Nonlinear Optimization

Marcos Raydan

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

FCT, UNL
Caparica, Portugal
March – June, 2020

Fundamental observations

- Nature **loves** to optimize, and the world is far from linear.
- Mankind **needs** to optimize.
- Scientists like to **build** parameterized **models** of natural phenomena, and try to match them to reality.
- This fitting of data to models occurs everywhere, and the best fit is, of course, an **optimization problem**.
- The solution of nonlinear optimization problems is one of the key components of modern computational and applied mathematics.

Examples in which optimization is an essential tool

- **Running a business:** to maximize profit, minimize loss, maximize efficiency, or minimize risk.
- **Engineering design:** minimize the weight of a bridge, or an airplane, or a tall building, and maximize the strength (avoiding resonance), within the design constraints.
- **Packing:** maximize the amount of food stored in a given container, or millions of transistors in a computer chip in a functional way.

What do they have in common?

We need to minimize (or maximize) an **objective function** by deciding the values of **free variables** from within a **feasible set**.

Optimization

General formulation

Minimize $f(x)$

subject to $x \in \Omega$

– f is the objective function, x is the vector of free variables, and Ω is the feasible (closed) set.

– Minimize $(-f(x)) \equiv$ Maximize $f(x)$.

– If Ω is equal to the whole space, we call it **unconstrained optimization**, if not, we call it **constrained optimization**.

– Constraints can be given by mathematical formulas (**equalities and/or inequalities**), or using "geometrical concepts" (subspaces, boxes, spheres, cones, etc.)

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 $Q^T Q = QQ^T = I$.

(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

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

$A \in \mathbb{R}^{n \times n}$ 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:

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

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

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

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 $z \in (x, x + p)$,

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

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

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

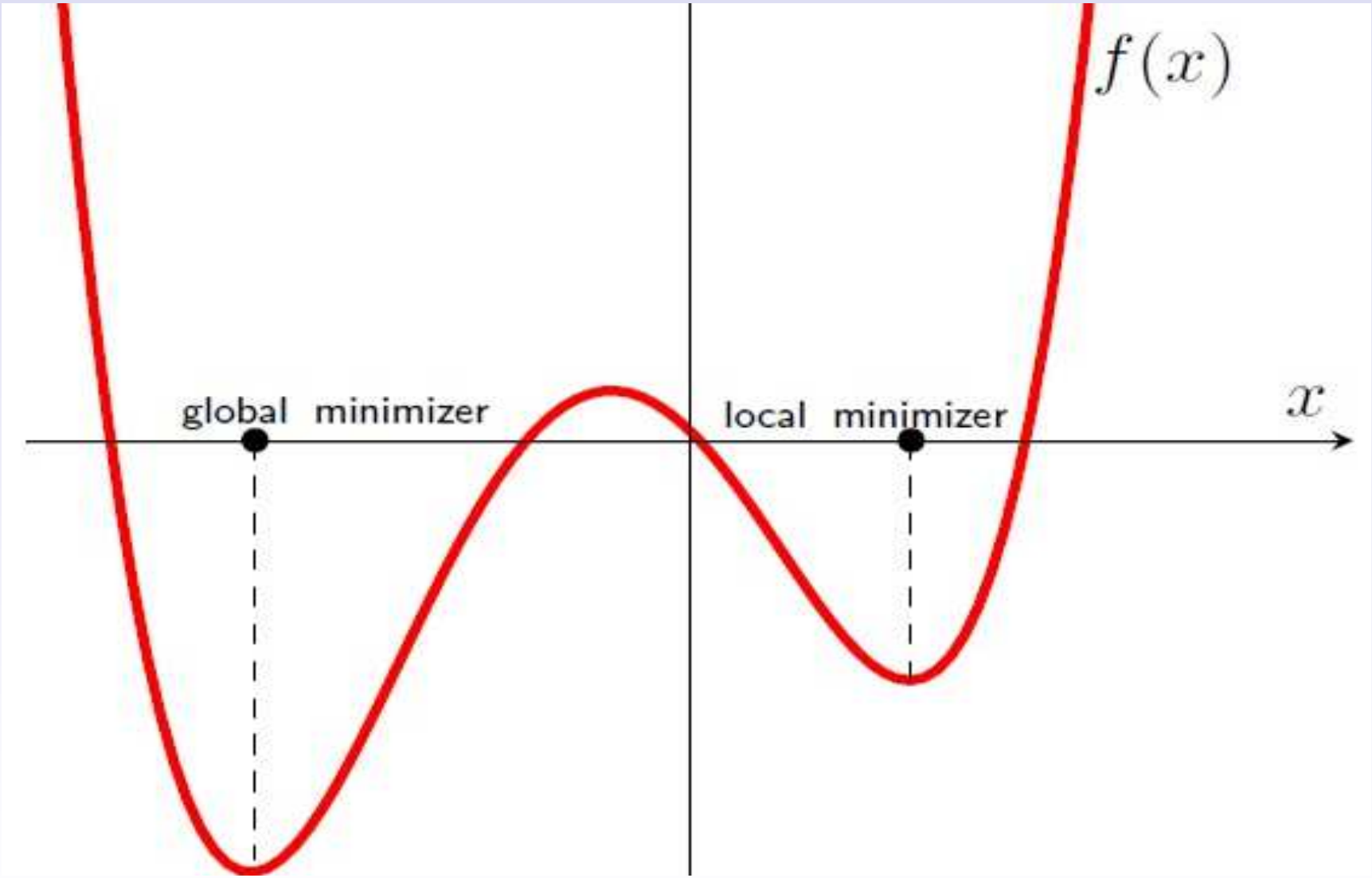
for all $x, y \in \Omega$.

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

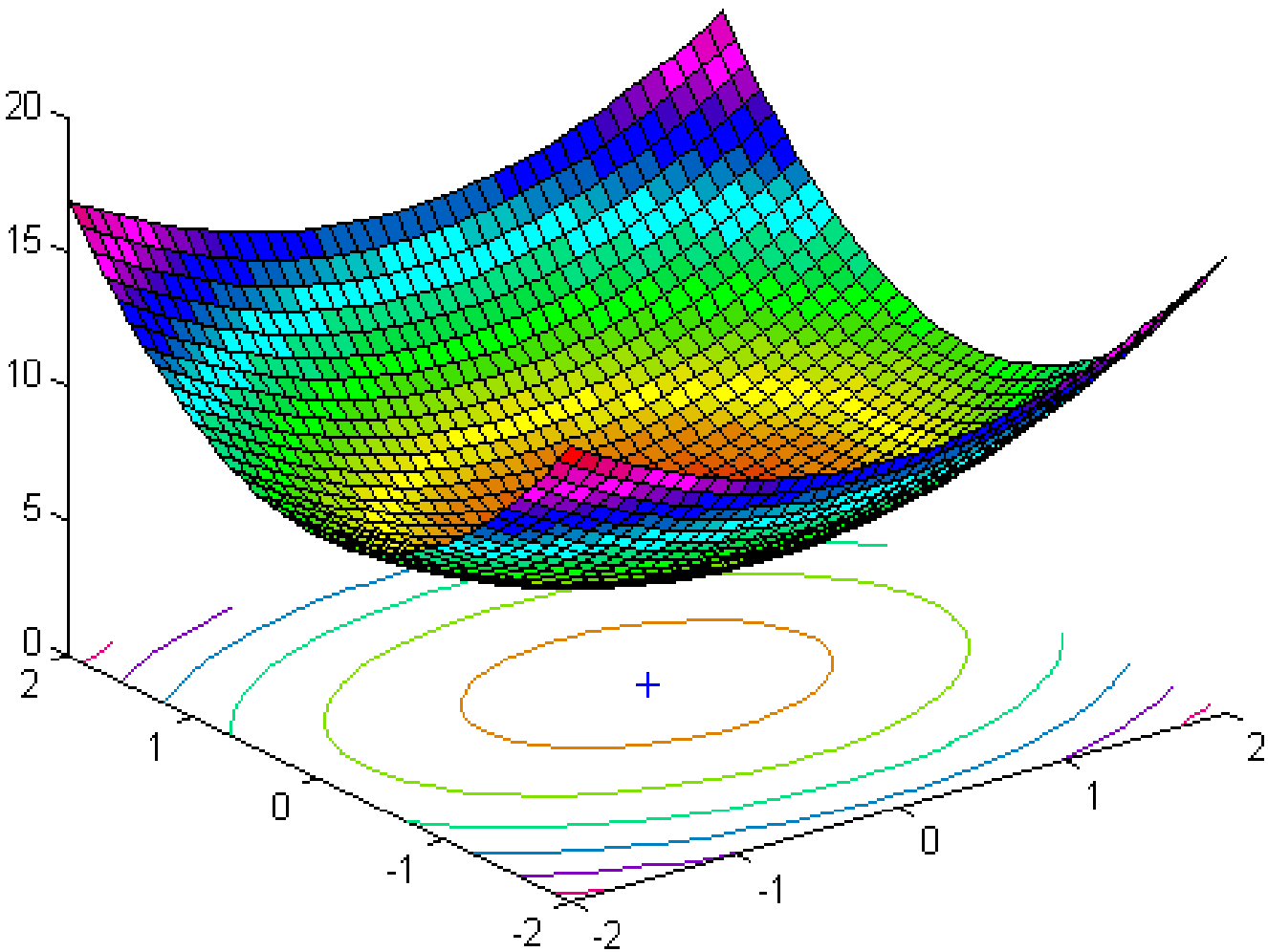
Theorem

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.

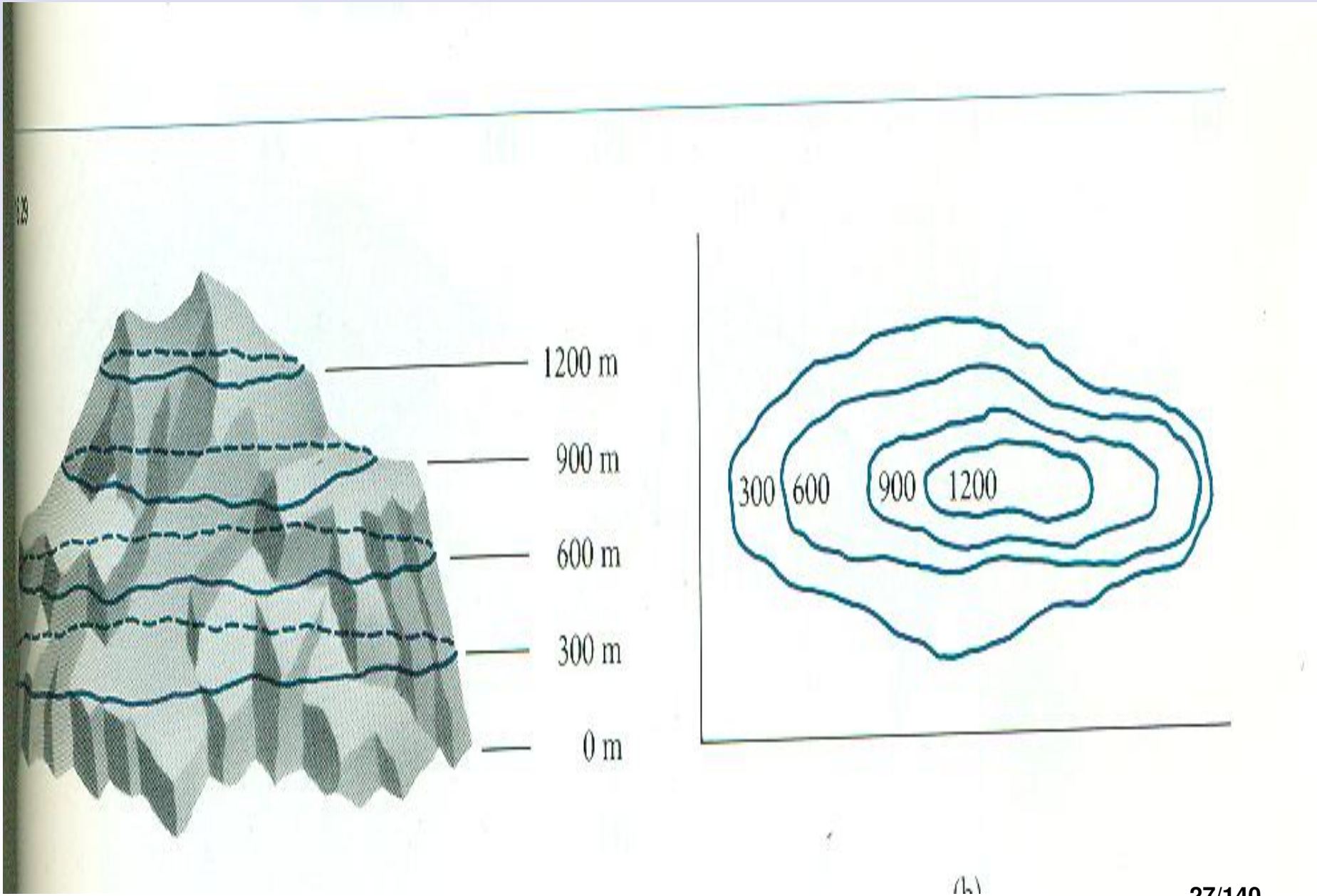
Minimizers in one dimension



Local minimizer and level curves



Level curves



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

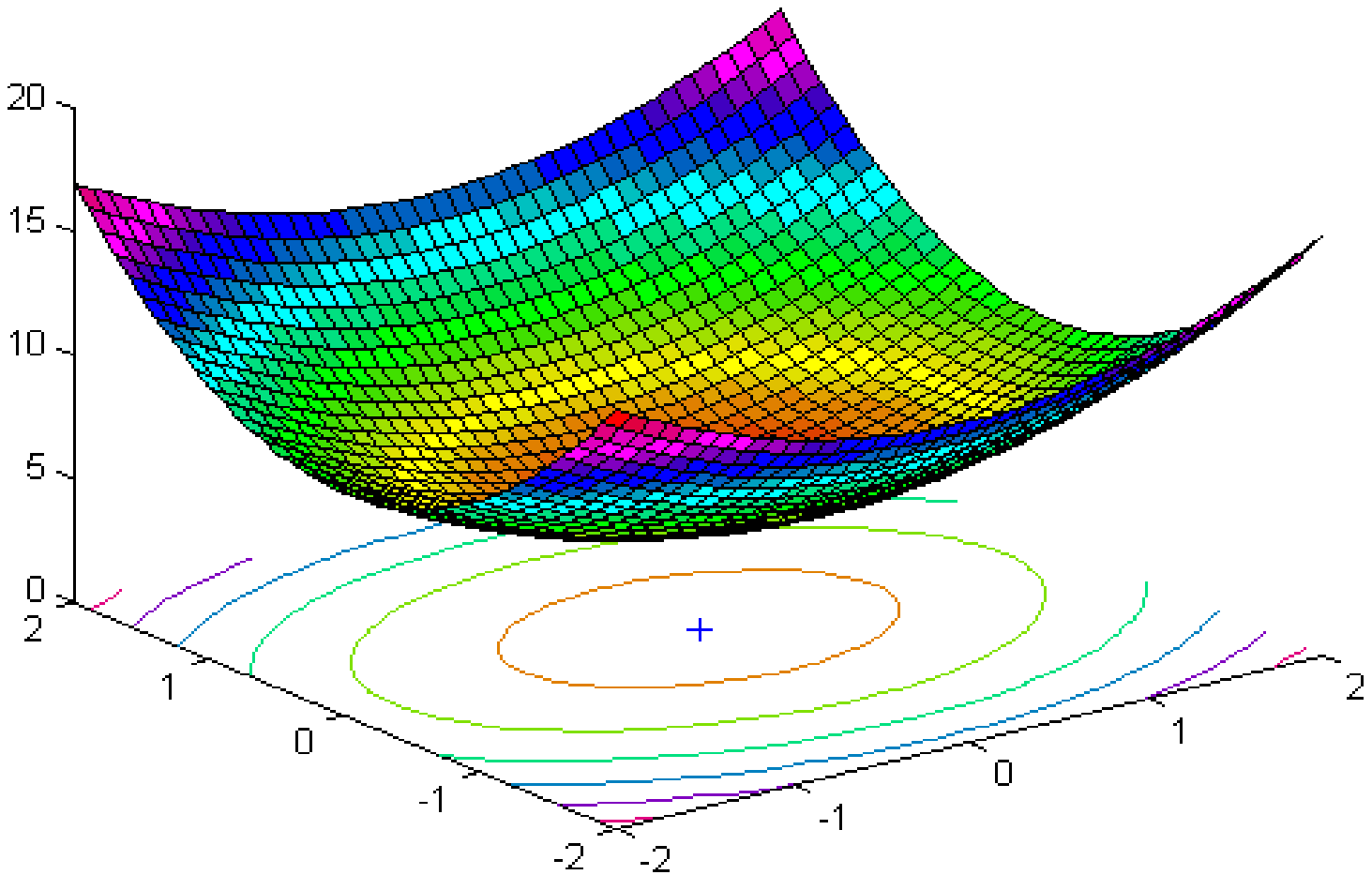
If A is positive semi-definite and x^* is a critical point of $q(x)$, then x^* is a global minimizer of $q(x)$.

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

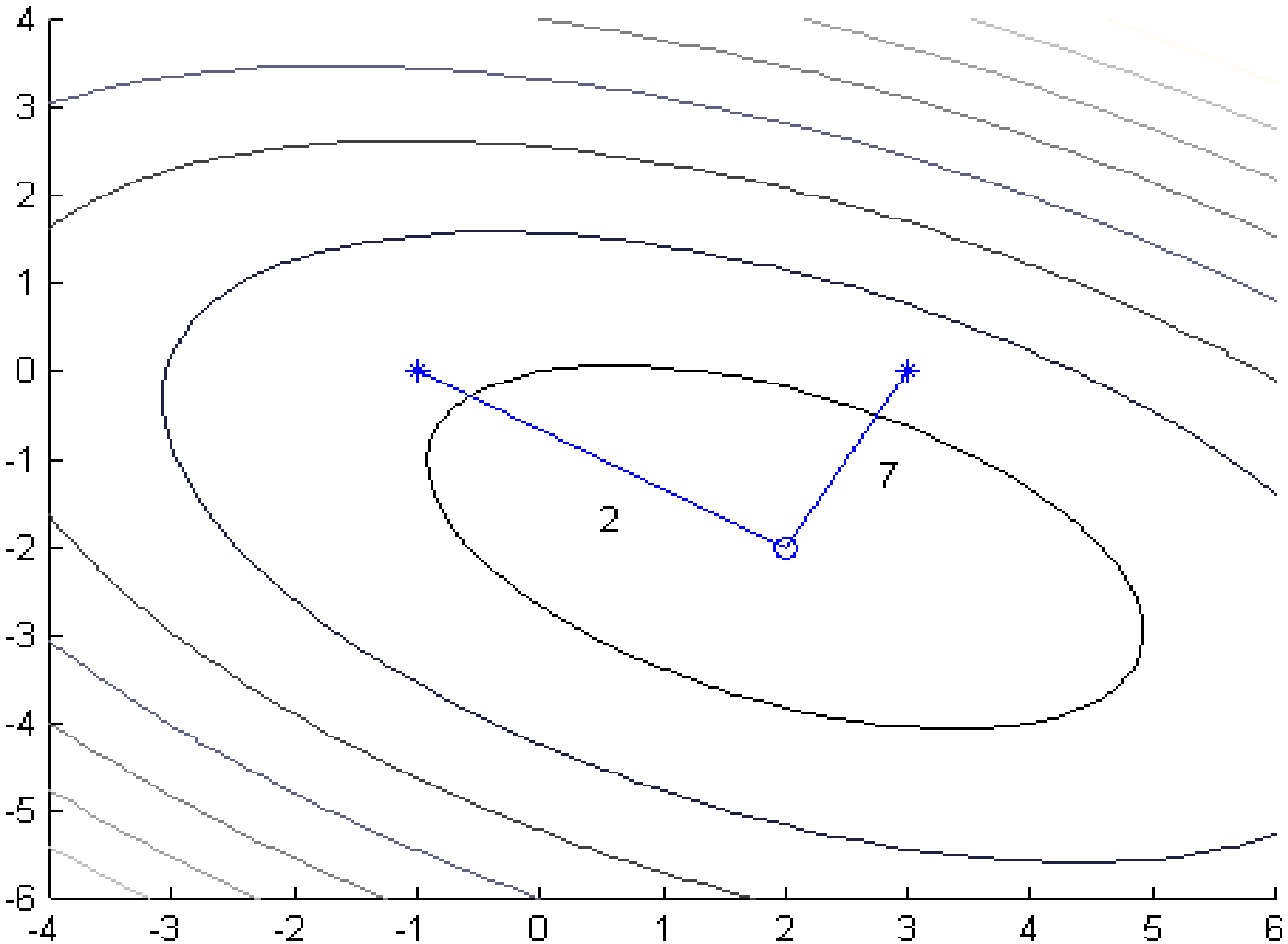
Why? If A is positive semi-definite then q is convex. If A is PD, q is strictly convex.

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

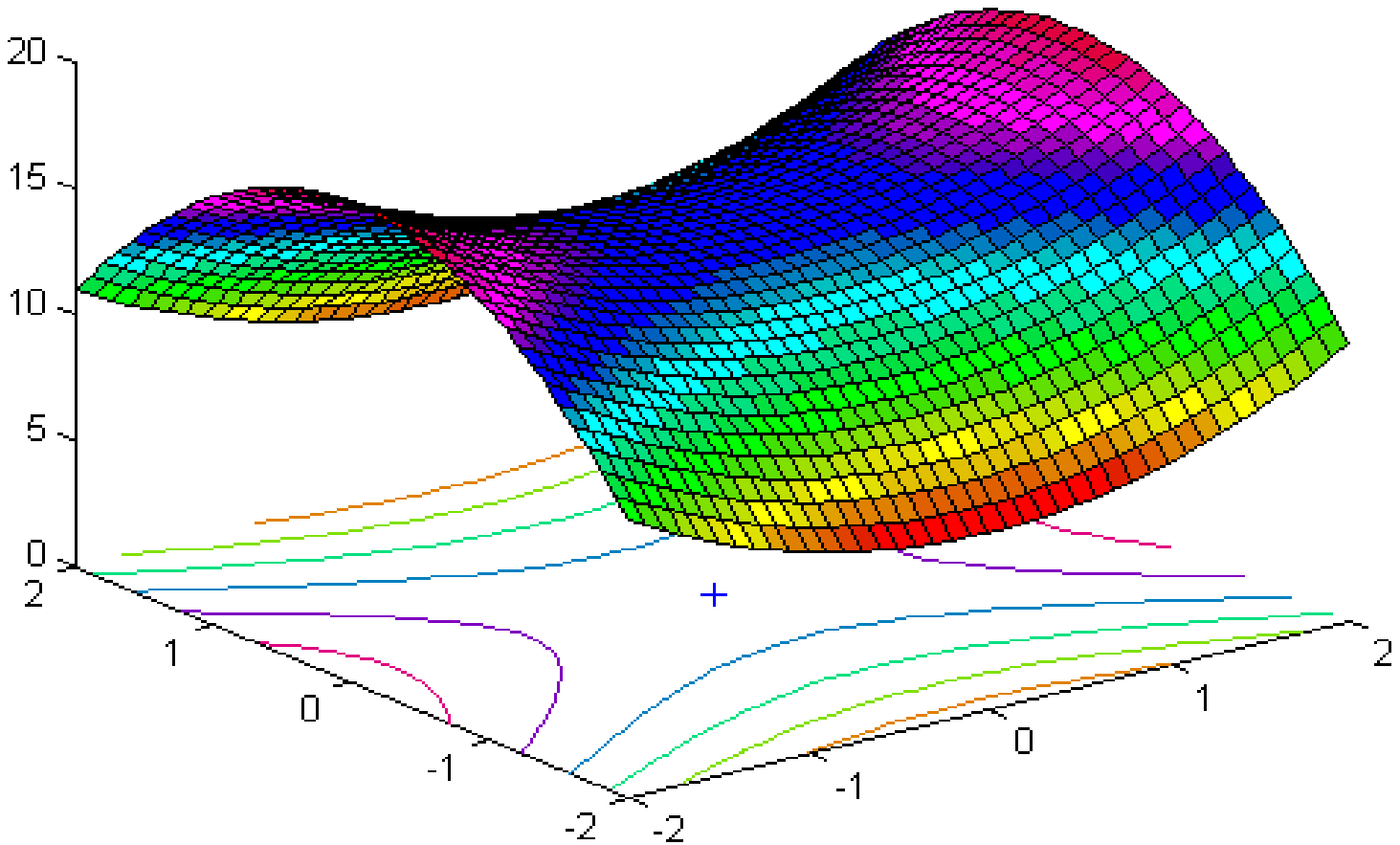
Strictly convex quadratic ($n = 2$)



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.

Iterative Methods (unconstrained)

$$\min_{x \in \mathbb{R}^n} f(x)$$

Descent Directions

Let $f : \mathbb{R}^n \to \mathbb{R}$. We say that a direction $d \in \mathbb{R}^n$ is a descent direction at $x \in \mathbb{R}^n$ if

$$\nabla f(x)^T d < 0.$$

Negative cosine $\Rightarrow$ the angle between $\nabla f(x)$ and d is greater than 90^0.

Theorem: The value of f decreases along d

Let $f : \mathbb{R}^n \to \mathbb{R}$, $f \in C^1(\mathbb{R}^n)$, $x \in \mathbb{R}^n$ such that $\nabla f(x) \neq 0$, and $d \in \mathbb{R}^n$ such that $\nabla f(x)^T d < 0$. Then, there exists $\bar{\alpha} > 0$ such that

$$f(x + \alpha d) < f(x) \quad \text{, for all } \alpha \in (0, \bar{\alpha}].$$

Descent Methods

Large family of iterative methods to find local minimizers:

$$x_{k+1} = x_k + \lambda_k d_k,$$

where d_k is a descent direction

λ_k is a step length that guarantees descent in f along d_k.

Different ways of choosing d_k and different ways of choosing λ_k produce different methods.

Friendly option (always available if $f \in C^1(\mathbb{R}^n)$):

$d_k = -\nabla f(x_k)$ (Cauchy, 1847).

Known as the negative gradient direction. It is the direction that guarantees the steepest local descent of f.

Classic Descent Methods

Cauchy's Method or steepest descent:

$$x_{k+1} = x_k - \lambda_k \nabla f(x_k),$$

where

$$\lambda_k = \text{argmin}_{\lambda > 0} f(x_k - \lambda \nabla f(x_k)).$$

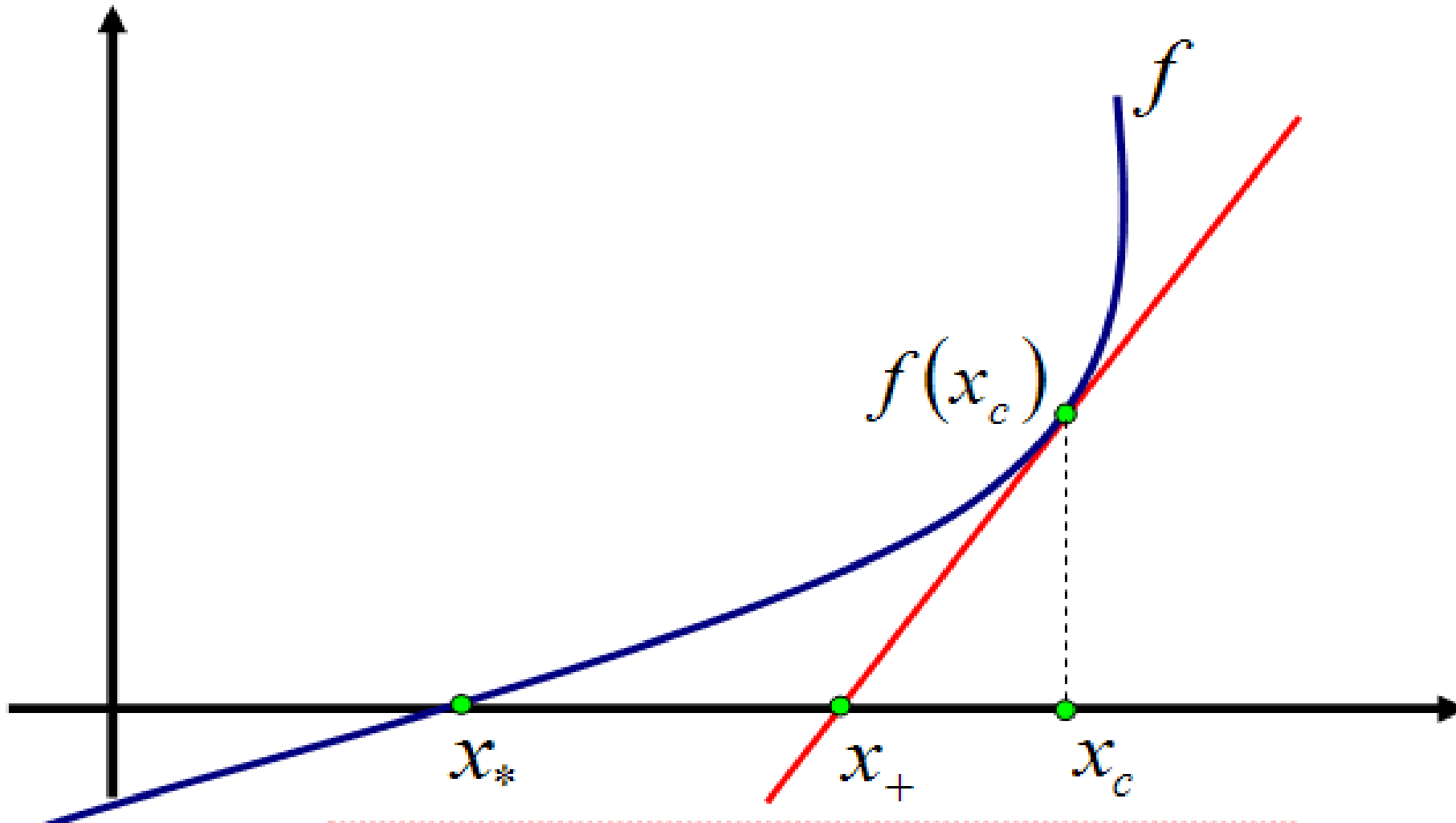
Newton's Method: Solve the following system for d_k

$$\nabla^2 f(x_k) d_k = -\nabla f(x_k).$$

If the Hessian at x_k is PD, then d_k in Newton's method is a descent direction:

$$\nabla f(x_k)^T d_k = -\nabla f(x_k)^T \nabla^2 f(x_k)^{-1} \nabla f(x_k) < 0.$$

Newton for $n = 1$



$$M_c(x) = f(x_c) + f'(x_c)(x - x_c)$$

Newton for quadratics

Consider Newton's method:

$$x_{k+1} = x_k - \lambda_k \nabla^2 f(x_k)^{-1} \nabla f(x_k),$$

when f is a strictly convex quadratic function. The exact solution is obtained at the first iteration (for $\lambda_0 = 1$), from any given initial guess x_0:

$$x_1 = x_0 - A^{-1}(Ax_0 - b) = A^{-1}b = x^*.$$

Cost? We need to solve a linear system of equations.

Cauchy for quadratics

Consider Cauchy's method in the strictly convex quadratic case:

$$x_{k+1} = x_k - \lambda_k g_k,$$

where $g_k = \nabla f(x_k) = Ax_k - b$.

The optimal step length is given by

$$\lambda_k = \frac{g_k^T g_k}{g_k^T A g_k}.$$

It follows that

$$E(x_{k+1}) \leq \left(\frac{\lambda_{\max}(A) - \lambda_{\min}(A)}{\lambda_{\max}(A) + \lambda_{\min}(A)} \right)^2 E(x_k),$$

where

$$E(x) = \frac{1}{2}(x - x^*)^T A(x - x^*).$$

Convergence to x^* is guaranteed from any given initial guess x_0

Cauchy for quadratics

$$E(x_{k+1}) \leq \left(\frac{\lambda_{\max}(A) - \lambda_{\min}(A)}{\lambda_{\max}(A) + \lambda_{\min}(A)} \right)^2 E(x_k).$$

Equivalently

$$\|x_k - x^*\|_A \leq \frac{\kappa_2(A) - 1}{\kappa_2(A) + 1} \|x_{k-1} - x^*\|_A,$$

where $\kappa_2(A) = \lambda_{\max}/\lambda_{\min}$, and $\|z\|_A^2 = z^T A z$.

If $\lambda_{\min}$ is not close to $\lambda_{\max}$ then the convergence is very slow.

Moreover, for all k

$$g_{k+1}^T g_k = 0.$$

These two facts together produce the so-called zig-zag behavior.

Cauchy for quadratics

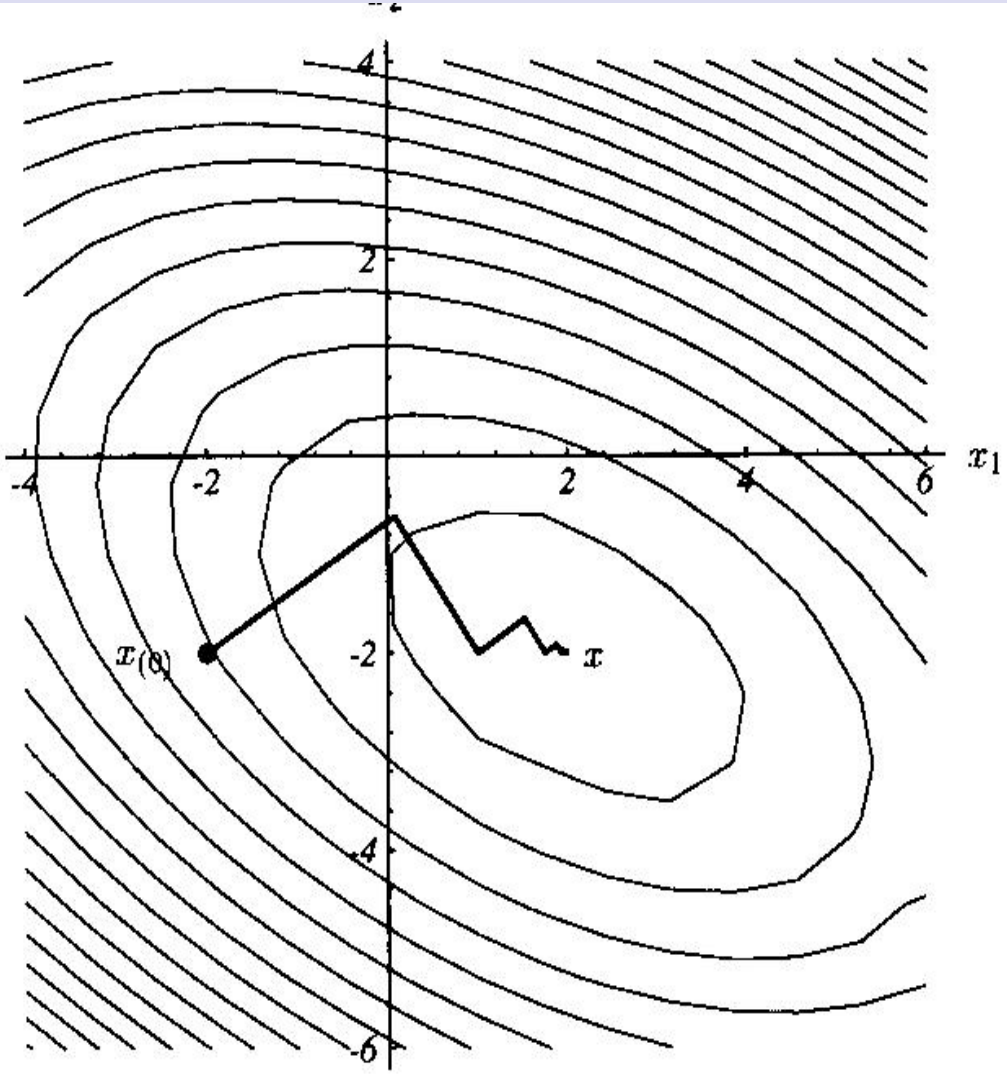


Figure 8: Here, the method of Steepest Descent starts at $[-2, -2]^T$ and converges at $[2, -2]^T$.

Taxonomy of speed of convergence

If $\{x_k\}$ converges to x^*, we need to monitor $e_k = x_k - x^*$.

We say that $\{e_k\}$ converges to 0 **q-order p** if there exist $c > 0$ and $k_0 \in \mathbb{N}$, such that

$$\|e_{k+1}\| \leqslant c\|e_k\|^p, \ \forall\, k \geqslant k_0.$$

If $p = 1$, we call it **q-linear** $(0 < c < 1)$.
Example: $\{10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}, \dots\}$

If $p = 2$, we call it **q-quadratic**. Example:
$\{10^{-1}, 10^{-2}, 10^{-4}, 10^{-8}, 10^{-16}, \dots\}$.

We say that $\{e_k\}$ converges to 0 **r-order p** if there exist $\{b_k\}$ and $k_0 \in \mathbb{N}$, such that

$$\|e_k\| \leqslant \|b_k\|, \ \forall\, k \geqslant k_0 \text{ and } \{b_k\} \text{ converges to 0 q-order p .}$$

Taxonomy of speed of convergence

Special cases:

We say that $\{e_k\}$ converges to 0 **q-superlinearly** if

$$\|e_{k+1}\| \leqslant c_k \|e_k\|, \ \forall\, k \geqslant k_0,$$

and $\{c_k\}$ converges monotonically to 0. Example:
$\{10^{-1}, 10^{-2}, 10^{-3}, 10^{-5}, 10^{-8}, \dots\}$

We say that $\{e_k\}$ converges to 0 **r-superlinearly** if there exist $\{b_k\}$ and $k_0 \in \mathbb{N}$, such that

$\|e_k\| \leqslant \|b_k\|, \ \forall\, k \geqslant k_0$ and $\{b_k\}$ converges to 0 q-superlinearly.

We say that $\{e_k\}$ converges to 0 **q-sublinearly** if

$$\|e_{k+1}\|/\|e_k\| < 1, \ \forall\, k \geqslant k_0,$$

and $\|e_{k+1}\|/\|e_k\|$ converges to 1.

Example of q-sublinearity: $\{1/k\}$.

Variants of Cauchy method for quadratics

Speed of convergence of Cauchy method: q-linear (slow, zig-zag).

How come? It involves an optimal step length on an optimal direction
$\Rightarrow g_{k+1}^T g_k = 0$. (On a 2-dim subspace; Akaike, 1959).

Variant: Random Cauchy ($\theta_k \in (0, 2)$ random) [COAP, 2002]

$$x_{k+1} = x_k - \theta_k \lambda_k g_k,$$

If $\theta_k = 1$ we recover Cauchy's method.
If $\theta_k = 2$, $f(x_{k+1}) = f(x_k)$.

Variant: Spectral Gradient (Barzilai-Borwein) [IMA J.N. A., 1988, 1993]

$$\theta_k = 1 \qquad \lambda_k = \frac{g_{k-1}^T g_{k-1}}{g_{k-1}^T A g_{k-1}}.$$

No zig-zag effect (Random is q-linear, and B-B is r-linear).

Variants of Cauchy method for quadratics

Let $f(x)$ be a strictly convex quadratic function, and x^* the unique minimizer of f.

Convergence theorem for spectral variant [IMA J.N. A., 1993]

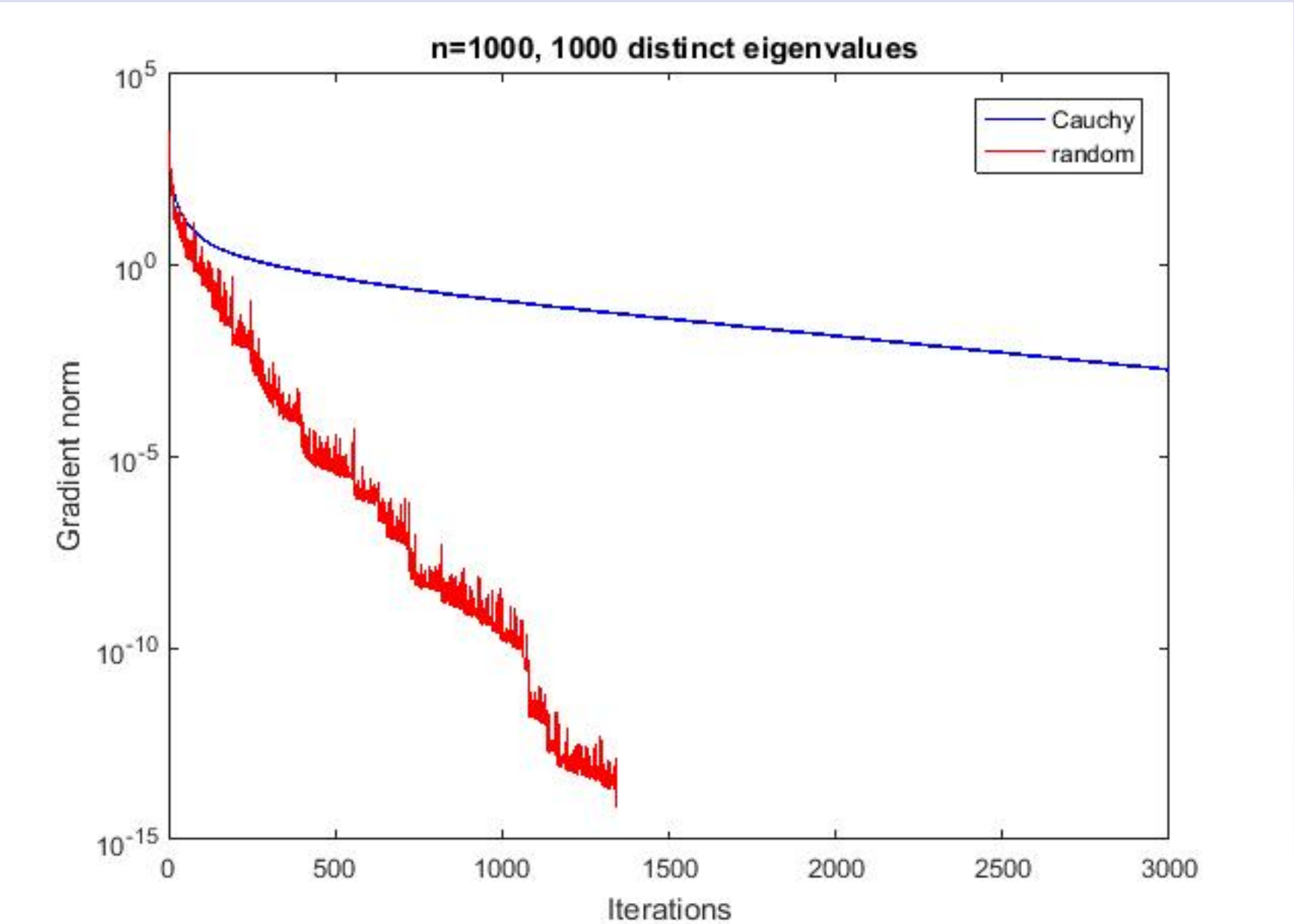
The sequence $\{x_k\}$ generated by the spectral gradient method converges to x^*.

For the spectral variant, the convergence is R-linear [2002]

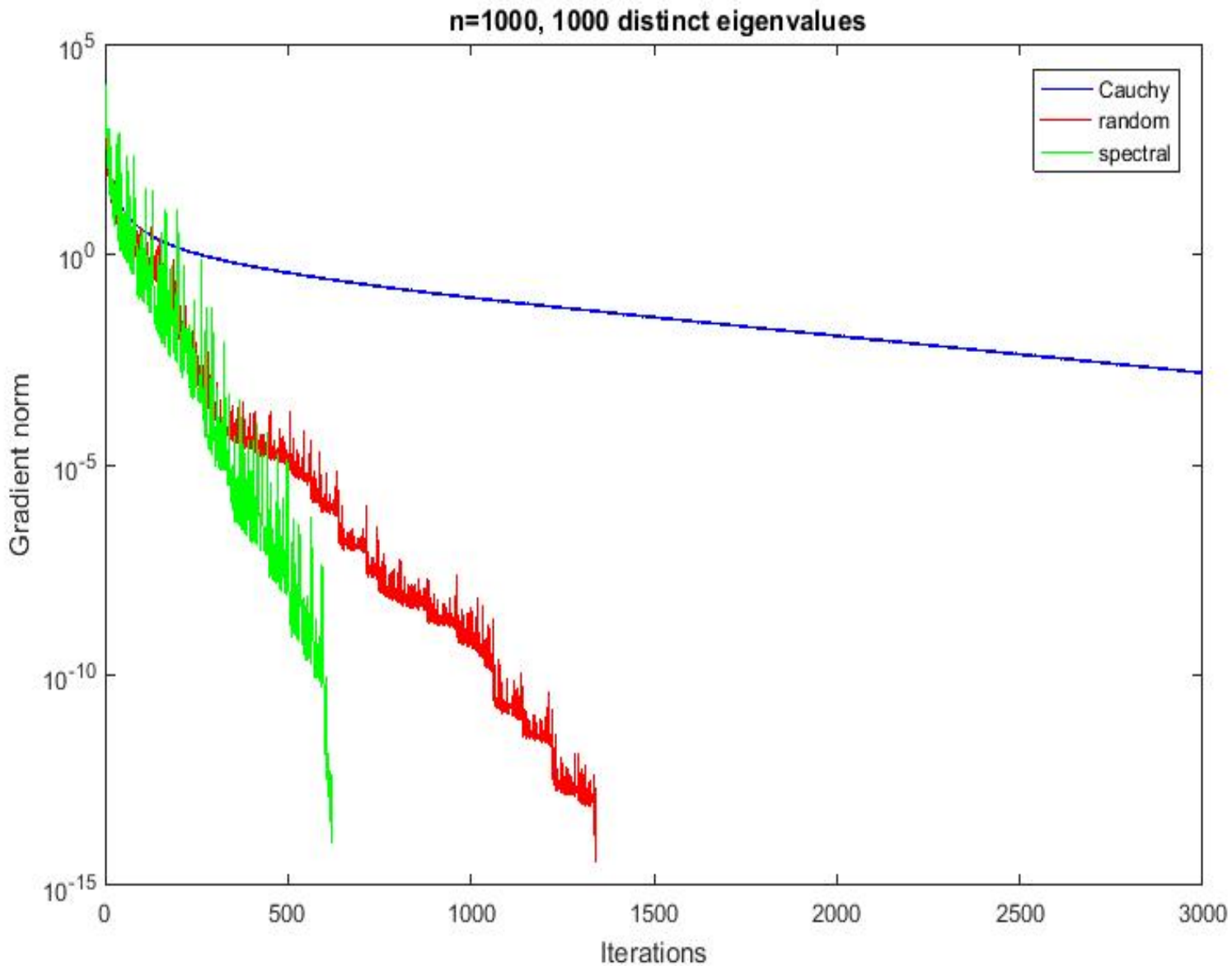
Convergence theorem for random variant [COAP, 2002]

If the sequence θ_k has an accumulation point $\bar{\theta} \in (0, 2)$ then the sequence $\{x_k\}$ generated by the random Cauchy gradient method converges to x^*.

Cauchy Vs. Random Cauchy



Cauchy Vs Random Cauchy Vs Spectral (B-B)



General case (non quadratic functions)

Problem:

$$\min_{x \in \mathbb{R}^n} f(x)$$

General family:

$$x_{k+1} = x_k + \lambda_k d_k,$$

where

$$d_k = -H_k^{-1} g_k,$$

H_k is nonsingular, symmetric, and approximates $\nabla^2 f(x_k)$.

Interesting cases:

$H_k = I$ (Cauchy and variants)

$H_k = \nabla^2 f(x)$ (Newton)

Quasi-Newton (secant): SR1, BFGS, DFP, etc. [Later]

Conjugate Gradients: FR, PR, PR+, etc. [Later]

Newton's method for non quadratic functions

Solve the following system for d_k

$$\nabla^2 f(x_k) d_k = -\nabla f(x_k),$$

and then set $x_{k+1} = x_k + d_k$.

Convergence result

If $\nabla^2 f(x^*)$ is nonsingular, $\nabla^2 f(x)$ is Lipschitz continuous in a neighborhood of x^*, and the initial guess x_0 is sufficiently close to x^*, then $\{x_k\}$ generated by Newton's method converges *q-quadratically* to x^*, i.e., there exist $c > 0$ and $\hat{k} \geq 0$ such that for all $k \geq \hat{k}$,

$$\|x_{k+1} - x^*\| \leq c\|x_k - x^*\|^2.$$

Variants of Newton's method

Modified Newton's method: Fix the Hessian at the first iteration and use it for all k

$$x_{k+1} = x_k - \nabla^2 f(x_0)^{-1} \nabla f(x_k).$$

Under the same hypothesis of Newton's convergence theorem, the iterations are well-defined and converges *q-linearly* to x^*.

Another variant: If the second derivatives of f are not available, or are too expensive to compute, they can be approximated by finite differences. A standard option is to approximate the jth column of $\nabla^2 f(x_k)$ by a forward difference quotient:

$$\nabla^2 f(x_k)_j \approx (\nabla f(x_k + h_k e_j) - \nabla f(x_k))/h_k,$$

where e_j denotes the jth unit vector and $h_k > 0$ is a suitable small number.

Newton's method with finite differences

Build an approximation A_k to the Hessian at x_k such that

$$(A_k)_j = (\nabla f(x_k + h_k e_j) - \nabla f(x_k))/h_k.$$

Convergence result

Under the same hypothesis of Newton's convergence theorem, if $0 < |h_k| \leq h$, then the sequence

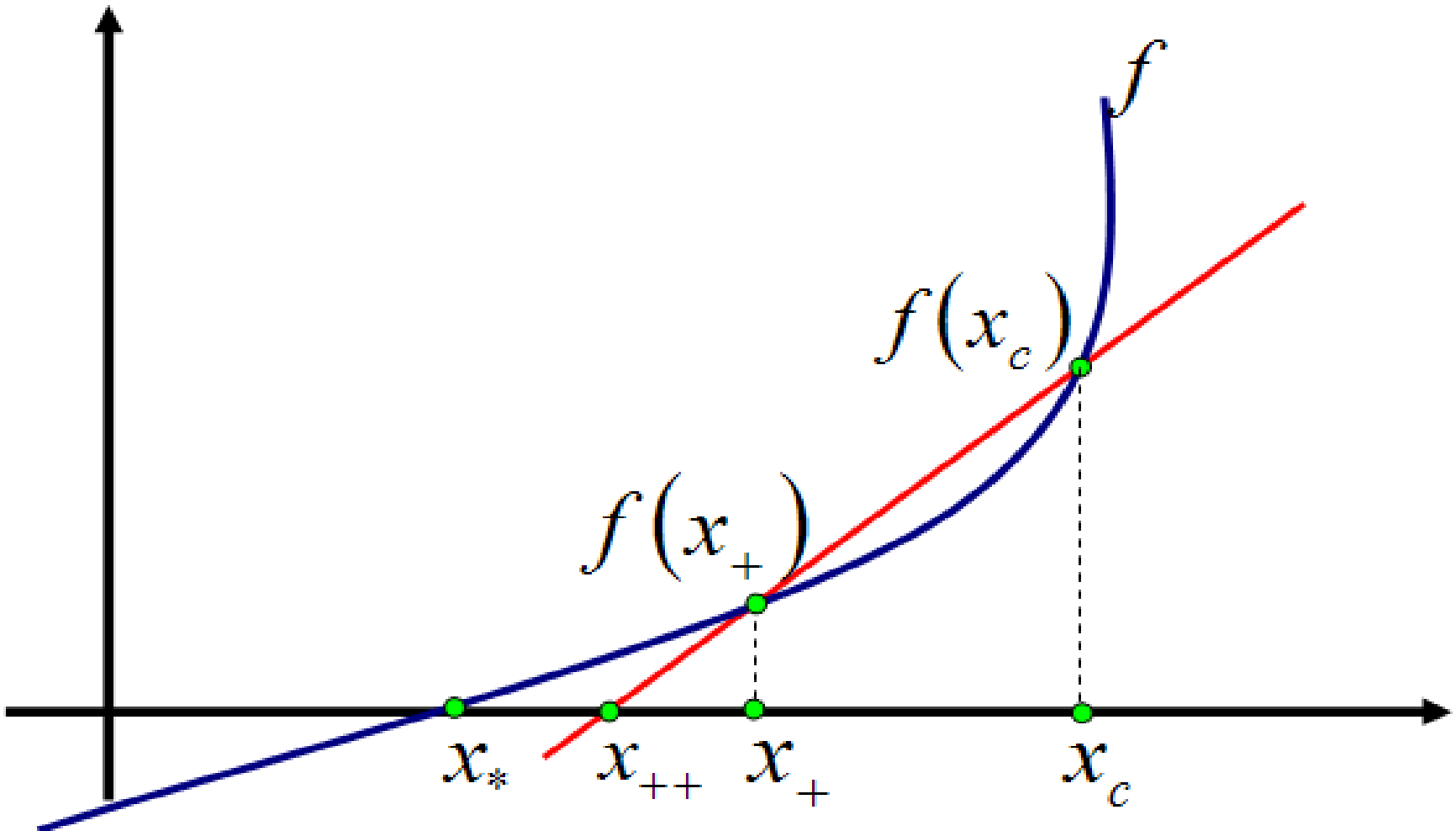
$$x_{k+1} = x_k - A_k^{-1} \nabla f(x_k),$$

where A_k is built using finite difference, is well-defined and converges q-linearly to x^*. Moreover, if $h_k \to 0$ then it converges q-superlinearly; and if $|h_k| \leq c_1 \|x_k - x_*\|$ or $|h_k| \leq c_2 \|\nabla f(x_k)\|$ for some $c_1 > 0$ and $c_2 > 0$, then it converges q-quadratically.

Notice that $(n+1)$ gradient evaluations are needed per iteration.

Quasi-Newton (or secant) methods

Secant for $n = 1$ (finding roots)



$$\hat{M}_+(x) = f(x_+) + a_+(x - x_+)$$

Secant for $n > 1$ (minimization)

If $n = 1$

$$x_{k+1} = x_k - \frac{f'(x_k)}{a_k} \quad \text{where} \quad a_k \approx f''(x_k),$$

given by $a_k = (f'(x_k) - f'(x_{k-1}))/(x_k - x_{k-1})$.

If we define $s_{k-1} = (x_k - x_{k-1})$, then

$$a_k(x_k - x_{k-1}) = a_k s_{k-1} = y_{k-1} \equiv f'(x_k) - f'(x_{k-1}).$$

If $n > 1$, we would like (secant equation):

$$A_k s_{k-1} = y_{k-1} \equiv g_k - g_{k-1}.$$

Given s_{k-1} and y_{k-1}, we need to build A_k.

n equations and n^2 unknowns $\Rightarrow$ infinitely many options !

Secant for $n > 1$ (Broyden, 1965)

Idea: Choose A_{k+1} as close as possible to A_k

Broyden's method (closest in Frobenius norm):

$$A_{k+1} = A_k + (y_k - A_k s_k) s_k^T / s_k^T s_k.$$

Satisfies the secant equation: $A_{k+1} s_k = y_k$.

Iterations: $x_{k+1} = x_k - A_k^{-1} \nabla f(x_k)$.

Local and q-superlinear convergence.

Second derivatives are not required.

Notice: A_{k+1} is equal to A_k plus a rank-one update.

Bad news: Symmetry is not preserved !

Convergence of quasi-Newton methods

Fact: $\{A_k\}$ does not converge to $H(x^*)$. However, they work!

First ingredient: (using any induced norm)

Definition: Bounded Deterioration (BD)

Let $H(x^*)$ be the Hessian matrix at the solution, $L > 0$ the Lipschitz constant for $H(x)$, and $e_k = x_k - x^*$. A secant (or quasi-Newton) method has the BD property if for all k,

$$\|A_{k+1} - H(x^*)\|_F \leq \|A_k - H(x^*)\|_F + \frac{L}{2}(\|e_k\|_2 + \|e_{k+1}\|_2).$$

Theorem

Any quasi-Newton method that has the BD property (e.g., Broyden's method), converges locally and q-linearly to x^*.

Convergence of quasi-Newton methods

Second ingredient: (using any induced norm)

Dennis-Moré (DM) condition (1973)

A quasi-Newton method has the DM condition if

$$\lim_{k\to\infty} \frac{\|(A_k - H(x^*))s_k\|}{\|s_k\|} = 0.$$

Theorem

Any quasi-Newton method that converges at least q-linearly (e.g., Newton and Broyden), and has the DM condition, converges q-superlinearly to x^*.

Moreover, any quasi-Newton method converges q-superlinearly if and only if it has the DM condition.

Secant for $n > 1$ (PSB)

Idea: Choose A_{k+1} symmetric and as close as possible to A_k.

PSB Method (limit of an alternating projection scheme):

$$A_{k+1} = A_k + \frac{(y_k - A_k s_k)s_k^T + s_k(y_k - A_k s_k)^T}{s_k^T s_k} - \frac{(y_k - A_k s_k)^T s_k}{(s_k^T s_k)^2}(s_k s_k^T).$$

Satisfies the secant equation.

Local and q-superlinear convergence.

It preserves symmetry.

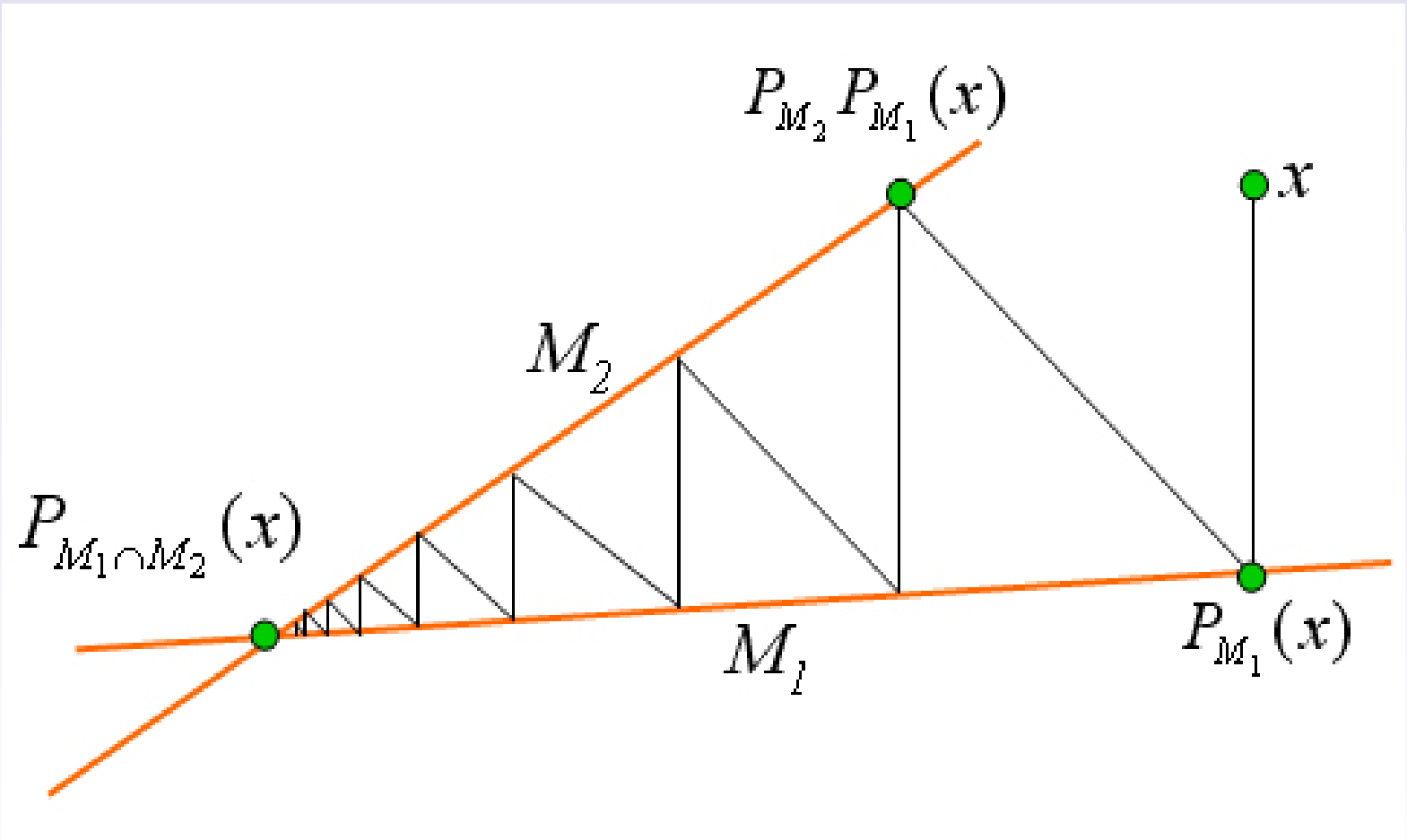
Notice: A_{k+1} is equal to A_k plus a rank-two update.

Bad news: PD is not preserved !

Alternating projection scheme for PSB

Set $M_1 = \{B : Bs_k = y_k\}$ and $M_2 = \{B : B = B^T\}$.

In Figure (below): $x = A_k$ and $P_{M_1 \bigcap M_2}(x) = A_{k+1}$.



Secant for $n > 1$ (BFGS and SR1)

Choose A_{k+1} symmetric and PD, and as close as possible to A_k.

BFGS Method (rank-two update):

$$A_{k+1} = A_k + \frac{y_k y_k^T}{y_k^T s_k} - \frac{(A_k s_k)(A_k s_k)^T}{s_k^T A_k s_k}.$$

SR1 Method (rank-one update):

$$A_{k+1} = A_k + \frac{(y_k - A_k s_k)(y_k - A_k s_k)^T}{(y_k - A_k s_k)^T s_k}.$$

Satisfy the secant equation and preserve symmetry.

Local and q-superlinear convergence.

Preserve PD (if $y_k^T s_k > 0$ for BFGS and $(y_k - A_k s_k)^T s_k > 0$ for SR1).

Bad news (for n large): They do not preserve the sparsity structure !

Secant: General Algorithm

Algorithm

1: Given $A_0 \in \mathbb{R}^{n \times n}$, $x_0 \in \mathbb{R}^n$
2: **for** $k = 0, 1, \cdots$ until convergence **do**
3: **Solve** $A_k s_k = -g_k$
4: **Set** $x_{k+1} = x_k + s_k$
5: **Set** $y_k = g_{k+1} - g_k$
6: **Build** $A_{k+1} = A_k + \textit{low rank update}$
7: **end for**

If we use Line Search globalization, Step 4 changes:
$x_{k+1} = x_k + \lambda_k s_k$,

where $\lambda_k > 0$ satisfies certain descent conditions (next topic!).

Line Search (LS) globalization strategies

Line Search (LS) globalization strategies

Let us recall our iteration to find local minimizers:

$$x_{k+1} = x_k + \lambda_k d_k,$$

where d_k is a descent direction (always possible):

$$d_k^T g_k < 0.$$

The steepness (quality) of the direction d_k is measured by

$$0 < \cos\theta_k = \frac{-g_k^T d_k}{\|g_k\| \, \|d_k\|} \leq 1.$$

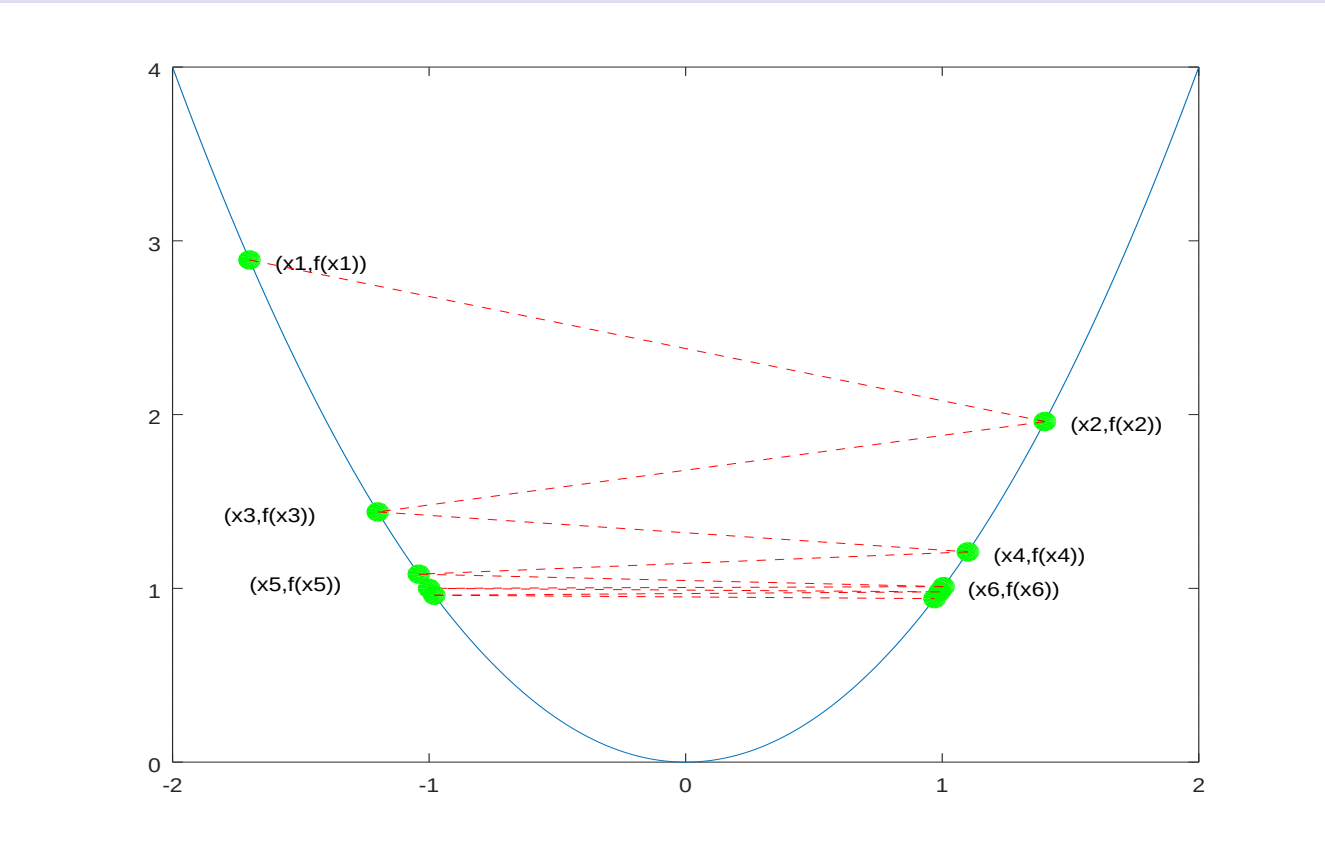
The computation of the step length $\lambda_k > 0$ is the line search, and may itself be an iteration.

Is it enough to choose $\lambda_k > 0$ such that

$$f(x_k + \lambda_k d_k) < f(x_k)?$$

Line Search (LS) globalization strategies

Not enough! Here, the iterates oscillate far from the minimizer

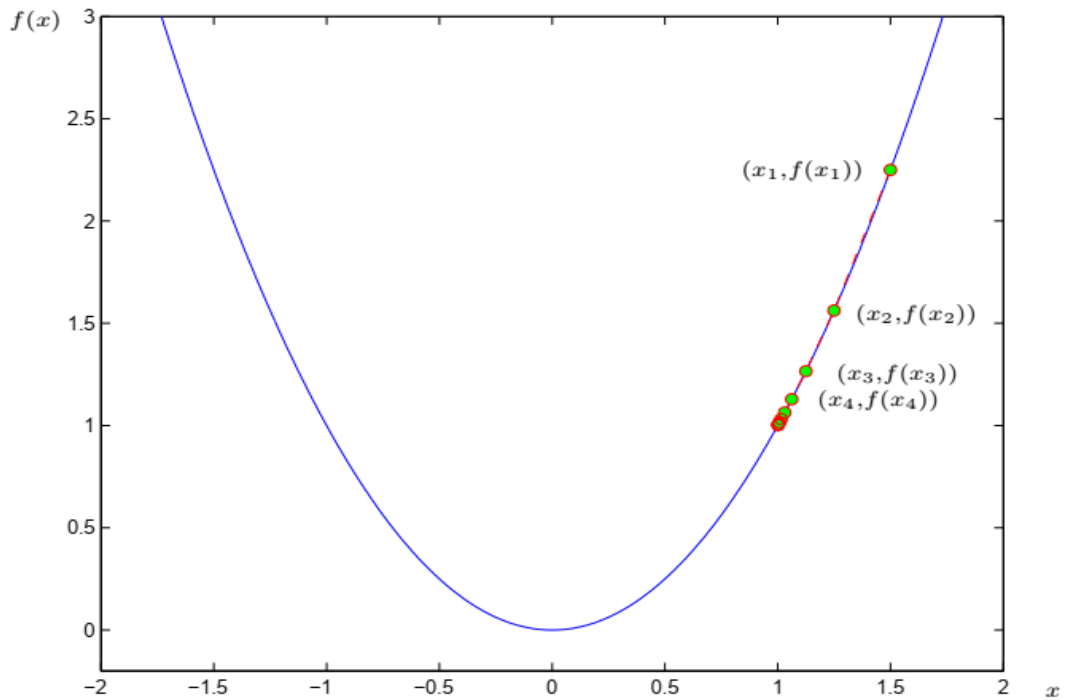


$f(x) = x^2, \quad x_0 = 2, \quad d_k = (-1)^{k+1}, \quad$ and $\quad \lambda_k = 2 + \frac{3}{2^{k+1}}$.

The steps are too long for the small decrease that they provide.

Line Search (LS) globalization strategies

Not Enough to avoid long steps!
Here, the iterates converge to the non-stationary value 1.



$f(x) = x^2, \quad x_0 = 2, \quad d_k = -1, \quad$ and $\lambda_k = \frac{1}{2^{k+1}}$.

The steps are too short for such a good descent direction.

Line Search (LS) globalization strategies

Bottom line:
A "naive" line search method can fail if it allows steps that are either too long or too short relative to the amount of decrease that might be obtained along a good quality direction.

How to relate a sufficient decrease with the direction?

Key option: Armijo condition

$$f(x_k + \lambda_k d_k) \quad \leq \quad f(x_k) + \alpha \lambda_k g_k^T d_k$$

where $\alpha > 0$ is chosen (fixed) small (e.g., $\alpha = 10^{-4}$)

The step is asked to give more than simply decrease in f.

Since $g_k^T d_k < 0$, the longer the step λ_k the larger the required decrease in f.

Line Search (LS) globalization strategies

The Armijo condition avoids steps that are too long.

How can we avoid also the ones that are too short?

Key option: Armijo-Goldstein (AG) condition

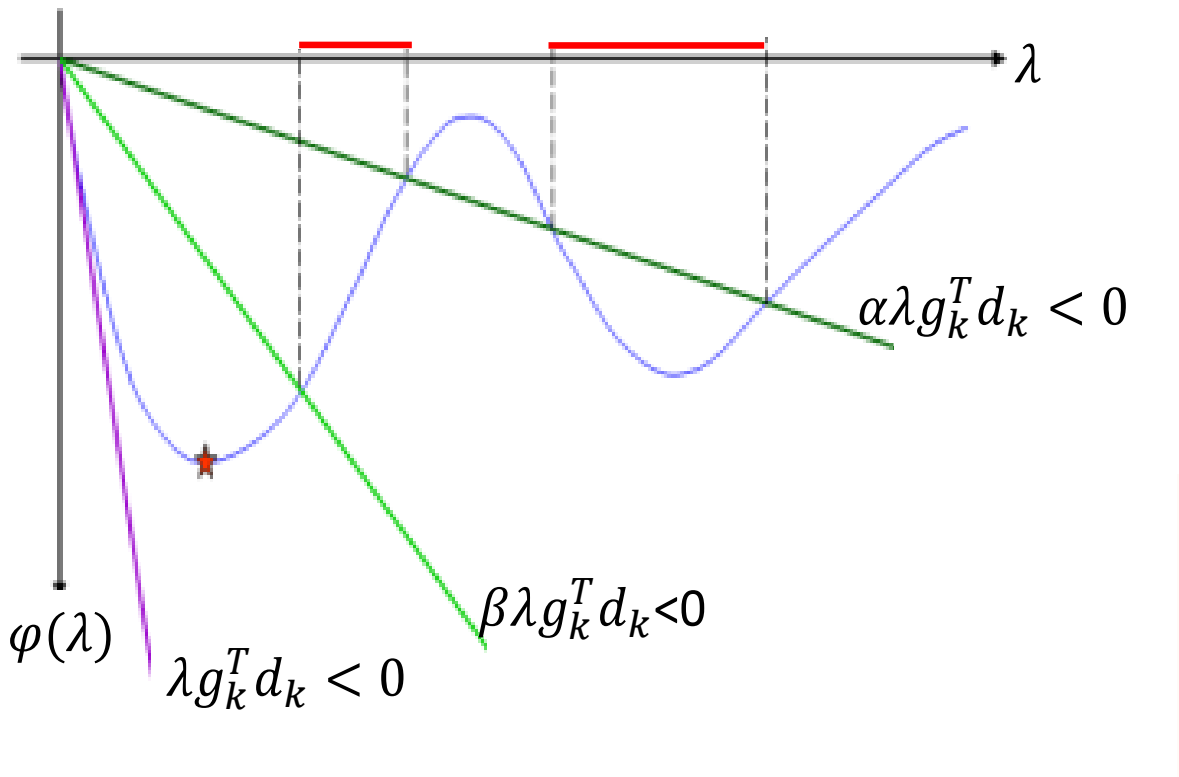
$$\beta\lambda_k g_k^T d_k \;\leq\; f(x_k+\lambda_k d_k)-f(x_k) \;\leq\; \alpha\lambda_k g_k^T d_k$$

where $0<\alpha<\beta<1$ (e.g., $\alpha=10^{-4}$ and $\beta=0.95$).

Define $\varphi(\lambda)=f(x_k+\lambda d_k)-f(x_k)$. Note: $\varphi'(0)=g_k^T d_k$

The slope of $\varphi(\lambda)$ at $\lambda=0$ is $g_k^T d_k<0$.

Line Search (LS) globalization strategies



The step lengths accepted by the AG condition are in red.

Line Search (LS) globalization strategies

Under standard assumptions it is always possible to find an interval for which the AG conditions are satisfied.

Lemma

Let $f : \mathbb{R}^n \to \mathbb{R}$, $f \in C^1(\mathbb{R}^n)$, $x \in \mathbb{R}^n$ and $d \in \mathbb{R}^n$ such that $\nabla f(x)^T d < 0$, $m \in \mathbb{R}$ such that $f(x + \lambda d) \geq m$ for all $\lambda > 0$, and $0 < \alpha < \beta < 1$. Then there exist $0 < \lambda_l < \lambda_u$ such that $(x + \lambda d)$ satisfies the AG conditions for $\lambda \in (\lambda_l, \lambda_u)$.

Two key questions:

How do we find λ_k that satisfies the AG conditions?

If we find it at each k, Does it guarantee global convergence?

Line Search (LS) globalization strategies

Practical option:

Backtracking strategy (based on Armijo condition)

Backtracking Algorithm

1: **Given** $d_k \in \mathbb{R}^n$, $\alpha \in (0,1)$, and $0 < l < u < 1$
2: **Set** $\lambda = \lambda_{init}$ (e.g. $\lambda_{init} = 1$)
3: **While** $f(x_k + \lambda d_k) > f(x_k) + \alpha \lambda g_k^T d_k$ **do**
4: $\lambda := \rho\lambda, \quad \rho \in [l, u]$
5: **End While**
6: **Set** $\lambda_k = \lambda$
7: **Set** $x_{k+1} = x_k + \lambda_k d_k$

Notice: Long steps are avoided by the Armijo condition
Short steps are avoided since the first allowable λ is accepted.

Typical values: $\alpha = 10^{-4}, \quad l = 0.1, \quad$ and $\quad u = 0.9$

Global convergence?

Key Lemma

Let f be bounded below in $\mathbb{R}^n$ and $C^1(\mathcal{N})$, where $\mathcal{N}$ is an open set that contains the level set $\mathcal{L} = \{x \ : \ f(x) \leq f(x_0)\}$. Let us assume that the gradient of f is Lipschitz, i.e., there exists $L > 0$ such that

$$\|g(x) - g(z)\|_2 \leq L\|x - z\|_2,$$

for all x and z in $\mathcal{N}$. Consider any iterative method of the form $x_{k+1} = x_k + \lambda_k d_k$, where d_k is a descent direction, and λ_k satisfies the AG conditions. Then,

$$\sum_{k \geq 1} \cos^2 \theta_k \ \|g_k\|_2^2 < \infty.$$

Line Search (LS) globalization strategies

Key Lemma (roughly speaking)

Under standard assumptions on f, if we use a method of the form $x_{k+1} = x_k + \lambda_k d_k$, where d_k is a descent direction, and λ_k satisfies the AG conditions. Then,

$$\sum_{k \geq 1} \cos^2 \theta_k \, \|g_k\|_2^2 < \infty.$$

If $\cos \theta_k \geq \rho > 0$, for all k, then

$$\lim_{k \to \infty} \|g_k\|_2 = 0.$$

If the d_k's do not tend to be orthogonal to the g_k's
$\Rightarrow$ convergence.

Convergence for special cases (LS)

Cauchy (variants) + AG: $\cos\theta_k = 1$
$\Rightarrow$ convergence.

Newton (quasi-Newton): $d_k = -H_k^{-1} g_k$.

Enough to force H_k to be PD, and $\kappa(H_k) \leq M$ for all k,

$$\kappa(H_k) = \lambda_{\max}(H_k)/\lambda_{\min}(H_k) \leq M.$$

$\Rightarrow$ convergence.

It can be accomplished by using a (dynamically) modified Cholesky factorization of H_k

Least-squares problems

Review: Linear least-squares problems

First, consider the linear case:

$$Ax = b,$$

but $b \notin range(A)$, e.g., $A \in \mathbb{R}^{m \times n}$, $m > n$ (No solutions!)

$$\min_{x \in \mathbb{R}^n} f(x),$$

$$f(x) = \frac{1}{2}\|Ax - b\|_2^2 = \frac{1}{2}(Ax - b)^T(Ax - b).$$

$$\nabla f(x) = A^T(Ax - b), \qquad \nabla^2 f(x) = A^T A$$

Note that: $A^T b \in range(A^T A)$; and $A^T A$ is P (semi) D.

Equivalent: Solve the Normal Equations

$$A^T Ax = A^T b$$

Nonlinear least-squares problems

Let $F : \mathbb{R}^n \to \mathbb{R}^m$, we want x^* that solves

$$\min_{x \in \mathbb{R}^n} f(x),$$

where

$$f(x) = \frac{1}{2}\|F(x)\|_2^2.$$

If the nonlinear system has a (or many) solution (s) x^*, then $F(x^*) = 0$, and so $f(x^*) = 0$.

If the nonlinear system has no solutions, we want $\nabla f(x^*) = 0$.

$$\nabla f(x) = J(x)^T F(x)$$

Notation: $J(x)$ is the Jacobian matrix of F at x.

Nonlinear least-squares problems

General method: $x_{k+1} = x_k + \lambda_k d_k$.

Cauchy:

$$d_k^C = -\nabla f(x_k) = -J(x_k)^T F(x_k),$$

Newton:

$$d_k^N = -\nabla^2 f(x_k)^{-1} \nabla f(x_k)$$

$$d_k^N = -[J(x_k)^T J(x_k) + T(x_k)]^{-1} J(x_k)^T F(x_k),$$

where

$$T(x_k) = \sum_{i=1}^{n} F_i(x_k) \nabla^2 F_i(x_k).$$

Observation: If $F(x^*) \approx 0$, close to $x^* \to T(x_k) \approx 0$.

Gauss-Newton:

$$d_k^{GN} = -(J(x_k)^T J(x_k))^{-1} J(x_k)^T F(x_k).$$

Theorem

If $J(x)$ has full-column rank, then d^C and d^{GN} are descent directions for $f(x) = \frac{1}{2}\|F(x)\|_2^2$.

Proof:

$$\nabla f(x_k)^T d_k^C = -F(x_k)^T J(x_k) J(x_k)^T F(x_k) = -\|J(x_k)^T F(x_k)\|_2^2 < 0$$

$$\nabla f(x_k)^T d_k^{GN} = -(F(x_k)^T J(x_k))(J(x_k)^T J(x_k))^{-1}(J(x_k)^T F(x_k)) < 0$$

Recall that if $J(x)$ has full-column rank, $J(x_k)^T J(x_k)$ is PD

Improvement on Gauss-Newton: Levenberg-Marquardt (LM)

The vector d_k^{LM} satisfies the linear system:

$$(J(x_k)^T J(x_k) + \gamma_k I) d_k^{LM} = -J(x_k)^T F(x_k),$$

for some $\gamma_k \geq 0$.

If $\gamma_k = 0$, $d_k^{LM} = d^{GN}$ (aggressive).

If $\gamma_k \to +\infty$, $d_k^{LM} \to d^C$ (conservative).

γ_k is chosen according to the previous reduction in $f(x)$

Low-cost methods for large n

$$\min_{x \in \mathbb{R}^n} f(x)$$

Low-cost methods for large n

Spectral gradient: $x_{k+1} = x_k - \lambda_k \nabla f(x_k)$

Recall the *secant equation*,

$$A_{k+1} s_k = y_k$$

Forcing $A_{k+1} = \alpha_{k+1} I$ and minimizing $\|\alpha_{k+1} s_k - y_k\|_2^2$:

$$\alpha_{k+1}^{S1} = \frac{s_k^T y_k}{s_k^T s_k}$$

Finally:

$$x_{k+1} = x_k - \frac{1}{\alpha_k^{S1}} \nabla f(x_k)$$

If f is a convex quadratic S1 has global convergence (1993):

$$\alpha_{k+1}^{S1} = \frac{s_k^T A s_k}{s_k^T s_k} = \frac{g_k^T A g_k}{g_k^T g_k}$$

Low-cost methods for large n

A second Spectral option: Minimize $\|s_k - \alpha_{k+1}^{-1} y_k\|_2^2$

$$\alpha_{k+1}^{S2} = \frac{y_k^T y_k}{s_k^T y_k}$$

Finally:

$$x_{k+1} = x_k - \frac{1}{\alpha_k^{S2}} \nabla f(x_k)$$

If f is a convex quadratic S2 has global convergence (1993):

$$\alpha_{k+1}^{S2} = \frac{s_k^T A^2 s_k}{s_k^T A s_k} = \frac{g_k^T A^2 g_k}{g_k^T A g_k}$$

Fact for all k (2002): $\lambda_{\min}(A) \leq \alpha_k^{S1} \leq \alpha_k^{S2} \leq \lambda_{\max}(A)$

Nonmonotone LS for the spectral gradient method

$$x_{k+1} = x_k - \lambda_k \nabla f(x_k)$$

$$\lambda_k^{S1} = \frac{s_{k-1}^T s_{k-1}}{s_{k-1}^T y_{k-1}} \quad \text{or} \quad \lambda_k^{S2} = \frac{s_{k-1}^T y_{k-1}}{y_{k-1}^T y_{k-1}}$$

They are r-linear (nonmonotone), and so a special LS needs to be applied until the next iterate satisfies:

$$f(x_{k+1}) \leq \max_{0 \leq j \leq M} f(x_{k-j}) - \gamma \lambda_k g_k^T g_k + \eta_k$$

where $M > 1$ is a positive integer, $0 < \gamma < 1$, $\eta_k > 0$ for all k, and $\sum_{k \geq 1} \eta_k < \infty$.

Global convergence [1997, 2008]: $\lim_{k \to \infty} \|g_k\| = 0$.

Low-cost methods for large n

An adaptive Spectral option (2006, 2018): ABBmin

$$x_{k+1} = x_k - \lambda_k^{AS} \nabla f(x_k)$$

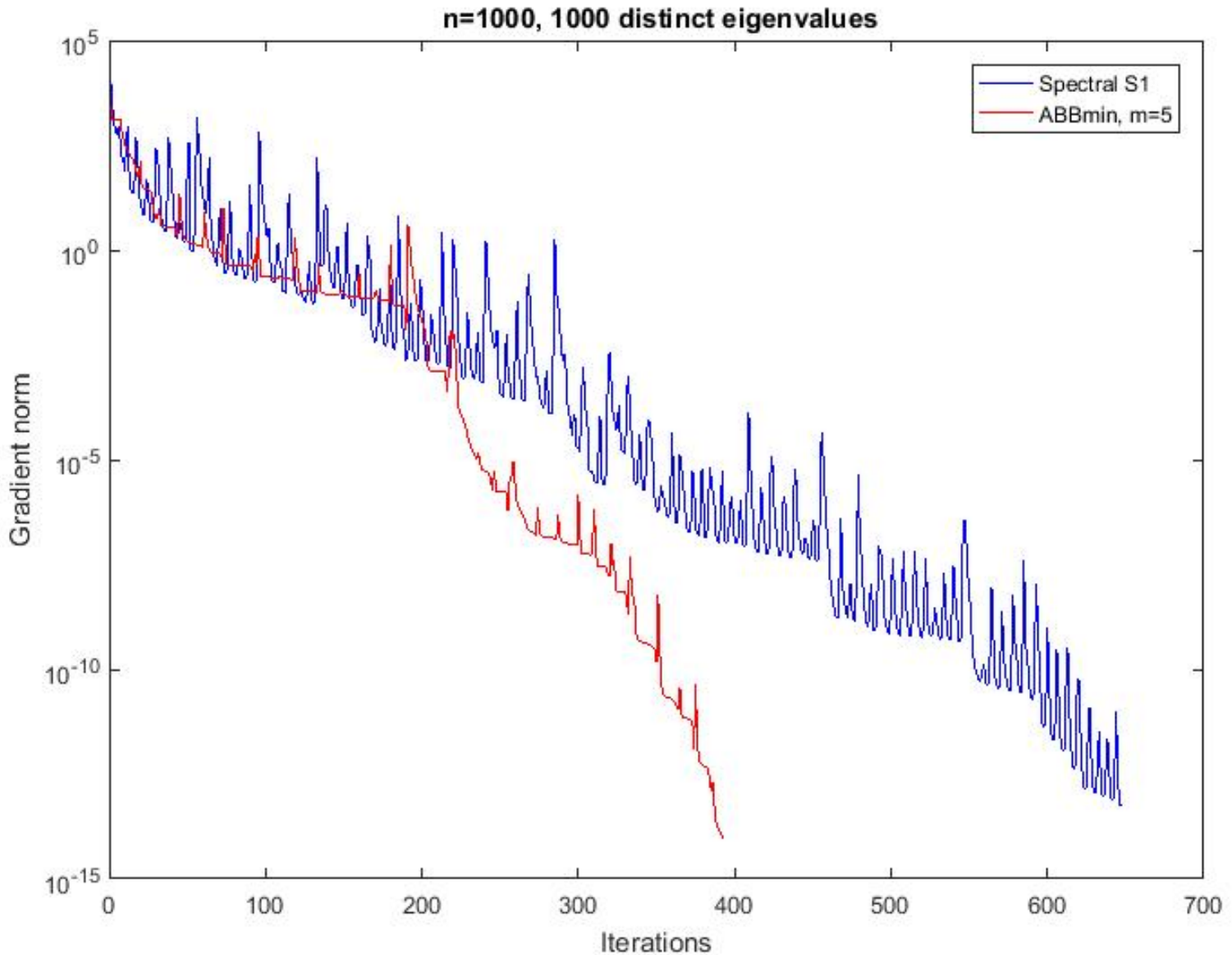
$$\lambda_k^{AS} = \begin{cases} \min\{\lambda_j^{S2} \ : \ j = \max\{1, k-m\}, \ldots, k\}, & \text{if} \quad \lambda_k^{S2}/\lambda_k^{S1} < \tilde{\tau} \\ \lambda_k^{S1}, & \text{else} \end{cases}$$

where $\tilde{\tau} \in (0,1)$ (in practice $\tilde{\tau} \approx 0.8$), m is a positive integer.

For convex quadratics, ABBmin has global convergence

For general functions it can be embedded into the previous Nonmonotone LS strategy (2018), and $\lim_{k\to\infty} \|g_k\| = 0$

ABBmin Vs Spectral



Minimization on convex sets

$$\min_{x \in \Omega} f(x)$$

Ω closed and convex

Optimality conditions on a set Ω

Definition: Given $x^* \in \Omega$, we say that ϕ is a feasible curve in Ω starting at x^* if it is continuously differentiable and

$$\phi : [0, t) \to \Omega \quad \text{for } t > 0, \quad \text{such that } \phi(0) = x^*$$

First order necessary condition:

If $x^* \in \Omega$ is a local minimizer and ϕ is a feasible curve in Ω starting at x^*, then

$$\nabla f(x^*)^T \phi'(0) \geq 0.$$

Let $\gamma(t) = f(\phi(t))$. Note: $\gamma'(0) \geq 0$ (i.e., f is increasing along ϕ).

$$0 \leq \gamma'(0) = \nabla f(\phi(t))\phi'(t)|_0 = \nabla f(x^*)^T \phi'(0).$$

Optimality conditions on convex sets

The set Ω is **closed** and **convex**, and so

If $x^* \in \Omega$ and $z \in \Omega$, for $t \in [0, 1]$

$$tz + (1 - t)x^* \in \Omega.$$

Then $\phi(t) = x^* + t(z - x^*)$ is a feasible curve, $t \in [0, 1]$. The condition $\nabla f(x^*)^T \phi'(0) \geq 0$ reduces to

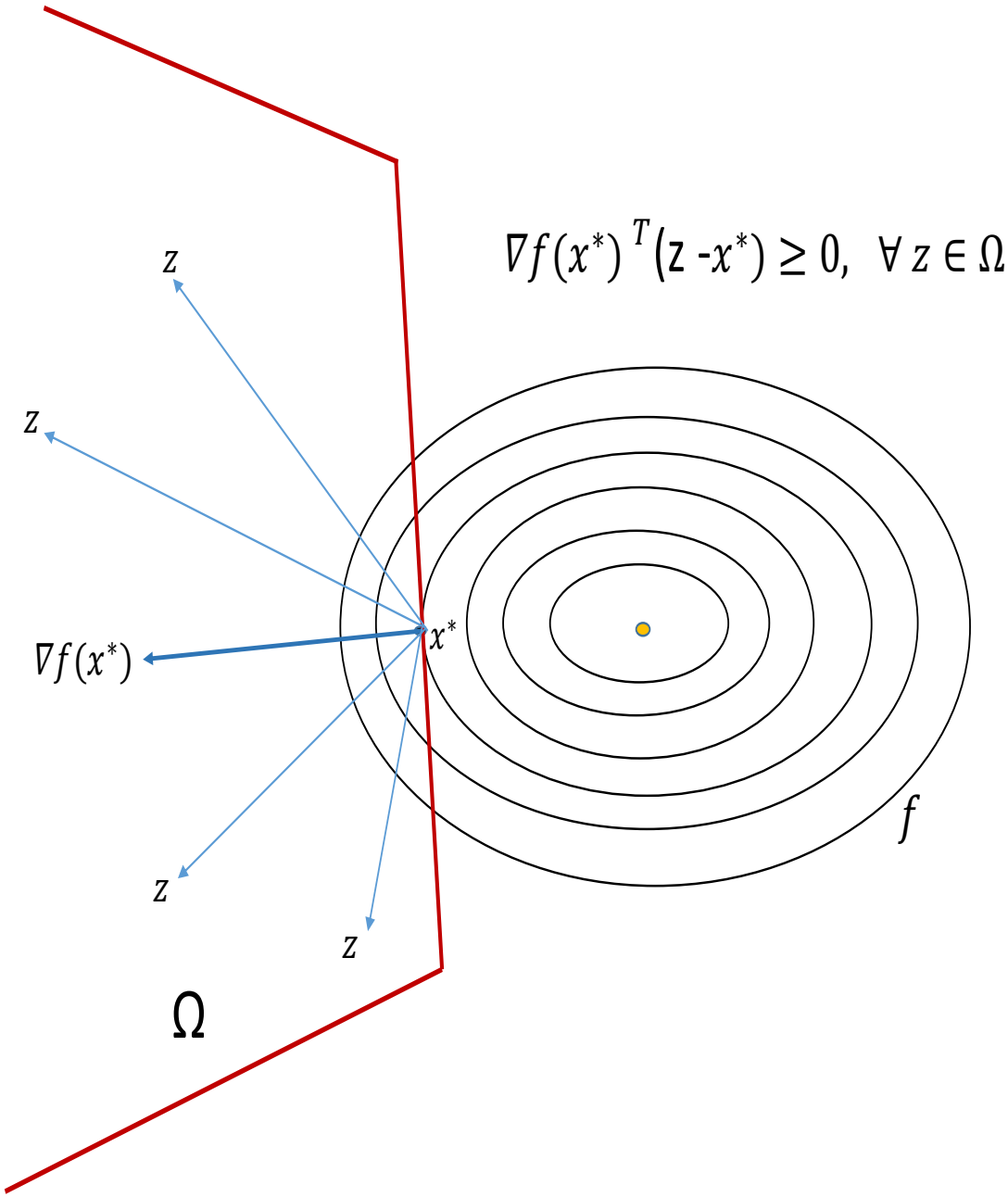
First order necessary condition:

$$\nabla f(x^*)^T (z - x^*) \geq 0 \quad \text{for all } z \in \Omega$$

If f is **strictly convex**, this condition is also **sufficient**.

Geometrical interpretation: **all feasible vectors** starting at x^* must form an angle of 90°, or less, with the gradient vector.

Optimality conditions on convex sets



Projecting onto convex sets

Problem: Given $x \in \mathbb{R}^n$ find $x^* \in \Omega$ such that

$$P_\Omega(x) \equiv x^* = \text{argmin}_{z \in \Omega} \|z - x\|_2.$$

Theorem

(a) For all $x \in \mathbb{R}^n$ there exists a unique $x^* = P_\Omega(x)$.

(b) (Kolmogorov) $x^* = P_\Omega(x)$ if and only if for all $z \in \Omega$

$$(x - x^*)^T (z - x^*) \leq 0.$$

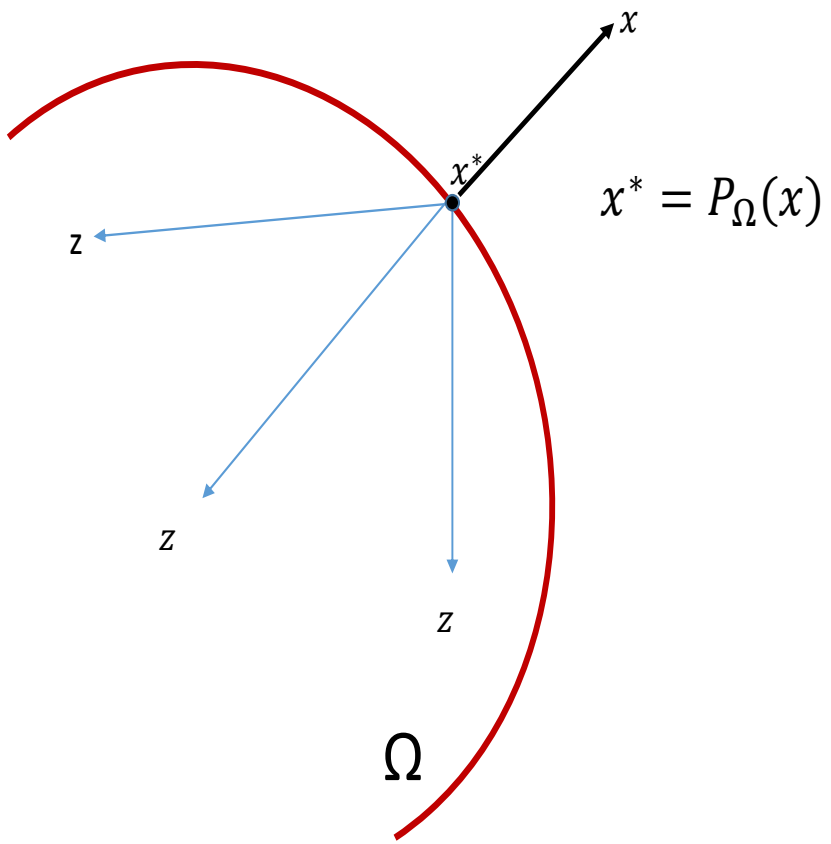
(c) (Non expansive) $\|P_\Omega(x) - P_\Omega(y)\|_2 \leq \|x - y\|_2$

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

Projecting onto convex sets

Kolmogorov Condition

$$(x - x^*)^T (z - x^*) \leq 0, \qquad \forall z \in \Omega$$



It is easy to project onto some important convex sets

Box: $B = \{x \in \mathbb{R}^n : l_i \leq x_i \leq u_i \text{ for all } i\}$,

where l and u are given vectors in $\mathbb{R}^n$.

$$(P_B(y))_i = \begin{cases} y_i & \text{if} \quad l_i \leq y_i \leq u_i \\ u_i & \text{if} \quad y_i > u_i \\ l_i & \text{if} \quad y_i < l_i. \end{cases}$$

Sphere: $C = \{x \in \mathbb{R}^n : \|x - a\|_2 \leq r\}$,

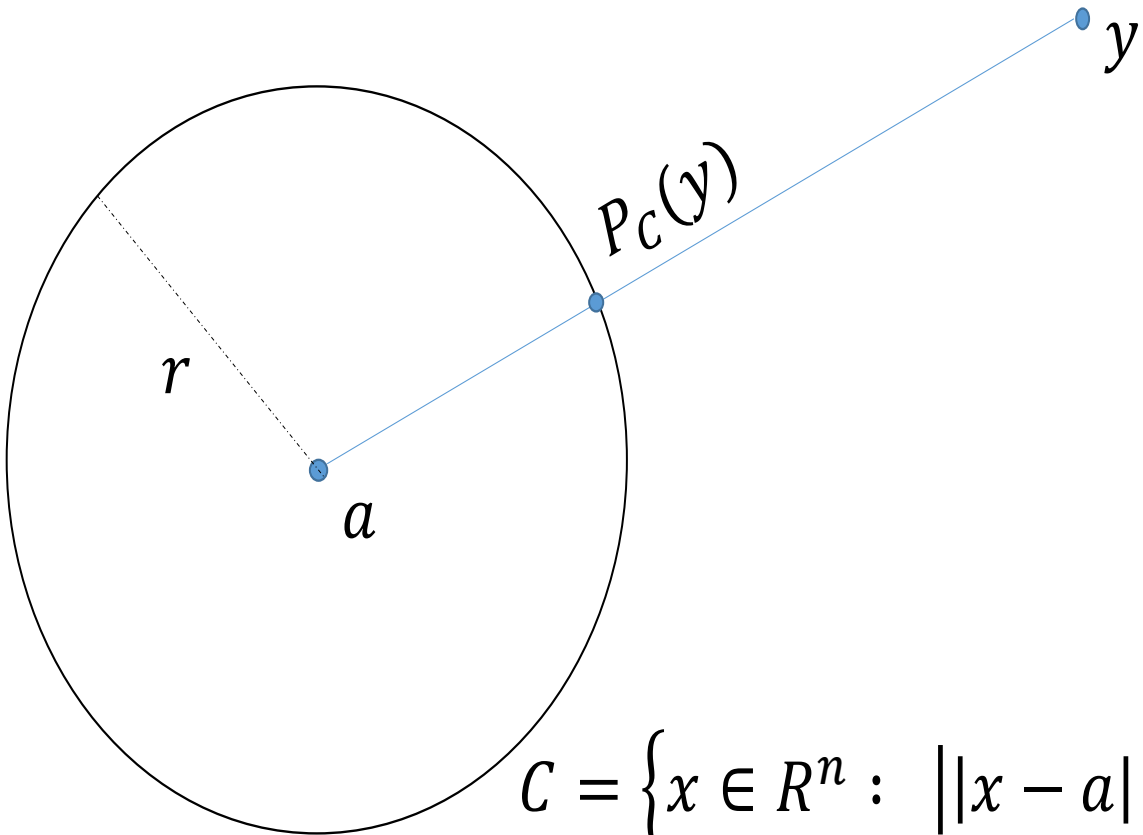
where $a \in \mathbb{R}^n$ is the center and $r > 0$ is the radius.

$$\text{if } \|y - a\|_2 > r, \qquad P_C(y) = a + \frac{r}{\|y - a\|_2}(y - a)$$

Ellipse: $\Omega = \{x \in \mathbb{R}^n : x^T A x \leq r\}$,

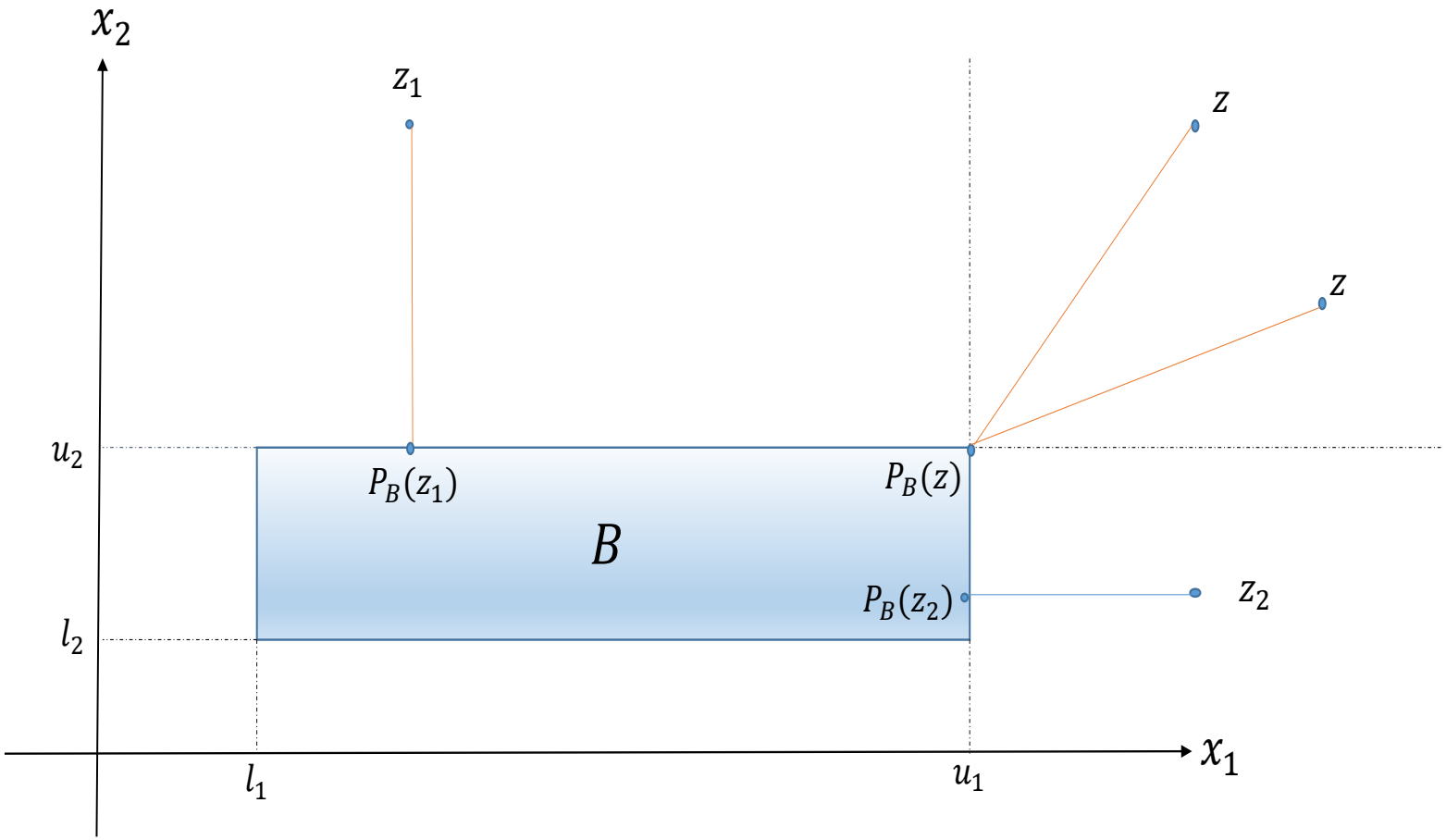
where A is symmetric and PD (not so easy!).

Projecting onto a sphere



$$C = \left\{ x \in R^n : \ \left|\left|x - a\right|\right|_2 \leq r \right\}$$

Projecting onto a box



$$B = \{x \in R^n : l \leq x \leq u\}$$

Projected Gradient-type (PG) Method

The natural extension of Cauchy Method:

$$x_{k+1} = P_\Omega(x_k - \lambda_k \nabla f(x_k))$$

where, $P_\Omega(z)$ denotes the projection of z onto Ω, and

$$\lambda_k = \text{argmin}_{\lambda>0} f(P_\Omega(x_k - \lambda \nabla f(x_k))).$$

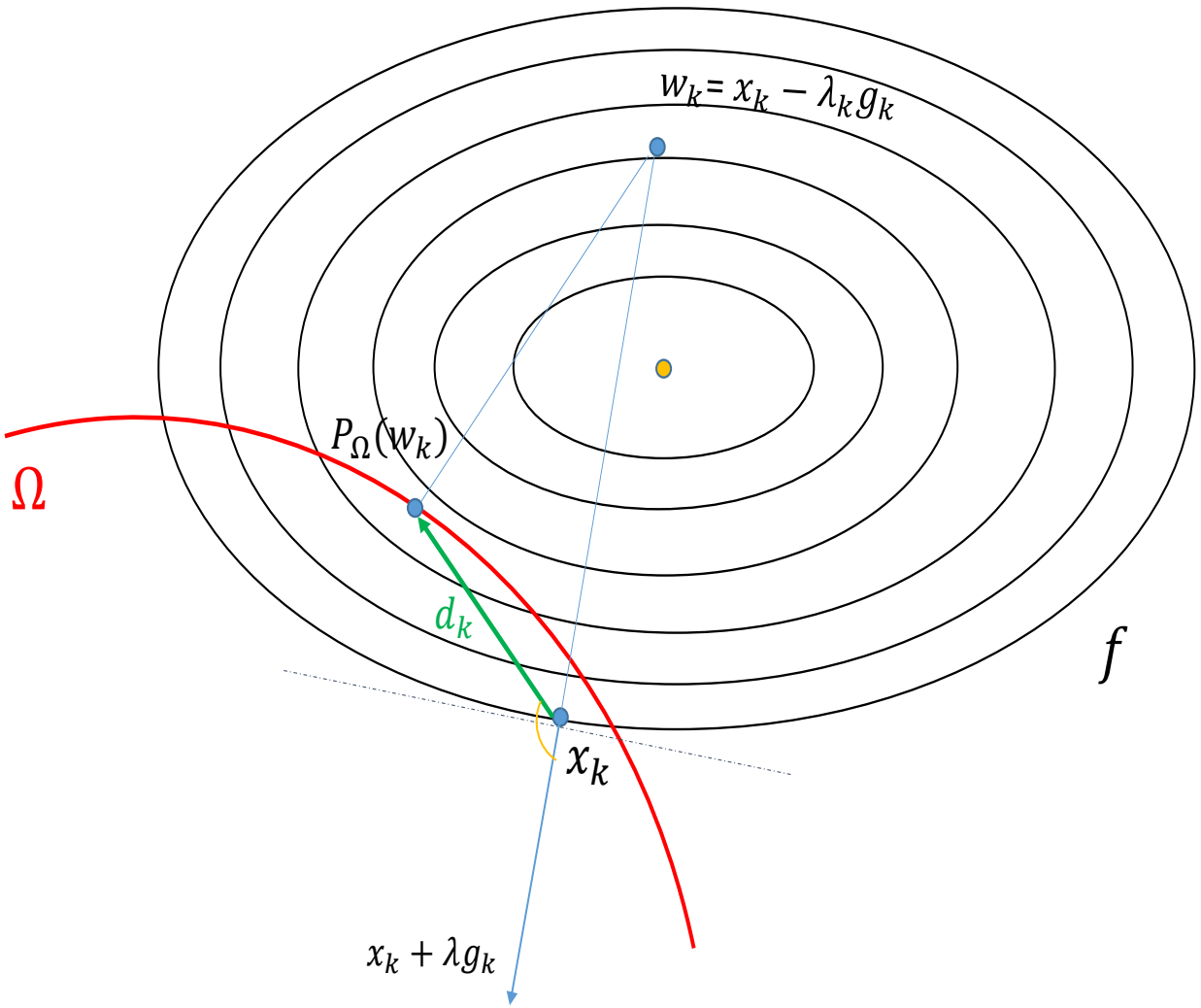
Another option (both are descent directions):

$$x_{k+1} = x_k + \alpha_k(P_\Omega(x_k - \lambda_k \nabla f(x_k)) - x_k)$$

Step length λ_k can be chosen as spectral S1 or S2, ABBmin...)

$0 < \alpha_k \leq 1$ is chosen using backtracking (e.g., non monotone)

Continuous Projected Gradient-type Algorithms



Practical optimality condition on convex sets

Recall the optimality condition:

$$\nabla f(x^*)^T(z - x^*) \geq 0 \quad \text{for all } z \in \Omega.$$

Using Kolmogorov's characterization we can write it as a practical condition (computable):

Take the one above, change the sign, multiply by $\lambda > 0$, add and subtract x^*, and we have for all $z \in \Omega$

$$(x^* - \lambda \nabla f(x^*) - x^*)^T(z - x^*) \leq 0.$$

Using now Kolmogorov characterization we get the practical optimality condition

$$x^* = P_\Omega(x^* - \lambda \nabla f(x^*)).$$

Algorithm PG (to get x_{k+1} and λ_{k+1})

Given $x_k \in \mathbb{R}^n$, $0 < tol \ll 1$, $0 < \gamma \ll 1$, $\eta_k > 0$ s.t. $\sum_{k\geq 1} \eta_k < \infty$.

Projected Gradient-type Algorithm (Non-monotone LS backtracking)

1: **If** $\|P_\Omega(x_k - \nabla f(x_k)) - x_k\|_2 \leq tol$, Stop at x_k

2: **Set** $d_k = P_\Omega(x_k - \lambda_k \nabla f(x_k)) - x_k$, and $\alpha = 1$

3: **Set** $x_+ = x_k + \alpha d_k$

4: **While** $f(x_+) > \max_{0\leq j\leq \min\{k,M-1\}} f(x_{k-j}) + \gamma\alpha d_k^T \nabla f(x_k) + \eta_k$

5: **Choose** $\alpha_{new} \in [0.1\alpha, 0.9\alpha]$

6: **Set** $\alpha = \alpha_{new}$ and $x_+ = x_k + \alpha d_k$

7: **End While**

8: **Set** $\alpha_k = \alpha$, $x_{k+1} = x_+$, $s_k = x_{k+1} - x_k$, $y_k = \nabla f(x_{k+1}) - \nabla f(x_k)$

9: **Compute Step Length** λ_{k+1} (S1 or S2 or ABBmin...)

Under standard assumptions: $\|P_\Omega(x_k - \nabla f(x_k)) - x_k\|_2 \to 0$

Linear equality constraints

$$\min_{Ax=b} f(x)$$

Linear equality constraints

Problem:

$$\min_{Ax=b} f(x),$$

where $f : \mathbb{R}^n \to \mathbb{R}$, $A \in \mathbb{R}^{m\times n}$, $1 \leq m < n$, $rank(A) = m$.

If $d \in null(A)$ and $\bar{x} \in \Omega = \{x \in \mathbb{R}^n : Ax = b\}$ then

$x = \bar{x} + \alpha d \in \Omega$, hence

$$d \text{ is feasible if and only if } d \in null(A)$$

The necessary condition says: $\nabla f(x^*)^T d \geq 0$, for all feasible d.
But $(-d)$ is also feasible:

First order necessary condition

$$\nabla f(x^*)^T d = 0, \text{ for all } d \in null(A).$$

Linear equality constraints

Let $\{z_1, z_2, \ldots, z_{n-m}\}$ be a basis for $null(A)$.

Let $Z \in \mathbb{R}^{n \times (n-m)}$ be the matrix with columns z_i's.

If $\bar{x} \in \Omega$,

$$\Omega = \{x \in \mathbb{R}^n : x = \bar{x} + Zw, w \in \mathbb{R}^{n-m}\}$$

First order necessary condition

$$Z^T \nabla f(x^*) = 0.$$

$\nabla f(x^*)$ is orthogonal to $null(A)$ and so

$\nabla f(x^*) \in range(A^T) \rightarrow \nabla f(x^*) = A^T \lambda$, for some vector $\lambda \in \mathbb{R}^m$.

(If $x \in null(A)$, $a_i^T x = 0$ for all i)

Notation: λ is the Lagrangian vector

Linear equality constraints

$$\min_{Ax=b} f(x)$$

Second order necessary condition

$$y^T \nabla^2 f(x^*) y \geq 0 \quad \text{for all} \quad y \in null(A).$$

Second order sufficient condition

If $Ax^* = b$, $Z^T \nabla f(x^*) = 0$ and

$$y^T \nabla^2 f(x^*) y > 0 \quad \text{for all} \quad y \in null(A),$$

then x^* is a local minimizer.

Linear equality constraints

Example:

$$\min_{x_1+x_2=2} (x_1^2 + x_2^2)$$

Ingredients:

$$\nabla f(x) = 2(x_1, x_2)^T; \quad a = (1,1)^T; \quad \nabla^2 f(x) = 2I$$

Hence, f is strictly convex.

Solution: $x^* = (1,1)^T$, $\nabla f(x^*) = (2,2)^T$, $\lambda^* = 2$.

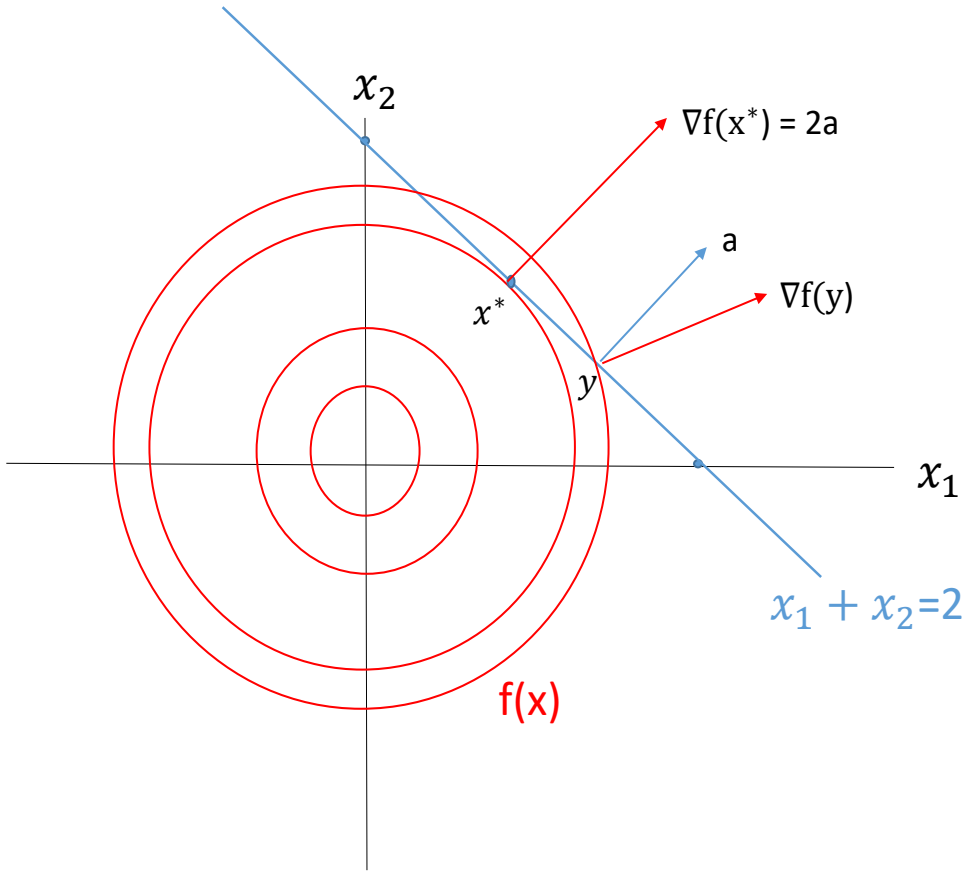
Linear equality constraints

$$\begin{cases} \text{Min } f(x) = x_1^2 + x_2^2 \\ \text{s.t. } x_1 + x_2 = 2 \end{cases}$$

$x^* = (1,1)^T$

$\nabla f(x^*) = (2,2)^T$

$a=(1,1)^T \qquad \lambda^* = 2$



Linear equality constraints

Natural choice for a descent direction:

$$d_k = -ZZ^T \nabla f(x_k) \in \mathbb{R}^{n-m}$$

Indeed: $d_k = Zw$ where $w = -Z^T \nabla f(x_k) \rightarrow d_k \in null(A)$.
Moreover,

$$-\nabla f(x_k)^T ZZ^T \nabla f(x_k) < 0.$$

Newton-type ideas

Motivation: minimize the Lagrangian function

$$\ell(x, \lambda) = f(x) + \lambda^T (Ax - b)$$

Linear equality constraints

$$\ell(x, \lambda) = f(x) + \lambda^T (Ax - b)$$

$$\nabla \ell(x, \lambda) = \begin{bmatrix} \nabla f(x) + A^T \lambda \\ Ax - b \end{bmatrix}$$

$$\nabla^2 \ell(x, \lambda) = \begin{bmatrix} \nabla^2 f(x) & A^T \\ A & 0 \end{bmatrix},$$

symmetric and indefinite.

We can apply Newton, Quasi-Newton, low-cost gradient-type methods, etc., to find saddle points of $\ell(x, \lambda)$.

For quasi-Newton methods, we only need to approximate the $(1, 1)$ matrix block: $\nabla^2 f(x_k)$.

Linear equality constraints

Newton's method: Solve for $(s_k, \delta_k)^T$

$$\begin{bmatrix} \nabla^2 f(x_k) & A^T \\ A & 0 \end{bmatrix} \begin{pmatrix} s_k \\ \delta_k \end{pmatrix} = - \begin{pmatrix} \nabla f(x_k) + A^T \lambda_k \\ Ax_k - b \end{pmatrix}.$$

Set $x_{k+1} = x_k + s_k$ and $\lambda_{k+1} = \lambda_k + \delta_k$.

Another option: Solve for $(s_k, \delta_k)^T$

$$\begin{bmatrix} \nabla^2 f(x_k) & A^T \\ -A & 0 \end{bmatrix} \begin{pmatrix} s_k \\ \delta_k \end{pmatrix} = - \begin{pmatrix} \nabla f(x_k) + A^T \lambda_k \\ b - Ax_k \end{pmatrix}.$$

In the second option, if $\nabla^2 f(x_k)$ is PD, then

the block matrix is not symmetric but PD !

Nonlinear equality constraints

$$\min_{h(x)=0} f(x)$$

Nonlinear equality constraints

Problem:

$$\min_{h(x)=0} f(x),$$

where $f : \mathbb{R}^n \to \mathbb{R}$, $h_i : \mathbb{R}^n \to \mathbb{R}$, $1 \leq i \leq m$, $h_i \in C^1$ for all i.
Notation: $h(x) = [h_1(x), \ldots, h_m(x)]^T$.

The linearly constrained approach is extended in a natural way, as long as the gradients of the constraints are LI at x^*:

First order necessary condition

If x^* is a local minimizer, and x^* is a regular point, i.e., if $\{\nabla h_1(x^*), \ldots, \nabla h_m(x^*)\}$ are LI, then there exist $\lambda_1, \ldots, \lambda_m$ (real numbers) such that

$$\nabla f(x^*) + \sum_{i=1}^{m} \lambda_i \nabla h_i(x^*) = 0.$$

Nonlinear equality constraints

In other words, If x^* is a local minimizer

$$\nabla f(x^*) \in span\{\nabla h_1(x^*), \ldots, \nabla h_m(x^*)\}.$$

Lagrangian function:

$$\ell(x, \lambda) = f(x) + \sum_{i=1}^{m} \lambda_i h_i(x) = f(x) + \lambda^T h(x)$$

Example:

$$\min_{x_1^2 + x_2^2 = 2} (x_1 + x_2)$$

Solution:

$x^* = (-1, -1)^T$, $\nabla f(x^*) = (1, 1)^T$, $\nabla h(x^*) = (-2, -2)^T$,

and $\lambda^* = 0.5$.

Nonlinear equality constraints

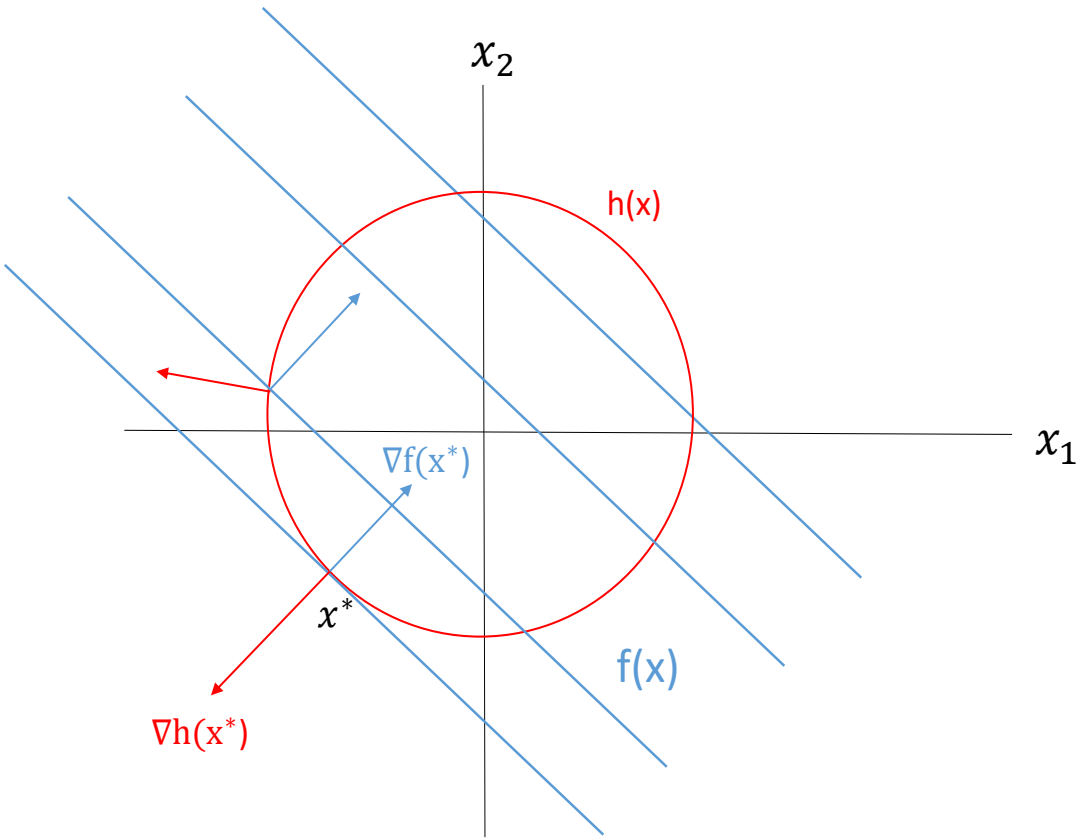
$$\begin{cases} \text{Min } f(x) = x_1 + x_2 \\ \text{s.t. } h(x) = x_1^2 + x_2^2 = 2 \end{cases}$$

$x^* = (-1, -1)^T$

$\nabla f(x^*) = (1,1)^T$

$\nabla h(x^*) = (-2, -2)^T$

$\lambda^* = 1/2$



Nonlinear equality constraints

Second order necessary conditions

If x^* is a local minimizer, and x^* is a regular point, then there exist $\lambda_1, \ldots, \lambda_m$ such that

$$\nabla f(x^*) + \sum_{i=1}^{m} \lambda_i \nabla h_i(x^*) = 0.$$

Moreover, if $f, h \in C^2$

$$y^T(\nabla^2 f(x^*) + \sum_{i=1}^{m} \lambda_i \nabla^2 h_i(x^*))y \geq 0,$$

for all y orthogonal to $span\{\nabla h_1(x^*), \ldots, \nabla h_m(x^*)\}$.

Nonlinear equality constraints

Second order sufficient conditions

If $h(x^*) = 0$, x^* is a regular point, there exist $\lambda_1, \ldots, \lambda_m$ such that

$$\nabla f(x^*) + \sum_{i=1}^{m} \lambda_i \nabla h_i(x^*) = 0,$$

and if

$$y^T(\nabla^2 f(x^*) + \sum_{i=1}^{m} \lambda_i \nabla^2 h_i(x^*))y > 0,$$

for all $y \neq 0$ orthogonal to $span\{\nabla h_1(x^*), \ldots, \nabla h_m(x^*)\}$, then x^* is a strict local minimizer.

Nonlinear equality constraints

$$\nabla^2 \ell(x, \lambda) = \begin{bmatrix} \nabla^2_{xx} \ell(x, \lambda) & J_h(x) \\ J_h(x_k)^T & 0 \end{bmatrix}$$

The m columns of $J_h(x)$ are the gradients $\nabla h_i(x)$, $1 \leq i \leq m$

Newton's method:

Solve for $(s_k, \delta_k)^T$

$$\begin{bmatrix} \nabla^2_{xx} \ell(x_k, \lambda_k) & J_h(x_k) \\ J_h(x_k)^T & 0 \end{bmatrix} \begin{pmatrix} s_k \\ \delta_k \end{pmatrix} = - \begin{pmatrix} \nabla f(x_k) + J_h(x_k) \lambda_k \\ h(x_k) \end{pmatrix}.$$

Set $x_{k+1} = x_k + s_k$ and $\lambda_{k+1} = \lambda_k + \delta_k$.

Nonlinear equality constraints

Direct solution of a saddle point linear system:

Solve for $(s, \delta)^T$

$$\begin{bmatrix} B & A^T \\ A & 0 \end{bmatrix} \begin{pmatrix} s \\ \delta \end{pmatrix} = \begin{pmatrix} y \\ h \end{pmatrix},$$

where B is nonsingular. Apply a 2×2 block Gaussian elimination:

$$\begin{bmatrix} B & A^T \\ 0 & S \end{bmatrix} \begin{pmatrix} s \\ \delta \end{pmatrix} = \begin{pmatrix} y \\ \widehat{h} \end{pmatrix},$$

where $\widehat{h} = h - AB^{-1}y$, and

$$S = -AB^{-1}A^T$$

$S \in \mathbb{R}^{m \times m}$ is known as the Schur's complement.

Now, using back substitution, solve for δ the system $S\delta = \widehat{h}$, and then solve for s the system $Bs = y - A^T\delta$.

Nonlinear equality constraints

If x^* is not regular, the existence of Lagrange multipliers cannot be guaranteed.

Example 1 (regular):

$$\min_{x_1+x_2+x_3=1} 0.5(x_1^2 + x_2^2 + x_3^2)$$

First order necessary condition: $x_i^* + \lambda^* = 0, \quad i = 1, 2, 3.$

Solution: $x^* = (1/3, 1/3, 1/3)^T, \quad \lambda^* = -1/3.$

Since the Hessian of f is I, then the solution is unique.

Nonlinear equality constraints

Example 2 (Not regular):

$$\min(x_1 + x_2),$$

Subject to $(x_1 - 1)^2 + x_2^2 - 1 = 0$, and $(x_1 - 2)^2 + x_2^2 - 4 = 0$.

Solution: $x^* = (0, 0)^T$, but $\nabla f(x^*) = (1, 1)^T$,

and $\nabla h_1(x^*) = (-2, 0)^T$, $\nabla h_2(x^*) = (-4, 0)^T$.

Note that $\nabla h_1(x^*)$ and $\nabla h_2(x^*)$ are linearly dependent, and $\nabla f(x^*)$ cannot be written as a combination of them.

Nonlinear equality constraints

The danger of eliminating variables in the nonlinear case:

$$\min_{(x-1)^3=y^2} (x^2 + y^2)$$

Temptation: substitute y^2, and consider

$$\min_{x\in\mathbb{R}^n} (x^2 + (x-1)^3)$$

The original problem has a unique solution $x^* = (1,0)^T$

The new (unconstrained) problem has no solutions !

When $x \to -\infty, \quad (x^2 + (x-1)^3) \to -\infty$.

What really happened when we substitute $y^2 = (x-1)^3$?

An implicit constraint was forgotten: $x \geq 1$.

Penalization and augmented Lagrangian methods

$$\min_{h(x)=0} f(x)$$

Penalization methods

$$\min_{h(x)=0} f(x)$$

Consider, for a given $\mu > 0$, the minimization of:

$$P(x, \mu) = f(x) + \frac{\mu}{2} h(x)^T h(x) = f(x) + \frac{\mu}{2} \sum_{i=1}^{n} h_i(x)^2.$$

The additional term $\frac{\mu}{2} h(x)^T h(x)$ is a penalization when $h(x) \neq 0$.

We call μ the penalty parameter. $P(x, \mu)$ is a merit function.

In practice we start with $\mu_0 > 0$ and iteratively solve an unconstrained problem at each k, choosing $\mu_{k+1} > \mu_k$, and x_0 at iteration $k + 1$ as the minimizer of the subproblem at iter k.

Penalization methods

Model Algorithm: Given $\mu_0 > 0$, $x_0 \in \mathbb{R}^n$, $k = 0$.

Step 1: $x_{k+1} = \text{argmin}_x P(x, \mu_k)$.

Using an unconstrained method, and x_k as the initial guess.

Step 2: Choose $\mu_{k+1} > \mu_k$, $k = k + 1$ and go to Step 1.

In addition to $p(x) = \frac{1}{2}\|h(x)\|_2^2$, another options for the penalization term:

$p(x) = \|h(x)\|_2$, and $p(x) = \|h(x)\|_1$

Lemma

If at every k, x_k is generated by the penalization method, i.e., x_k is the global minimizer of $P(x, \mu_k) = f(x) + \mu\, p(x)$, then:

$$P(x_k, \mu_k) \leq P(x_{k+1}, \mu_{k+1})$$

$$p(x_{k+1}) \leq p(x_k)$$

$$f(x_k) \leq f(x_{k+1}).$$

Moreover, if x^* is the global minimizer of the original problem, it follows that:

$$f(x_k) \leq P(x_k, \mu_k) \leq f(x^*).$$

Theorem

If at every k, x_k is the global minimizer of $P(x, \mu_k)$, and if $\mu_k \to \infty$, then every limit point of the sequence $\{x_k\}$ is a solution of the original problem.

Moreover, if $p(x) = \frac{1}{2}\|h(x)\|_2^2$ and we assume regularity of the constraints, then $\mu_k \, h_i(x_k) \to \lambda_i^*$ (Lagrange multipliers).

In many applications, the Lagrange multipliers at the solution are important for sensitivity analysis.

Example to illustrate the difficulties for large values of $\mu > 0$:

$$\min_{x_1+x_2=2} (x_1^2 + x_2^2)$$

$$P(x, \mu) = x_1^2 + x_2^2 + \frac{\mu}{2}(x_1 + x_2 - 2)^2$$

$$\nabla_x P(x, \mu) = [2x_1 + \mu(x_1 + x_2 - 2),\ 2x_2 + \mu(x_1 + x_2 - 2)]^T$$

$$\nabla_x^2 P(x, \mu) = \begin{bmatrix} 2+\mu & \mu \\ \mu & 2+\mu \end{bmatrix}$$

Eigenvalues of $\nabla_x^2 P(x, \mu)$: 2 and $2 + 2\mu$.

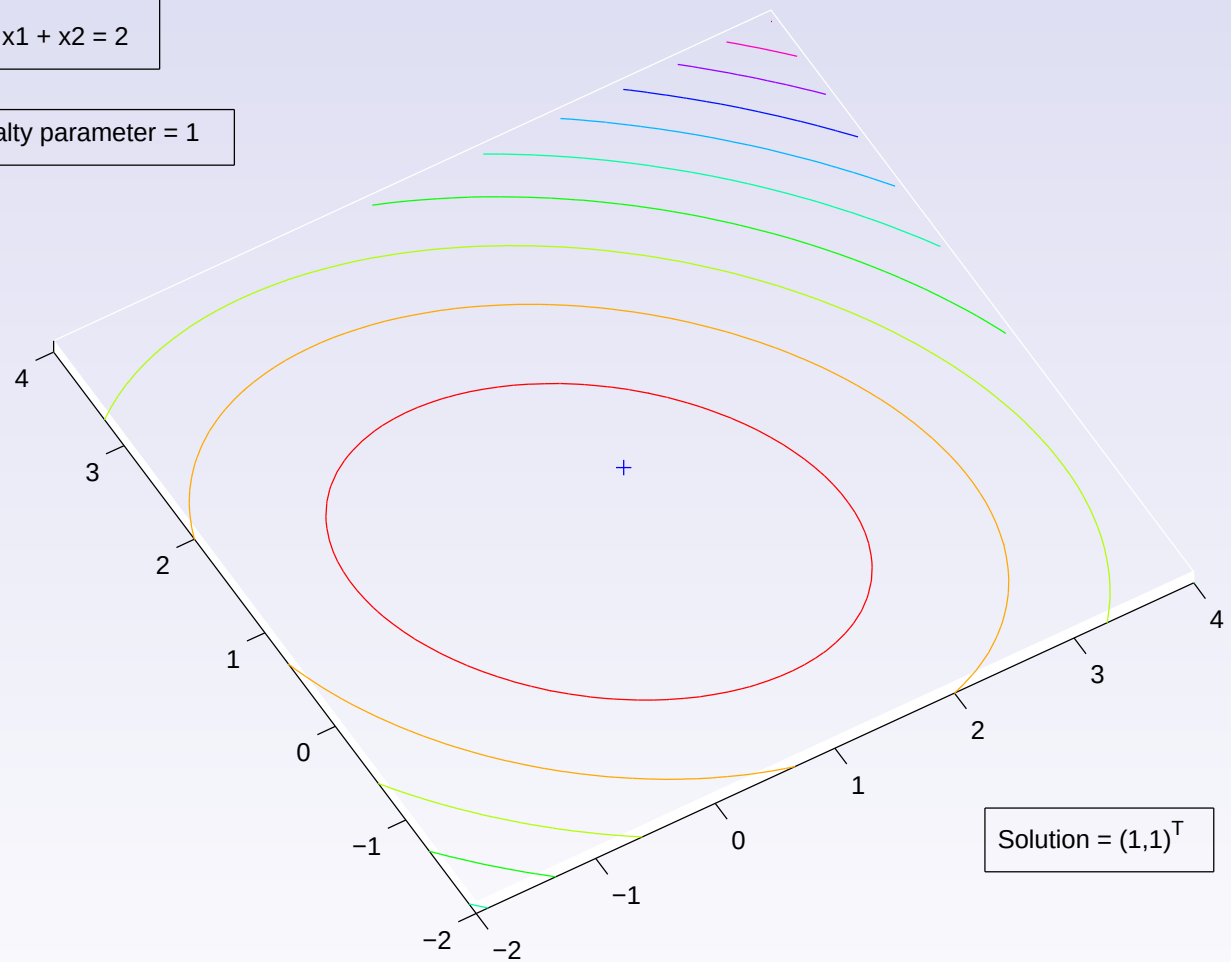
$\kappa(\nabla_x^2 P(x, \mu)) = (1 + \mu)$, it tends to ∞ if μ tends to ∞.

Penalization methods (2-dimensional example)

Min $x1^2 + x2^2$

S.t: $x1 + x2 = 2$

Penalty parameter = 1



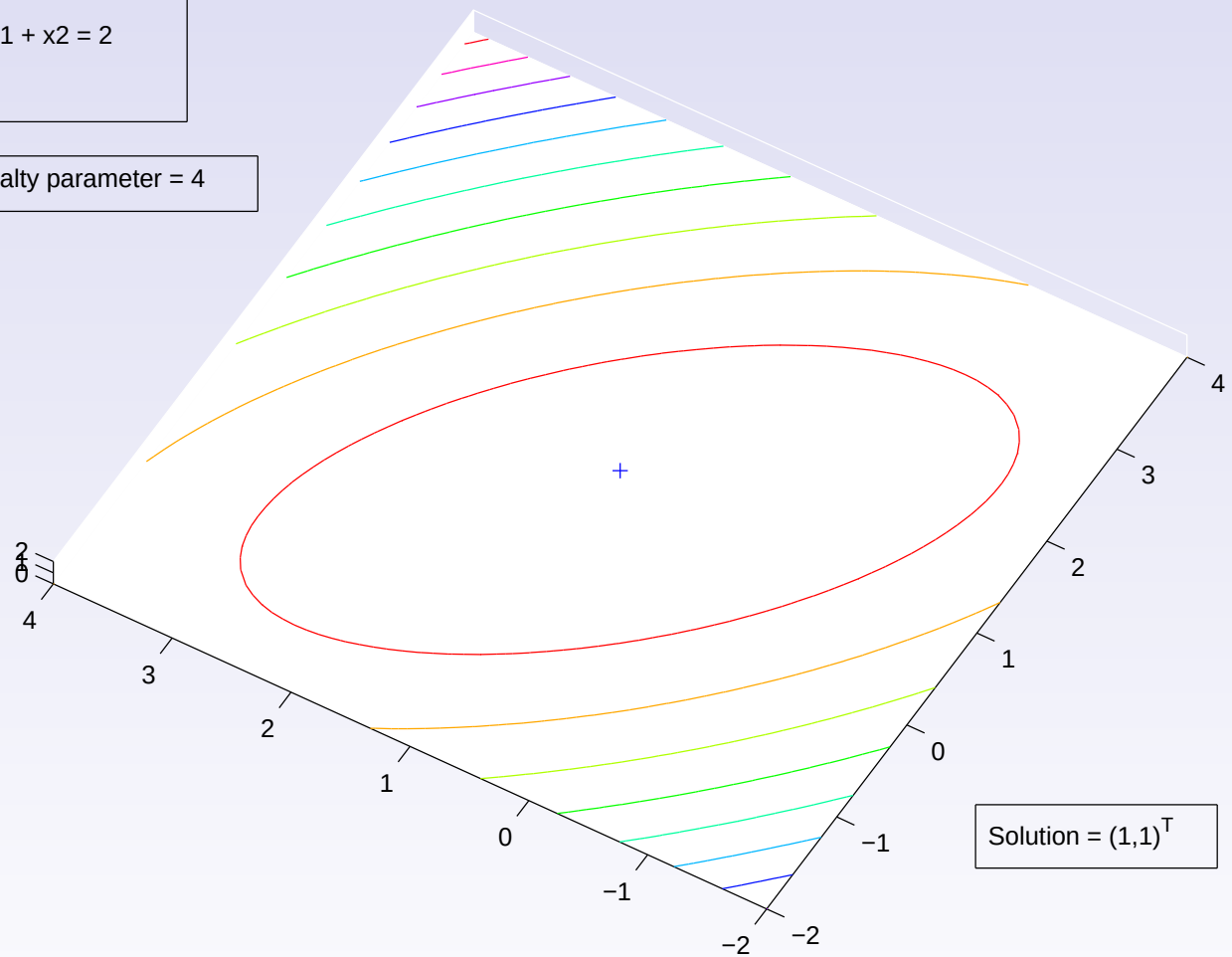
Solution = $(1,1)^T$

Penalization methods (2-dimensional example)

Min $x1^2 + x2^2$

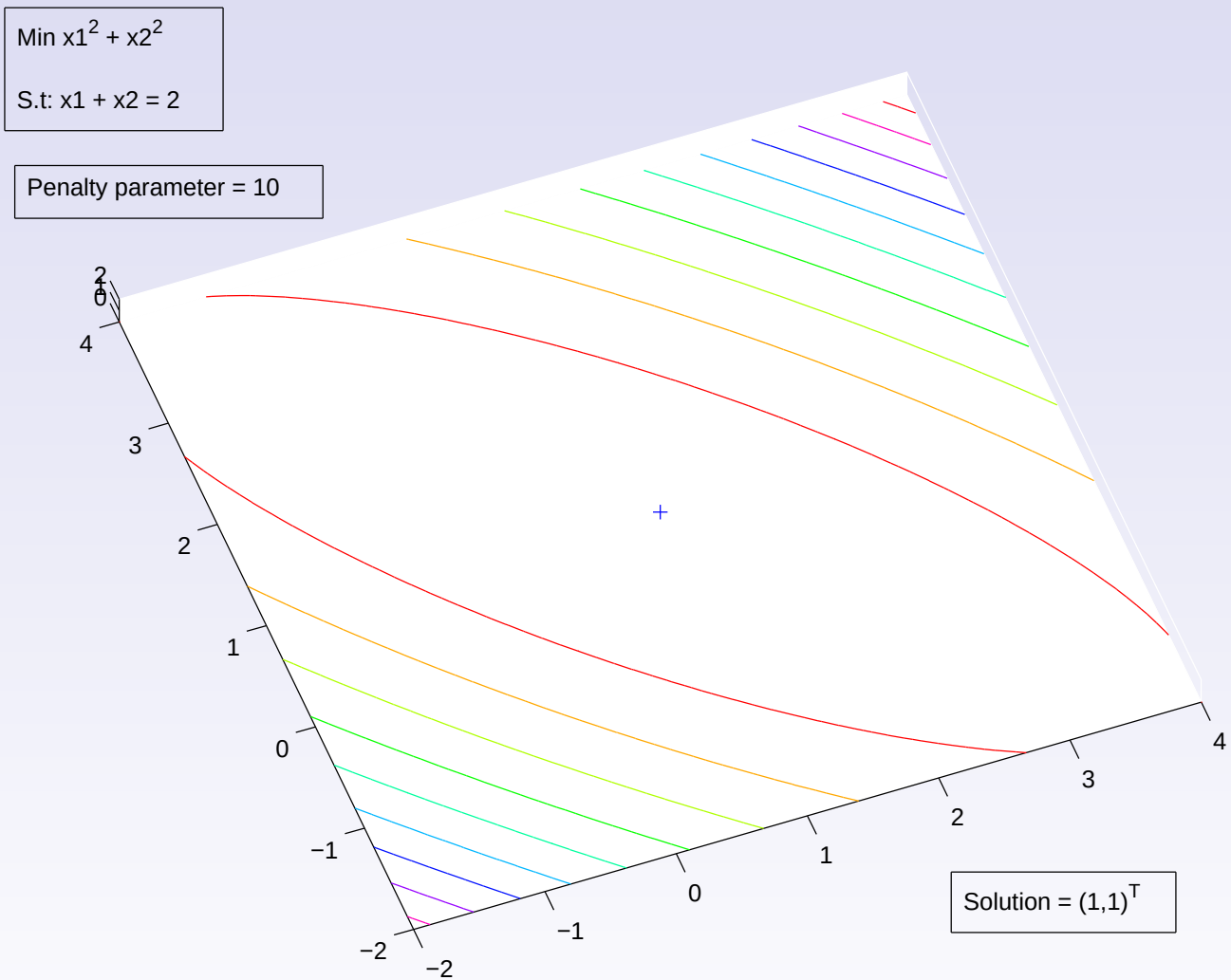
S.t: $x1 + x2 = 2$

Penalty parameter = 4



Solution = $(1,1)^T$

Penalization methods (2-dimensional example)



Augmented Lagrangian methods

$$\min_{h(x)=0} f(x)$$

Consider, for a given $\mu > 0$, the minimization of:

$$L(x, \lambda, \mu) = f(x) + \lambda^T h(x) + \frac{\mu}{2} h(x)^T h(x).$$

The additional term $\frac{\mu}{2} h(x)^T h(x)$ is a penalization when $h(x) = 0$ is not satisfied, and the Lagrangian term $\lambda^T h(x)$ avoids the use of "very" large values of μ. In here, the vector λ is used as a parameter.

We call μ the penalty parameter. $L(x, \lambda, \mu)$ is also a merit function.

In practice we start with $\mu_0 > 0$ and iteratively solve an unconstrained problem at each k, choosing $\mu_{k+1} > \mu_k$, and x_0 at $k + 1$ as the minimizer of the subproblem at k.

Now we also need an estimate of the Lagrangian vector λ_k.

Augmented Lagrangian methods

Model Algorithm: Given $\mu_0 > 0$, $\lambda_0 \in \mathbb{R}^m$, $x_0 \in \mathbb{R}^n$, $k = 0$.

Step 1: $x_{k+1} = \text{argmin}_x L(x, \lambda_k, \mu_k)$.

Using an unconstrained method, and x_k as the initial guess.

Step 2: $\lambda_{k+1} =$ updating formula (x_k, λ_k, μ_k)

Step 3: Choose $\mu_{k+1} > \mu_k$, $k = k + 1$ and go to Step 1.

Updating formula ??

We need a lemma:

Lemma (from linear algebra)

Let $B = B^T \in \mathbb{R}^{n\times n}$ such that $z^T Bz > 0$ for all $z \in null(A)$, $z \neq 0$, and $A \in \mathbb{R}^{m\times n}$. Then, there exists $\bar{\lambda} \geq 0$ such that

$$B + \lambda A^T A \text{ is PD for all } \lambda \geq \bar{\lambda}.$$

Theorem (motivates the use of the Augmented Lagrangian)

If x^* satisfies the sufficient optimality conditions for the problem $\min_{h(x)=0} f(x)$, and $\lambda^* \in \mathbb{R}^m$ is the vector of Lagrange multipliers at x^*, then there exists $\bar{\mu} \geq 0$ such that the function

$$L(x) = f(x) + (\lambda^*)^T h(x) + \frac{\mu}{2} h(x)^T h(x)$$

has at x^* a local strict minimizer for all $\mu \geq \bar{\mu}$.

Augmented Lagrangian methods

The previous theorem motivates the use of the Augmented Lagrangian function:

If we know the exact Lagrange multipliers at the solution, then it suffices to have a finite positive μ (sufficiently large) to find the solution of the original constrained problem, by solving only one unconstrained problem.

Difficulty: We do not have access (in advance) to the Lagrange multipliers!

We can estimate them in an iterative way ("during the fly").

Augmented Lagrangian methods

Notice that solving $\min_{h(x)=0} f(x)$ is equivalent to solving

$$\min_{h(x)=0} f(x) + \lambda^T h(x)$$

for any vector λ (e.g., any Lagrange multipliers approximation).

If we now add a penalization term, we have

$$\min[f(x) + \lambda^T h(x) + \frac{\mu}{2} h(x)^T h(x)]$$

that for each vector λ is a different problem. Forcing the gradient to zero:

$$\nabla f(x) + J_h(x)(\lambda + \mu h(x)) = 0.$$

Recalling the first order conditions, it seems natural to use $(\lambda + \mu h(x))$ to estimate the multipliers.

Augmented Lagrangian methods

Algorithm: Given $\mu_0 > 0$, $\lambda_0 \in \mathbb{R}^m$, $x_0 \in \mathbb{R}^n$, $k = 0$.

Step 1: $x_{k+1} = \text{argmin}_x L(x, \lambda_k, \mu_k)$.

Using an unconstrained method, and x_k as the initial guess.

Step 2: Choose $\mu_{k+1} \geq \mu_k$

Step 3: $\lambda_{k+1} = \lambda_k + \mu_{k+1} h(x_{k+1})$, $k = k + 1$ and go to Step 1.

Usually μ_k is increased when at some k we have observed a significant reduction in the value of $L(x, \lambda_k, \mu_k)$.

Another formula for λ_{k+1} (Step 3):

Solve $J_h(x_{k+1})\lambda = -\nabla f(x_{k+1})$ in the least-squares sense.