

Strong convexity criteria. Optimality conditions. Lagrange function.

Daniil Merkulov

Optimization for ML. Faculty of Computer Science. HSE University



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 1: 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 $\forall x \in \text{int}(S) \neq \emptyset$:

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

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

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

Tools for discovering convexity

- Definition (Jensen's inequality)

Tools for discovering convexity

- Definition (Jensen's inequality)
- Differential criteria of convexity

Tools for discovering convexity

- Definition (Jensen's inequality)
- Differential criteria of convexity
- Operations, that preserve convexity

Tools for discovering convexity

- Definition (Jensen's inequality)
- Differential criteria of convexity
- Operations, that preserve convexity
- Connection with epigraph

The function is convex if and only if its epigraph is a convex set.

Tools for discovering convexity

- Definition (Jensen's inequality)
- Differential criteria of convexity
- Operations, that preserve convexity
- Connection with epigraph

The function is convex if and only if its epigraph is a convex set.

- Connection with sublevel set

If $f(x)$ - is a convex function defined on the convex set $S \subseteq \mathbb{R}^n$, then for any β sublevel set \mathcal{L}_β is convex.

The function $f(x)$ defined on the convex set $S \subseteq \mathbb{R}^n$ is closed if and only if for any β sublevel set \mathcal{L}_β is closed.

Tools for discovering convexity

- Definition (Jensen's inequality)
- Differential criteria of convexity
- Operations, that preserve convexity
- Connection with epigraph

The function is convex if and only if its epigraph is a convex set.

- Connection with sublevel set

If $f(x)$ - is a convex function defined on the convex set $S \subseteq \mathbb{R}^n$, then for any β sublevel set \mathcal{L}_β is convex.

The function $f(x)$ defined on the convex set $S \subseteq \mathbb{R}^n$ is closed if and only if for any β sublevel set \mathcal{L}_β is closed.

- Reduction to a line

$f : S \rightarrow \mathbb{R}$ 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.

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



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 3: 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,

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

or equivalently,

$$\begin{aligned} f(x) - f(x_0) - \frac{\mu}{2}(1 - \lambda)\|x - x_0\|^2 &\geq \frac{1}{\lambda}[f(\lambda x + (1 - \lambda)x_0) - f(x_0)] = \\ &= \frac{1}{\lambda}[f(x_0 + \lambda(x - x_0)) - f(x_0)] = \frac{1}{\lambda}[\lambda\langle \nabla f(x_0), x - x_0 \rangle + o(\lambda)] = \\ &= \langle \nabla f(x_0), x - x_0 \rangle + \frac{o(\lambda)}{\lambda}. \end{aligned}$$

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:

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

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

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

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

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 $\forall x \in \text{int}(S) \neq \emptyset$:

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

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 $\forall x \in \text{int}(S) \neq \emptyset$:

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

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 $\text{int}X \neq \emptyset$. 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,

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

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 $x_k \rightarrow x$ as $k \rightarrow \infty$. 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,

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

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 $\forall x \in \text{ri}(S)$.^a

Proper convex function

Function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ 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.

^aPlease, read here about difference between interior and relative interior.

Convexity and continuity

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

Proper convex function

Function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ 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.

^aPlease, read here about difference between interior and relative interior.

Closed function

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



Figure 4: 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.

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 $\alpha_i \geq 0$; $\sum_{i=1}^n \alpha_i = 1$ (probability simplex)

For the infinite dimension case:

$$f\left(\int_S x p(x) dx\right) \leq \int_S f(x) p(x) dx$$

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

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 $\alpha_i \geq 0$; $\sum_{i=1}^n \alpha_i = 1$ (probability simplex)

For the infinite dimension case:

$$f\left(\int_S x p(x) dx\right) \leq \int_S f(x) p(x) dx$$

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

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

Operations that preserve convexity

- Non-negative sum of the convex functions:

$$\alpha f(x) + \beta g(x), (\alpha \geq 0, \beta \geq 0).$$



Figure 5: Pointwise maximum (supremum) of convex functions is convex

Operations that preserve convexity

- 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.



Figure 5: Pointwise maximum (supremum) of convex functions is convex

Operations that preserve convexity

- 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.
- 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.

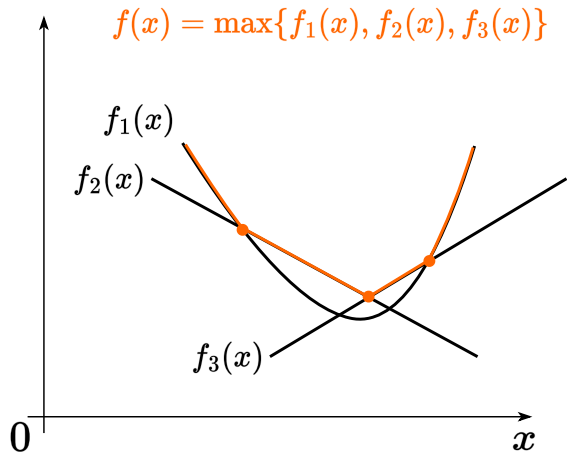


Figure 5: Pointwise maximum (supremum) of convex functions is convex

Operations that preserve convexity

- 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.
- 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.
- If $f(x, y)$ is convex on x for any $y \in Y$:
 $g(x) = \sup_{y \in Y} f(x, y)$ is convex.



Figure 5: Pointwise maximum (supremum) of convex functions is convex

Operations that preserve convexity

- 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.
- 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.
- If $f(x, y)$ is convex on x for any $y \in Y$:
 $g(x) = \sup_{y \in Y} f(x, y)$ is convex.
- If $f(x)$ is convex on S , then $g(x, t) = tf(x/t)$ - is convex with $x/t \in S, t > 0$.



Figure 5: Pointwise maximum (supremum) of convex functions is convex

Operations that preserve convexity

- 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.
- 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.
- If $f(x, y)$ is convex on x for any $y \in Y$:
 $g(x) = \sup_{y \in Y} f(x, y)$ 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 $f_1 : S_1 \rightarrow \mathbb{R}$ and $f_2 : S_2 \rightarrow \mathbb{R}$, where $\text{range}(f_1) \subseteq S_2$. If f_1 and f_2 are convex, and f_2 is increasing, then $f_2 \circ f_1$ is convex on S_1 .



Figure 5: Pointwise maximum (supremum) of convex functions is convex

Maximum eigenvalue of a matrix is a convex function

Example

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

Other forms of convexity

- Log-convexity: $\log f$ is convex; Log convexity implies convexity.

Other forms of convexity

- Log-convexity: $\log f$ is convex; Log convexity implies convexity.
- Log-concavity: $\log f$ concave; **not** closed under addition!

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, \dots, x_n

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, \dots, x_n
- Operator convexity: $f(\lambda X + (1 - \lambda)Y)$

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, \dots, x_n
- Operator convexity: $f(\lambda X + (1 - \lambda)Y)$
- Quasiconvexity: $f(\lambda x + (1 - \lambda)y) \leq \max\{f(x), f(y)\}$

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, \dots, x_n
- Operator convexity: $f(\lambda X + (1 - \lambda)Y)$
- Quasiconvexity: $f(\lambda x + (1 - \lambda)y) \leq \max\{f(x), f(y)\}$
- Pseudoconvexity: $\langle \nabla f(y), x - y \rangle \geq 0 \longrightarrow f(x) \geq f(y)$

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, \dots, 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: $f : \mathbb{Z}^n \rightarrow \mathbb{Z}$; “convexity + matroid theory.”

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

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

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

It is interesting, that 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 $\mu > 0$,

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

It is interesting, that Gradient Descent algorithm has

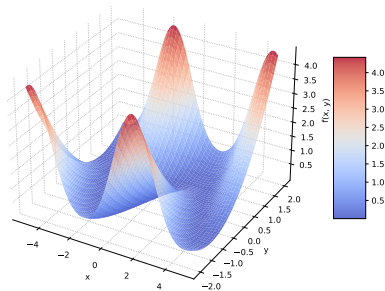
The following functions satisfy the PL-condition, but are not convex. [🔗Link to the code](#)

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



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

Non-convex PL function



Convex optimization problem



Note, that there is an agreement in notation of mathematical programming. The problems of the following type are called **Convex optimization problem**:

$$\begin{aligned} f_0(x) &\rightarrow \min_{x \in \mathbb{R}^n} \\ \text{s.t. } f_i(x) &\leq 0, \quad i = 1, \dots, m \\ Ax &= b, \end{aligned} \quad (\text{COP})$$

where all the functions $f_0(x), f_1(x), \dots, f_m(x)$ are convex and all the equality constraints are affine. It sounds a bit strange, but not all convex problems are convex optimization problems.

$$f_0(x) \rightarrow \min_{x \in S}, \quad (\text{CP})$$

where $f_0(x)$ is a convex function, defined on the convex set S . The necessity of affine equality constraint is essential.

Figure 8: The idea behind the definition of a convex optimization problem

Linear Least Squares aka Linear Regression



Figure 9: 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:

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

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 ¹

1. Is this problem convex? Strongly convex?

Linear Least Squares aka Linear Regression ¹

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


¹Take 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 strong convexity of the objective function by adding an l_2 -penalty, also known as Tikhonov regularization, l_2 -regularization, or weight decay.

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

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

Take a look at the code

Neural networks?

Background

$$f(x) \rightarrow \min_{x \in S}$$

- A set S is usually called a **budget set**.

We say that the problem has a solution if the budget set **is not empty**: $x^* \in S$, in which the minimum or the infimum of the given function is achieved.

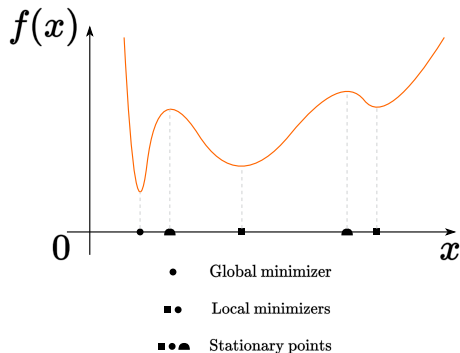


Figure 10: Illustration of different stationary (critical) points

Background

$$f(x) \rightarrow \min_{x \in S}$$

- A set S is usually called a **budget set**.

We say that the problem has a solution if the budget set is **not empty**: $x^* \in S$, in which the minimum or the infimum of the given function is achieved.

- A point x^* is a **global minimizer** if $f(x^*) \leq f(x)$ for all x .

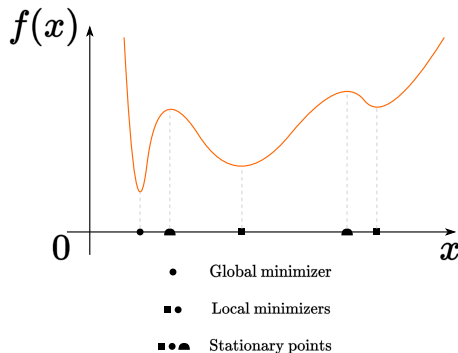


Figure 10: Illustration of different stationary (critical) points

Background

$$f(x) \rightarrow \min_{x \in S}$$

- A set S is usually called a **budget set**.

We say that the problem has a solution if the budget set is **not empty**: $x^* \in S$, in which the minimum or the infimum of the given function is achieved.

- A point x^* is a **global minimizer** if $f(x^*) \leq f(x)$ for all x .
- A point x^* is a **local minimizer** if there exists a neighborhood N of x^* such that $f(x^*) \leq f(x)$ for all $x \in N$.

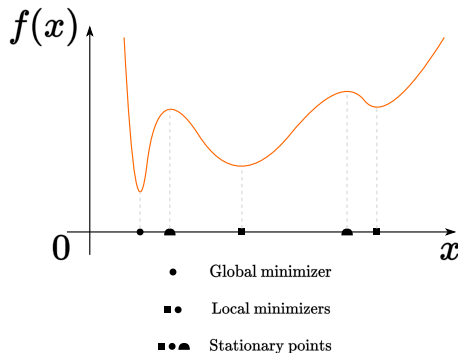


Figure 10: Illustration of different stationary (critical) points

Background

$$f(x) \rightarrow \min_{x \in S}$$

- A set S is usually called a **budget set**.

We say that the problem has a solution if the budget set is **not empty**: $x^* \in S$, in which the minimum or the infimum of the given function is achieved.

- A point x^* is a **global minimizer** if $f(x^*) \leq f(x)$ for all x .
- A point x^* is a **local minimizer** if there exists a neighborhood N of x^* such that $f(x^*) \leq f(x)$ for all $x \in N$.
- A point x^* is a **strict local minimizer** (also called a **strong local minimizer**) if there exists a neighborhood N of x^* such that $f(x^*) < f(x)$ for all $x \in N$ with $x \neq x^*$.

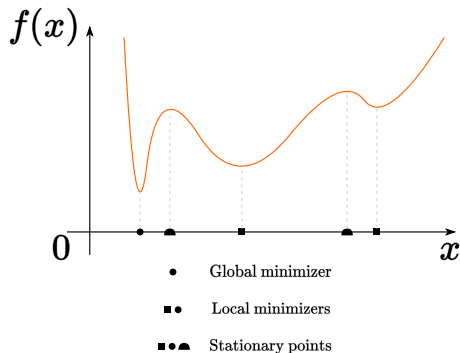


Figure 10: Illustration of different stationary (critical) points

Background

$$f(x) \rightarrow \min_{x \in S}$$

- A set S is usually called a **budget set**.

We say that the problem has a solution if the budget set is **not empty**: $x^* \in S$, in which the minimum or the infimum of the given function is achieved.

- A point x^* is a **global minimizer** if $f(x^*) \leq f(x)$ for all x .
- A point x^* is a **local minimizer** if there exists a neighborhood N of x^* such that $f(x^*) \leq f(x)$ for all $x \in N$.
- A point x^* is a **strict local minimizer** (also called a **strong local minimizer**) if there exists a neighborhood N of x^* such that $f(x^*) < f(x)$ for all $x \in N$ with $x \neq x^*$.
- We call x^* a **stationary point** (or critical) if $\nabla f(x^*) = 0$. Any local minimizer must be a stationary point.

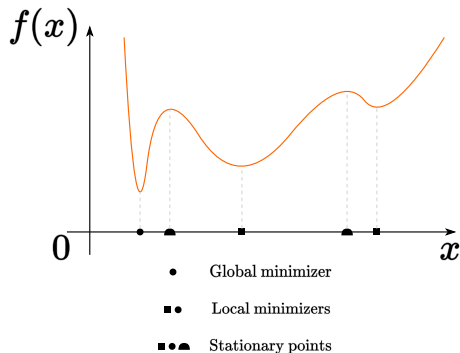


Figure 10: Illustration of different stationary (critical) points

Extreme value (Weierstrass) theorem

Theorem

Let $S \subset \mathbb{R}^n$ be a compact set and $f(x)$ a continuous function on S . So that, the point of the global minimum of the function $f(x)$ on S exists.

Extreme value (Weierstrass) theorem

Theorem

Let $S \subset \mathbb{R}^n$ be a compact set and $f(x)$ a continuous function on S . So that, the point of the global minimum of the function $f(x)$ on S exists.

GOOD NEWS EVERYONE!



Figure 11: A lot of practical problems are theoretically solvable

Extreme value (Weierstrass) theorem

Theorem

Let $S \subset \mathbb{R}^n$ be a compact set and $f(x)$ a continuous function on S . So that, the point of the global minimum of the function $f(x)$ on S exists.

GOOD NEWS EVERYONE!



Figure 11: A lot of practical problems are theoretically solvable

Taylor's Theorem

Suppose that $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is continuously differentiable and that $p \in \mathbb{R}^n$. Then we have:

$$f(x + p) = f(x) + \nabla f(x + tp)^T p \quad \text{for some } t \in (0, 1)$$

Moreover, if f is twice continuously differentiable, we have:

$$\nabla f(x + p) = \nabla f(x) + \int_0^1 \nabla^2 f(x + tp) p dt$$

$$f(x + p) = f(x) + \nabla f(x)^T p + \frac{1}{2} p^T \nabla^2 f(x + tp) p$$

for some $t \in (0, 1)$.

Unconstrained optimization

First-Order Necessary Conditions

If x^* is a local minimizer and f is continuously differentiable in an open neighborhood, then

$$\nabla f(x^*) = 0 \tag{1}$$

Proof

Suppose for contradiction that $\nabla f(x^*) \neq 0$. Define the vector $p = -\nabla f(x^*)$ and note that

$$p^T \nabla f(x^*) = -\|\nabla f(x^*)\|^2 < 0$$

Because ∇f is continuous near x^* , there is a scalar $T > 0$ such that

$$p^T \nabla f(x^* + tp) < 0, \text{ for all } t \in [0, T]$$

For any $\bar{t} \in (0, T]$, we have by Taylor's theorem that

$$f(x^* + \bar{t}p) = f(x^*) + \bar{t}p^T \nabla f(x^* + tp), \text{ for some } t \in (0, \bar{t})$$

Therefore, $f(x^* + \bar{t}p) < f(x^*)$ for all $\bar{t} \in (0, T]$. We have found a direction from x^* along which f decreases, so x^* is not a local minimizer, leading to a contradiction.

Second-Order Sufficient Conditions

Theorem

Suppose that $\nabla^2 f$ is continuous in an open neighborhood of x^* and that

$$\nabla f(x^*) = 0 \quad \nabla^2 f(x^*) \succ 0.$$

Then x^* is a strict local minimizer of f .

Proof Because the Hessian is continuous and positive definite at x^* , we can choose a radius $r > 0$ such that $\nabla^2 f(x)$ remains positive definite for all x in the open ball $B = \{z \mid \|z - x^*\| < r\}$. Taking any nonzero vector p with $\|p\| < r$, we have $x^* + p \in B$ and so

$$\begin{aligned} f(x^* + p) &= f(x^*) + p^T \nabla f(x^*) + \frac{1}{2} p^T \nabla^2 f(z) p \\ &= f(x^*) + \frac{1}{2} p^T \nabla^2 f(z) p \end{aligned}$$

where $z = x^* + tp$ for some $t \in (0, 1)$. Since $z \in B$, we have $p^T \nabla^2 f(z) p > 0$, and therefore $f(x^* + p) > f(x^*)$, giving the result.

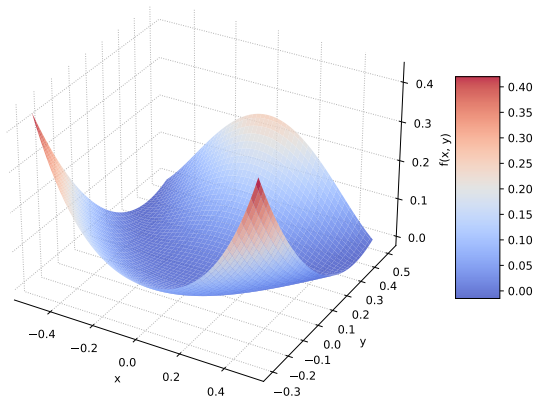
Peano counterexample

Note, that if $\nabla f(x^*) = 0, \nabla^2 f(x^*) \succeq 0$, i.e. the hessian is positive *semidefinite*, we cannot be sure if x^* is a local minimum.

$$f(x, y) = (2x^2 - y)(x^2 - y)$$

One can verify, that $\nabla^2 f(0, 0) = 0$ and $\nabla f(0, 0) = (0, 0)$, but $(0, 0)$ is not a local minimizer. Although the surface does not have a local minimizer at the origin, its intersection with any vertical plane through the origin (a plane with equation $y = mx$ or $x = 0$) is a curve that has a local minimum at the origin. In other words, if a point starts at the origin $(0, 0)$ of the plane, and moves away from the origin along any straight line, the value of $(2x^2 - y)(x^2 - y)$ will increase at the start of the motion. Nevertheless, $(0, 0)$ is not a local minimizer of the function, because moving along a parabola such as $y = \sqrt{2}x^2$ will cause the function value to decrease.

Non-convex PL function



General first order local optimality condition

Direction $d \in \mathbb{R}^n$ is a feasible direction at $x^* \in S \subseteq \mathbb{R}^n$ if small steps along d do not take us outside of S .

Consider a set $S \subseteq \mathbb{R}^n$ and a function $f: \mathbb{R}^n \rightarrow \mathbb{R}$. Suppose that $x^* \in S$ is a point of local minimum for f over S , and further assume that f is continuously differentiable around x^* .

- Then for every feasible direction $d \in \mathbb{R}^n$ at x^* it holds that $\nabla f(x^*)^\top d \geq 0$.

$$\nabla f(x^*)^\top (x - x^*) \geq 0, \forall x \in S.$$

$$f(x) = x_1 + x_2 \rightarrow \min_{x_1, x_2 \in \mathbb{R}^2}$$

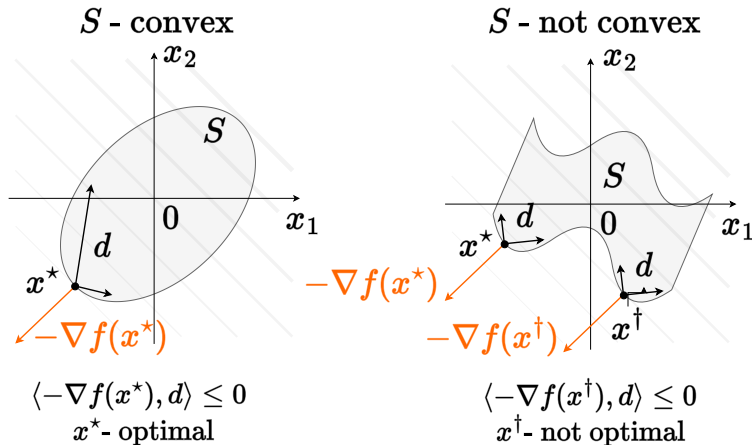


Figure 12: General first order local optimality condition

General first order local optimality condition

Direction $d \in \mathbb{R}^n$ is a feasible direction at $x^* \in S \subseteq \mathbb{R}^n$ if small steps along d do not take us outside of S .

Consider a set $S \subseteq \mathbb{R}^n$ and a function $f: \mathbb{R}^n \rightarrow \mathbb{R}$. Suppose that $x^* \in S$ is a point of local minimum for f over S , and further assume that f is continuously differentiable around x^* .

1. Then for every feasible direction $d \in \mathbb{R}^n$ at x^* it holds that $\nabla f(x^*)^\top d \geq 0$.
2. If, additionally, S is convex then

$$\nabla f(x^*)^\top (x - x^*) \geq 0, \forall x \in S.$$

$$f(x) = x_1 + x_2 \rightarrow \min_{x_1, x_2 \in \mathbb{R}^2}$$

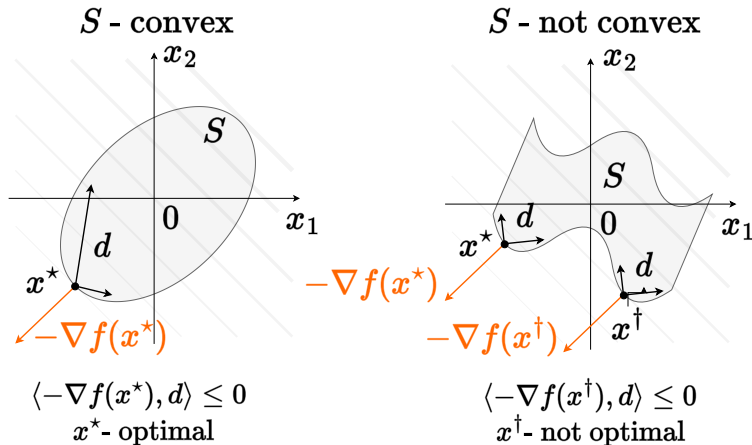


Figure 12: General first order local optimality condition

Convex case

It should be mentioned, that in the **convex** case (i.e., $f(x)$ is convex) necessary condition becomes sufficient.

One more important result for convex unconstrained case sounds as follows. If $f(x) : S \rightarrow \mathbb{R}$ - convex function defined on the convex set S , then:

- Any local minima is the global one.

Convex case

It should be mentioned, that in the **convex** case (i.e., $f(x)$ is convex) necessary condition becomes sufficient.

One more important result for convex unconstrained case sounds as follows. If $f(x) : S \rightarrow \mathbb{R}$ - convex function defined on the convex set S , then:

- Any local minima is the global one.
- The set of the local minimizers S^* is convex.

Convex case

It should be mentioned, that in the **convex** case (i.e., $f(x)$ is convex) necessary condition becomes sufficient.

One more important result for convex unconstrained case sounds as follows. If $f(x) : S \rightarrow \mathbb{R}$ - convex function defined on the convex set S , then:

- Any local minima is the global one.
- The set of the local minimizers S^* is convex.
- If $f(x)$ - strictly or strongly convex function, then S^* contains only one single point $S^* = \{x^*\}$.

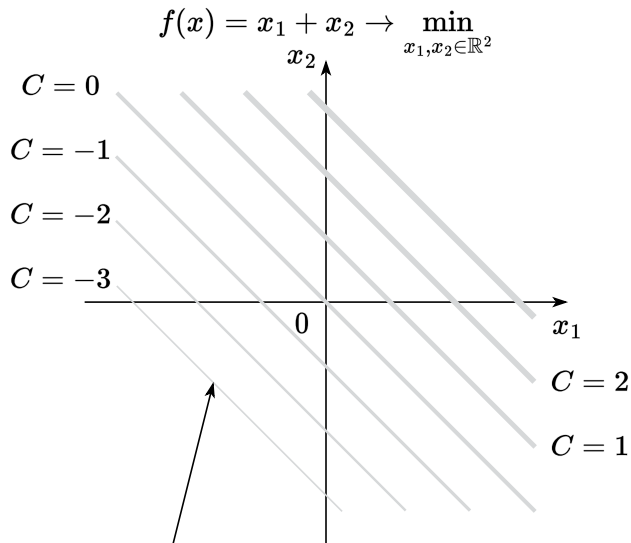
Optimization with equality conditions

Things are pretty simple and intuitive in unconstrained problem. In this section we will add one equality constraint, i.e.

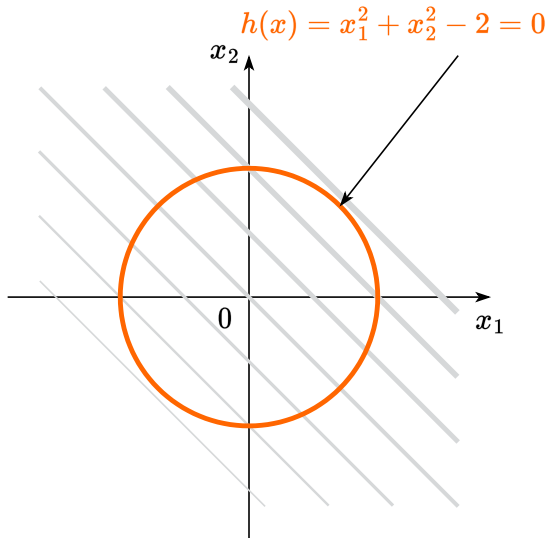
$$\begin{aligned} f(x) &\rightarrow \min_{x \in \mathbb{R}^n} \\ \text{s.t. } h(x) &= 0 \end{aligned}$$

We will try to illustrate approach to solve this problem through the simple example with $f(x) = x_1 + x_2$ and $h(x) = x_1^2 + x_2^2 - 2$.

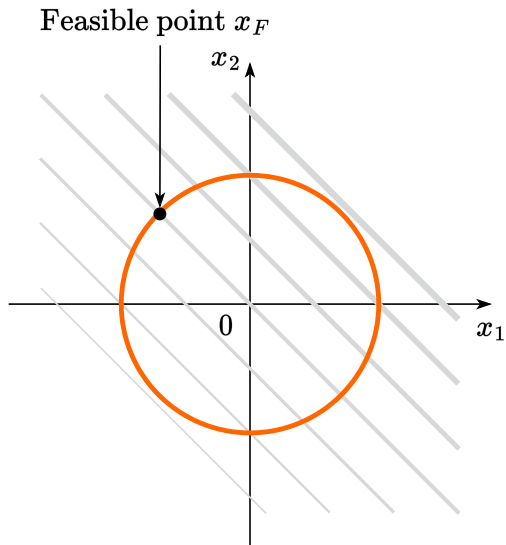
Optimization with equality conditions



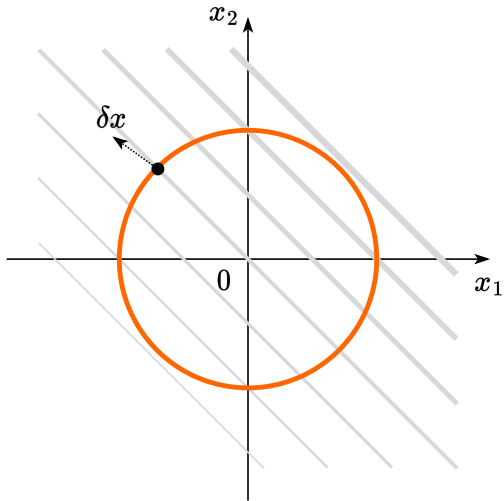
Optimization with equality conditions



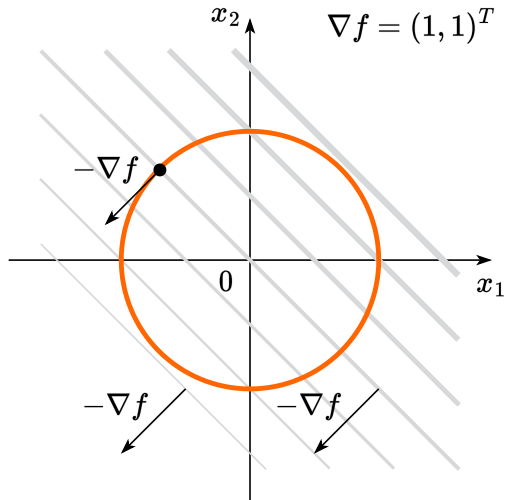
Optimization with equality conditions



Optimization with equality conditions

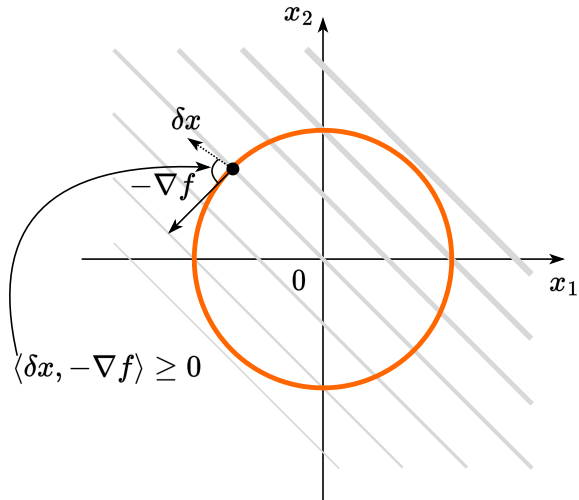


Optimization with equality conditions

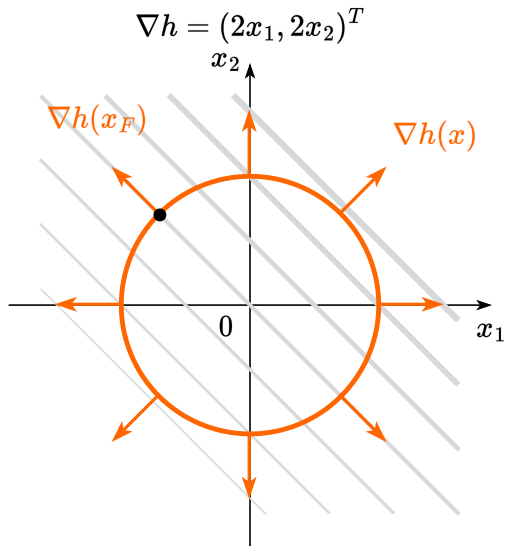


Optimization with equality conditions

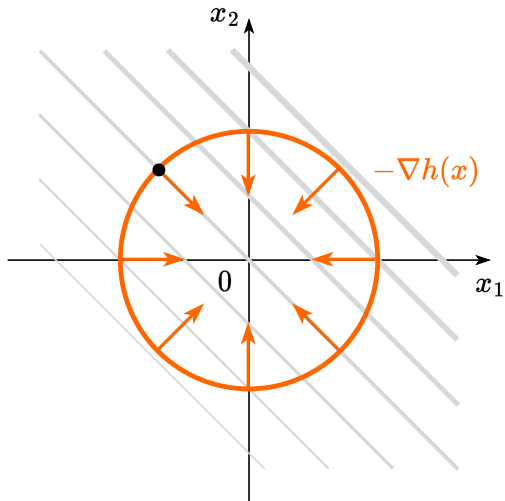
We want: $f(x_F + \delta x) \leq f(x_F)$



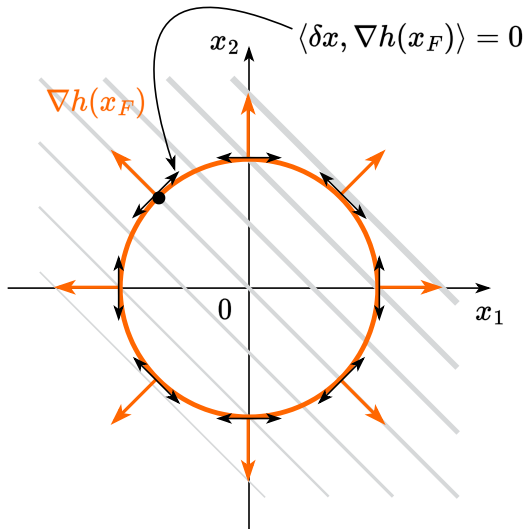
Optimization with equality conditions



Optimization with equality conditions



Optimization with equality conditions



Optimization with equality conditions

Generally: in order to move from x_F along the budget set towards decreasing the function, we need to guarantee two conditions:

$$\langle \delta x, \nabla h(x_F) \rangle = 0$$

$$\langle \delta x, -\nabla f(x_F) \rangle > 0$$

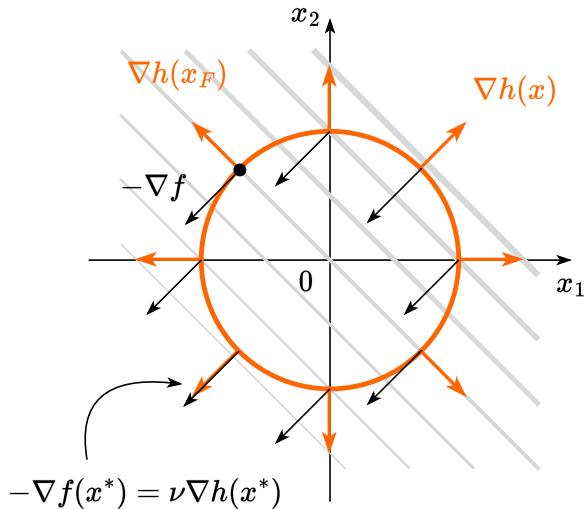
Let's assume, that in the process of such a movement we have come to the point where

$$-\nabla f(x) = \nu \nabla h(x)$$

$$\langle \delta x, -\nabla f(x) \rangle = \langle \delta x, \nu \nabla h(x) \rangle = 0$$

Then we came to the point of the budget set, moving from which it will not be possible to reduce our function. This is the local minimum in the constrained problem :)

Optimization with equality conditions



Lagrangian

So let's define a Lagrange function (just for our convenience):

$$L(x, \nu) = f(x) + \nu h(x)$$

Then if the problem is *regular* (we will define it later) and the point x^* is the local minimum of the problem described above, then there exist ν^* :

Necessary conditions

$$\nabla_x L(x^*, \nu^*) = 0 \text{ that's written above}$$

$$\nabla_\nu L(x^*, \nu^*) = 0 \text{ budget constraint}$$

We should notice that $L(x^*, \nu^*) = f(x^*)$.

Equality constrained problem

$$\begin{aligned} f(x) &\rightarrow \min_{x \in \mathbb{R}^n} \\ \text{s.t. } h_i(x) &= 0, \quad i = 1, \dots, p \end{aligned} \quad (\text{ECP})$$

Solution

$$L(x, \nu) = f(x) + \sum_{i=1}^p \nu_i h_i(x) = f(x) + \nu^\top h(x)$$

Let $f(x)$ and $h_i(x)$ be twice differentiable at the point x^* and continuously differentiable in some neighborhood x^* . The local minimum conditions for $x \in \mathbb{R}^n, \nu \in \mathbb{R}^p$ are written as

ECP: Necessary conditions

$$\nabla_x L(x^*, \nu^*) = 0$$

$$\nabla_\nu L(x^*, \nu^*) = 0$$

Depending on the behavior of the Hessian, the critical points can have a different character.

Linear Least Squares

Example

Pose the optimization problem and solve them for linear system $Ax = b$, $A \in \mathbb{R}^{m \times n}$ for three cases (assuming the matrix is full rank):

- $m < n$

Linear Least Squares

Example

Pose the optimization problem and solve them for linear system $Ax = b$, $A \in \mathbb{R}^{m \times n}$ for three cases (assuming the matrix is full rank):

- $m < n$
- $m = n$

Linear Least Squares

Example

Pose the optimization problem and solve them for linear system $Ax = b$, $A \in \mathbb{R}^{m \times n}$ for three cases (assuming the matrix is full rank):

- $m < n$
- $m = n$
- $m > n$