Nonlinear Optimization

Marcos Raydan

Centro de Matemática e Aplicações (CMA)
FCT, Universidade Nova de Lisboa

FCT, UNL
Caparica, Portugal
March – June, 2020

Part I

Prerequisites

Basic Linear Algebra Concepts

$\mathbb{R}^n$ is the vector space of real vectors (columns) of dimension n.

$x \in \mathbb{R}^n$ is a matrix with n rows and one column.

x^T is a matrix with one row and n columns.

Let $A \in \mathbb{R}^{m \times n}$ be a matrix with m rows and n columns ($m \geq n$). The product $y = Ax$ produces y as a linear combination of the columns of A:

$$y = Ax = \sum_{j=1}^{n} x_j A_j$$

Equivalently

$$y_i = \sum_{j=1}^{n} a_{ij} x_j.$$

If $x, y \in \mathbb{R}^n$, $x^T y$ is a scalar; xy^T is a rank-one $n \times n$ matrix.

If $z \in \mathbb{R}^n$, $(xy^T)z = (y^T z)x$.

Basic Linear Algebra Concepts

$range(A) = \{y \in \mathbb{R}^m : Ax = y \text{ for some } x \in \mathbb{R}^n\}$;

$rank(A) = \dim range(A)$.

$null(A) = \{x \in \mathbb{R}^n : Ax = 0\}$

$range(A) =$ is the subspace spanned by the columns of A.

$$rank(A) + \dim null(A) = n.$$

$A \in \mathbb{R}^{n \times n}$ is *invertible* or *nonsingular* if $rank(A) = n$. Equivalently:

(a) there exists $A^{-1} \in \mathbb{R}^{n \times n}$ such that $AA^{-1} = A^{-1}A = I$ (Identity),
(b) the rows of A are linearly independent,
(c) the linear system $Ax = b$ has a unique solution for each $b \in \mathbb{R}^n$,
(d) $range(A) = \mathbb{R}^n$,
(e) the unique solution of $Ax = 0$ is $x = 0$,
(f) $null(A) = \{0\}$,
(g) the scalar 0 is NOT an eigenvalue of A,
(h) $det(A) \neq 0$.

Basic Linear Algebra Concepts

$A \in \mathbb{R}^{n \times n}$, $x \in \mathbb{C}^n$, $x \neq 0$ is an *eigenvector* of A, and $\lambda \in \mathbb{C}$ its associated *eigenvalue*, if

$$Ax = \lambda x.$$

Eigenvectors associated with distinct eigenvalues are linearly independent (LI).

If there exists a set of n LI eigenvectors, we have the spectral decomposition $A = X \Lambda X^{-1}$, $\Lambda = diag(\lambda_i)$, $X = [x_1, \ldots, x_n]$

Two key properties:

$$\det(A) = \prod_{j=1}^{n} \lambda_j,$$

$$\text{trace}(A) = \sum_{j=1}^{n} \lambda_j,$$

where $\text{trace}(A) = \sum_{j=1}^{n} a_{jj}$.

Basic Linear Algebra Concepts

The vectors $x, y \in \mathbb{R}^n$ are *orthogonal* if $x^T y = 0$.

$Q \in \mathbb{R}^{n \times n}$ is *orthogonal* if $Q^T = Q^{-1}$, i.e., if $Q^T Q = Q Q^T = I$.

(a) $(Qx)^T (Qy) = x^T y$.

(b) $\|Qx\|_2 = \|x\|_2$,

(c) (Pythagorean Theorem) if x is orthogonal to y, $\|x + y\|_2^2 = \|x\|_2^2 + \|y\|_2^2$,

(d) (Cauchy-Schwarz inequality) $|x^T y| \leq \|x\|_2 \|y\|_2$,

(e) (Parallelogram Law) $\|x + y\|_2^2 + \|x - y\|_2^2 = 2\|x\|_2^2 + 2\|y\|_2^2$.

Cosine of the angle θ between x and y

$$\cos\theta(x, y) = \frac{x^T y}{\|x\|_2 \|y\|_2}$$

Basic Linear Algebra Concepts

$A \in \mathbb{R}^{n \times n}$ is *symmetric* if $a_{ij} = a_{ji} \Rightarrow A^T = A$. In that case:

Theorem

A has n real eigenvalues $\lambda_1, \cdots, \lambda_n$, and associated real eigenvectors $x_1, \cdots, x_n$ that form an orthogonal basis of $\mathbb{R}^n$.

Moreover, if $x^T Ax > 0$ for all vector $x \neq 0$, then we say that A is *positive definite (PD)*.

Positive semi-definite if $x^T Ax \geq 0$ for all vector $x \in \mathbb{R}^n$.

Negative (semi) definite (ND) if $-A$ is positive (semi) definite.

Indefinite if A is neither P semi D nor N semi D.

Theorem

Given a symmetric $A \in \mathbb{R}^{n \times n}$, the following statements are equivalent:

(a) $x^T Ax > 0$ for all vector $x \neq 0$,

(b) all the eigenvalues of A are real positive numbers,

(c) there exists $W \in \mathbb{R}^{n \times n}$, nonsingular, such that $A = W^T W$,

(d) there exists $B \in \mathbb{R}^{n \times n}$, positive definite, such that $A = B^2$,

(e) for any nonsingular matrix X, $X^T AX$ is positive definite,

(f) all the principal sub-matrices of A are positive definite.

Vector norms

A *norm* is a function $\|\cdot\| : \mathbb{R}^n \to \mathbb{R}$ that assigns a nonnegative real value (length) to each vector.

For vectors x and y and scalar β, a norm must satisfy:

(1) $\|x\| \geq 0$, and $\|x\| = 0$ if and only if $x = 0$,
(2) $\|\beta x\| = |\beta|\|x\|$,
(3) $\|x + y\| \leq \|x\| + \|y\|$.

Three important norms in optimization:

$$\begin{aligned} \|x\|_1 &= \sum_{i=1}^{n} |x_i|, \\ \|x\|_2 &= \left(\sum_{i=1}^{n} x_i^2 \right)^{1/2} = \sqrt{x^T x}, \\ \|x\|_\infty &= \max_{1 \leq i \leq n} |x_i|. \end{aligned}$$

Induced Matrix Norms

If $\|\cdot\|$ is a vector norm, then it induces a matrix norm $\|\cdot\|$, as follows

$$\|A\| = \sup_{x \neq 0} \frac{\|Ax\|}{\|x\|} = \max_{\|x\|=1} \|Ax\|.$$

$$\|A\|_1 = \max_{1 \leq j \leq n} \|A_j\|_1, \quad \text{(columns)}$$

$$\|A\|_\infty = \max_{1 \leq i \leq m} \|a_i^T\|_1, \quad \text{(rows)}$$

$$\|A\|_2 = \sqrt{\lambda_{\max}(A^T A)}.$$

Property of any induced matrix norm:

$$\|Ax\| \leq \|A\|\|x\|.$$

Frobenius Norm

An important not induced norm: Frobenius norm:

$$\|A\|_F = \left(\sum_{i=1}^{m} \sum_{j=1}^{n} |a_{ij}|^2 \right)^{1/2} = \left(\sum_{j=1}^{n} \|A_j\|_2^2 \right)^{1/2}.$$

$$\|A\|_F = \sqrt{\text{trace}(A^T A)} = \sqrt{\text{trace}(A A^T)}.$$

It is not induced. However:

$$\|Ax\|_2 \leq \|A\|_F \|x\|_2.$$

$$\|AB\|_F \leq \|A\|_F \|B\|_F.$$

$$\cos(A, B) = \frac{\text{trace}(A^T B)}{\|A\|_F \|B\|_F}$$

Multivariable Calculus (brief review)

A continuous function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be *continuously differentiable at* $x \in \mathbb{R}^n$, if $(\partial f/\partial x_i)(x)$ exists and is continuous, for $i = 1, \cdots, n$.

The gradient vector of f at x (notation):

$$\nabla f(x) = [(\partial f/\partial x_1)(x), \cdots, (\partial f/\partial x_n)(x)]^T.$$

If $x = x(y)$ and y is a vector, Chain Rule:

$$\nabla_y f(x(y)) = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i} \nabla x_i(y).$$

Directional derivative

Lemma

Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable in an open convex set $D \subset \mathbb{R}^n$. Then for all $x \in D$ and $p \in \mathbb{R}^n$, $p \neq 0$, the directional derivative in the direction of p, defined by

$$\frac{\partial f}{\partial p}(x) = \lim_{\epsilon \to 0} \frac{f(x + \epsilon p) - f(x)}{\epsilon},$$

exists and equals $\nabla f(x)^T p$.

Key lemma to find gradients!

Example 1: If $f(x) = x^T x$, $\nabla f(x) = 2x$.

Example 2: If $f(x) = x^T A x$, $A^T = A$, $\nabla f(x) = 2Ax$

Multivariable versions of FTC and MVT

Theorem

Let $f \in C^1(D)$, $x \in D$ and $(x+p) \in D$ for some vector $p \neq 0$, then (FTC)

$$f(x+p) - f(x) = \int_0^1 \nabla f(x+tp)^T p \, dt.$$

Moreover, there exists $z \in (x, x+p)$ such that (MVT)

$$f(x+p) - f(x) = \nabla f(z)^T p.$$

Second order derivatives

The function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be *twice continuously differentiable at* $x \in \mathbb{R}^n$, if $(\partial^2 f / \partial x_i \partial x_i)(x)$ exists and is continuous, for $1 \leq i, j \leq n$.

The Hessian of f at x:

$$\nabla^2 f(x) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix},$$

is a symmetric matrix.

Second directional derivative

Lemma

Let $f : \mathbb{R}^n \to \mathbb{R}$ be twice continuously differentiable in an open convex set $D \subset \mathbb{R}^n$. Then for all $x \in D$ and $p \in \mathbb{R}^n$, $p \neq 0$, the second directional derivative of f at x in the direction of p, defined by

$$\frac{\partial^2 f}{\partial p^2}(x) = \lim_{\epsilon \to 0} \frac{\frac{\partial f}{\partial p}(x + \epsilon p) - \frac{\partial f}{\partial p}(x)}{\epsilon}$$

exists and equals $p^T \nabla^2 f(x) p$.

Taylor's Theorem

Under the same hypothesis, there exists $z \in (x, x + p)$,

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

Part II

Optimality Conditions

Convexity

A set Ω in a vector space is convex if

$$(1-\alpha)x + \alpha y \in \Omega$$

for all $x \in \Omega$, $y \in \Omega$ and $0 < \alpha < 1$.

A function f over a convex set Ω is convex if for all $x, y \in \Omega$, and $\alpha \in [0, 1]$, it holds that

$$f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha)f(y).$$

If for all $\alpha \in (0, 1)$ and $x \neq y$ the inequality is strict, then we say that f is *strictly convex*. We say that f is concave if $-f$ es convex.

Unconstrained case: optimality conditions

Definition: x^* is a local minimizer of f if there exists $\epsilon > 0$ such that $f(x^*) \leq f(x)$ for all $\|x - x^*\| \leq \epsilon$.

First order necessary condition:

If $f : \mathbb{R}^n \to \mathbb{R}$, $f \in C^1(D)$, where D is open and convex, and $x^* \in D$ is a local minimizer of f, then $\nabla f(x^*) = 0$.

Second order necessary condition:

If $f : \mathbb{R}^n \to \mathbb{R}$, $f \in C^2(D)$, where D is open and convex, and $x^* \in D$ is a local minimizer of f, then $\nabla f(x^*) = 0$ and $\nabla^2 f(x^*)$ is positive semi-definite.

Unconstrained case: optimality conditions

Sufficient condition:

If $f : \mathbb{R}^n \to \mathbb{R}$, $f \in C^2(D)$, where D is open and convex, and if

$\nabla f(x^*) = 0$ and $\nabla^2 f(x^*)$ is positive definite for $x^* \in D$

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

Example 1: $f(x) = x^T Ax$ where A is PD.

Example 2: $f(x) = \sum_{i=1}^{n} (e^{x_i} - x_i)$.

Convexity using derivatives

Theorem

If f is continuously differentiable, then f is convex over a convex set Ω if and only if

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

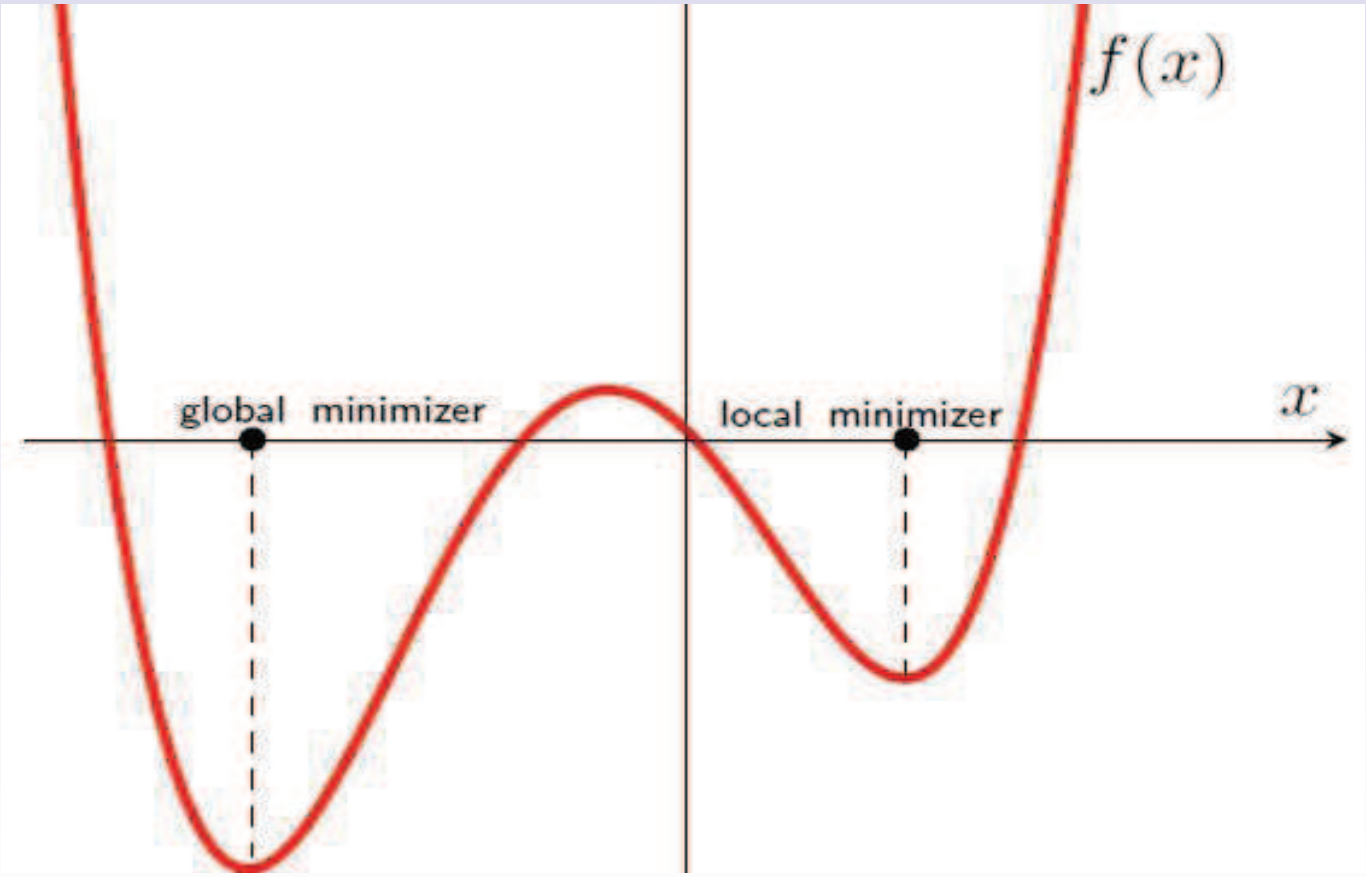
for all $x, y \in \Omega$.

Moreover, if f is C^2, using Taylor's Theorem we can establish the following characterization of convexity (very convenient!)

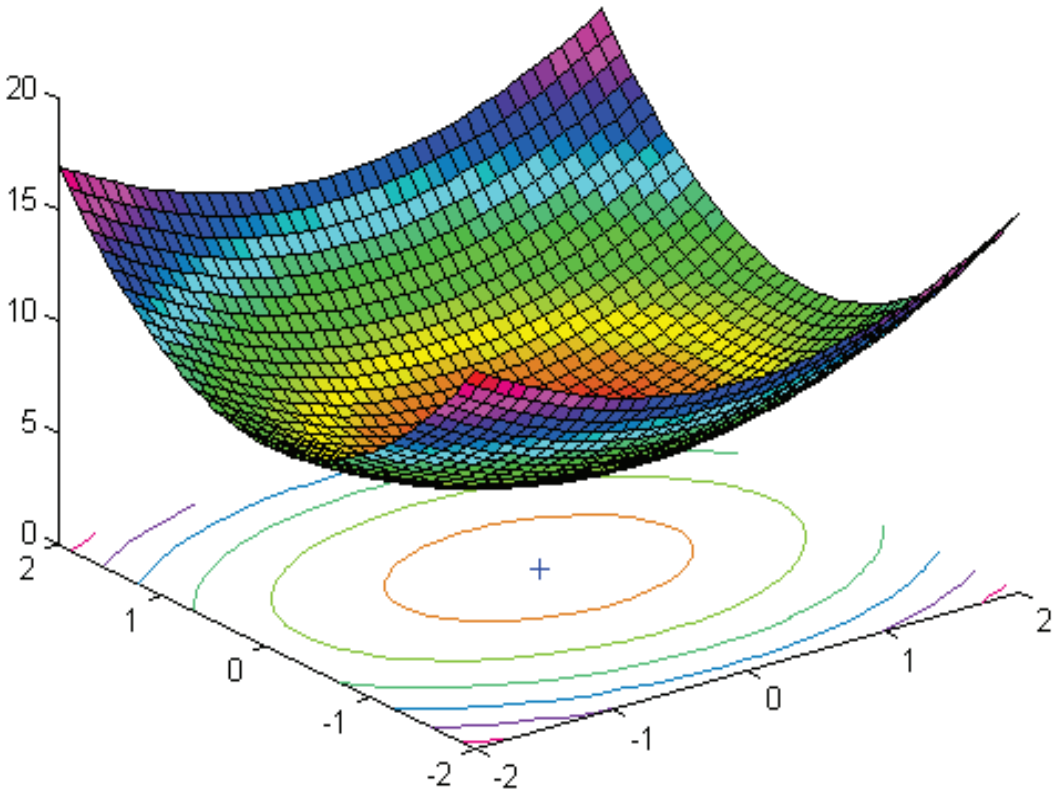
Theorem

If f is twice continuously differentiable then f is convex, over a convex Ω, if and only if $\nabla^2 f(x)$ is positive semi-definite for all x in Ω.

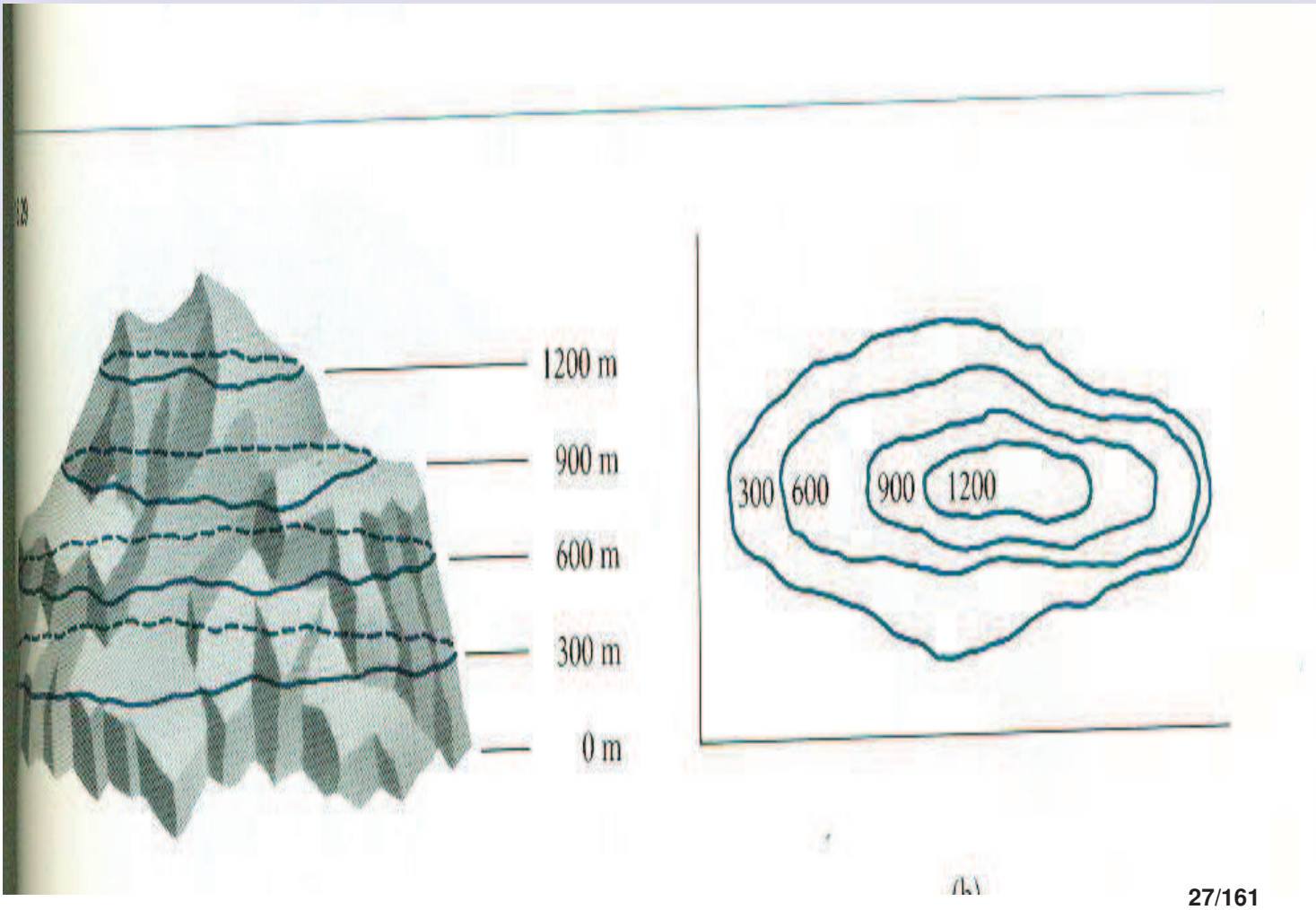
Minimizers in one dimension



Local minimizer and level curves



Level curves



Part III

Quadratic Functions

Quadratic Functions

Polynomials of degree 2 in n variables. Example (3 variables):

$$q(x_1, x_2, x_3) = x_1^2 - 3x_3^2 + 2x_1x_2 - 5x_2x_3 + x_1 - x_3 - 4\ ,$$

Locally, via Taylor, they approximate general smooth functions.

All quadratic functions can be written as:

$$q(x) = \frac{1}{2}x^T Ax - b^T x + c,$$

where $A \in \mathbb{R}^{n\times n}$ is symmetric, the vectors $x, b \in \mathbb{R}^n$ and $c \in \mathbb{R}$.

Quadratic Functions

In our example

$$q(x_1, x_2, x_3) = x_1^2 - 3x_3^2 + 2x_1x_2 - 5x_2x_3 + x_1 - x_3 - 4\,,$$

we have that

$$A = \begin{pmatrix} 2 & 2 & 0 \\ 2 & 0 & -5 \\ 0 & -5 & -6 \end{pmatrix}\,, \quad b = \begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix}\,, \quad c = -4\,.$$

To obtain c, b, and A we need to compute the gradient and the Hessian of $q(x)$.

Critical points for quadratics

Gradient and Hessian

If $q(x) = \frac{1}{2}x^T Ax - b^T x + c$, where $A^T = A$, then $\nabla q(x) = Ax - b$ and the matrix $\nabla^2 q(x) = A$ for all $x \in \mathbb{R}^n$.

The critical points or stationary points x^* (i.e., $\nabla f(x^*) = 0$) of $q(x)$ are solutions of the linear system

$$Ax = b.$$

Do they always have critical points?

Critical points for quadratics

Theorem

The quadratic function $q(x)$ has critical points if and only if

$$b \in range(A).$$

And it has a unique critical point if and only if

A is nonsingular.

Classification of critical points for quadratics

There are three options: The system $Ax = b$ has no solutions, one solution or infinite solutions.

If $b \notin range(A)$ then $q(x)$ has no critical points, i.e., the gradient is not zero for all $x \in \mathbb{R}^n$.

If A is nonsingular, then $q(x)$ has a unique critical point $x^* = A^{-1}b$, that could be a maximizer, a minimizer, or a saddle point.

If the linear system has infinite solutions, then $q(x)$ has infinite critical points and they are all of the same kind.

Classification of critical points for quadratics

Convex case

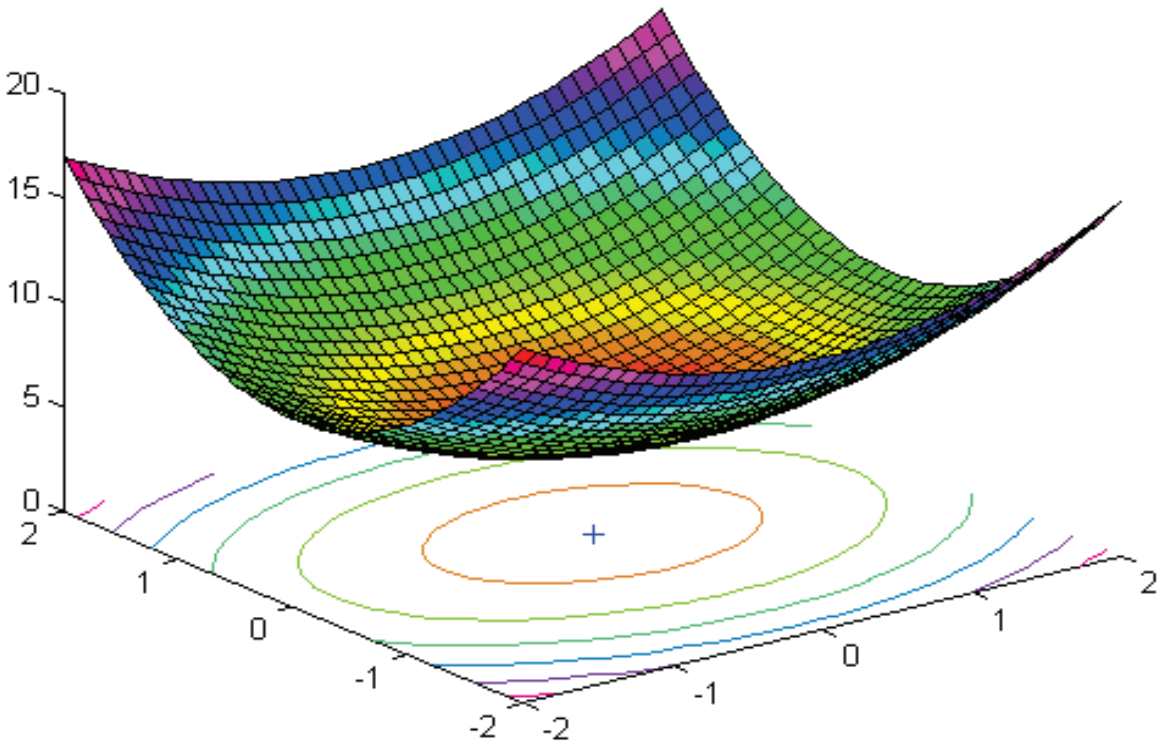
If A is positive semi-definite and x^* is a critical point of $q(x)$, then x^* is a global minimizer of $q(x)$.

Moreover, if A is PD then x^* is an isolated global minimizer (unique).

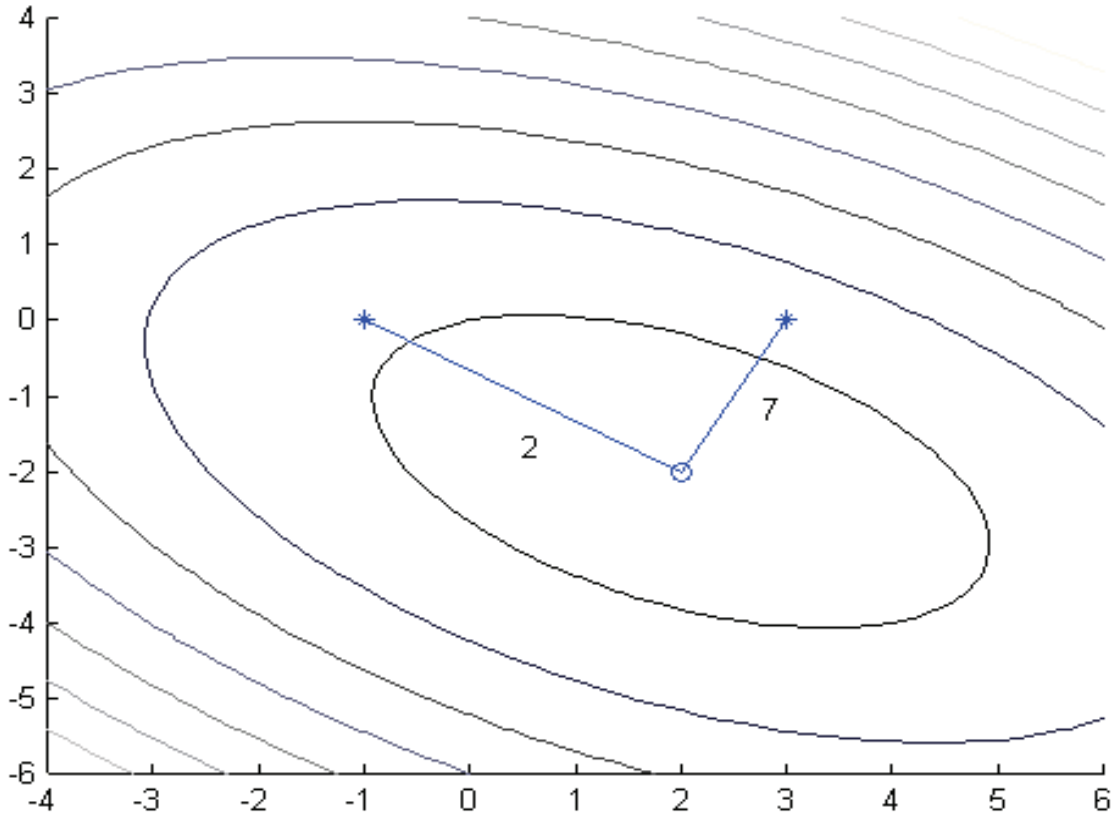
Why? If A is positive semi-definite then q is convex. If A is PD, q is strictly convex.

If A is negative semi-definite, then all critical points are maximizers, and if A is indefinite, they are all saddle points.

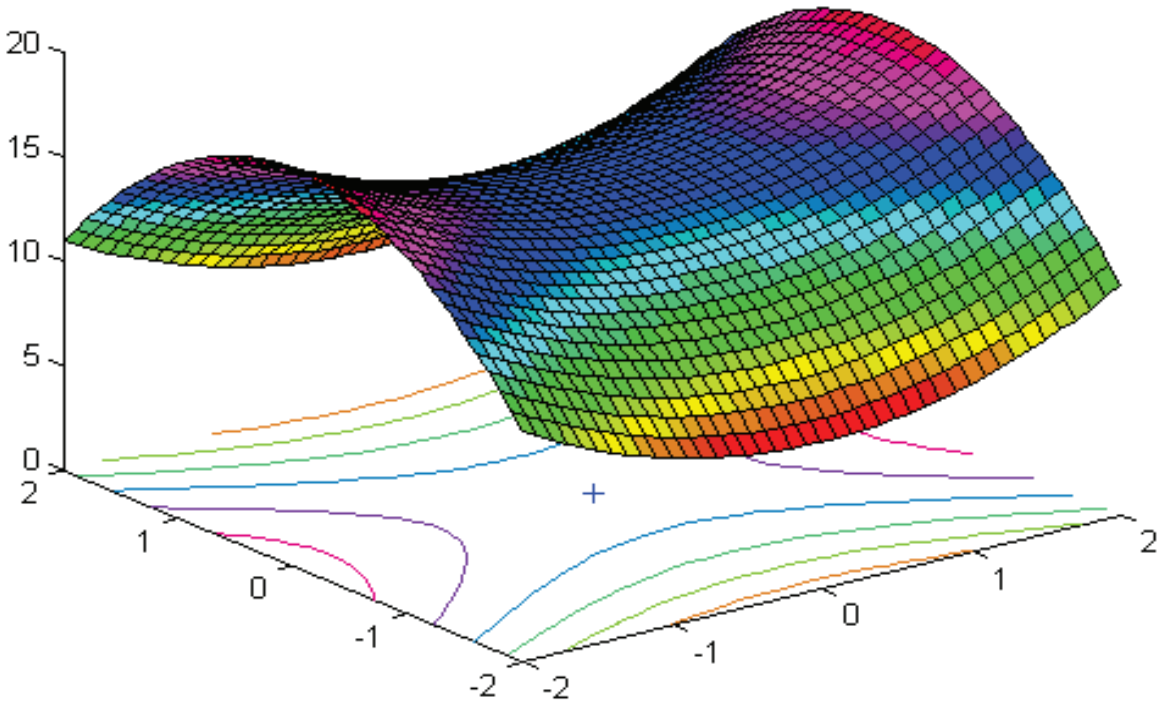
Strictly convex quadratic ($n = 2$)



Strictly convex quadratic ($\lambda_1 = 2$ and $\lambda_2 = 7$)



Saddle point ($n = 2$)



Level sets

Convexity

If f is a convex function and $M \in [-\infty, +\infty]$, then the level sets $\{x \ : \ f(x) < M\}$ and $\{x \ : \ f(x) \leq M\}$ are convex sets.

It is not true in the other direction: if a function has convex level sets for all $M \in [-\infty, +\infty]$, it is not necessarily a convex function. Example: $f(x) = x^3$.

If q is a quadratic and A is PD, the level sets are concentric ellipsoids, and the unique minimizer is at the center of all the ellipsoids. The principal axis are the eigenvectors of A.