,

Proof of second-order differential criterion of strong convexity

Sufficiency: Using Taylor's formula with the Lagrange remainder and the target inequality, we obtain for $x + y \in X$:

$$f(x+y) - f(x) - \langle \nabla f(x), y \rangle = \frac{1}{2} \langle y, \nabla^2 f(x + \alpha y) y \rangle \geq \frac{\mu}{2} \|y\|^2,$$

where $0 \leq \alpha \leq 1$. Therefore,

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

Consequently, by the first-order criterion of strong convexity, the function $f(x)$ is strongly convex with a constant μ. It is important to mention, that $\mu = 0$ stands for the convex case and corresponding differential criterion.

Convex and concave function

Example

Show, that $f(x) = c^\top x + b$ is convex and concave.

Simplest strongly convex function

Example

Show, that $f(x) = x^\top Ax$, where $A \succeq 0$ - is convex on $\mathbb{R}^n$. Is it strongly convex?

Convexity and continuity

Let $f(x)$ - be a convex function on a convex set $S \subseteq \mathbb{R}^n$. Then $f(x)$ is continuous $\forall x \in$ **ri**(S). [1]

Proper convex function

Function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be **proper convex function** if it never takes on the value $-\infty$ and not identically equal to ∞.

Indicator function

$$\delta_S(x) = \begin{cases} \infty, & x \in S, \\ 0, & x \notin S, \end{cases}$$

is a proper convex function.

Convexity and continuity

Let $f(x)$ - be a convex function on a convex set $S \subseteq \mathbb{R}^n$. Then $f(x)$ is continuous $\forall x \in \textbf{ri}(S)$. [1]

Proper convex function

Function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be **proper convex function** if it never takes on the value $-\infty$ and not identically equal to ∞.

Indicator function

$$\delta_S(x) = \begin{cases} \infty, & x \in S, \\ 0, & x \notin S, \end{cases}$$

is a proper convex function.

Closed function

Function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be **closed** if for each $\alpha \in \mathbb{R}$, the sublevel set is closed.
Equivalently, if the epigraph is closed, then the function f is closed.

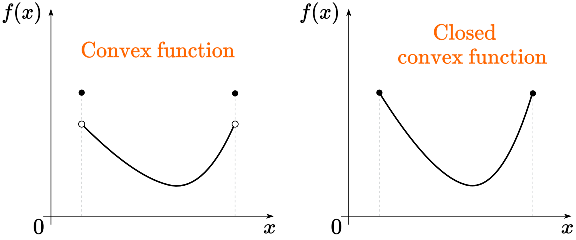


Figure 16: The concept of a closed function is introduced to avoid such breaches at the border.

Facts about convexity

- $f(x)$ is called (strictly, strongly) concave if the function $-f(x)$ - is (strictly, strongly) convex.
- Jensen's inequality for the convex functions:

$$f\left(\sum_{i=1}^n \alpha_i x_i\right) \leq \sum_{i=1}^n \alpha_i f(x_i)$$

for $\alpha_i \geq 0; \quad \sum\limits_{i=1}^n \alpha_i = 1$ (probability simplex)

For the infinite dimension case:

$$f\left(\int\limits_S xp(x)dx\right) \leq \int\limits_S f(x)p(x)dx$$

If the integrals exist and $p(x) \geq 0, \quad \int\limits_S p(x)dx = 1$.

- If the function $f(x)$ and the set S are convex, then any local minimum $x^* = \arg\min\limits_{x \in S} f(x)$ will be the global one. Strong convexity guarantees the uniqueness of the solution.

Other forms of convexity

- Log-convexity: $\log f$ is convex; Log convexity implies convexity.
- Log-concavity: $\log f$ concave; **not** closed under addition!
- Exponential convexity: $[f(x_i + x_j)] \succeq 0$, for $x_1, \ldots, x_n$
- Operator convexity: $f(\lambda X + (1-\lambda)Y)$
- Quasiconvexity: $f(\lambda x + (1-\lambda)y) \leq \max\{f(x), f(y)\}$
- Pseudoconvexity: $\langle \nabla f(y), x - y\rangle \geq 0 \longrightarrow f(x) \geq f(y)$
- Discrete convexity: $f : \mathbb{Z}^n \to \mathbb{Z}$; “convexity + matroid theory.”

Polyak- Lojasiewicz condition. Linear convergence of gradient descent without convexity

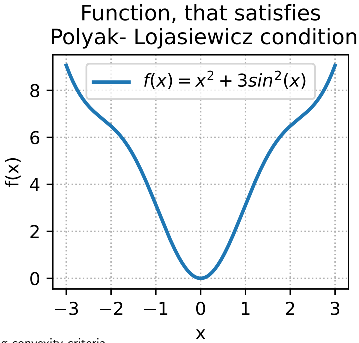
PL inequality holds if the following condition is satisfied for some $\mu > 0$,

$$\|\nabla f(x)\|^2 \geq 2\mu(f(x) - f^*) \forall x$$

It is interesting, that the Gradient Descent algorithm has

The following functions satisfy the PL condition but are not convex. Link to the code

$$f(x) = x^2 + 3\sin^2(x)$$



Polyak- Lojasiewicz condition. Linear convergence of gradient descent without convexity

PL inequality holds if the following condition is satisfied for some $\mu > 0$,

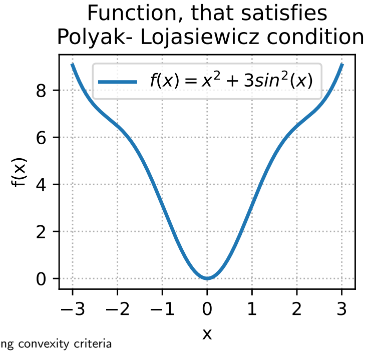
$$\|\nabla f(x)\|^2 \geq 2\mu(f(x) - f^*) \forall x$$

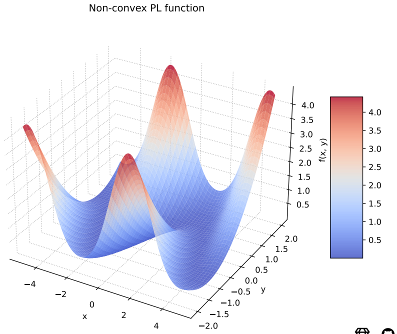
It is interesting, that the Gradient Descent algorithm has

The following functions satisfy the PL condition but are not convex. Link to the code

$$f(x) = x^2 + 3\sin^2(x)$$

$$f(x,y) = \frac{(y - \sin x)^2}{2}$$





Convexity in ML

Linear Least Squares aka Linear Regression

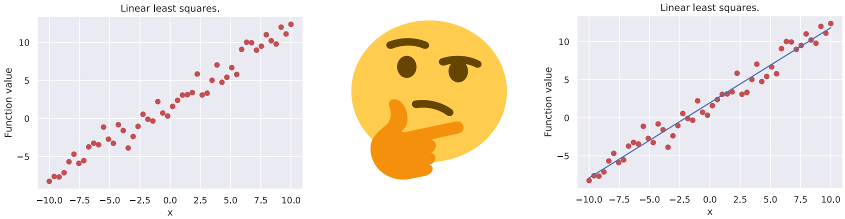


Figure 19: Illustration

In a least-squares, or linear regression, problem, we have measurements $X \in \mathbb{R}^{m \times n}$ and $y \in \mathbb{R}^m$ and seek a vector $\theta \in \mathbb{R}^n$ such that $X\theta$ is close to y. Closeness is defined as the sum of the squared differences:

$$\sum_{i=1}^{m} (x_i^\top \theta - y_i)^2 = \|X\theta - y\|_2^2 \to \min_{\theta \in \mathbb{R}^n}$$

For example, we might have a dataset of m users, each represented by n features. Each row $x_i^\top$ of X is the features for user i, while the corresponding entry y_i of y is the measurement we want to predict from $x_i^\top$, such as ad spending. The prediction is given by $x_i^\top \theta$.

Linear Least Squares aka Linear Regression [2]

1. Is this problem convex? Strongly convex?

Linear Least Squares aka Linear Regression [2]

1. Is this problem convex? Strongly convex?
2. What do you think about the convergence of Gradient Descent for this problem?

[2] Take a look at the 🐍 example of real-world data linear least squares problem

l_2-regularized Linear Least Squares

In the underdetermined case, it is often desirable to restore the strong convexity of the objective function by adding an l_2-penality, also known as Tikhonov regularization, l_2-regularization, or weight decay.

$$\|X\theta - y\|_2^2 + \frac{\mu}{2}\|\theta\|_2^2 \to \min_{\theta \in \mathbb{R}^n}$$

Note: With this modification, the objective is μ-strongly convex again.

Take a look at the 🐍code

Most important difference between convexity and strong convexity

$$f(x) = \frac{1}{2m}\|Ax - b\|_2^2 + \frac{\mu}{2}\|x\|_2^2 \to \min_{x \in \mathbb{R}^n}, \quad A \in \mathbb{R}^{m \times n}, b \in \mathbb{R}^m$$

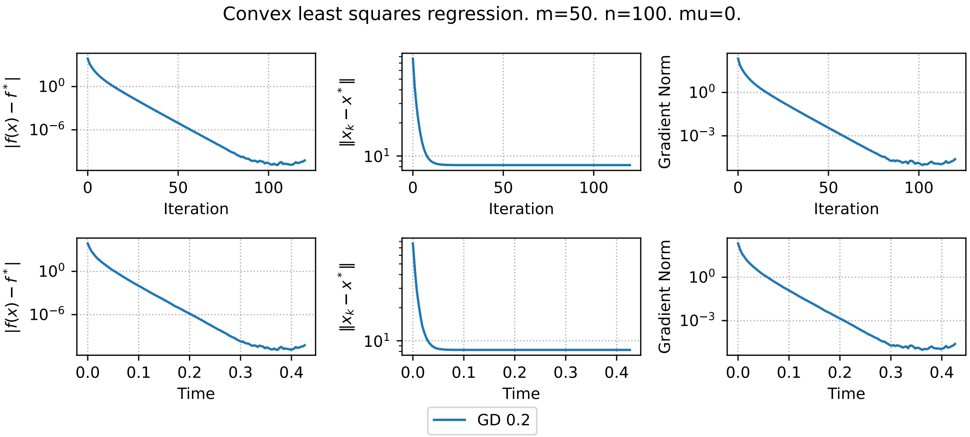


Figure 20: Convex problem does not have convergence in domain

Most important difference between convexity and strong convexity

$$f(x) = \frac{1}{2m}\|Ax - b\|_2^2 + \frac{\mu}{2}\|x\|_2^2 \to \min_{x \in \mathbb{R}^n}, \quad A \in \mathbb{R}^{m \times n}, b \in \mathbb{R}^m$$

Strongly convex least squares regression. m=50. n=100. mu=0.1.

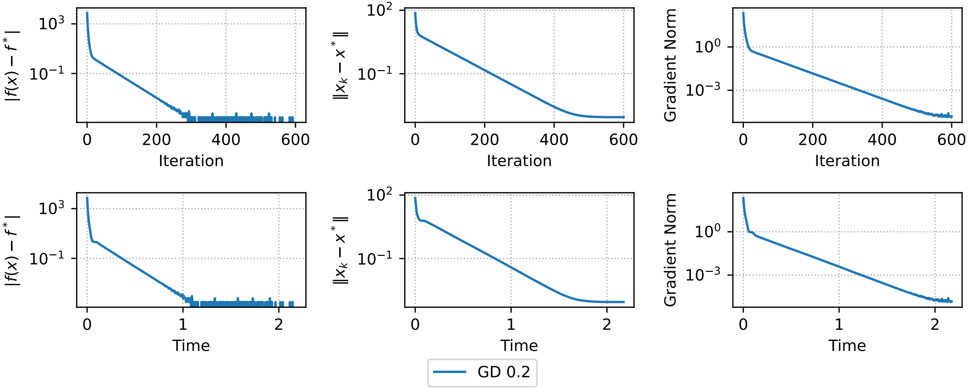


Figure 21: But if you add even small amount of regularization, you will ensure convergence in domain

Most important difference between convexity and strong convexity

$$f(x) = \frac{1}{2m}\|Ax - b\|_2^2 + \frac{\mu}{2}\|x\|_2^2 \to \min_{x \in \mathbb{R}^n}, \quad A \in \mathbb{R}^{m \times n}, b \in \mathbb{R}^m$$

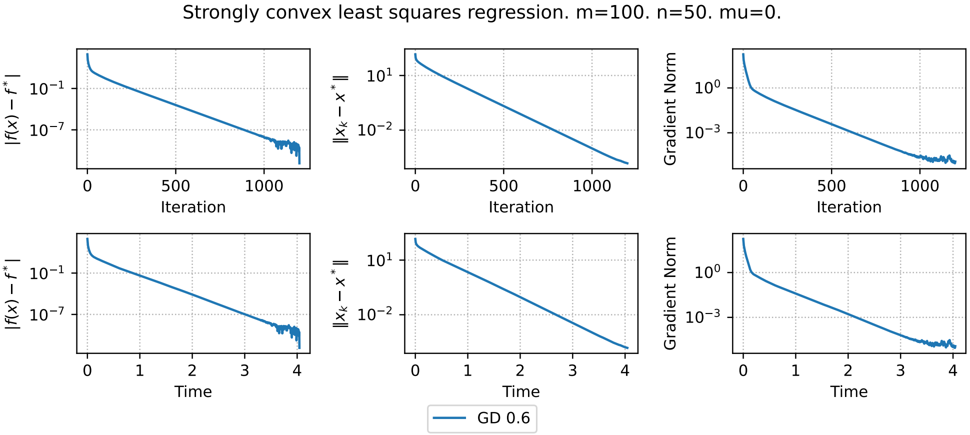


Figure 22: Another way to ensure convergence in the previous problem is to switch the dimension values

You have to have strong convexity (or PL) to ensure convergence with a high precision

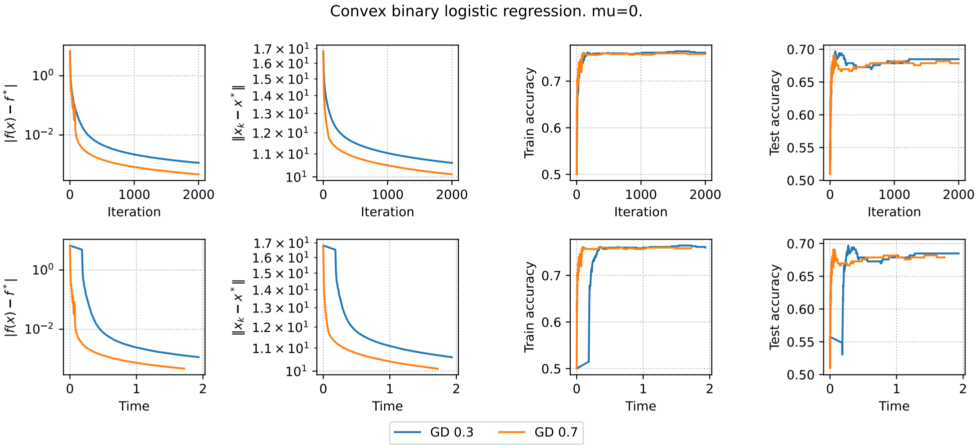


Figure 23: Only small precision is achievable with sublinear convergence

You have to have strong convexity (or PL) to ensure convergence with a high precision

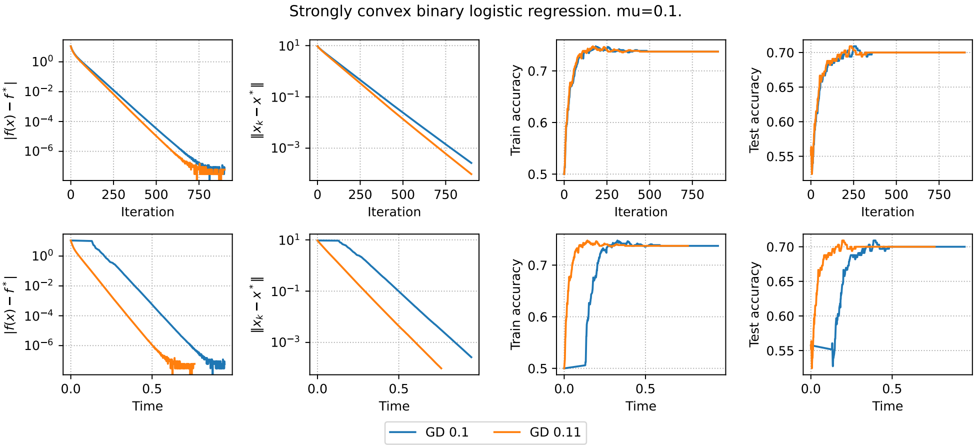


Figure 24: Strong convexity ensures linear convergence

Any local minimum is a global minimum for Deep Linear Networks [3]

We consider the following optimization problem:

$$\min_{W_1,\dots,W_L} L(W_1,\dots,W_L) = \frac{1}{2}\|W_L W_{L-1} \cdots W_1 X - Y\|_F^2,$$

where

$X \in \mathbb{R}^{d_x \times n}$ is the data/input matrix,

$Y \in \mathbb{R}^{d_y \times n}$ is the "label"/output matrix.

ℹ Theorem

Let $k = \min(d_x, d_y)$ be the "width" of the network, and define

$$V = \{(W_1,\dots,W_L) \mid \operatorname{rank}(\Pi_i W_i) = k\}.$$

Then, every critical point of $L(W)$ in V is a global minimum, while every critical point in the complement V^c is a saddle point.

[3] Global optimality conditions for deep neural networks