Statistics and optimization
Formalization and basic properties

Sylvain Sardy

Université de Genève

1. Problem statement

We want to minimize a function

$$\begin{aligned} f: \quad X \subset \mathbb{R}^N &\to \mathbb{R} \\ \mathbf{x} &\to f(\mathbf{x}) \end{aligned}$$

where:

- f is the *cost function*
- X is the *feasible set*.

Notation: we want to solve $\min_{\mathbf{x} \in X} f(\mathbf{x})$.

A solution is

$$\mathbf{x}^* \in \arg\min_{\mathbf{x} \in X} f(\mathbf{x}).$$

If the solution is unique then

$$\mathbf{x}^* = \arg\min_{\mathbf{x} \in X} f(\mathbf{x})$$

Property: maximizing f is equivalent to minimizing $-f$.

Goal: find a

- *global solution* $\mathbf{x}^* \in X$: a point such that

$$f(\mathbf{x}^*) \leq f(\mathbf{x}) \quad \forall \mathbf{x} \in X$$

- or at least a *local solution* $\tilde{x} \in X$: a point such that

$$\exists \epsilon > 0 : \; f(\tilde{\mathbf{x}}) \leq f(\mathbf{x}) \quad \forall \mathbf{x} \in X \quad \text{such that} \quad \|\mathbf{x} - \tilde{\mathbf{x}}^*\| < \epsilon.$$

Optimization problems are classified in several groups:

- discrete (e.g., combinatorial route planning, matching, and integer programming) versus continuous
- univariate versus multivariate
- unconstrained ($X = \mathbb{R}^{\mathrm{N}}$) versus constrained (e.g., $\{\mathbf{x} \in \mathbb{R}^{\mathrm{P}} : \ \|\mathbf{x}\|_1 \leq \mathrm{C}\}$).

  The constraints can be for instance:

  - equality or inequality constraints. Note that:
    - $g_i(\mathbf{x}) = c_i$ is equivalent to two inequality constraints.
    - $g_i(\mathbf{x}) \leq c_i$ is equivalent to $g_i(\mathbf{x}) + s_i = c_i$ and $s_i \geq 0$; this last variable is called a slack variable.
  - linear or non-linear constraints.
  - box constraints if $X = I_1 \times \ldots \times I_N$, where I_n's are closed intervals

- linear programming: f is linear and X is specified by linear inequality constraints.
  The optimal solutions can be found by searching among the finite set of extreme points of the polyhedron X, for instance with the simplex method.
- convex programming: f is convex and X is defined by linear equality constraints and convex inequality constraints of the form $h_j(\mathbf{x}) \leq 0$ with h_j convex for $j = 1, \ldots, J$.
- nonlinear programming
- differentiable versus nondifferentiable optimization

2. Existence

Proposition: (Weierstrass' Theorem)
Let $f : X \subset \mathbb{R}^N \to \mathbb{R}$ be lower semicontinuous in X.
If either one of the following conditions holds:

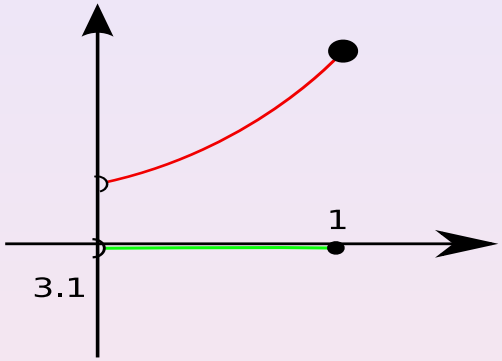
- X is compact
- X is closed and f is coercive
- there exists γ such that the level set

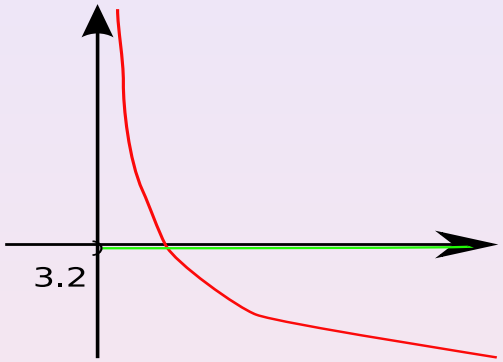
$$\{\mathbf{x} \in X : \ f(\mathbf{x}) \leq \gamma\}$$

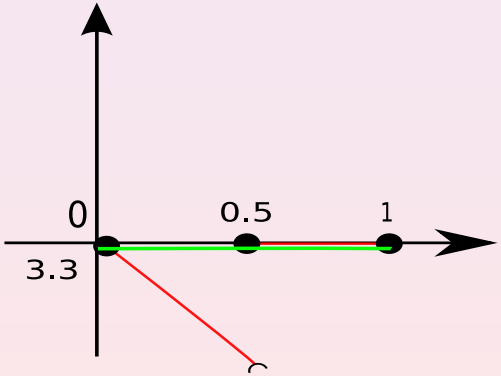
  is nonempty and compact

then there exists $\mathbf{x}^* \in X$ such that $f(\mathbf{x}^*) = \min_{\mathbf{x} \in X} f(\mathbf{x})$.

Counter-example 2.1: $f(x) = \exp(x)$ on $D = (0, 1]$.
Counter-example 2.2: $f(x) = -\log(x)$ on $D = R^{+*}$.
Counter-example 2.3: $f(x) = -x \cdot 1(x < 0.5)$ on $D = [0, 1]$ (upper semicontinuous)







3. Optimality conditions

Necessary optimality conditions
Let $\mathbf{x}^\star$ be an unconstrained local minimum of $f : \mathbb{R}^N \to \mathbb{R}$, and assume that f is continuously differentiable in an open set S containing $\mathbf{x}^\star$. Then

$$\nabla f(\mathbf{x}^\star) = \mathbf{0}. \quad \text{(First order necessary condition)}$$

If moreover f is twice continuously differentiable within S, then

$$\nabla^2 f(\mathbf{x}^\star) \geq 0^\dagger. \quad \text{(Second order necessary condition)}$$

[† positive semidefinite]

Proof.

Propositions for convex case

Let $f : X \to \mathbb{R}$ be convex over the convex set X.

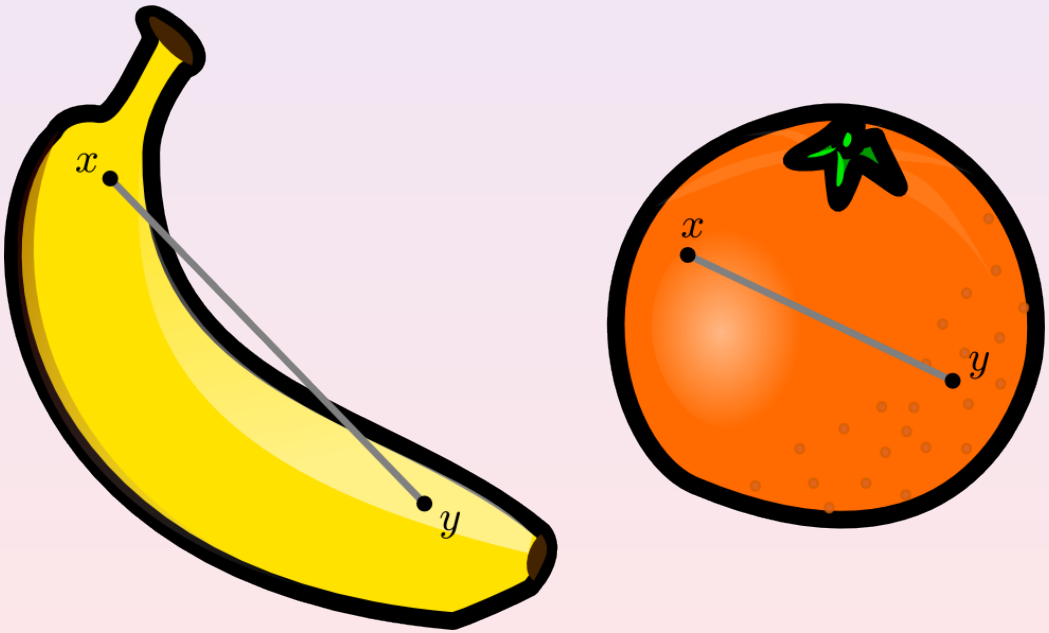
- A **local minimum** of f over X is also a **global minimum** over X.
- If the convexity is strict, then there exists at most one global minimum of f over X.
- If f is continuously differentiable and X is open, then $\nabla f(\mathbf{x}^\star) = \mathbf{0}$ is a **necessary and sufficient condition** for $\mathbf{x}^\star \in X$ to be a global minimum of f over X.

Proof: postponed.

4. Convex set and function

Definition of convex set: A set $C \subset R^N$ is convex if for every $\mathbf{x} \in C$ and $\mathbf{y} \in C$, the line segment joining $\mathbf{x}$ and $\mathbf{y}$ also lies in C, i.e., for every $\lambda \in [0, 1]$, then $\lambda\mathbf{x} + (1 - \lambda)\mathbf{y} \in C$.

Example: a banana is not convex, an orange is.



Property: Let C_i be convex sets.
Then the intersection $\bigcap C_i$ is also convex.

Recall: an *hyperplane* is a set of the form $\{\mathbf{x} \in \mathbb{R}^{\mathrm{N}} : \ \mathbf{a}^{\mathrm{T}}\mathbf{x} = \mathrm{b}\}$ for $\mathbf{a} \neq \mathbf{0}$ and $b \in \mathbb{R}$.
An hyperplane divides $\mathbb{R}^{\mathrm{N}}$ into two *halfspaces*, e.g., the closed halfspace $\{\mathbf{x} \in \mathbb{R}^{\mathrm{N}} : \ \mathbf{a}^{\mathrm{T}}\mathbf{x} \geq \mathrm{b}\}$.
Halfspaces are convex.

Applications in Statistics:

- Total variation density estimation:

$$\min_{\mathbf{f}>\mathbf{0}} -\sum_{n=1}^{N} \log f_n + \lambda \sum_{n=1}^{N-1} |f_{n+1} - f_n| \quad \text{s.t.} \quad \mathbf{w}^T\mathbf{f} = 1.$$

- Poisson GLM for a regression matrix $X_{N\times P}$.
  The set $\{\boldsymbol{\alpha} \in \mathbb{R}^{\mathrm{P}} : \ \boldsymbol{\mu} = \mathrm{X}\boldsymbol{\alpha} > \mathbf{0}\}$ is convex.

Definition

A convex combination of $\mathbf{x}_1, \ldots, \mathbf{x}_N$ is a point of the form

$$\lambda_1 \mathbf{x}_1 + \ldots + \lambda_N \mathbf{x}_N$$

with $\lambda_n \geq 0$ for all $n \in \{1, 2, \ldots, N\}$ and $\sum_{n=1}^{N} \lambda_n = 1$.

Definition

The *convex hull* of a set C is the set of all convex combinations of points in C. Convex hulls are convex.

Example: Poisson GLM with log or identity link for extrapolation.

Property

The image of a convex set under an affine transformation is convex. Same for the inverse image.

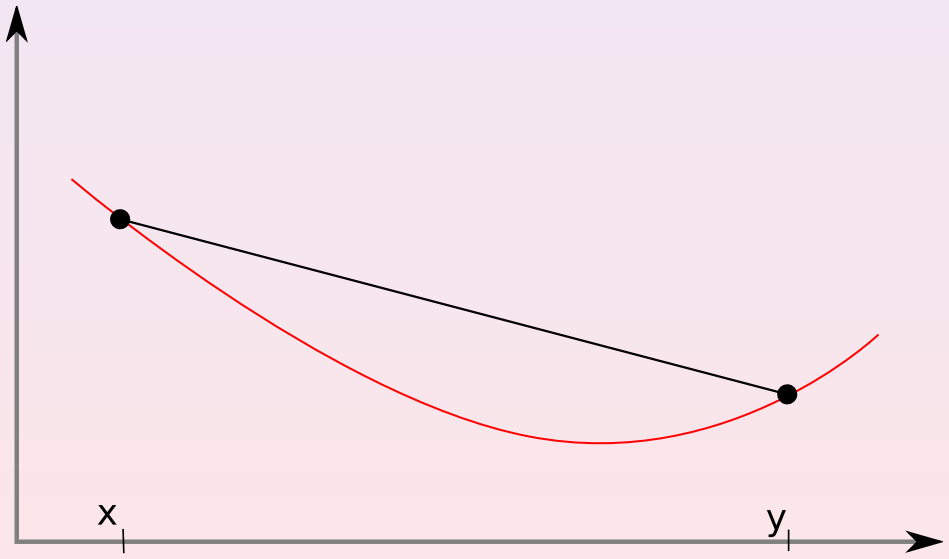
Definition of convex function

A function $f(x)$ defined on a convex set C is:

- strictly convex on C if

$$f(\lambda \mathbf{x} + (1-\lambda)\mathbf{y}) < \lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{y})$$

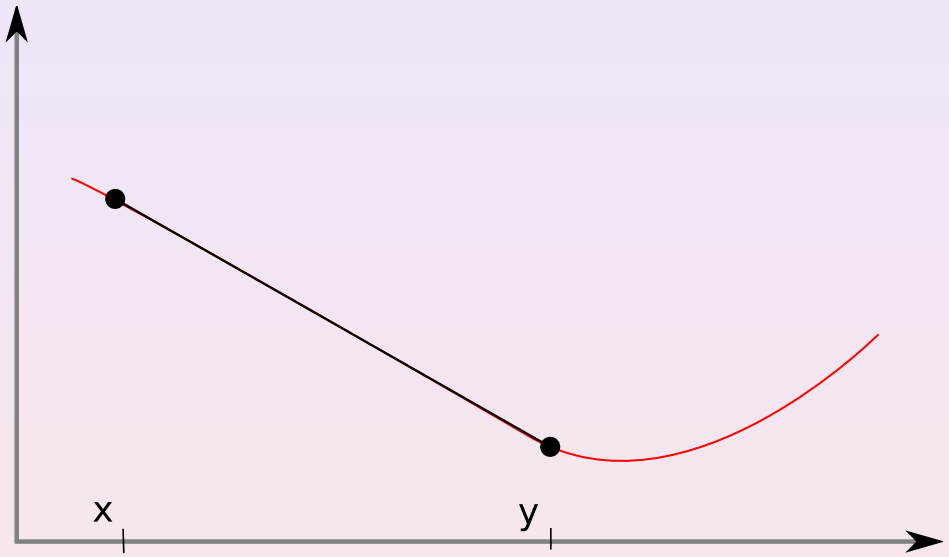
for all $\mathbf{x}$, $\mathbf{y}$ in C with $\mathbf{x} \neq \mathbf{y}$ and all $0 < \lambda < 1$.



- convex on C if

$$f(\lambda \mathbf{x} + (1-\lambda)\mathbf{y}) \leq \lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{y})$$

for all $\mathbf{x}$, $\mathbf{y}$ in C and all $0 \leq \lambda \leq 1$.



Equivalently, a function is convex iff its epigraph

$$\mathrm{epi}(f) = \{(\mathbf{x}, y) : \ \mathbf{x} \in \mathbb{R}^{\mathrm{N}}, \ \mathrm{y} \in \mathbb{R}, \ \mathrm{y} \geq \mathrm{f}(\mathbf{x})\} \subseteq \mathbb{R}^{\mathrm{N}+1}$$

is a convex set.

Convexity in Statistics: many statistical estimators solve

$$\min_{\mathbf{x}} \operatorname{loss}(x; Y) + \lambda \operatorname{pen}(x)$$

Examples:

- Gaussian negative log-likelihood:

$$\operatorname{loss}(x; Y) = \frac{1}{2}(x - Y)^2$$

- ℓ_1 and ℓ_2 penalties:

$$\operatorname{pen}(x) = |x| \quad \text{and} \quad \operatorname{pen}(x) = x^2$$

- Poisson negative log-likelihood:

$$-l(x; Y) = -Y \log x + x$$

- Support vector machine regression loss:

$$\operatorname{loss}(x, Y) = \begin{cases} 0 & |x - Y| \leq \tau \\ |x - Y| - \tau & \text{else} \end{cases}$$

- Support vector machine classification loss:

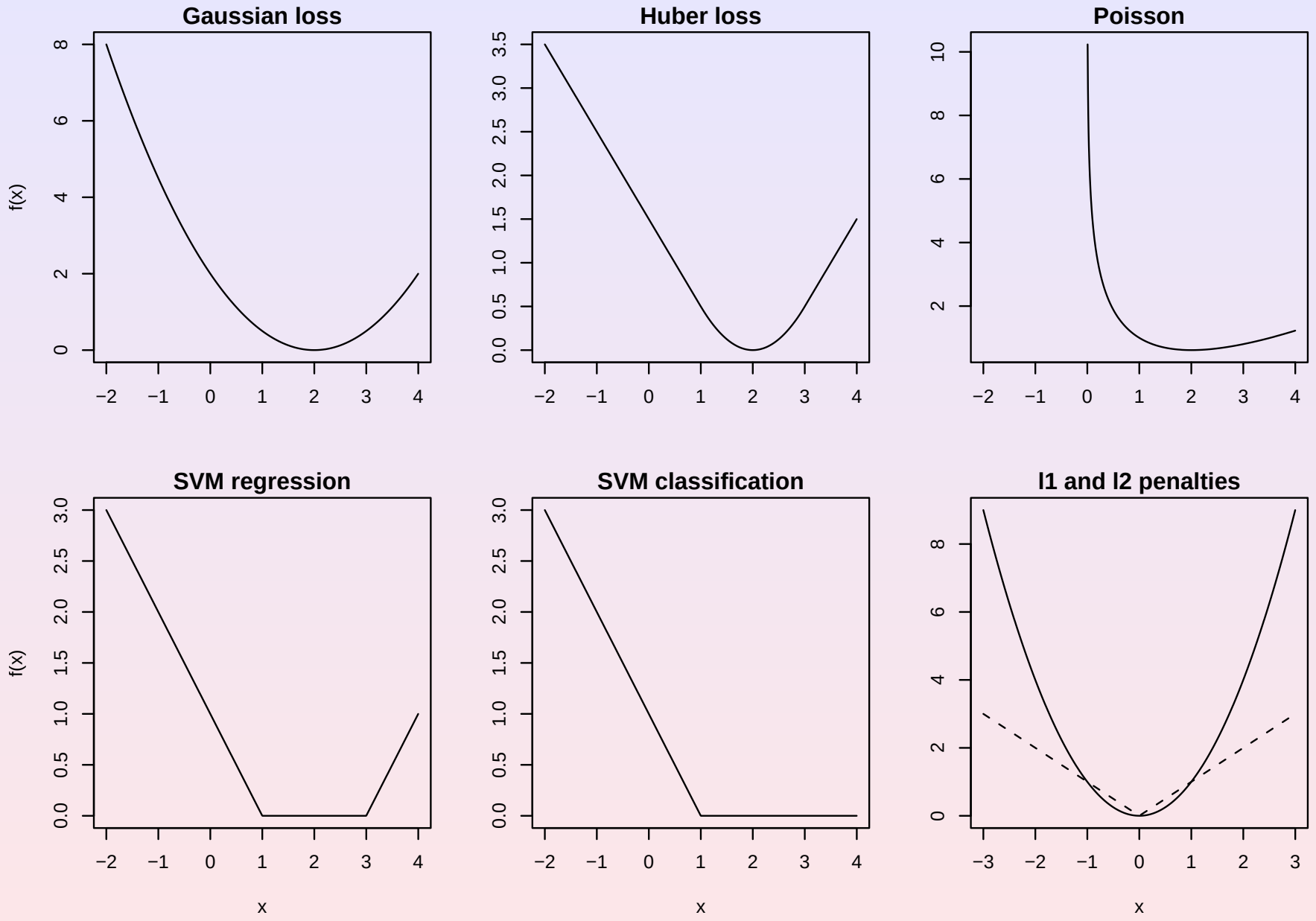
$$\operatorname{loss}(x, Y) = (1 - Yx)_+$$

- Robust regression with Huber loss function

$$\mathrm{loss}(x, Y) = \begin{cases} (x - Y)^2/2 & \text{if } |x - Y| \leq \tau \\ \tau|x - Y| - \tau^2/2 & \text{else} \end{cases}$$

- Total variation penalty for segmentation

$$\mathrm{pen}(\mathbf{x}) = \sum_{n=1}^{N-1} |x_{n+1} - x_n|$$


Gaussian loss
Huber loss
Poisson
SVM regression
SVM classification
l1 and l2 penalties
f(x)
x

Properties:

- a linear function is convex.
- any vector norm is convex.
- the weighted sum of convex functions, with positive weights, is convex. Moreover if at least one of the functions is strictly convex, then the weighted sum is strictly convex.
- if f is convex, then $f(A\mathbf{x} + \mathbf{b})$ is convex. If moreover $\ker(A) = \{\mathbf{0}\}$ and f is strictly convex, then $f(A\mathbf{x} + \mathbf{b})$ is strictly convex.
- a convex function on C open and convex is continuous.

Applications: ridge regression, lasso, total variation, soft-waveshrink, support vector machine, etc...

Proposition: Let $C \in \mathbb{R}^{\mathrm{N}}$ be a convex set and let $f : \mathbb{R}^{\mathrm{N}} \to \mathbb{R}$ be differentiable on C.

- The function f is convex on C iff

$$f(\mathbf{z}) \geq f(\mathbf{x}) + (\mathbf{z} - \mathbf{x})^{\mathrm{T}} \nabla f(\mathbf{x}), \quad \forall \mathbf{x}, \mathbf{z} \in C \tag{1}$$

- If the inequality (1) is strict whenever $\mathbf{x} \neq \mathbf{z}$, then f is strictly convex over C.

Proof.

Propositions Let $C \subset \mathbb{R}^N$ be a convex set, let $f : \mathbb{R}^N \to \mathbb{R}$ be twice continuously differentiable on C, and let Q be a real symmetric $N \times N$ matrix.

- If $\nabla^2 f(\mathbf{x}) \geq 0$ on C, then f is convex on C.
  Proof.
- If $C = \mathbb{R}^N$, then the converse is true.
  Consequence: the quadratic function $f(\mathbf{x}) = \mathbf{x}^T Q \mathbf{x}$ is convex iff Q is positive semidefinite.
- If $\nabla^2 f(\mathbf{x}) > 0$ on C, then f is strictly convex on C.
  Converse is false: take $f(x) = x^4$.
- The quadratic function $f(\mathbf{x}) = \mathbf{x}^T Q \mathbf{x}$ is strictly convex iff Q is positive definite.
  Proof.

We can now prove the Propositions on page 9.

Use of optimality conditions

In the absence of convexity, we have the following sufficiency conditions for local optimality.

Second order Sufficient optimality conditions
Let $f : \mathbb{R}^{\mathrm{N}} \to \mathbb{R}$ be twice continuously differentiable in an open set S. Suppose that a vector $\mathbf{x}^\star \in S$ satisfies

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

Then $\mathbf{x}^\star$ is a strict unconstrained local minimum of f.

In practice, to solve unconstrained optimization:

- 1. Check twice continuous differentiability of cost function.
  2. Find all stationary points.
  3. Check positive semidefiniteness at these points; disregard those where it is not positive semidefinite.
  4. Check positive definiteness at remaining points. Those are strict local minima.
- 1. Check continuous differentiability of cost function.
  2. Check existence of a solution.
  3. Find all stationary points.
  4. Declare as global minimum the one(s) with smallest cost.

Solving

$$\nabla f(\mathbf{x}^\star) = \mathbf{0}$$

is often as difficult as finding $\mathbf{x}^\star$ such that

$$f(\mathbf{x}^\star) \leq f(\mathbf{x}) \quad \forall \mathbf{x} \in \mathbb{R}^n.$$

The condition $\nabla f(\mathbf{x}^\star) = \mathbf{0}$ is used to check optimality.

Iterative algorithms terminate when optimality conditions hold up to an approximation level

$$\text{Stop at } k\text{th iteration when } \|\nabla f(\mathbf{x}^k)\| / \|\nabla f(\mathbf{x}^0)\| < \epsilon,$$

for some small $\epsilon > 0$.