Numerisches Rechnen

Optimierung

M. Grepl

J. Berger & R. O'Connor

Institut für Geometrie und Praktische Mathematik
RWTH Aachen

Wintersemester 2015/16

Problem Statement

Nonlinear Programming

We consider the problem

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

where

- $f : \mathbb{R}^n \to \mathbb{R}$ is a continuous (and usually differentiable) function of n variables $x \in \mathbb{R}^n$
- $X = \mathbb{R}^n$ or (more generally) X is a subset of $\mathbb{R}^n$.

- If $X = \mathbb{R}^n$, the problem is called unconstrained
- If f is linear and X is polyhedral, the problem is a linear programming problem. Otherwise it is a nonlinear programming problem.

Problem Statement – Constrained

Constrained Problem

We consider the problem

$$\begin{aligned} \min_{x \in \mathbb{R}^n} \quad & f(x) \\ \text{subject to} \quad & h(x) = 0, \\ & g(x) \leq 0 \end{aligned}$$

where $f : \mathbb{R}^n \to \mathbb{R}$, $h : \mathbb{R}^n \to \mathbb{R}^m$, and $g : \mathbb{R}^n \to \mathbb{R}^r$ are continuously differentiable functions.

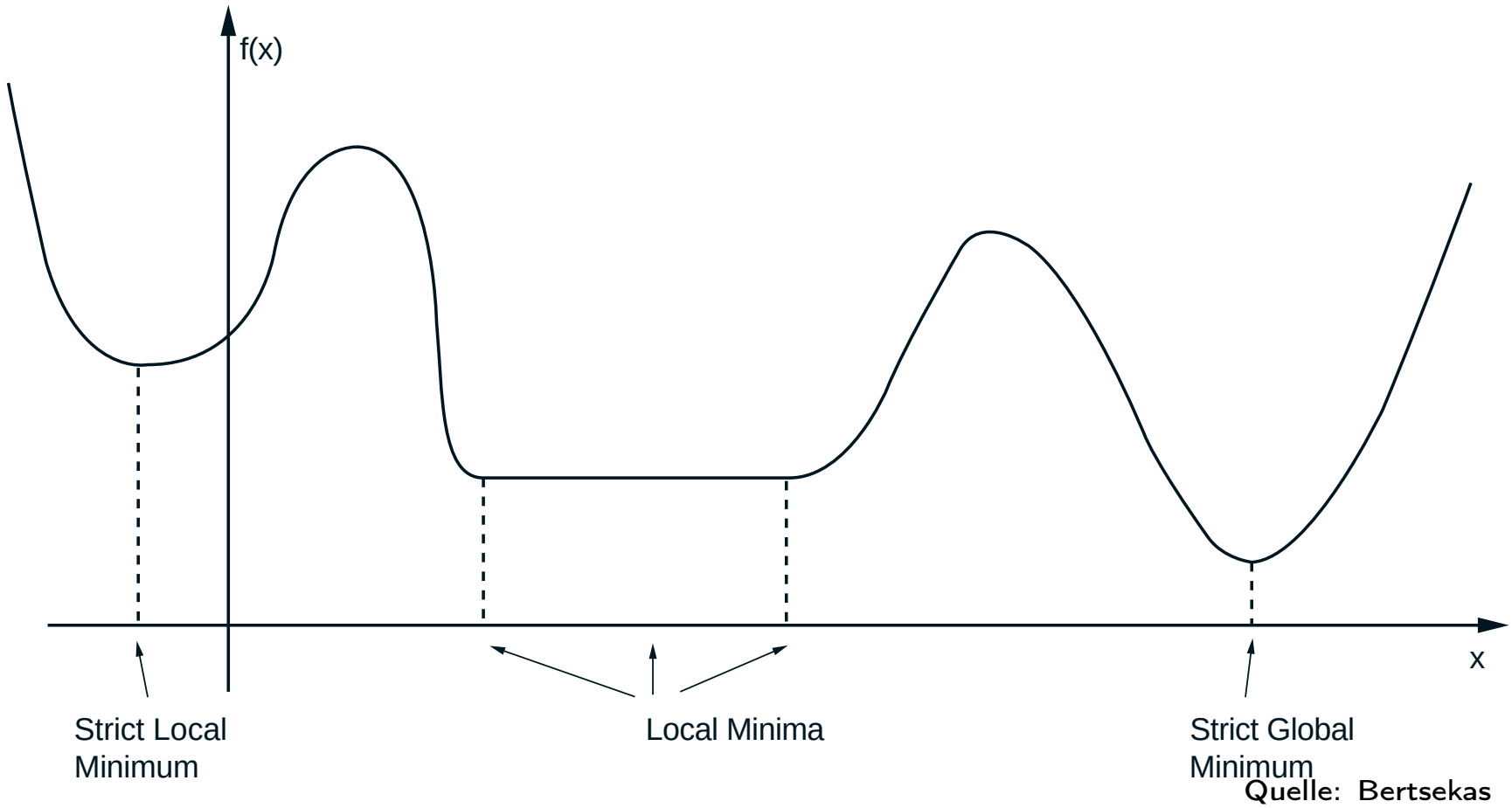
Here

- $h = (h_1, h_2, \ldots, h_m)$ are the equality constraints, and
- $g = (g_1, g_2, \ldots, g_r)$ are the inequality constraints.

Two Main Issues

- Characterization of minima
  - Necessary conditions
  - Sufficient conditions
  - Lagrange multiplier theory
  - Sensitivity
  - Duality
- Computation by iterative algorithms
  - Iterative descent
  - Approximation methods
  - Dual and primal-dual methods

Local and Global Minima



Quelle: Bertsekas

Unconstrained local and global minima in one dimension

Local and Global Minima

Definitions

- A point x^* is an unconstrained local minimum of f if there exists and $\epsilon > 0$ such that
$$f(x^*) \leq f(x), \quad \forall x \text{ with } \|x - x^*\| < \epsilon.$$
- A point x^* is a strict unconstrained local minimum of f if there exists and $\epsilon > 0$ such that
$$f(x^*) < f(x), \quad \forall x \neq x^* \text{ with } \|x - x^*\| < \epsilon.$$
- A point x^* is a global minimum of f if
$$f(x^*) \leq f(x), \quad \forall x \in \mathbb{R}^n.$$

Taylor Expansion

Taylor's Theorem

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

$$f(x+p) = f(x) + \nabla f(x+tp)^T p,$$

for some $t \in [0,1]$.

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

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

for some $t \in [0,1]$, and that

$$f(x+p) = f(x) + p^T \nabla f(x) + \tfrac{1}{2} p^T \nabla^2 f(x)\, p + O(\|p\|^3).$$

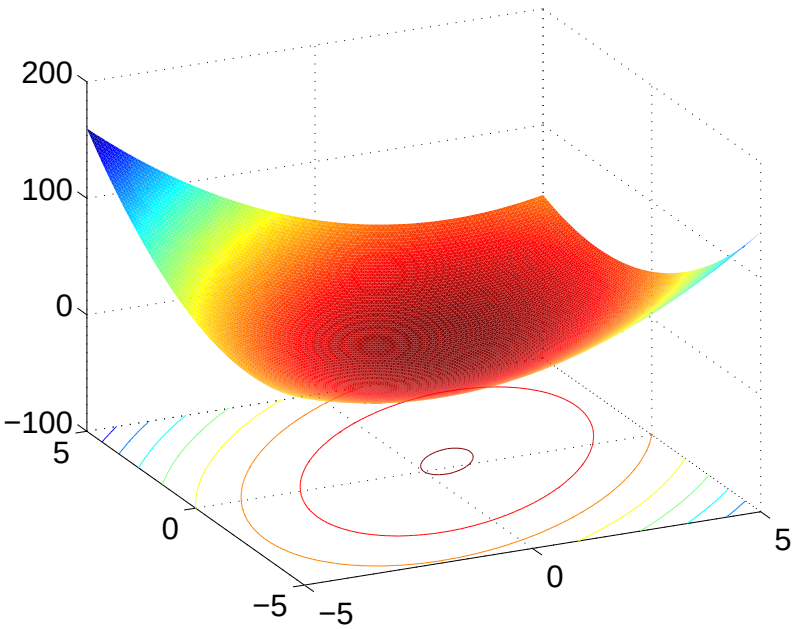
Special Case: Quadratic Cost Functions

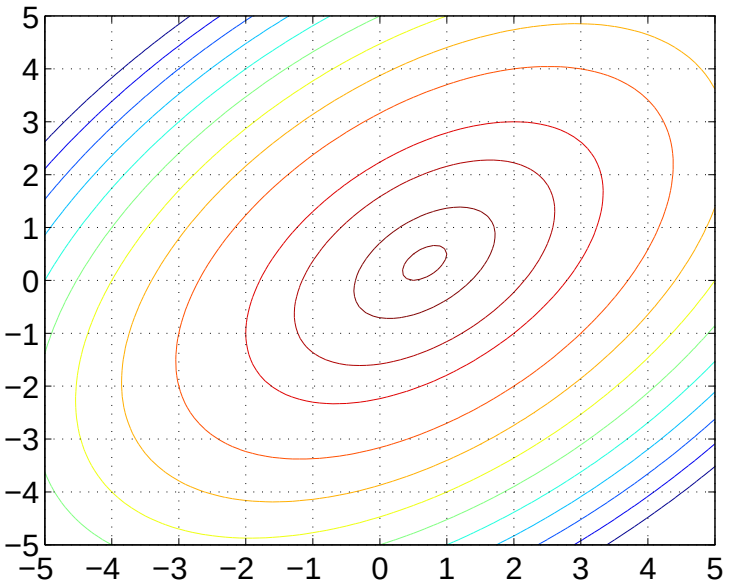
Given $\boldsymbol{A} \in \mathbb{R}^{n \times n}$, $\boldsymbol{b} \in \mathbb{R}^n$, and $\boldsymbol{c} \in \mathbb{R}$, we define the quadratic form

$$\boldsymbol{f(x) = \frac{1}{2} x^T A x - x^T b + c}$$

where $\boldsymbol{x} \in \mathbb{R}^n$.

Example in $\mathbb{R}^2$





Quadratic Cost Functions: Geometric Meaning

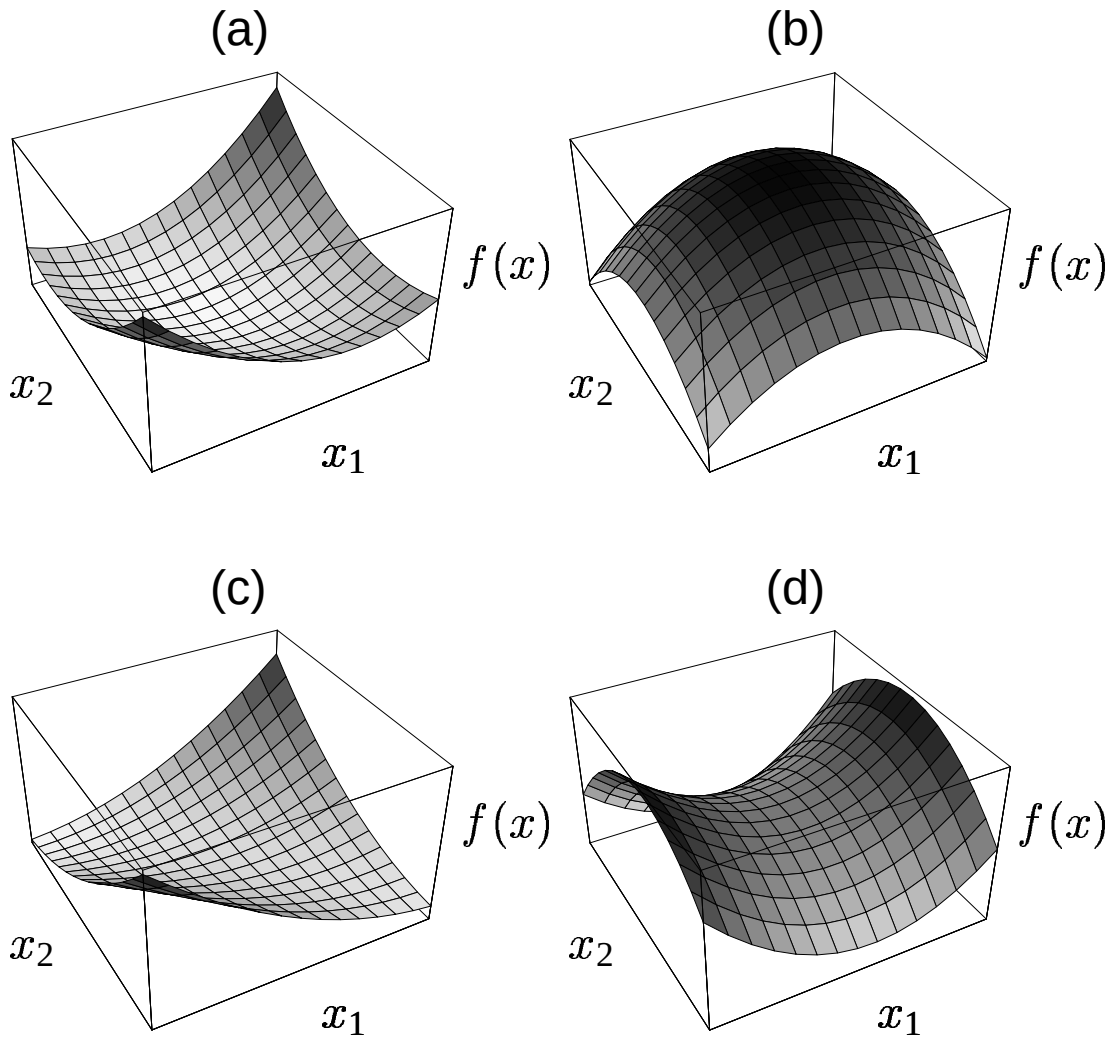
Quadratic forms for $\boldsymbol{A}$:

(a) positive-definite

(b) negative-definite

(c) indefinite (and positive-indefinite)

(d) indefinite



Quelle: J.R. Shewchuk

Quadratic Cost Functions: An Important Property

If f is a quadratic form and A is s.p.d., then

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

solves the linear system $Ax = b$, that is

$$Ax^* = b \quad \Leftrightarrow \quad x^* = \arg\min_{x \in \mathbb{R}^n} f(x).$$

Proof

- Second Order Sufficient Condition
- Consider pertubation $x + p$, $p \in \mathbb{R}^n$

Note:

- Solution of linear system $Ax = b$ "reduces" to solving an optimization problem ($\rightarrow$ Conjugate Gradient Method).

Gradient and Directional Derivative

Recall that the derivative of $f(x, y) : \mathbb{R}^2 \to \mathbb{R}$ along the

- x-direction (i.e. keeping y constant) is $\frac{\partial f}{\partial x}$
- y-direction (i.e. keeping x constant) is $\frac{\partial f}{\partial y}$

and the gradient is given by $\nabla f = [\frac{\partial f}{\partial x} \ \frac{\partial f}{\partial y}]^T$.

Directional Derivative

The directional derivative of a function $f : \mathbb{R}^n \to \mathbb{R}$ in the direction p is given by

$$\nabla_p f(x) := \lim_{\epsilon \to 0} \frac{f(x + \epsilon p) - f(x)}{\epsilon}.$$

For f continuously differentiable, we have

$$\nabla_p f(x) = \nabla f(x)^T p.$$

Perpendicularity of $\nabla f(x)$ to the level curve

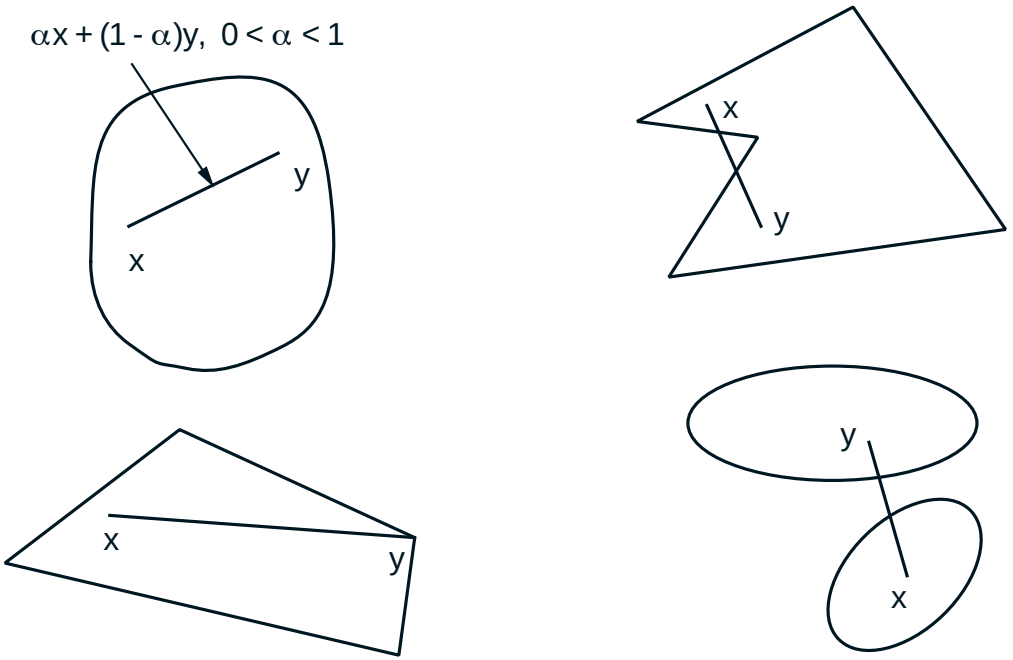
- Consider any point x_0 and the level curve of f through x_0.
- Then the gradient of f at x_0, $\nabla f(x_0)$, is perpendicular to the tangent direction of the contour at x_0.
- Two ways to see this:
  - Intuitively, the value of f does not change along the tangent, so the gradient must be in the perpendicular direction.
  - More formally, along the tangent, the value of f does not change since we are moving along the contour.
    $\Rightarrow$ The directional derivative is zero along p, if p is in the direction of the tangent, that is
    $$\nabla_p f(x) = p^T \nabla f(x) = 0$$
    and thus $\nabla f(x)$ is perpendicular to p.

Convex Sets

Definition

Let C be a subset of $\mathbb{R}^n$. We say that C is convex if

$$\alpha x + (1-\alpha)y \in C, \quad \forall x, y \in C, \ \forall \alpha \in [0,1].$$



Convex Sets

Nonconvex Sets

Quelle: Bertsekas

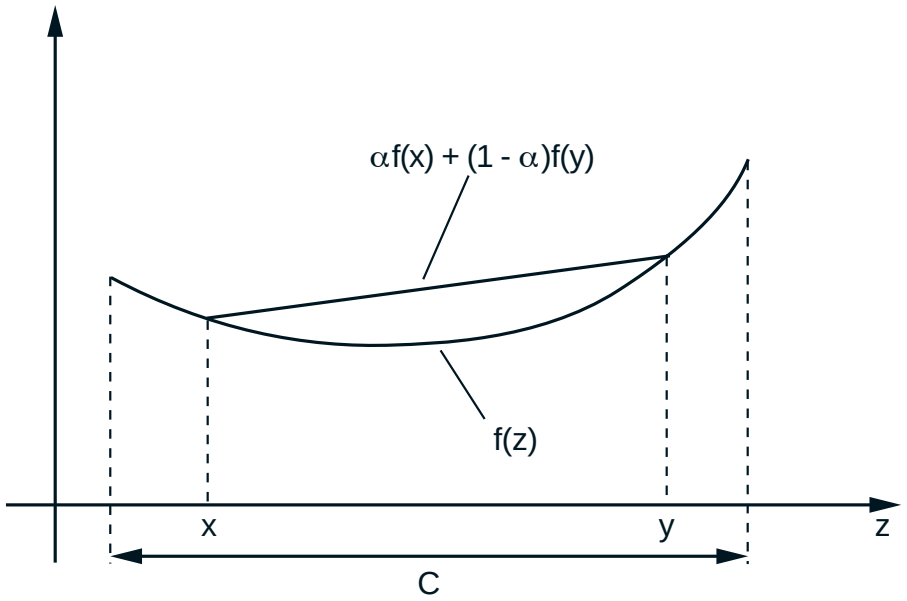
Convex Functions

Definition

Let C be a convex subset of $\mathbb{R}^n$. A function $f : C \to \mathbb{R}$ is called convex if

$$f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha) f(y), \quad \forall x, y \in C, \ \forall \alpha \in [0,1].$$

The function f is called concave if $-f$ is convex.



Quelle: Bertsekas

Necessary Optimality Conditions

First Order Necessary Conditions

Let x^* be an unconstrained local minimum of $f : \mathbb{R}^n \to \mathbb{R}$, and assume that f is continuously differentiable in an open neighbourhood of x^*, then

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

Second Order Necessary Conditions

Let x^* be an unconstrained local minimum of $f : \mathbb{R}^n \to \mathbb{R}$, and assume that f is twice continuously differentiable in an open neighbourhood of x^*, then

$$\nabla f(x^*) = 0$$

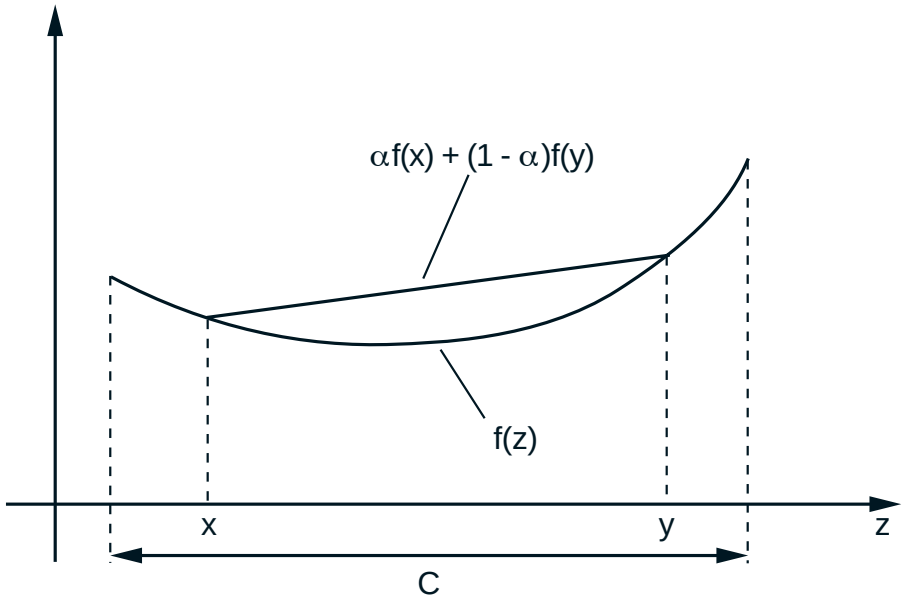
and

$$\nabla^2 f \text{ is positive semidefinite.}$$

The Case of a Convex Cost Function

Convex Cost Function

When f is convex, any local minimum x^* is a global minimum of f. If in addition f is differentiable, then any stationary point x^*, i.e., where $\nabla f(x^*) = 0$, is a global minimum of f.



Quelle: Bertsekas

Sufficient Optimality Conditions

Second Order Sufficient Conditions

Let $f : \mathbb{R}^n \to \mathbb{R}$ be twice continuously differentiable in an open neighbourhood of x^* and suppose that x^* satisfies the conditions

$$\nabla f(x^*) = 0$$

and

$$\nabla^2 f \text{ is positive definite.}$$

Then x^* is a strict unconstrained local minimum of f.

In particular, there exists scalars $\gamma > 0$ and $\epsilon > 0$ such that

$$f(x) \geq f(x^*) + \frac{\gamma}{2}\|x - x^*\|^2, \quad \forall x \text{ with } \|x - x^*\| < \epsilon.$$

Topics

Unconstrained Optimization:

- Gradient Methods
- Newton's Method and Variations
- Least-Squares Problems

References:

- J. Nocedal, S.J. Wright. Numerical Optimization (Second Edition). Springer Verlag
- D.P. Bertsekas. Nonlinear Programming (Second Edition). Athena Scientific
- ...

Existence of Minima

Consider

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

Two possibilities:

- The set $\{f(x)|x \in X\}$ is unbounded below and there is not optimal solution
- The set $\{f(x)|x \in X\}$ is bounded below
  - A global minimum exists if f is continuous and X is compact (Weierstrass theorem)
  - A global minimum exists if X is closed, and f is continuous and coercive, that is, $f(x) \to \infty$ when $\|x\| \to \infty$.

Principal Gradient Methods

Gradient Method

We choose the iteration

$$x^{k+1} = x^k + \alpha^k d^k, \quad k = 0, 1, \ldots$$

where, if $\nabla f(x^k) \neq 0$, the direction d^k satsifies

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

and α^k is a positive stepsize.

Goal: Choose direction d^k and stepsize α^k such that

$$f(x^{k+1}) = f(x^k + \alpha^k d^k) < f(x^k), \quad k = 0, 1, \ldots$$

Principal Gradient Methods

Definition

Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable, a direction d is called gradient related if

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

Note: If d^k is gradient related, then

$$f(x^{k+1}) = f(x^k + \alpha\, d^k) < f(x^k), \quad k = 0, 1, \ldots$$

for α sufficiently small. Proof ...

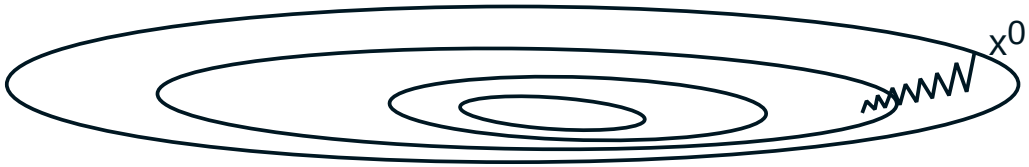
Principal Example

Given a positive definite matrix D^k, we choose the direction $d^k = -D^k \nabla f(x^k)$ to obtain

$$x^{k+1} = x^k - \alpha^k\, D^k\, \nabla f(x^k), \quad k = 0, 1, \ldots.$$

Descent Directions

- **Steepest Descent**
  - Choose $D^k = I$, where I is the $n \times n$ identity matrix
  - Iteration
$$x^{k+1} = x^k - \alpha^k \nabla f(x^k)$$
  - Problem: often leads to slow convergence



Quelle: Bertsekas

- **Newton's Method**
  - Choose $D^k = (\nabla^2 f(x^k))^{-1}$
  - Iteration
$$x^{k+1} = x^k - \alpha^k (\nabla^2 f(x^k))^{-1} \nabla f(x^k)$$

Descent Directions

- Newton's Method – Idea: Minimize quadratic approximation of f around x^k at each iteration

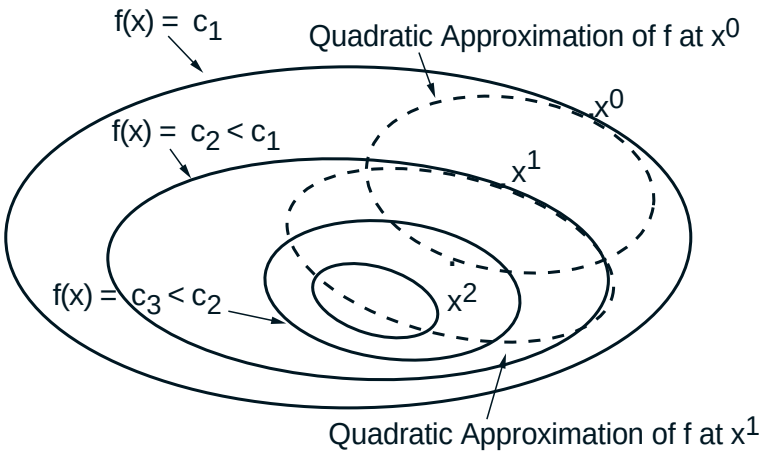
$$f^k(x) = f(x^k) + \nabla f(x^k)^T (x - x^k) + \tfrac{1}{2}(x - x^k)^T \nabla^2 f(x^k)\,(x - x^k).$$

Next iterate x^{k+1} is minimum of $f^k(x)$, i.e., $\nabla f^k(x) \stackrel{!}{=} 0$,

$$\nabla f(x^k) + \nabla^2 f(x^k)\,(x - x^k) = 0$$

and thus

$$x^{k+1} = x^k - (\nabla^2 f(x^k))^{-1} \nabla f(x^k)$$



Quelle: Bertsekas

Descent Directions

- **Diagonally Scaled Steepest Descent**
  - Choose $D^k = \mathrm{diag}(d_i^k),\ i = 1, \ldots, n$, where

$$d_i^k \approx \left(\frac{\partial^2 f(x^k)}{(\partial x_i)^2} \right)^{-1}$$

- **Modified Newton's Method**
  - Choose $D^k = (\nabla^2 f(x^0))^{-1},\ k = 0, 1, \ldots$
  - Hessian $\nabla^2 f(x^0)$ needs to be positive definite
  - Variation: Recompute Hessian every $p > 1$ iterations
- **Discretized Newton's Method**
  - Choose $D^k = (H(x^k))^{-1},\ k = 0, 1, \ldots$
  - $H(x^k)$ is spd approximation of $\nabla^2 f(x^k)$
  - Finite difference approximations of second derivative based on first dervatives of f

Descent Directions

- **Quasi-Newton methods**
  Approximate Hessian matrix using rank-one updates, i.e., adjust D^k at each step in order to approximate the Newton direction.
  Most popular: Broyden-Fletcher-Goldfarb-Shanno (BFGS) method
- **Gauss-Newton Method** (Least Squares Problem)

$$\min_{x \in \mathbb{R}^n} f(x) = \tfrac{1}{2}\|g(x)\|^2 = \tfrac{1}{2} \sum_{i=1}^{m} (g_i(x))^2$$

Other choices, where d^k is not expressed as $d^k = -D^k \nabla f(x^k)$

- Conjugate Gradient Method
- Coordinate Descent Method

Stepsize Selection

- **Minimization Rule**
  - Choose α^k such that
$$f(x^k + \alpha^k d^k) = \min_{\alpha \geq 0} f(x^k + \alpha d^k)$$
  - Minimize cost function along the direction d^k.
- **Limited Minimization Rule**
  - Choose α^k such that
$$f(x^k + \alpha^k d^k) = \min_{\alpha \in [0,s]} f(x^k + \alpha d^k)$$
    where $s > 0$ is a fixed scalar

Note: Minimization and limited minimization rules typically require one-dimensional line search algorithms (e.g. quadratic or cubic interpolation method, Golden Section method) which are solved approximately.

Stepsize Selection

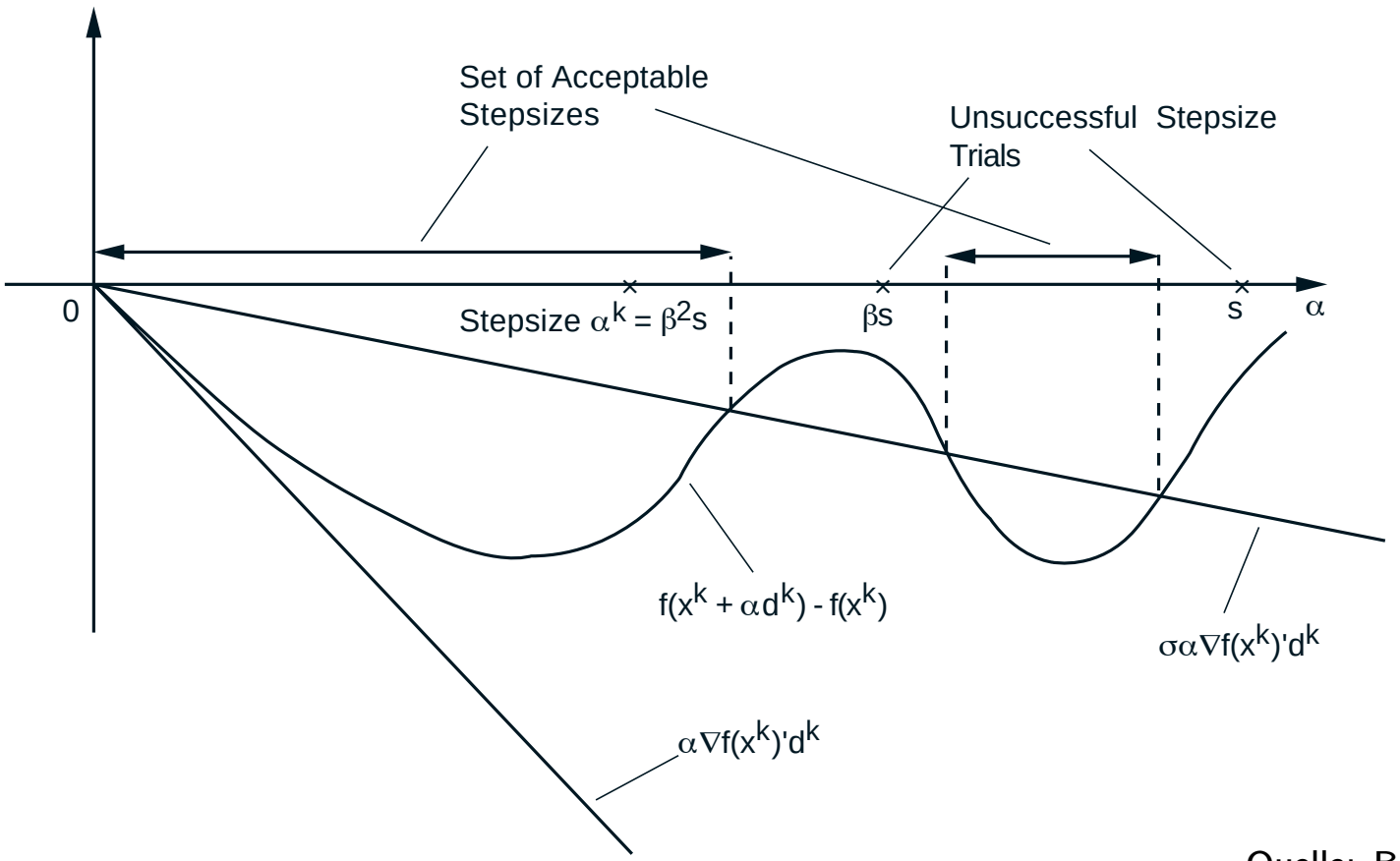
- **Constant Stepsize**
  - Select a fixed stepsize $s > 0$ and
$$\alpha^k = s$$
  - If stepsize is too large, divergence will occur. If stepsize is too small, convergence will be very slow.
- **Diminishing Stepsize**
  - Choose α^k such that
$$\alpha^k \to 0 \quad \text{and} \quad \sum_{k=0}^{\infty} \alpha^k = \infty$$
  - Good theoretical convergence properties, but convergence tends to be slow

Stepsize Selection

- **Successive Stepsize Reduction – Armijo Rule**
  - Choose fixed scalars s, β, and σ, with $0 < \beta < 1$, and $0 < \sigma < 1$, and set
    $$\alpha^k = \beta^{m_k} s,$$
    where m_k is the first nonnegative integer m for which
    $$f(x^k) - f(x^k + \beta^m s d^k) \geq -\sigma \beta^m s \nabla f(x^k)^T d^k \quad .$$
  - Typical values: $s = 1$, $\sigma \in [10^{-5}, 10^{-1}]$, $\beta \in [0.1, 0.5]$
- **Goldstein Rule**
  - Select fixed scalar $\sigma \in (0, 0.5)$, choose α^k to satisfy
    $$\sigma \leq \frac{f(x^k + \alpha^k d^k) - f(x^k)}{\alpha^k \nabla f(x^k)^T d^k} \leq 1 - \sigma.$$

Stepsize Selection

Line search by the Armijo rule



Quelle: Bertsekas

Convergence Results

Constant and Diminishing Stepsizes

Let $\{x^k\}$ be a sequence generated by a gradient method $x^{k+1} = x^k + \alpha^k d^k$, where $\{d^k\}$ is gradient related. Assume that for some constant $L > 0$, we have

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|, \quad \forall x, y \in \mathbb{R}^n.$$

Assume that either

- there exists a scalar ϵ such that for all k

$$0 < \epsilon \leq \alpha^k \leq \frac{(2-\epsilon)|\nabla f(x^k)^T d^k|}{L\|d^k\|^2}$$

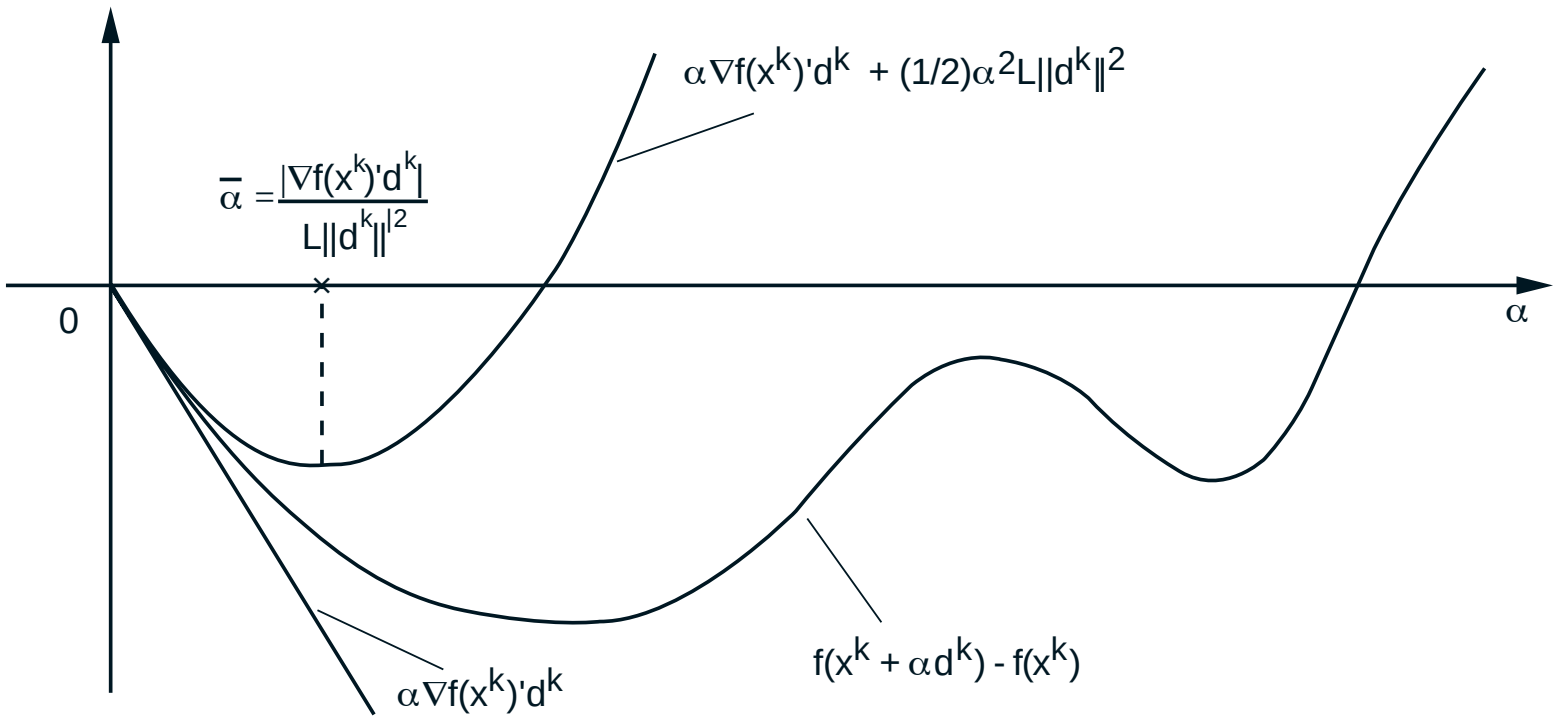
  or

- $\alpha^k \to 0$ and $\sum_{k=0}^{\infty} \alpha^k = \infty$.

Then either $f(x^k) \to -\infty$ or else $\{f(x^k)\}$ converges to a finite value and $\nabla f(x^k) \to 0$.

Constant and Diminishing Stepsizes

Idea of convergence proof



Quelle: Bertsekas

Convergence Results

Minimization, Armijo, and Goldstein Rule

Let $\{x^k\}$ be a sequence generated by a gradient method $x^{k+1} = x^k + \alpha^k d^k$, and assume that $\{d^k\}$ is gradient related and α^k is chosen by the minimization rule, or the limited minimization rule, or the Armijo, or the Goldstein rule. Then every limit point of $\{x^k\}$ is a stationary point.

Rate of Convergence

- Quadratic Model Analysis: $f(x) = \frac{1}{2}x^T Q x$, with $Q > 0$
- Steepest Descent

$$x^{k+1} = x^k - \alpha^k \nabla f(x^k) = (I - \alpha^k Q)x^k$$

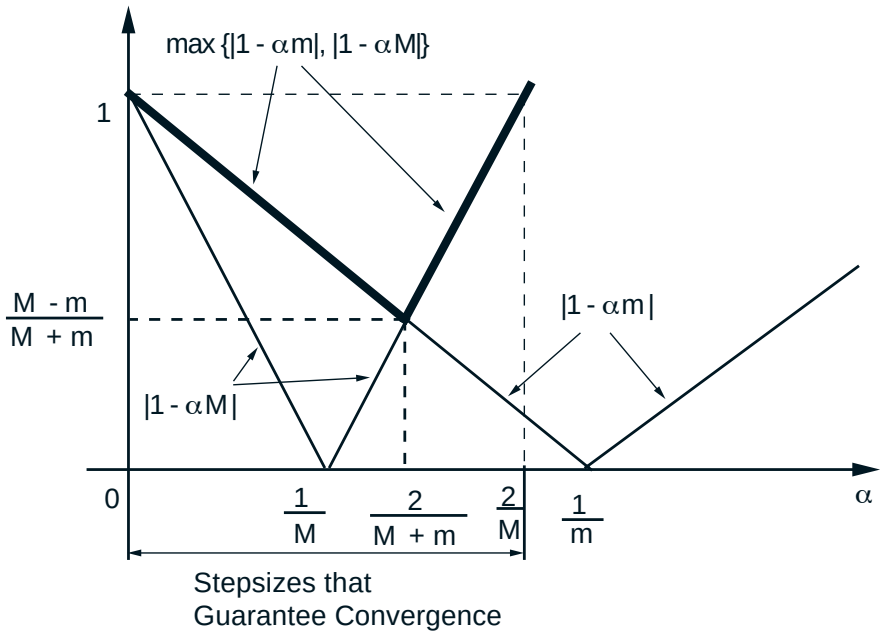
We obtain

$$\frac{\|x^{k+1}\|}{\|x^k\|} \leq \frac{M-m}{M+m}$$

For **minimization stepsize**

$$\frac{f(x^{k+1})}{f(x^k)} \leq \left(\frac{M-m}{M+m}\right)^2$$

Condition number of Q is $\frac{M}{m}$



Quelle: Bertsekas

Gauss-Newton

Least Squares Problem

$$\min_{x \in \mathbb{R}^n} f(x) = \tfrac{1}{2}\|g(x)\|^2 = \tfrac{1}{2}\sum_{i=1}^{m} (g_i(x))^2$$

Gauss-Newton:

- Given a point x^k, linearize g to obtain
$$\tilde{g}(x, x^k) = g(x^k) + \nabla g(x^k)^T (x - x^k)$$
- Minimize norm of linearized function $\tilde{g}$
$$x^{k+1} = \arg\min_{x \in \mathbb{R}^n} \tfrac{1}{2}\|\tilde{g}(x, x^k)\|^2$$
- Iteration
$$x^{k+1} = x^k - (\nabla g(x^k) \nabla g(x^k)^T)^{-1} \nabla g(x^k)\, g(x^k)$$
assuming $\nabla g(x^k) \nabla g(x^k)^T$ is invertible.

Modified Gauss-Newton Method

- Often implemented in the modified form

$$x^{k+1} = x^k - \alpha^k (\nabla g(x^k) \nabla g(x^k)^T + \Delta^k)^{-1} \nabla g(x^k)\, g(x^k)$$

where Δ^k is a diagonal matrix such that

$$\nabla g(x^k) \nabla g(x^k)^T + \Delta^k \ : \text{ positive definite}$$

to

  - ensure descent
  - deal with case if $\nabla g(x^k) \nabla g(x^k)^T$ singular
  - enhance convergence if $\nabla g(x^k) \nabla g(x^k)^T$ nearly singular

- Levenberg-Marquardt method:

$$\Delta^k = \mu\, I,$$

where $\mu \in \mathbb{R}^+$ and I is the identity matrix.

Neural Networks – Example

Neural network training problem ($m = 5$, weights u_0, u_1)

$$\frac{1}{2} \sum_{i=1}^{5} (z_i - \phi(u_1 y_i + u_0))^2$$