Certificat Big Data
Introduction to Numerical Optimization

Sixin Zhang
with Serge Gratton

INP ENSEEIHT

sixin.zhang@toulouse-inp.fr

Outline

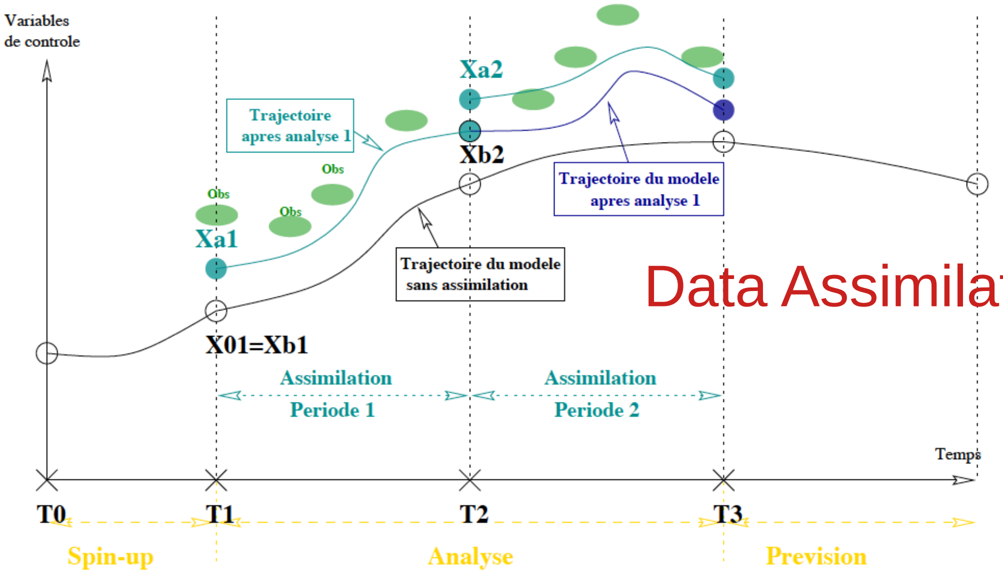
- Introduction
  - Motivation
  - Preliminary knowledge
- Basic theory of Optimization
- Optimization methods without constraint
- Optimization methods with constraints

Reference

- J. Gergaud, S. Gratton, D. Ruiz. **Optimisation numérique : aspects théoriques et algorithmes**, Polycopié du cours d'Optimisation, ENSEEIHT - Sciences du numérique.
- M. Bierlaire. **Introduction à l'optimisation différentiable**, Presses polytechniques et universitaires romandes, 2006.
- J. Nocedal, S. Wright. **Numerical Optimization**, Springer Series in Operations Research, 2006.

Introduction : Optimization in real-world problems

- **Predict dynamics of atmosphere and ocean**
  - How to combine “optimally” the information from observation and model?



Data Assimilation

Introduction : Optimization in real-world problems

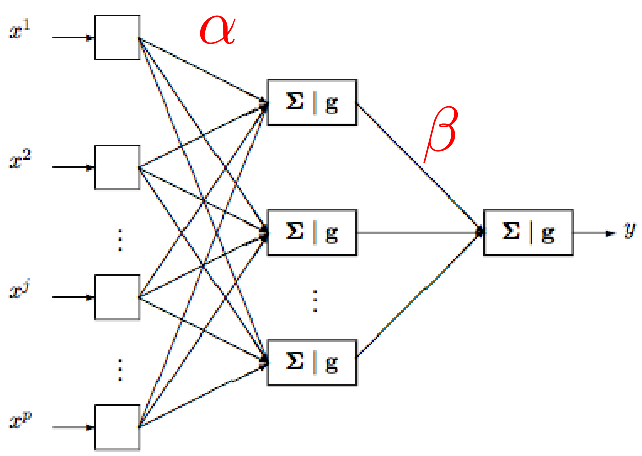
- **Machine learning**
  - Input vector: $x = (x_i)_{i \leq p} \in \mathbb{R}^p$
  - Output value: $y = f(x, \alpha, \beta) \in \mathbb{R}$
  - Supervised learning: optimize the *parameters* to fit observed data points

e.g. Observe $\{(x_n, y_n)\}_{n \leq N}$

$$\text{Objective: } \min_{\alpha,\beta} \frac{1}{N} \sum_{n \leq N} (y_n - f(x_n, \alpha, \beta))^2$$

Least-square optimization problem

Parameters



Wikistat: Réseaux de neurones

Introduction : Optimization in real-world problems

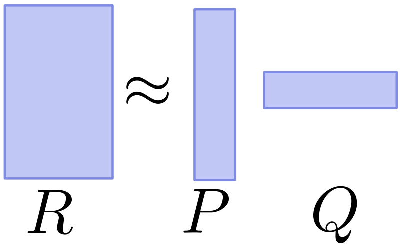
- **Recommendation** (film, music, book, etc)
  - Data: uses provide ratings of products +/-/?
  - Format: (user,product,rating)
  - **Question**: predict unobserved ratings (?)
- A low-rank matrix model
  - Approximate the matrix R by a low-rank matrix R',
    - Represent R' by PQ so that rank(R') is small.

R product

	+		-	+	
	+		?	?	
user	+		-	?	
	+		-	?	
	+		-	+	

Objective: $\min_{P,Q} \sum_{(i,j) observed} ([R]_{i,j} - [PQ]_{i,j})^2$

Constrainted optimizatoin problem



Preliminary: Linear algebra

- **Definition:** Positive definite and semi-definite matrix

Let A be a symmetric matrix

- A is positive semi-definite if $\forall x \in \mathbb{R}^n, x^\intercal A x \geq 0$
- A is positive definite if $\forall x \in \mathbb{R}^n, x \neq 0, x^\intercal A x > 0$

- **Theorem:** equivalent conditions

For a symmetric matrix A

- A is positive semi-definite iff all the eigenvalues of A are ≥ 0
- A is positive definite iff all the eigenvalues of A are > 0

Preliminary: Calculus

- **Definition**: Gradient of a real-valued differentiable function f
  - In dimension 1

$$\forall x \in \mathbf{R},\ f'(x) = \lim_{\delta \to 0} \frac{f(x+\delta)-f(x)}{\delta}$$

$$\Rightarrow \text{ If } \delta \approx 0, \text{ then } f(x+\delta) \approx f(x) + \delta f'(x)$$

  - In dimension n

$$\forall x \in \mathbf{R}^n,\ h \in \mathbf{R}^n,\ \nabla f(x)^T h = \lim_{\delta \to 0} \frac{f(x+\delta h)-f(x)}{\delta}$$

$$\Rightarrow \text{ If } \delta \approx 0, \text{ then } f(x+\delta h) \approx f(x) + \delta \nabla f(x)^T h$$

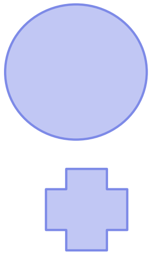
$$Gradient : \nabla f(x)$$

Preliminary: Convex set and convex function

- **Definition**: Convex set
  - Let E be a vector space. A subset C of E is **convex** if

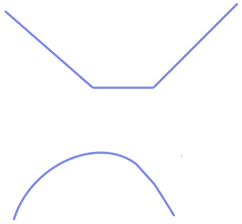
$$\forall (x, y) \in C^2, \forall \alpha \in [0, 1], \alpha x + (1 - \alpha) y \in C$$

  - In other words, the line connecting x and y is also in the set C



- **Definition**: Convex function
  - Let f be a function: C → R. It is convex in a **convex** domain C if

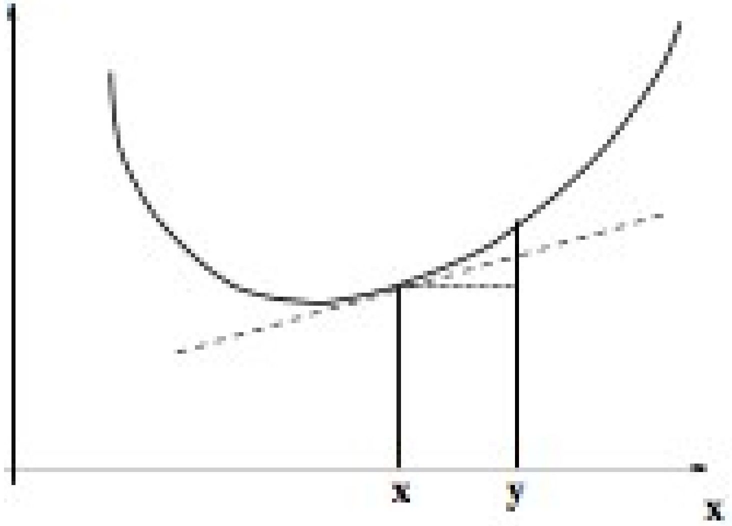
$$\forall (x, y) \in C^2, \forall \alpha \in [0, 1],$$
$$f(\alpha x + (1 - \alpha) y) \leq \alpha f(x) + (1 - \alpha) f(y)$$



Preliminary: Convex set and convex function

- Geometric interpretation

$$\forall (x,y) \in C^2,\ f(y) - f(x) \geq f'(x)(y-x)$$



Q: What if =?

Preliminary: Convex set and convex function

- **Definition**: Strictly convex function
  - Let f be a function: C → R. It is **strictly convex** in convex C if

$$\forall (x,y) \in C^2, x \neq y, \forall \alpha \in [0,1],$$
$$f(\alpha x + (1-\alpha)y) < \alpha f(x) + (1-\alpha) f(y)$$

  - If f is strictly convex, then f is convex
  - If f is convex on an open set C, then f is also continuous on C.

Q: Is $f(x)=x^4$ strictly convex?

Preliminary: Convex set and convex function

- **Theorem**: Convexity and **first-order derivative**

Let $\Omega \in E$ be an open set in a normed vector space E *and* $\mathrm{C} \in \Omega$ is a convex subset of Ω.

Assume $\mathrm{f} : \Omega \to \mathbb{R}$ is differentiable on Ω, then we have

- f is **convex** on C if and only if

$$\forall (x, y) \in C^2,\ f(y) - f(x) \geq f'(x)(y - x)$$

- f is **strictly convex** on C if and only if

$$\forall (x, y) \in C^2,\ x \neq y,\ f(y) - f(x) > f'(x)(y - x)$$

Preliminary: Convex set and convex function

- Theorem: Convexity and **second-order derivative**

$$\mathit{Let}\ \ \Omega \in E \text{ be an open set in } \mathbb{R}^n \ \mathit{and}\, \mathrm{C} \in \Omega \text{ be a convex subset of } \Omega.$$

$$\mathit{Assume}\ \ \mathrm{f} : \Omega \rightarrow \mathbb{R} \text{ is twice differentiable on } \Omega, \text{ then we have}$$

- f is convex on C if and only if

$$\forall (x, y) \in C^2,\ f''(x)(y-x, y-x) \geq 0$$

- Equivalent condition if $C = E = \mathbb{R}^n$

$$\forall (x, h) \in (\mathbb{R}^n)^2, f''(x)(h, h) = h^\intercal \nabla^2 f(x) h \geq 0$$

Hessian $\nabla^2 f(x)$ is positive semi-definite.

Outline

- Introduction
- **Basic theory of Optimization**
  - Existence of solutions
  - Uniqueness of the solution
- Optimization methods without constraint
- Optimization methods with constraints

Problem definition

- Minimize a real-valued function f

$$(P) \quad \min_{x \in C} f(x) \qquad\qquad C \subset \mathbb{R}^n$$

- If C is empty, (P) has no solution.
- If C is finite, (P) has at least one solution.
- Next, consider non-empty C having infinite elements

Compact and closed case

- Assume **C is compact** and non-empty

$$(P) \quad \min_{x \in C} f(x) \qquad\qquad C \subset \mathbb{R}^n$$

- **Theorem**

f is continous on non-empty compact C

$\implies$ (P) admits at least one solution.

Q: What if C is not compact?

Ex: $f(x) = 1/x$, $C = (0, \infty)$, $f(x) > 0$, no minimal solution exists on C.

Compact and closed case

- Assume C is closed and non-empty
- **Definition (coercive)**

$$f \text{ is coercive if } f(x) \to \infty \text{ when } \|x\| \to \infty$$

Theorem

$$f \text{ is continous on non-empty closed } C \text{ and } f \text{ is coercive}$$
$$\implies \text{(P) admits at least one solution}$$

Q: $f(x) = \sin(x)x$, $C = [0, 10^{10}]$, does $f(x)$ admit a minimal solution on C?

Uniqueness of solution: convex case

- **Theorem (convex f)**

Assume C is a convex subset of $\mathbb{R}^n$, and f is convex on C, then the solution set of (P) is either empty or convex.

- **Theorem (strictly convex f)**

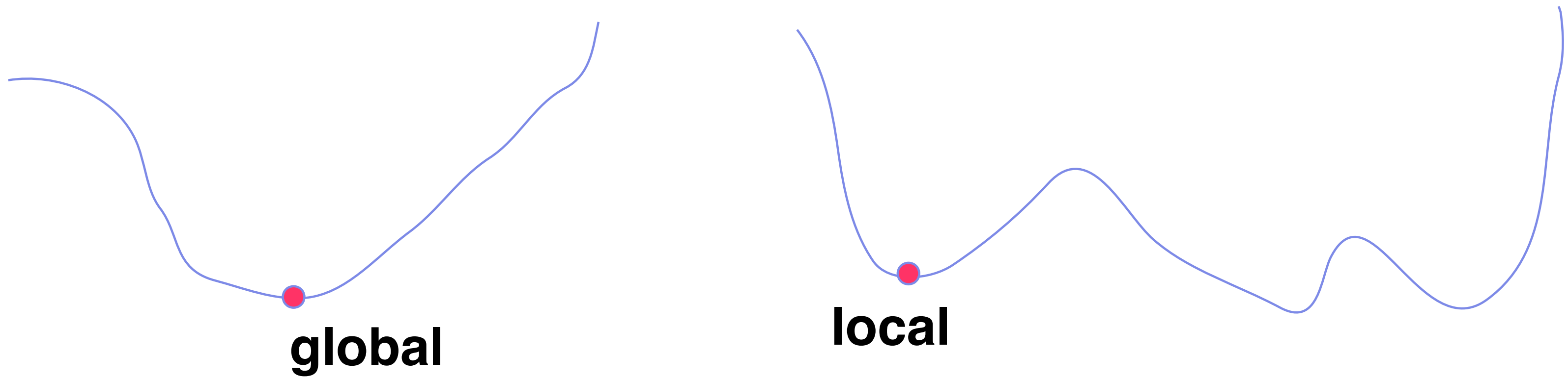
Assume C is a convex subset of $\mathbb{R}^n$, and f is strictly convex on C, then the solution set of (P) has at most one element.

Uniqueness of solution: convex case

- **Theorem** (local optimum is global optimum)

> Assume C is a convex subset of $\mathbb{R}^n$, and f is convex on C, then any local optimum of f is also a global optimum of f

Definition (local and global optimum)



Outline

- Introduction
- Basic theory of Optimization
- **Optimization methods without constraint**
  - Optimality conditions
  - Numerical algorithms
  - Convergence guarantee of algorithms
- Optimization methods with constraints

Problem definition

- Minimize a real-valued function f

$$(P_{sc}) \quad \min_{x \in O} f(x) \quad \text{open set } O \subset \mathbb{R}^n$$

- **Definition (local optimum)**

We call x^* is a local optimum of f if

$$\exists \epsilon > 0, s.t. \forall x \in B(x^*, \epsilon), \quad f(x^*) \leq f(x)$$

Note: $B(x, r)$ is a open ball of radius r centered at x

Necessary conditions of optimality

- **Theorem**: First-order conditions

$$\text{Let } x^* \in O. \text{ Assume } f \text{ is differentiable at } x^*. \text{ Then}$$
$$\text{x}^* \text{ is a } \textbf{local minimum} \text{ of } f \implies \nabla f(x^*) = 0$$

This condition is not true if O is not open (see optimization with constraints)

- **Definition**: critical point

$$\text{We call } x \in O \text{ is a } \textbf{criticial point} \text{ of } f \text{ if } \nabla f(x) = 0$$

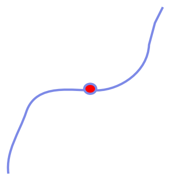
Necessary conditions of optimality

- **Theorem:** Second-order conditions

Let $x^* \in O$. Assume f is twice differentiable at x^*. Then
x* is a **local minimum** of $f \implies \nabla^2 f(x^*)$ is positive semi-definite

- Positive semi-definite is necessary, but not sufficient

Ex: $f(x) = x^3$, $f'(0) = 0$, $f''(0) \geq 0$, but 0 is not a local optimum



Sufficient conditions of optimality

- **Theorem:** First-order conditions

> Let $x^* \in O$. Assume $O \subset \mathbb{R}^n$ is open and convex,
> f is convex on O and differentiable at x^*. Then
> $\nabla f(x^*) = 0 \Longrightarrow x^*$ is a **global minimum** of f

Remark: this is very particular as f is convex.

Sufficient conditions of optimality

- **Theorem**: Second-order conditions

Let $x^* \in O$ such that $\nabla f(x^*) = 0$.

Assume f is twice differentiable at x^*, then

- If $\nabla^2 f(x^*)$ is positive definite $\Rightarrow x^*$ is a local minimum of f
- If f is twice differentiable over O, and

  $\exists \epsilon > 0$ such that $B(x^*, \epsilon) \subset O$, and $\forall x \in B(x^*, \epsilon)$,

  $\nabla^2 f(x)$ is positive semi-definite

  $\Rightarrow x^*$ is a local minimum of f

Analytical solutions

- General strategy to solve $(P_{sc}) \quad \min_{x \in O} f(x) \quad \text{open set } O \subset \mathbb{R}^n$
  - Demonstrate the existence (and uniqueness) of the solutions
  - Find critical points

$$\text{Find } x^* \in O \text{ such that } \nabla f(x^*) = 0.$$

  - Stop in some particular case

    e.g. f is convex on convex O: all the critical points are global optima
  - Search for local optima among all the critical points
    - Use second-order conditions $\text{Is } \nabla^2 f(x^*) \text{ positive definite?}$

Analytical solutions

- Example: minimize a strictly convex quadratic function

$$\min_{x\in\mathbb{R}^n} f(x) = \frac{1}{2}x^\intercal A x - b^\intercal x + c$$

with (symmetric) poisitive definite A, $b \in \mathbb{R}^n$, and $c \in \mathbb{R}$

- This problem admits a unique solution x^*
  - Existence: f is continous on $\mathbb{R}^n$ (closed,non-empty), and coercive (due to A positive definite)
  - Uniqueness: f is strictly convex on convex $\mathbb{R}^n$
- The solution solves a linear system: $Ax^* = b$

Numerical solutions

- Beyond quadratic function, it is non-trivial to find analytical solutions.
- Numerical methods allow to
  - **Find critical points**
    - Linear system (Ax=b): matrix factorization (LU, Cholesky), iterative methods (steepest descent, conjugate gradient)
    - Non-linear system: iterative methods (Newton, non-linear conjugate gradient)
  - Challenges: Cost and time of computations? Precision of solutions? Convergence? Find all the critical points?

Numerical solutions

- Numerical methods allow to
  - **Check optimality of critical points**: study eigenvalues of Hessian
    - Iterative methods (QR, power method)
    - Challenges: Cost and time of computations? Precision of solutions? Convergence?
  - Consequently, in many cases, we can only find **approximate** critical points or local optima.
  - We shall study several classical numerical algorithms for this purpose.

Gradient descent algorithm

- **Definition**: Descent direction

> Let $x \in O$. Assume f is differentiable at x.
>
> We say that d is a descent direction at x if $\nabla f(x)^{\intercal} d < 0$

Remark**:** It only makes sense to discuss descent directions at non-critical points

$$\text{If } d = -\nabla f(x) \neq 0, \text{ then}$$
$$\nabla f(x)^* d = -\|\nabla f(x)\|^2 < 0.$$

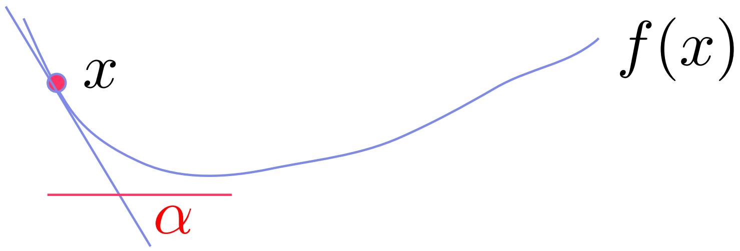
=> **Existence of steepest descent direction**

Gradient descent algorithm

- **Proposition**: descent direction allows to decrease f

Assume f is continously differentiable on O. Let $x \in O$ and $d \in \mathbb{R}^n$. If d is a descent direction of f at x, then there exists $\eta > 0$ such that

$$\forall \alpha \in (0, \eta],\ x + \alpha d \in O \text{ and } f(x + \alpha d) < f(x)$$



Gradient descent algorithm

- Base algorithm

1. Initialize $x = x_0$.
2. For $k = 0, 1, 2, \cdots$ do
3. Calculate a descent direction d_k such that $\nabla f(x_k)^\intercal d_k < 0$
4. Compute a step-size $\alpha_k > 0$
5. Update $x_{k+1} = x_k + \alpha_k d_k$
6. Check stopping criteria
7. Endfor

- Steepest descent direction $d_k = -\nabla f(x_k)$

Gradient descent algorithm

- Search for step-sizes 4. Compute a step-size $\alpha_k > 0$
- Stopping criteria 6. Check stopping criteria

- Gradient vanishing: $\|\nabla f(x_k)\| \leq \epsilon_1(\|\nabla f(x_0)\| + \eta)$
- Stagnation: $\|x_{k+1} - x_k\| \leq \epsilon_2(\|x_k\| + \eta)$
- Maximal number of iterataions K: $k \leq K$.

Gradient descent algorithm: Quadratic example

- Quadratic function

$$\min_{x \in \mathbb{R}^n} f(x) = \frac{1}{2} x^\intercal A x - b^\intercal x + c$$

with (symmetric) poisitive definite A, $b \in \mathbb{R}^n$, and $c \in \mathbb{R}$

- Steepest descent direction $\quad d_k = -\nabla f(x_k) = -(Ax_k - b)$

- Optimal step size: $\quad \min_\alpha \phi(\alpha) = f(x_k + \alpha d_k)$

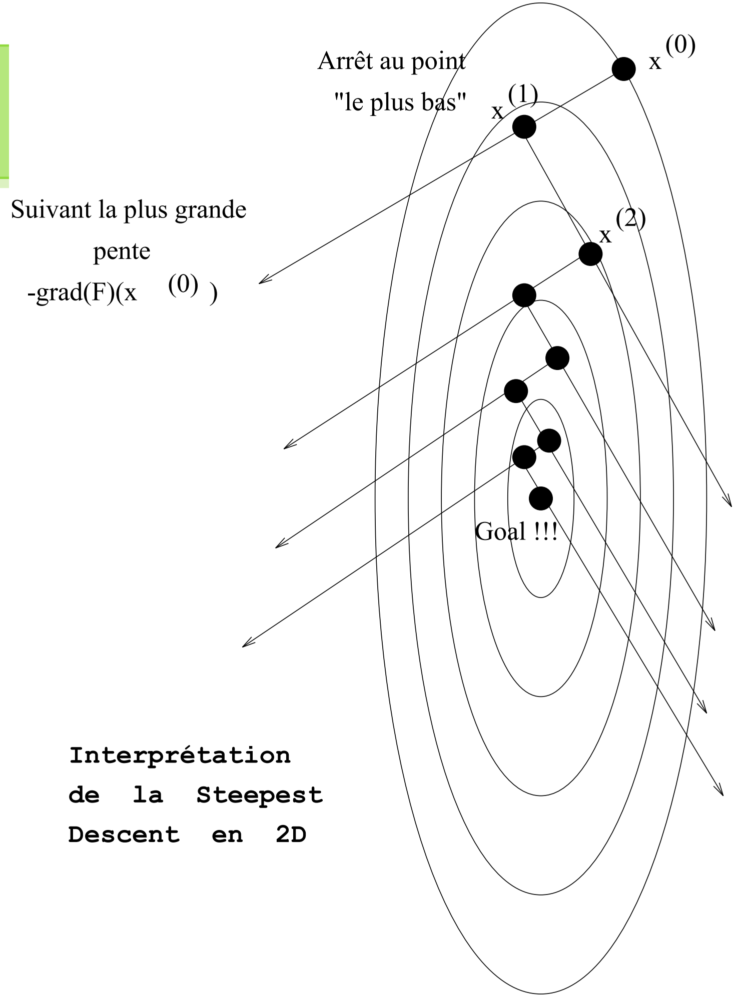
$$\phi'(\alpha) = \nabla f(x_k + \alpha d_k)^\intercal d_k = 0 \Leftrightarrow \alpha = \frac{d_k^\intercal d_k}{d_k^\intercal A d_k}$$

$$\phi''(\alpha) = d_k^\intercal \nabla^2 f(x_k + \alpha d_k) d_k = d_k^\intercal A d_k > 0 \quad \text{if} \quad d_k \neq 0$$

Quadratic example

- Steepest descent

1. Initialize $x = x_0$.
2. For $k = 0, 1, 2, \cdots$ do
3. Calculate $d_k = b - Ax_k$
4. Compute step-size $\alpha_k = \frac{d_k^\top d_k}{d_k^\top A d_k}$
5. Update $x_{k+1} = x_k + \alpha_k d_k$
6. Check stopping criteria
7. Endfor



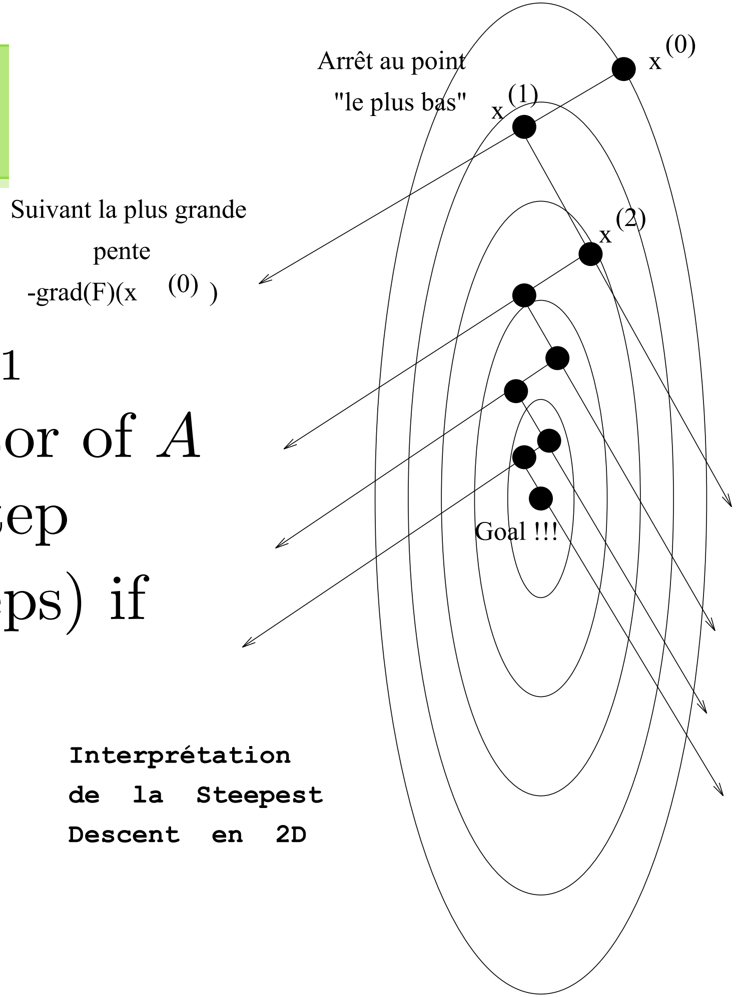
Interprétation de la Steepest Descent en 2D

Quadratic example

- Some properties
  - $\forall$k =0,1,2,$\cdots$, d_k is orthogonal to d_{k+1}
  - If $x^* - x_0 = \beta u$ where u is a eigenvector of A then the algorithm converges in one step
  - Very slow convergence (need many steps) if

$$\kappa_2(A) = \frac{\lambda_{\max}(A)}{\lambda_{\min}(A)} \text{ is large}$$

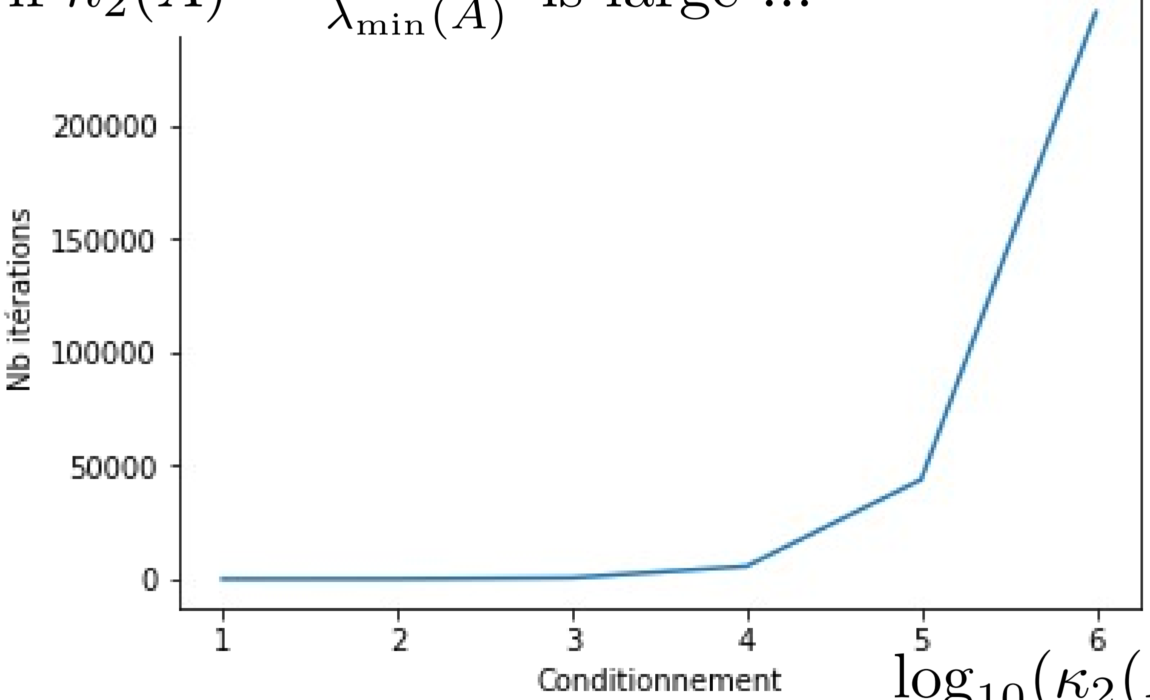
$\kappa_2(A)$: condition number of A



Interprétation de la Steepest Descent en 2D

Quadratic example

What if $\kappa_2(A) = \frac{\lambda_{\max}(A)}{\lambda_{\min}(A)}$ is large ...



$\log_{10}(\kappa_2(A))$

Newton's Method

- Application of the Newton method to find a root of an equation

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

- Let $x_k \in \mathbb{R}^n$. Assume m is a local approximation of f near x_k,

$$m(x) = f(x_k) + \nabla f(x_k)^\intercal (x - x_k) + \frac{1}{2}(x - x_k)^\intercal \nabla^2 f(x_k)(x - x_k)$$

If $\nabla^2 f(x_k)$ is positive definite, then the minimum of m is

$$x^* = x_k - \nabla^2 f(x_k)^{-1} \nabla f(x_k)$$

Descent direction $\quad \mathrm{d}_k = -\nabla^2 f(x_k)^{-1} \nabla f(x_k)$

Newton's Method

- Basic idea (assume invertible and positive definite Hessian)

1. Initialize $x = x_0$.
2. For $k = 0, 1, 2, \cdots$ do
3. Calculate a descent direction $d_k = -\nabla^2 f(x_k)^{-1} \nabla f(x_k)$
4. Set the step-size $\alpha_k = 1$ (constant step-size version)
5. Update $x_{k+1} = x_k + \alpha_k d_k$
6. Check stopping criteria
7. Endfor

- In practice, find d_k by solving $\nabla^2 f(x_k) d_k = -\nabla f(x_k)$

Newton's Method

- **Theorem**: local convergence with constant step-size

Let $x^* \in O$, with open and convex O, and assume

- f is twice continously differentiable on O
- $x \mapsto \nabla^2 f(x)$ is Lipschitz continuous on O

  (there is $\gamma > 0$ s.t. $\|\nabla^2 f(y) - \nabla^2 f(x)\| \leq \gamma \|y - x\|$)
- $\nabla f(x^*) = 0$ and $\nabla^2 f(x^*)$ is positive definite

Then there exists $(\delta, K) \in \mathbb{R}^2_+$ such that

$$\|x_0 - x^*\| \leq \delta \Rightarrow \|x_{k+1} - x^*\| \leq K \|x_k - x^*\|^2$$

Moreoever, if $\delta K < 1$, then x_k is (quadratically) convergent

Non-linear least-square problem

- Problem

$$\min_{x \in \mathbb{R}^n} f(x) = \frac{1}{2}\|F(x)\|^2$$

with $F : \mathbb{R}^n \to \mathbb{R}^p$ continuously differentiable on $\mathbb{R}^n$.

- **Definition**: Jacobian

$$J_F(x) = \frac{\partial F}{\partial x} \in \mathbb{R}^{p \times n}$$

Let $J_F(x)$ be the Jacobian matrix of F evaluated at x

- $f(x+d) = f(x) + J_F(x)d + o(\|d\|)$
- $J_F(x)$ is continous on $\mathbb{R}^n$

Gauss-Newton method

$$\min_{x \in \mathbb{R}^n} f(x) = \frac{1}{2}\|F(x)\|^2$$

- For a non-linear least-square problem, Hessian can be approximated by the Jacobian near global optimum by

$$\nabla^2 f(x_k) \approx J_F(x_k)^\intercal J_F(x_k)$$

- Newton method → Gauss-Newton method

3. Calculate a descent direction $d_k = -(J_F(x_k)^\intercal J_F(x_k))^{-1} \nabla f(x_k)$

- In practice, find d_k by solving $J_F(x_k)^\intercal J_F(x_k) d_k = -\nabla f(x_k)$

Gauss-Newton method

$$\min_{x\in\mathbb{R}^n} f(x) = \frac{1}{2}\|F(x)\|^2$$

- Interpretation: Linearization of F near x_k

$$(P_k) \qquad \min_{d\in\mathbb{R}^n} g_k(d) = \frac{1}{2}\|F(x_k) + J_F(x_k)d\|^2$$

- (P_k) is a quadratic problem
- (P_k) optimal solution results in the Gauss-Newton direction

Optimal d_k: $J_F(x_k)^\intercal J_F(x_k)d_k = -J_F(x_k)^\intercal F(x_k) = -\nabla f(x_k)$

- If $\mathrm{rank} J_F(x_k)$ is n, then (P_k) admits a unique solution

Gauss-Newton method

$$\min_{x\in\mathbb{R}^n} f(x) = \frac{1}{2}\|F(x)\|^2$$

- **Theorem**: local convergence

Let $x^* \in O$, with open and convex O, and assume

- f is twice continously differentiable on O
- $J_F(x^*)$ has rank n
- $\nabla f(x^*) = 0$ and $\nabla^2 f(x^*)$ is positive definite

Then there exists $\delta \in \mathbb{R}_+$ such that

$$\|x_0 - x^*\| \leq \delta \Rightarrow \|x_k - x^*\| \to 0, \quad k \to \infty$$

Example: Convergence of Newton's method?

- A non-linear least-square problem
  - Estimate parameters of Michaelis-Menten kinetics (models of enzyme kinetics in biology)

$$\mathrm{V(S)} = \mathrm{V}_{\max} \frac{S}{K_m + S}$$

  - Use p observations (S,V(S)) at S = S_i , i=1,…,p

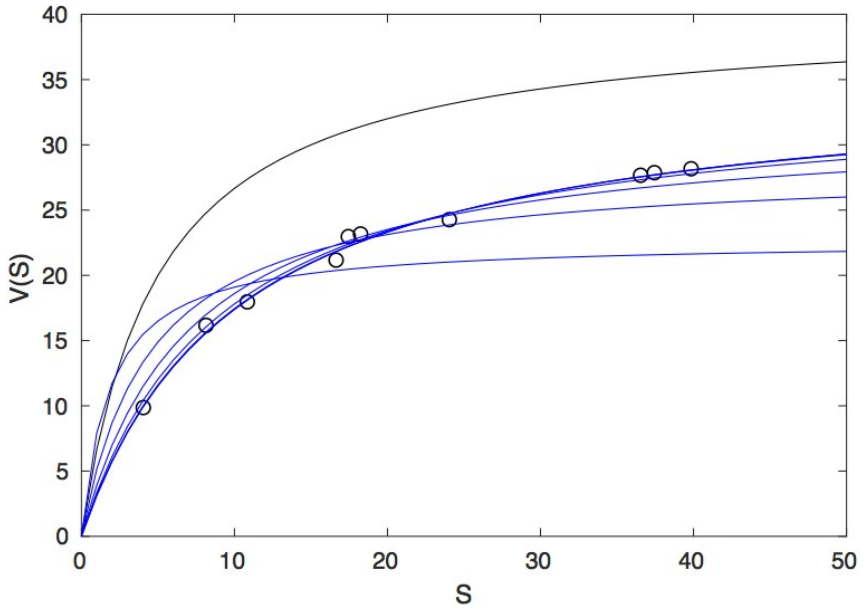
$$\min_{(V_{\max}, K_m) \in \mathbb{R}^2} f(V_{\max}, K_m) = \frac{1}{2} \sum_{i=1}^{p} (V(S_i) - V_{\max} \frac{S_i}{K_m + S_i})^2$$

  - Apply Newton's method to minimize f

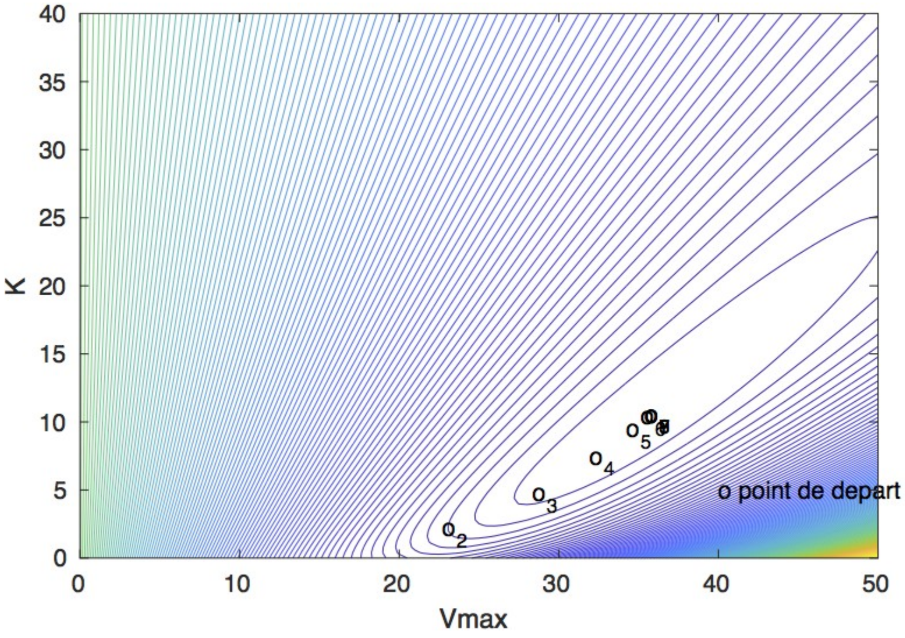
Example: Convergence of Newton's method?

- Convergence : depends on initialization $x_0 = [40,5]$

Model fit to observations



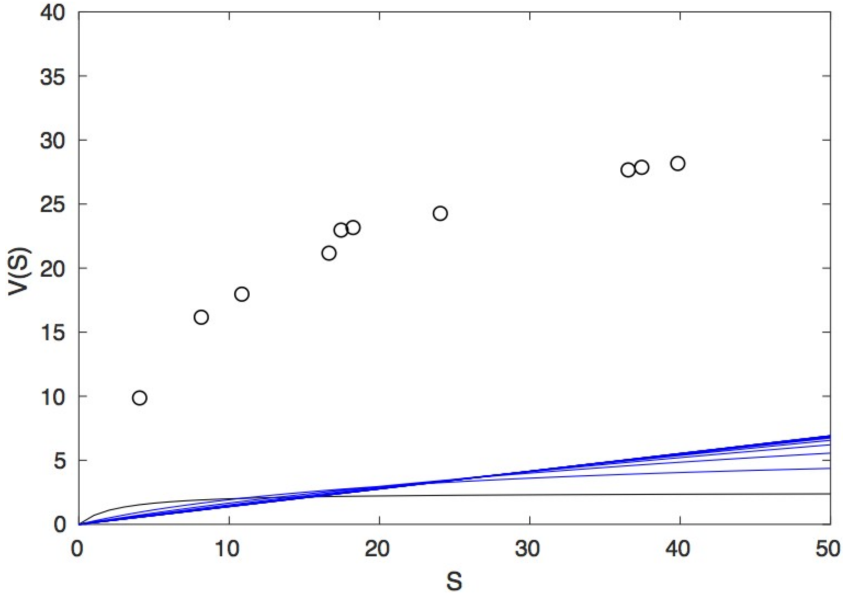
Level set of f and iterations



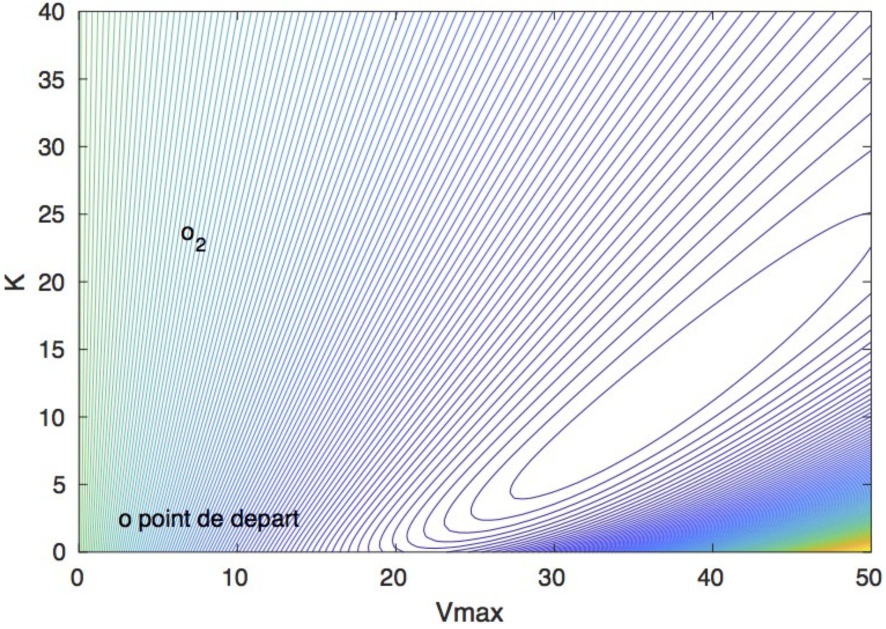
Example: Convergence of Newton's method?

- Divergence : depends on initialization x_0= [2.5,2.5]

Model fit to observations



Level set of f and iterations



Globalization of descent methods

- Problem: achieve global convergence to critical points

$$\forall x_0 \in O, \text{ the sequence } (x_k) \text{ converges towards to a critical point of } f$$

- Classical strategies
  - **Line search**: find suitable step size α_k
  - Trust-region methods
  - Regularization methods

Line search

- Idea: search along descent direction to minimize f

If d is a descent direction of f at x, then there exists $\eta > 0$ such that

$$\forall \alpha \in (0, \eta],\ x + \alpha d \in O \text{ and } f(x + \alpha d) < f(x)$$

- Line search: naive strategy

$$\text{Given a direction } d, \text{ compute } \alpha \text{ such that } f(x + \alpha d) < f(x)$$

Gradient descent with line search

- Base algorithm with line search

1. Initialize $x = x_0$.
2. For $k = 0, 1, 2, \cdots$ do
3. Calculate a descent direction d_k such that $\nabla f(x_k) d_k < 0$
4. Compute a step-size α_k such that $f(x_k + \alpha_k d_k) < f(x_k)$
5. Update $x_{k+1} = x_k + \alpha_k d_k$
6. Check stopping criteria
7. Endfor

Gradient descent with line search

- A decreasing sequence is not always optimal
- Example $f(x) = x^2$

1. Initialize $x = x_0 = 2$.
2. For $k = 0, 1, 2, \cdots$ do
3. Calculate a descent direction $d_k = -1$
4. Compute a step-size $\alpha_k = 2^{-(k+1)}$
5. Update $x_{k+1} = x_k + \alpha_k d_k$
6. Check stopping criteria
7. Endfor

$$x_k = 1 + 2^{-k} \to 1$$

1 is not a critical point of f

Gradient descent with line search

- Wolfe conditions to guarantee global convergence

Let $\beta_1 \in (0,1)$, $\beta_2 \in (\beta_1, 1)$ and d be a descent direction of f at x

We say $\alpha > 0$ satisfies Wolfe conditions if:

- Sufficient decrease: $f(x + \alpha d) \leq f(x) + \beta_1 \alpha \nabla f(x)^\intercal d$
- Sufficient progress: $\nabla f(x + \alpha d)^\intercal d \geq \beta_2 \nabla f(x)^\intercal d$

Gradient descent with line search

- **Theorem**: existence of good step-size

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a differentiable function, $x \in \mathbb{R}^n$
and d is a descent direction. Assume f is bounded below along d,

$$\exists c \in R, \forall \alpha \geq 0, f(x + \alpha d) \geq c$$

Then

- $\forall \beta_1 \in (0,1)$, $\exists \eta > 0$ s.t. sufficient descrease cond. holds if $\alpha \in (0, \eta)$
- $\forall \beta_1 \in (0,1)$, $\forall \beta_2 \in (\beta_1, 1)$, $\exists \alpha > 0$ s.t. Wolfe conditions hold

Gradient descent with line search

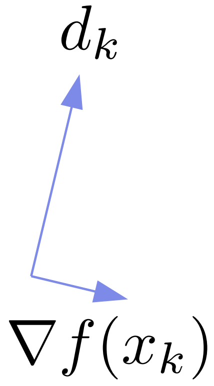
- **Theorem**: global convergence

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a continously differentiable function

- f is bounded below
- $x \mapsto \nabla f(x)$ is Lipschitz continuous

Then the gradient descent algoirthm with line search which satifsies Wolfe conditions at each step results in

$$\lim_{k\to\infty} \nabla f(x_k) = 0 \qquad \text{or} \qquad \lim_{k\to\infty} \frac{\nabla f(x_k)^\intercal d_k}{\|\nabla f(x_k)^\intercal\| \|d_k\|} = 0$$



Gradient descent with line search

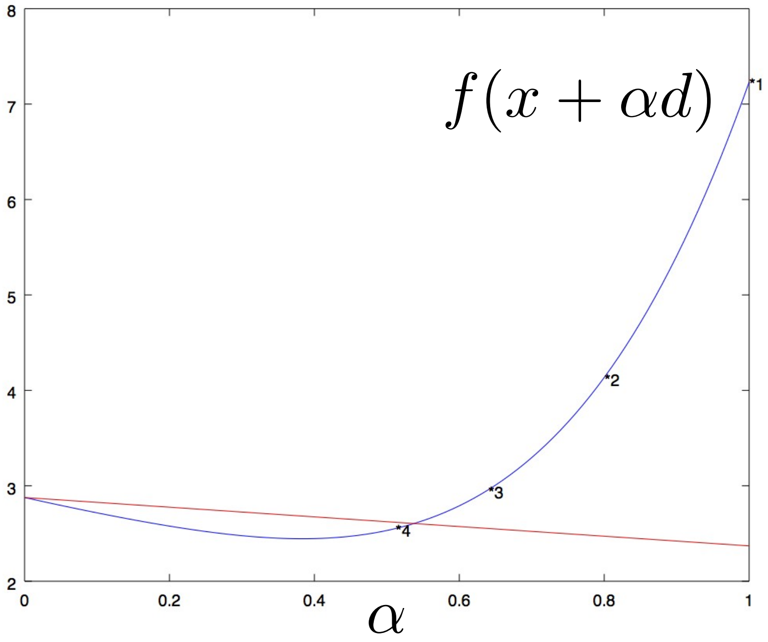
- Backtracking line search

Input: x, descent direction d, $\beta_1 \in (0,1)$, $\rho \in (0,1)$
1. Initialize $\alpha_0 > 0$
2. For $k = 0, 1, 2, \cdots$ do
3. If α_k verifies the first Wolfe condition, stop
4. Calculate $\alpha_{k+1} = \rho\alpha_k$
5. Endfor

- This approach is simple, and it requires no gradients of f.
- But the second Wolfe condition is not always true.

Backtracking line search

- Sufficient decrease: $f(x+\alpha d) \leq f(x) + \beta_1 \alpha \nabla f(x)^\intercal d$



Gradient descent with line search

- Bi-section line search for Wolfe conditions

Input: x, descent direction d, $\beta_1 \in (0,1)$, $\beta_2 \in (\beta_1, 1)$

1. Initialize $\alpha_0 > 0$, $a = 0$, $b = \infty$
2. For $k = 0, 1, 2, \cdots$ do
3. If α_k satisfies two Wolfe conditions, stop
4. If α_k does not satisfy the first Wolfe condition,

$$b = \alpha_k, \alpha_{k+1} = \frac{b+a}{2}$$

else (α_k does not satisfy the second Wolfe condition),

$$a = \alpha_k, \alpha_{k+1} = \begin{cases} 2a \text{ if } b = \infty \\ \frac{a+b}{2} \text{ if } b < \infty \end{cases}$$

6. Endfor

Outline

- Introduction
- Basic theory of Optimization
- Optimization methods without constraint
- **Optimization methods with constraints**
  - Optimality conditions
  - Numerical algorithms

Optimization methods with constraints

- Minimize a real-valued function under a constraint set

$$(P) \quad \min_{x \in C} f(x) \quad C \subset \mathbb{R}^n$$

- Various forms of constraints
  - C is a closed set def. by equality or inequality equations

$$C = \{x \in \mathbb{R}^n | h(x) = 0, g(x) \leq 0\}$$

  - C is an open set, and f is differentiable on $\mathbb{R}^n$, then

$$x^* \text{ is a local optimum of } (P) \Rightarrow \nabla f(x^*) = 0$$

Not true for a closed C

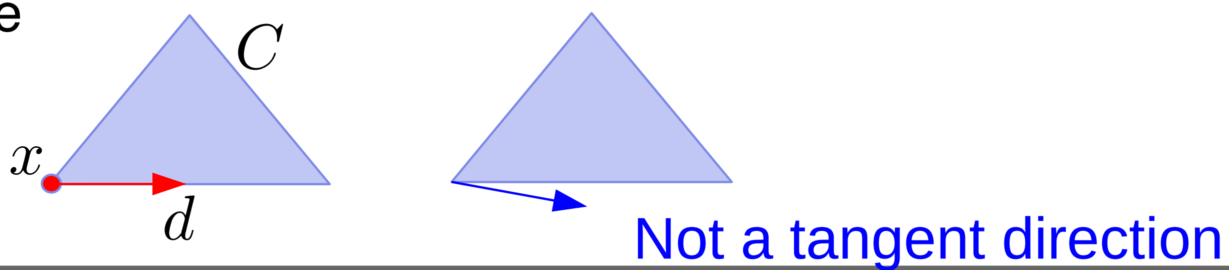
Necessary conditions of optimality

- **Definition**: tangent direction

Let $x \in C \subset \mathbb{R}^n$. $d \in \mathbb{R}^n$ is a **tangent direction** of C at x if there exists a sequence $(\alpha_k, d_k) \in \mathbb{R}^+ \times \mathbb{R}^n$ such that

$$\begin{aligned} \forall k \in \mathbb{N}, &\quad x_k = x + \alpha_k d_k \in C \\ d_k \to d, &\quad k \to \infty \\ \alpha_k \to 0, &\quad k \to \infty \end{aligned}$$

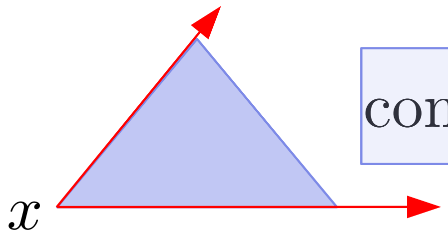
- Example



Necessary conditions of optimality

- **Definition**: tangent cone

Let $x \in C \subset \mathbb{R}^n$. The **tangent cone** $T(C, x)$ of C at x is the set of all the tangent directions of C at x.



cone: $d \in T(C, x) \Rightarrow \alpha d \in T(C, x)$ for all $\alpha \geq 0$

- **Theorem**: local optimality and tangent cone

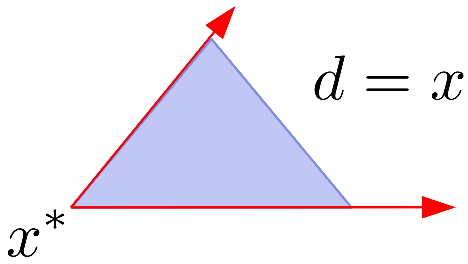
Let f be a differentiable function on $\mathbb{R}^n$. If $x^* \in C$ is a local optimum of (P), then $\forall d \in T(C, x^*), \nabla f(x^*)^\intercal d \geq 0$

Necessary conditions of optimality

- **Special case**: C is convex

Let f be a differentiable function on $\mathbb{R}^n$ and C be a convex set. If $x^* \in C$ is a local optimum of (P), then

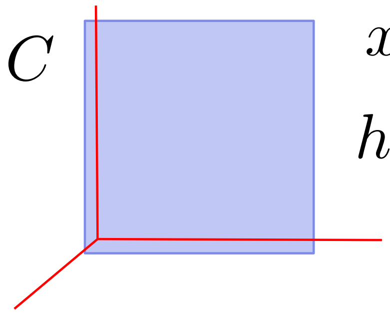
$$\forall x \in C, \nabla f(x^*)^\intercal (x - x^*) \geq 0$$



$$d = x - x^* \in T(C, x^*)$$

Equality constraints

- Consider $(P_h) \quad \min_{x \in C} f(x) \qquad C = \{x \in \mathbb{R}^n | h(x) = 0\}$
- Specified by a **vector-valued function** $h : \mathbb{R}^n \to \mathbb{R}^p$
- Example



$$x = (x_1, x_2, x_3)$$
$$h(x) = x_1 = 0$$

- **Qualifications of constraints**: when a tangent cone T(C,x) equals to

$$\{d \in \mathbb{R}^n | \nabla h(x)^\intercal d = 0\}$$

How to solve optimization with equality constraints?

- Introducing Lagrange multiplier

$$\begin{aligned} L \quad : \quad & \mathbb{R}^n \times \mathbb{R}^p \to \mathbb{R} \\ & (x, \lambda) \mapsto f(x) + \lambda^\intercal h(x) \end{aligned}$$

- Theorem (KKT, Karush-Kuhn-Tucker)

For the problem (P_h), if the following conditions hold

- f and h are continuously differentiable near x^*
- x^* is a local optimum of (P_h)
- $T(C, x^*) = \{d \in \mathbb{R}^n | \nabla h(x^*)^\intercal d = 0\}$

then $\exists \lambda^* \in \mathbb{R}^p$ s.t. $\nabla_x L(x^*, \lambda^*) = 0, h(x^*) = 0$

Example

- Quadratic problem with affine constraints

$$\min_{x \in \mathbb{R}^n} f(x) = \frac{1}{2} x^\intercal A x - b^\intercal x + c \quad \text{s.t.} \;\; Ex = d$$

where A is a positive definite matrix, E has full rank $p \leq n$

- This problem has a unique solution

Existence: f is continous and coercive on closed and non-empty C.
Uniquenes: f is strictly convex $(\forall x \in \mathbb{R}^n, \nabla^2 f(x) = A)$ on convex C.

The solution x^* satisfies a linear system:

$$Ax^* + E^\intercal \lambda^* = b, \quad Ex^* = d$$

Second-order optimality conditions

- Theorem (KKT, Karush-Kuhn-Tucker)

For the problem (P_h), if the following conditions hold

- f and h are **twice** continuously differentiable near x^*
- x^* is a local optimum of (P_h)
- $T(C, x^*) = \{d \in \mathbb{R}^n | \nabla h(x^*)^\intercal d = 0\}$

then $\exists \lambda^* \in \mathbb{R}^p$ s.t. $\nabla_x L(x^*, \lambda^*) = 0, h(x^*) = 0,$ and

$$\forall d \in T(C, x^*), \quad d^\intercal \nabla^2_{xx} L(x^*, \lambda^*) d \geq 0$$

Sufficient optimality conditions

- Special case: **Affine constraints and convex f**
  - affine: $h(x) = Ex - d$
- **Theorem:** sufficient conditions

For the problem (P_h), if the following conditions hold

- f is convex on C, h is affine
- f is continuously differentiable near x^*

Then x^* is a local (global) optimum of (P_h)

$$\iff \exists \lambda^* \in \mathbb{R}^p \text{ s.t. } \nabla_x L(x^*, \lambda^*) = 0, h(x^*) = 0$$

Analytical solution: general idea

- Assume f and h are differentiable
  - Demonstrate the existence and unicity of the solutions of (P_c)
  - Find solutions by solving

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

  - Check constraint qualifications
  - Stop in some particular cases
    - If h is affine and f is convex
  - Find other solutions and check the second order optimality condition

$$\forall d \in T(C, x^*), \quad d^\intercal \nabla^2_{xx} L(x^*, \lambda^*) d \geq 0$$

Numerical solution

- **Basic idea**: transform a problem with constraints into a problem without constraints, by adding penalties
- Lagrange method (a max-min game):

$$\max_{\lambda} \min_{x} f(x) + \lambda^{\intercal} h(x)$$

- Minimal of x does not always exist: add a quadratic penalty (ADMM)

$$f(x) + \lambda^{\intercal} h(x) + \frac{\mu}{2} \|h(x)\|^2$$

$\mu > 0$: encourage that $h(x) \approx 0$

- What if using only the quadratic penalty?