Optimisation

Lecture 11 - The Gradient Method

Fall semester - 2024

Dr. Eng. Valentin Leplat
Innopolis University
November 9, 2024

Outline

Introduction

Line-search methods

Objective: find an optimal solution of the problem

$$\min_{x \in \mathbb{R}^n} f(x) \tag{1}$$

The iterative algorithms that we will consider are of the form

$$x_{k+1} := x_k + t_k d_k, \quad k = 0, 1, ...$$

with:

- d_k - direction
- t_k - stepsize.

We will limit ourselves to descent directions.

Definition

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a continuously differentiable function over $\mathbb{R}^n$. A vector $d \neq 0 \in \mathbb{R}^n$ is called a descent direction of f at x if the directional derivative $\frac{\partial f}{\partial d}(x)$ is negative, meaning that

$$\frac{\partial f}{\partial d}(x) = \nabla f(x)^T d < 0.$$

The Descent Property of Descent Directions

Lemma

Let f be a continuously differentiable function over $\mathbb{R}^n$, and let $x \in \mathbb{R}^n$. Suppose that d is a descent direction of f at x. Then there exists $\epsilon > 0$ such that

$$f(x + td) < f(x)$$

for any $t \in (0, \epsilon]$.

Proof. Since $\frac{\partial f}{\partial d}(x) < 0$, it follows from the definition of the directional derivative that

$$\lim_{t \to 0^+} \frac{f(x + td) - f(x)}{t} = \frac{\partial f}{\partial d}(x) < 0 \tag{2}$$

Therefore, $\exists \epsilon > 0$ such that

$$\frac{f(x + td) - f(x)}{t} < 0 \tag{3}$$

for any $t \in (0, \epsilon]$, which readily implies the desired result.

Schematic Descent Direction Method

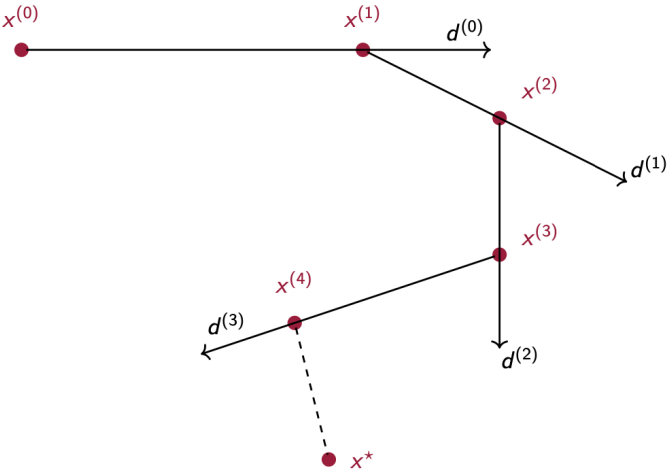
Initialization: pick $x_0 \in \mathbb{R}^n$ arbitrarily.
General step: for any $k = 0, 1, 2, ...$ set

1. pick a descent direction d_k.
2. find a step size t_k satisfying $f(x_k + t_k d_k) < f(x_k)$.
3. set $x_{k+1} = x_k + t_k d_k$.
4. if a stopping criteria is satisfied, then STOP and x_{k+1} is the output.

Of course, many details are missing in the above schematic algorithm:

- What is the starting point?
- How to choose the descent direction?
- What stepsize should be taken?
- What is the stopping criteria?

Schematic Descent Direction Method



Stepsize Selection Rules

- **constant stepsize** - $t_k = \bar{t}$ for any k.
- **exact stepsize**[1] - t_k is a minimizer of f along the ray $x_k + td_k$:

$$t_k = \text{argmin}_{t \geqslant 0} f(x_k + td_k) \tag{4}$$

- **backtracking**[2] - The method requires three parameters: $s > 0$, $\beta \in (0, 1)$ and $\gamma \in (0, 1)$. Here we start with an initial stepsize $t_k = s$.
  While
  $$f(x_k) - f(x_k + t_k d) < -\beta t_k \nabla f(x_k)^T d_k$$
  set $t_k := \gamma t_k$.

Sufficient Decrease Property:

$$f(x_k) - f(x_k + t_k d) \geqslant -\beta t_k \nabla f(x_k)^T d_k (\geqslant 0) \tag{5}$$

A stepsize t_k for which Inequality (5) holds for some $\beta \in (0, 1)$ is said to satisfy the *Armijo rule*.

[1] exact line search
[2] inexact line search

Exact Line Search for Quadratic Functions

$f(x) = x^T Ax + 2b^T x + c$ where A is an $n \times n$ positive definite matrix, $b \in \mathbb{R}^n$ and $c \in \mathbb{R}$. Let $x \in \mathbb{R}^n$ and let $d \in \mathbb{R}^n$ be a descent direction of f at x. The objective is to find a solution to

$$\text{argmin}_{t \geqslant 0} f(x + td)$$

The Gradient Method

The Gradient Method - Taking the Direction of Minus the Gradient

- In the gradient method $d_k := -\nabla f(x_k)$
- This is a descent direction as long as $\nabla f(x_k) \neq 0$ since

$$\frac{\partial f}{\partial d_k}(x_k) = -\nabla f(x_k)^T \nabla f(x_k) = -\|\nabla f(x_k)\|_2^2 < 0.$$

- In addition for being a descent direction, minus the gradient is also the steepest direction method.

Lemma

Let f be a continuously differentiable function and let $x \in \mathbb{R}^n$ be a non-stationary point ($\nabla f(x) \neq 0$). Then an optimal solution of

$$\min_{\{d \| \|d\|_2 = 1\}} \frac{\partial f}{\partial d}(x)$$

is $d = -\frac{\nabla f(x)}{\|\nabla f(x)\|_2}$.

Proof let as an exercise.

The Gradient Method

Input: $\epsilon > 0$ - tolerance parameter.
Initialization: pick $x_0 \in \mathbb{R}^n$ arbitrarily.
General step: for any $k = 0, 1, 2, ...$ execute the following steps:

1. pick a stepsize t_k by a line search procedure on the function

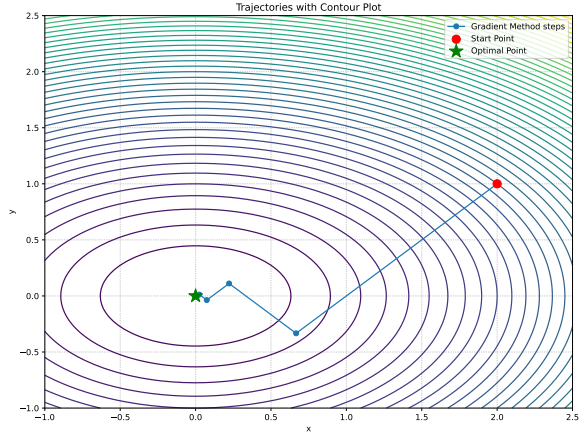
$$g(t) = f(x_k - t\nabla f(x_k))$$

2. Set $x_{k+1} = x_k - t_k \nabla f(x_k)$.
3. If $\|\nabla f(x_{k+1})\| \leqslant \epsilon$, then STOP and x_{k+1} is the output.

Numerical Example

$$\min x^2 + 2y^2$$

$x_0 = (2, 1)^T, \epsilon = 10^{-5}$, exact line search/step size.



▸ 13 iterations until convergence.

The Zig-Zag Effect

Lemma

Let $\{x_k\}_{k\geqslant 0}$ be the sequence generated by the gradient method with exact line search for solving a problem of minimizing a continuously differentiable function f . Then for any $k = 0, 1, 2, ...$

$$(x_{k+2} - x_{k+1})^T (x_{k+1} - x_k) = 0$$

Proof.

- $x_{k+1} - x_k = -t_k \nabla f(x_k), x_{k+2} - x_{k+1} = -t_{k+1} \nabla f(x_{k+1})$.
- Therefore,we need to prove that $(\nabla f(x_{k+1}))^T \nabla f(x_k) = 0$
- $t_k \in \text{argmin}_{t\geqslant 0}\{g(t) := f(x_k - t\nabla f(x_k))\}$
- Hence, $g^{'}(t_k) = 0$.
- $-\nabla f(x_k)^T \nabla f(x_k - t_k \nabla f(x_k)) = 0$.
- Finally, $\nabla f(x_k)^T \nabla f(x_{k+1}) = 0$

Numerical Example - Constant Step size, $\bar{t} = 0.1$

$$\min x^2 + 2y^2$$

$x_0 = (2,1)^T, \epsilon = 10^{-5}$, $\bar{t} = 0.1$.

```
iter_number =   1 norm_grad = 4.000000 fun_val = 3.280000
iter_number =   2 norm_grad = 2.937210 fun_val = 1.897600
iter_number =   3 norm_grad = 2.222791 fun_val = 1.141888
      :                 :                    :
iter_number =  56 norm_grad = 0.000015 fun_val = 0.000000
iter_number =  57 norm_grad = 0.000012 fun_val = 0.000000
iter_number =  58 norm_grad = 0.000010 fun_val = 0.000000
```

- quite a lot of iterations...

Numerical Example - Constant Step size, $\bar{t} = 10$

$$\min x^2 + 2y^2$$

$x_0 = (2, 1)^T, \epsilon = 10^{-5}$, $\bar{t} = 10$.

```
iter_number =   1 norm_grad = 1783.488716 fun_val = 476806.000000
iter_number =   2 norm_grad = 656209.693339 fun_val = 56962873606.0
iter_number =   3 norm_grad = 256032703.004797 fun_val = 8318300807
     :                :                                      :
iter_number = 119 norm_grad = NaN fun_val = NaN
```

- The sequence diverges:(
- Important question: how can we choose the constant stepsize so that convergence is guaranteed?

Lipschitz Continuity of the Gradient

Definition

Let f be a continuously differentiable function over $\mathbb{R}^n$. We say that f has a Lipschitz gradient if there exists $L \geqslant 0$ for which

$$\|\nabla f(x) - \nabla f(y)\| \leqslant L\|x - y\| \quad \text{for any } x, y \in \mathbb{R}^n.$$

L is called the Lipschitz constant.

- If ∇f is Lipschitz with constant L, then it is also Lipschitz with constant $\tilde{L}$ for all $\tilde{L} \geqslant L$.
- The class of functions with Lipschitz gradient with constant L is denoted by $C_L^{1,1}(\mathbb{R}^n)$ or just $C_L^{1,1}$.
- **Linear functions** - Given $c \in \mathbb{R}^n$, the function $f(x) = c^T x$ is in $C_0^{1,1}$.
- **Quadratic functions** - Let A be a symmetric $n \times n$ matrix, $b \in \mathbb{R}^n$ and $c \in \mathbb{R}$. Then the function $f(x) = x^T Ax + 2b^T x + c$ is a $C_L^{1,1}$ function. The smallest Lipschitz constant of ∇f is $2\|A\|_2$ – why ? *In class*

Equivalence to Boundedness of the Hessian

Theorem

Let f be a twice continuously differentiable function over $\mathbb{R}^n$. Then the following two claims are equivalent:

1. $f \in C_L^{1,1}(\mathbb{R}^n)$
2. $\|\nabla^2 f(x)\| \leqslant L$.

Proof on pages 73,74 of the book of Amir Beck [3]

Example: $f(x) = \sqrt{1+x^2} \in C^{1,1}$.
In class

[3] Amir Beck. *Introduction to Nonlinear Optimization: Theory, Algorithms, and Applications with MATLAB*. SIAM, 2014. Full pdf

Outline

Theorem - Global convergence of the Gradient Method

Let $\{x_k\}_{k\geqslant 0}$ be the sequence generated by GM for solving

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

with one of the following stepsize strategies:

1. constant stepsize $\bar{t} \in (0, \frac{2}{L})$,
2. exact line search,
3. backtracking procedure with parameters $s > 0$ and $\beta, \gamma \in (0, 1)$.

Assume that

- $f \in C_L^{1,1}(\mathbb{R}^n)$.
- f is bounded below over $\mathbb{R}^n$(there exists $m \in \mathbb{R}$ such that $f(x) > m$ for all $x \in \mathbb{R}^n$).

Then

1. for any k, $f(x_{k+1}) < f(x_k)$ unless $\nabla f(x_k) = 0$.
2. $\nabla f(x_k) \to 0$ as $k \to \infty$.

Convergence of the Gradient Method

Proof, see Theorem 4.25 in the book.
About the Theorem:

- it means that the iterates x_k converge to a *stationary point*.
- Global convergence is sometimes used to describe algorithms that converge regardless of their starting point (potentially starting faraway from a stationary point $x^\star$). Global convergence should not be confused with *global minimum*: an algorithm can be globally convergent without necessarily converging to the global minimum.
- It is not so restrictive to require f to be bounded below: if this is not the case, then the optimization problem is ill-posed.

Two Numerical Examples - Backtracking

$$\min x^2 + 2y^2$$

$x_0 = (2,1)^T, \epsilon = 10^{-5}$, $s = 2$, $\beta = 1/4$ and $\gamma = 1/2$.

Demo

- fast convergence
- no real advantage to exact line search.

ANOTHER EXAMPLE:

$$\min 0.01x^2 + y^2$$

$x_0 = (2,1)^T, \epsilon = 10^{-5}$, $s = 2$, $\beta = 1/4$ and $\gamma = 1/2$.

Demo Much more iterations required !

Important Question: Can we detect key properties of the objective function that imply slow/fast convergence?

Outline

Kantorovich Inequality

Lemma

Let A be a positive definite $n \times n$ matrix. Then for any $0 \neq x \in \mathbb{R}^n$ the inequality

$$\frac{(x^T x^2)}{(x^T A x)(x^T A^{-1} x)} \geqslant \frac{4\lambda_{\max}(A)\lambda_{\min}(A)}{(\lambda_{\max}(A) + \lambda_{\min}(A))^2} \tag{6}$$

Proof

- Denote $M = \lambda_{\max}(A)$ and $m = \lambda_{\min}(A)$.
- The eigenvalues of the matrix $A + MmA^{-1}$ are $\lambda_i(A) + \frac{Mm}{\lambda_i(A)}$
- The maximum of the 1-D function $\phi(t) = t + Mm/t$ over $[m, M]$ is attained at the endpoints m and M with a corresponding value of $M + m$.
- Thus, the eigenvalues of $A + MmA^{-1}$ are smaller than $(M + m)$, i.e. $(M + m)I \succeq A + MmA^{-1}$
- $x^T Ax + Mm(x^T A^{-1} x) \leqslant (M + m)(x^T x)$, therefore,

$$(x^T Ax)[Mm(x^T A^{-1} x)] \leqslant \frac{1}{4}[x^T Ax + Mm(x^T A^{-1} x)]^2 \leqslant \frac{1}{4}(M + m)^2 (x^T x)^2$$

Minimizing $x^T Ax$

Theorem

Let $\{x_k\}_{k\geqslant 0}$ be the sequence generated by the gradient method with exact linesearch for solving the problem

$$\min_{x\in\mathbb{R}^n} x^T Ax \quad (A > 0). \tag{7}$$

Then for any $k = 0, 1, ...$:

$$f(x_{k+1}) \leqslant \left(\frac{M-m}{M+m}\right)^2 f(x_k) \tag{8}$$

with $M = \lambda_{\max}(A)$ and $m = \lambda_{\min}(A)$ (> 0).

Proof

- $x_{k+1} := x_k - t_k d_k$
  with $t_k = \frac{d_k^T d_k}{2 d_k^T A d_k}$, and $d_k = 2Ax_k$

Minimizing $x^T Ax$

- $$\begin{aligned} f(x_{k+1}) &= x_{k+1}^T A x_{k+1} = (x_k - t_k d_k)^T A (x_k - t_k d_k) \\ &= x_k^T A x_k - 2 t_k d_k^T A x_k + t_k^2 d_k^T A d_k \\ &= x_k^T A x_k - t_k d_k^T d_k + t_k^2 d_k^T A d_k \end{aligned}$$

- Plugging in the expression for t_k

$$\begin{aligned} f(x_{k+1}) &= x_k^T A x_k - \frac{1}{4} \frac{(d_k^T d_k)^2}{(d_k^T A d_k)} \\ &= x_k^T A x_k \left(1 - \frac{1}{4} \frac{(d_k^T d_k)^2}{(d_k^T A d_k)(x_k^T A x_k)} \right) \\ &= x_k^T A x_k \left(1 - \frac{(d_k^T d_k)^2}{(d_k^T A d_k)(4 x_k^T A A^{-1} A x_k)} \right) \\ &= x_k^T A x_k \left(1 - \frac{(d_k^T d_k)^2}{(d_k^T A d_k)(d_k^T A^{-1} d_k^T)} \right) \underset{\text{Kantor.}}{\leqslant} \left(1 - \frac{4Mm}{(M+m)^2} \right) f(x_k) \leqslant \left(\frac{M-m}{M+m} \right)^2 f(x_k) \end{aligned}$$

The Condition Number

Definition

Let A be an $n \times n$ positive definite matrix. Then the condition number of A is defined by

$$\kappa(A) := \frac{\lambda_{\max}(A)}{\lambda_{\min}(A)}$$

- matrices (or quadratic functions) with large condition number are called ill-conditioned.
- matrices with small condition number are called well-conditioned.
- large (reps. small) condition number implies large (reps. small) number of iterations of the gradient method.
- For a non-quadratic function, the asymptotic rate of convergence of x_k to a stationary point $x^\star$ is usually determined by the condition number of $\nabla^2 f(x^\star)$.
- Massive impact on the *Sensitivity of Solutions to Linear Systems* $Ax = b$ ($A \succ 0$), that is when $b \to b + \Delta b$, then the solution $x = A^{-1}b \to x + \Delta x$ s.t. $\Delta x = A^{-1}\Delta b$ and $\frac{\|\Delta x\|}{\|x\|} \leqslant \kappa(A) \frac{\|\Delta b\|}{\|b\|}$

A Severely Ill-Condition Function - Rosenbrock

$$\min f(x_1, x_2) = 100(x_2 - x_1^2)^2 + (1 - x_1)^2$$

- optimal solution:$(x_1, x_2) = (1, 1)$, optimal value: 0.
- $$\nabla f(x_1, x_2) = \begin{pmatrix} -400x_1(x_2 - x_1^2) - 2(1 - x_1) \\ 200(x_2 - x_1^2) \end{pmatrix}$$
$$\nabla^2 f(x_1, x_2) = \begin{pmatrix} -400x_2 + 1200x_1^2 + 2 & -400x_1 \\ -400x_1 & 200 \end{pmatrix}$$
- $$\nabla^2 f(1, 1) = \begin{pmatrix} 802 & -400 \\ -400 & 200 \end{pmatrix}$$

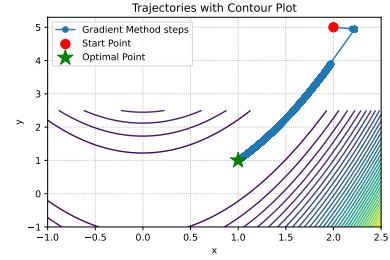
  condition number: 2508.

Solution of the Rosenbrock Problem with the Gradient Method

$x_0 = (2; 5)$, $s = 2$, $\beta = 0.25$, $\gamma = 0.5$, $\epsilon = 10^{-5}$, backtracking stepsize selection.

Demo

Solution of the Rosenbrock Problem with the Gradient Method



6890(!!!) iterations.

Extending the Gradient method

Scaled Gradient Method

- Consider the minimization problem

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

- For a given nonsingular matrix $S \in \mathbb{R}^{n \times n}$, we make the linear change of variables $x = Sy$, and obtain the equivalent problem

$$(\mathrm{P'}) \ : \min_{y \in \mathbb{R}^n} g(y) = f(Sy)$$

- Since $\nabla g(y) = S^T \nabla f(Sy) = S^T \nabla f(x)$, the gradient method for (P') is

$$y_{k+1} := y_k - t_k S^T \nabla f(Sy_k)$$

- Multiplying the latter equality by S from the left, and using the notation $x_k = Sy_k$:

$$x_{k+1} := x_k - t_k S S^T \nabla f(x_k)$$

- Defining $D = SS^T$, we obtain the scaled gradient method:

$$x_{k+1} := x_k - t_k D \nabla f(x_k)$$

Scaled Gradient Method

- $D \succ 0$, so the direction $d_k = -D\nabla f(x_k)$ is a descent direction: $\frac{\partial f}{\partial d_k}(x_k) = -\nabla f(x_k)^T D \nabla f(x_k) \leqslant 0$.

We also allow different scaling matrices at each iteration.

Scaled Gradient Method

Input: $\epsilon > 0$ - tolerance parameter.
Initialization: pick $x_0 \in \mathbb{R}^n$ arbitrarily.
General step: for any $k = 0, 1, 2, ...$ execute the following steps:

1. pick a scaling matrix $D_k \succ 0$.
2. pick a stepsize t_k by a line search procedure on the function

$$g(t) = f(x_k - tD_k\nabla f(x_k))$$

3. Set $x_{k+1} = x_k - t_k D_k \nabla f(x_k)$.
4. If $\|\nabla f(x_{k+1})\| \leqslant \epsilon$, then STOP and x_{k+1} is the output.

Choosing the Scaling Matrix D_k

- The scaled gradient method with scaling matrix D is equivalent to the gradient method employed on the function $g(y) = f(D^{1/2}y)$.
- Note that the gradient and Hessian of g are given by

$$\begin{aligned} \nabla g(y) &= D^{1/2}\nabla f(D^{1/2}y) = D^{1/2}\nabla f(x) \\ \nabla^2 g(y) &= D^{1/2}\nabla^2 f(D^{1/2}y)D^{1/2} = D^{1/2}\nabla^2 f(x)D^{1/2} \end{aligned} \tag{9}$$

- The objective is usually to pick D_k so as to make $D^{1/2}\nabla^2 f(x_k)D^{1/2}$ as well-conditioned as possible.
- A well-known choice (Newton's method): $D_k = [\nabla^2 f(x_k)]^{-1}$
- diagonal scaling : D_k is picked to be diagonal. For example,

$$[D_k]_{ii} = \left(\frac{\partial^2 f(x_k)}{\partial x_i^2}\right)^{-1}$$

- Diagonal scaling can be very effective when the decision variables are of different magnitudes.

Nesterov Accelerated Gradient Method

Unlike the basic gradient method (figure on the left), we will also use the information from the previous iteration (figure on the right):

Nesterov Accelerated Gradient Method ($f \in C_L^{1,1}(\mathbb{R}^n)$)

Initialization: pick $x_0 \in \mathbb{R}^n$ arbitrarily, set some $\alpha_0 \in (0,1)$, and set $y_0 = x_0$.
General step: for any $k = 0, 1, 2, \ldots$ execute the following steps:

1. Compute $f(y_k)$ and $\nabla f(y_k)$. Set $x_{k+1} \leftarrow y_k - \frac{1}{L}\nabla f(y_k)$
2. Compute $\alpha_{k+1} \in (0,1)$ from Equation

$$\alpha_{k+1}^2 = (1 - \alpha_{k+1})\alpha_k^2 \tag{10}$$

Set $\beta_k = \frac{\alpha_k(1-\alpha_k)}{\alpha_k^2 + \alpha_{k+1}}$ and $y_{k+1} \leftarrow x_{k+1} + \beta_k(x_{k+1} - x_k)$

Nesterov Accelerated Gradient Method

We omit the details in this slide, if requested, the exact forms of convergence rates can be provided :).

- For functions $f \in C_L^{1,1}(\mathbb{R}^n)$ and convex: convergence in $O(\frac{1}{k^2})$ instead of $O(\frac{1}{k})$ (see slide 46) !!

Nesterov Accelerated Gradient Method

We omit the details in this slide, if requested, the exact forms of convergence rates can be provided :).

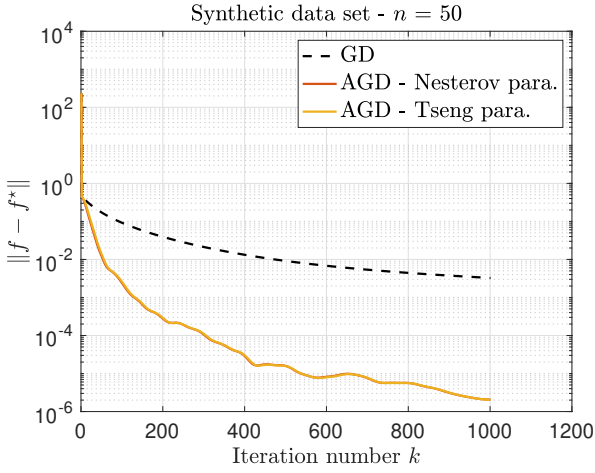
- For functions $f \in C_L^{1,1}(\mathbb{R}^n)$ and convex: convergence in $O(\frac{1}{k^2})$ instead of $O(\frac{1}{k})$ (see slide 46) !!

This is a significant acceleration, and it is optimal in the sense that no other first-order method can guarantee a faster convergence rate!

See (Y.Nesterov, 2018) book, section 2.2 ("Optimal Methods") for more details.

Nesterov Accelerated Gradient Method

Test case: Given $A \in \mathbb{R}^{n \times n}$ and $b \in \mathbb{R}^n$, we want to solve: $\min_x \frac{1}{2}\|Ax - b\|_2^2$.
Code Link, Code Ocean Link



Conclusions

Summary

We have seen

- the **Descent Direction Method**, and three stepsize selection rules: *constant*, *exact line search* and *backtracking* (trade off between two first approaches to satisfy the *sufficient decrease condition*).
- The **Gradient Method** - $d_k = -\nabla f(x_k)$:
  1. The *Zig-Zag Effect* when exact line search used.
  2. *Global* convergence of the Gradient Method: the iterates x_k converge to a stationary point (for $k \to \infty$), regardless of their starting point.
  3. The case of minimizing $x^T Ax$ ($A \succ 0$): linear rate of convergence with exact line search.
- **Extensions** of the Gradient Method:
  1. The *Scaled* Gradient Method: make the problem well-conditioned to reduce number of iterations.
  2. *Nesterov Accelerated* Gradient Method: optimal method for minimizing L-smooth convex functions.

Preparations for the next lecture

- Review the lecture :).
- Implement the Gradient Method (GM) with the different stepzise selection rules, and the extensions of GD (Scaled and Accelerated) and test on small dimension problems.

Appendix A - Linear convergence and convergence rate

Let $\{x_0, x_1, x_2, ...\}$ converge to $x^\star$. How fast does this sequence approach $x^\star$?
To answer this question, let's first introduce the notion of convergence rate μ.

If there exists a number $\mu \in]0, 1[$ such that

$$\lim_{k \to \infty} \frac{\|x_{k+1} - x^\star\|}{\|x_k - x^\star\|} = \mu,$$

then we say that the sequence converges **linearly**[a] with a rate μ. This means that, for k **sufficiently large**, we have

$$\|x_{k+1} - x^\star\| \leqslant \mu \|x_k - x^\star\| \text{ and } \|x_k - x^\star\| \leqslant \mu^k \|x_0 - x^\star\|$$

[a] Note, that sometimes this type of convergence is also called exponential or geometric.

The quantity $\|x_k - x^\star\|$ is the error e_k of the iterate k with respect to the sought solution $x^\star$. The linear convergence property can be rewritten as follows: $e_{k+1} \leqslant \mu e_k$.

Linear convergence: illustration

From the property of linear convergence $e_{k+1} \leqslant \mu e_k$, we deduce that

$$\log(e_k) \approx k \log(\mu) + \log(e_0),$$

which is of the form $\log(e_k) \approx ak + b$, hence the name linear convergence.

Linear convergence: illustration

From the property of linear convergence $e_{k+1} \leqslant \mu e_k$, we deduce that

$$\log(e_k) \approx k \log(\mu) + \log(e_0),$$

which is of the form $\log(e_k) \approx ak + b$, hence the name linear convergence.

Example.

Let the initial iterate be $x_0 = 2.1$ and the following sequence converging to $x^\star = 2 + \frac{\sqrt{2}}{2} \approx 2.707106781186$:

$$x_{k+1} = -(x_k)^2 + 5x_k - 3.5.$$

Linear convergence: illustration

From the property of linear convergence $e_{k+1} \leqslant \mu e_k$, we deduce that

$$\log(e_k) \approx k \log(\mu) + \log(e_0),$$

which is of the form $\log(e_k) \approx ak + b$, hence the name linear convergence.

Example.

Let the initial iterate be $x_0 = 2.1$ and the following sequence converging to $x^\star = 2 + \frac{\sqrt{2}}{2} \approx 2.707106781186$:

$$x_{k+1} = -(x_k)^2 + 5x_k - 3.5.$$

k	x_k	$e_k = \lvert x_k - x^\star \rvert$	e_k/e_{k-1}
0	2.1		
1	2.59		
2	2.7419		
3	2.69148439		
4	2.7133337283863		
5	2.7044887203327		
6	2.7081843632566		
7	2.7066592708954		
8	2.7072919457529		
9	2.7070300492259		
10	2.7071385587175		
11	2.7070936174924		

Linear convergence: illustration

From the property of linear convergence $e_{k+1} \leqslant \mu e_k$, we deduce that

$$\log(e_k) \approx k \log(\mu) + \log(e_0),$$

which is of the form $\log(e_k) \approx ak + b$, hence the name linear convergence.

Example.

Let the initial iterate be $x_0 = 2.1$ and the following sequence converging to $x^\star = 2 + \frac{\sqrt{2}}{2} \approx 2.707106781186$:

$$x_{k+1} = -(x_k)^2 + 5x_k - 3.5.$$

k	x_k	$e_k = \lvert x_k - x^\star \rvert$	e_k/e_{k-1}
0	2.1	$6.0710678118654e-01$	
1	2.59	$1.1710678118654e-01$	
2	2.7419	$3.4793218813453e-02$	
3	2.69148439	$1.5622391186545e-02$	
4	2.7133337283863	$6.2269471997806e-03$	
5	2.7044887203327	$2.6180608537584e-03$	
6	2.7081843632566	$1.0775820701116e-03$	
7	2.7066592708954	$4.4751029112610e-04$	
8	2.7072919457529	$1.8516456642725e-04$	
9	2.7070300492259	$7.6731960601428e-05$	
10	2.7071385587175	$3.1777530955956e-05$	
11	2.7070936174924	$1.3163694111639e-05$	

Linear convergence: illustration

From the property of linear convergence $e_{k+1} \leqslant \mu e_k$, we deduce that

$$\log(e_k) \approx k \log(\mu) + \log(e_0),$$

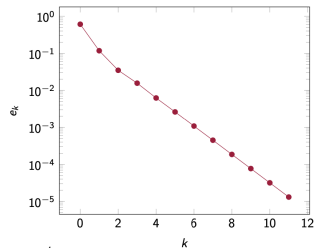
which is of the form $\log(e_k) \approx ak + b$, hence the name linear convergence.

Example.

Let the initial iterate be $x_0 = 2.1$ and the following sequence converging to $x^\star = 2 + \frac{\sqrt{2}}{2} \approx 2.707106781186$:

$$x_{k+1} = -(x_k)^2 + 5x_k - 3.5.$$

k	x_k	$e_k = \lvert x_k - x^\star \rvert$	e_k/e_{k-1}
0	2.1	$6.0710678118654e-01$	
1	2.59	$1.1710678118654e-01$	
2	2.7419	$3.4793218813453e-02$	
3	2.69148439	$1.5622391186545e-02$	
4	2.7133337283863	$6.2269471997806e-03$	
5	2.7044887203327	$2.6180608537584e-03$	
6	2.7081843632566	$1.0775820701116e-03$	
7	2.7066592708954	$4.4751029112610e-04$	
8	2.7072919457529	$1.8516456642725e-04$	
9	2.7070300492259	$7.6731960601428e-05$	
10	2.7071385587175	$3.1777530955956e-05$	
11	2.7070936174924	$1.3163694111639e-05$	



Linear convergence: illustration

From the property of linear convergence $e_{k+1} \leqslant \mu e_k$, we deduce that

$$\log(e_k) \approx k \log(\mu) + \log(e_0),$$

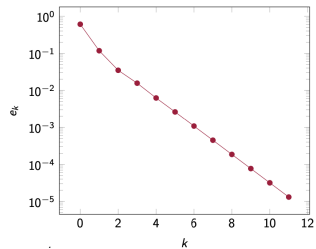
which is of the form $\log(e_k) \approx ak + b$, hence the name linear convergence.

Example.

Let the initial iterate be $x_0 = 2.1$ and the following sequence converging to $x^\star = 2 + \frac{\sqrt{2}}{2} \approx 2.707106781186$:

$$x_{k+1} = -(x_k)^2 + 5x_k - 3.5.$$

k	x_k	$e_k = \lvert x_k - x^\star \rvert$	e_k/e_{k-1}
0	2.1	$6.0710678118654e-01$	
1	2.59	$1.1710678118654e-01$	0.192893218813452
2	2.7419	$3.4793218813453e-02$	0.297106781186555
3	2.69148439	$1.5622391186545e-02$	0.449006781186490
4	2.7133337283863	$6.2269471997806e-03$	0.398591171186609
5	2.7044887203327	$2.6180608537584e-03$	0.420440509572765
6	2.7081843632566	$1.0775820701116e-03$	0.411595501519640
7	2.7066592708954	$4.4751029112610e-04$	0.415291144441310
8	2.7072919457529	$1.8516456642725e-04$	0.413766052086330
9	2.7070300492259	$7.6731960601428e-05$	0.414398726937717
10	2.7071385587175	$3.1777530955956e-05$	0.414136830427411
11	2.7070936174924	$1.3163694111639e-05$	0.414245339887626



Linear convergence: illustration

From the property of linear convergence $e_{k+1} \leqslant \mu e_k$, we deduce that

$$\log(e_k) \approx k \log(\mu) + \log(e_0),$$

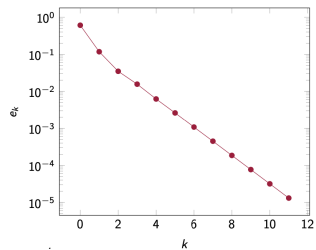
which is of the form $\log(e_k) \approx ak + b$, hence the name linear convergence.

Example.

Let the initial iterate be $x_0 = 2.1$ and the following sequence converging to $x^\star = 2 + \frac{\sqrt{2}}{2} \approx 2.707106781186$:

$$x_{k+1} = -(x_k)^2 + 5x_k - 3.5.$$

k	x_k	$e_k = \lvert x_k - x^\star \rvert$	$e_k/e_{k-1} \approx \mu$
0	2.1	$6.0710678118654e-01$	
1	2.59	$1.1710678118654e-01$	0.192893218813452
2	2.7419	$3.4793218813453e-02$	0.297106781186555
3	2.69148439	$1.5622391186545e-02$	0.449006781186490
4	2.7133337283863	$6.2269471997806e-03$	0.398591171186609
5	2.7044887203327	$2.6180608537584e-03$	0.420440509572765
6	2.7081843632566	$1.0775820701116e-03$	0.411595501519640
7	2.7066592708954	$4.4751029112610e-04$	0.415291144441310
8	2.7072919457529	$1.8516456642725e-04$	0.413766052086330
9	2.7070300492259	$7.6731960601428e-05$	0.414398726937717
10	2.7071385587175	$3.1777530955956e-05$	0.414136830427411
11	2.7070936174924	$1.3163694111639e-05$	0.414245339887626



Superlinear convergence and order of convergence

What to say when the limit below is 0 ?

$$\lim_{k \to \infty} \frac{\|x_{k+1} - x^\star\|}{\|x_k - x^\star\|} = 0.$$

- Convergence is then said to be **superlinear**..

Superlinear convergence and order of convergence

What to say when the limit below is 0 ?

$$\lim_{k \to \infty} \frac{\|x_{k+1} - x^\star\|}{\|x_k - x^\star\|} = 0.$$

- Convergence is then said to be **superlinear**..
- To distinguish suites that converge superlinearly, we introduce an additional notion: the order of convergence p.

Superlinear convergence and order of convergence

What to say when the limit below is 0 ?

$$\lim_{k \to \infty} \frac{\|x_{k+1} - x^\star\|}{\|x_k - x^\star\|} = 0.$$

- Convergence is then said to be **superlinear**..
- To distinguish suites that converge superlinearly, we introduce an additional notion: the order of convergence p.

Superlinear convergence and order of convergence

What to say when the limit below is 0 ?

$$\lim_{k\to\infty} \frac{\|x_{k+1} - x^\star\|}{\|x_k - x^\star\|} = 0.$$

- Convergence is then said to be **superlinear**..
- To distinguish suites that converge superlinearly, we introduce an additional notion: the order of convergence p.

If there exists a number $p > 1$ and a constant $C > 0$ such that

$$\lim_{k\to\infty} \frac{\|x_{k+1} - x^\star\|}{\|x_k - x^\star\|^p} = C,$$

then p is called the order of convergence of the sequence and C is the error constant. This means that, for k sufficiently large, we have

$$\|x_{k+1} - x^\star\| \leqslant C \|x_k - x^\star\|^p .$$

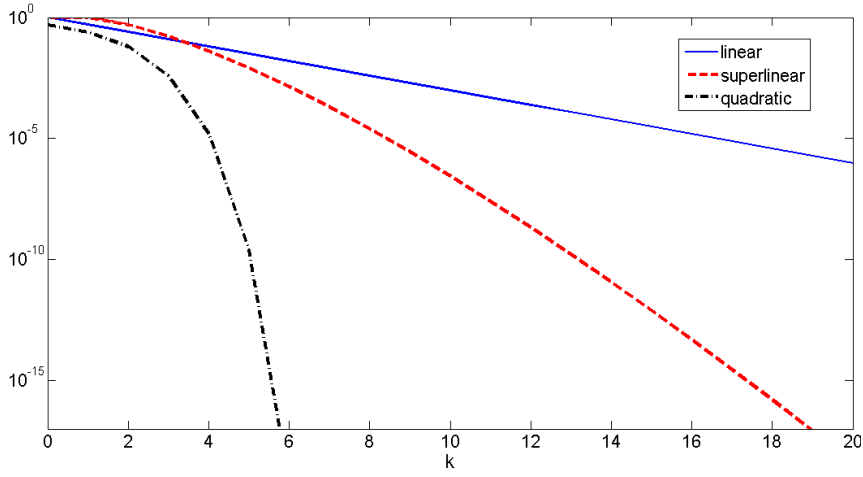
(convergence **quadratic** if $p = 2$, convergence **cubic** if $p = 3, \cdots$)

Superlinear convergence and order of convergence

- The rate of convergence μ and the order of convergence p are two numbers used to describe the speed of different types of convergence.
- A sequence has either a rate of convergence if convergence is linear or an order of convergence if convergence is superlinear, but not both.
- However, the lower the rate or the higher the order, the faster the convergence.

Rate of convergence of a sequence: examples

Difference in behavior between different types of convergence



Some fun here by Daniil Merkulov (Skoltech)

Appendix B - The convex case for GD

Theorem 1 - *Global* convergence of GD for Smooth Convex Function minimization

Let f be a convex differentiable function that has a minimizer $x^\star$ and whose gradient is L-Lipschitz continuous. The gradient method with constant step size $\alpha_k = \frac{1}{L}$ satisfies

$$f(x_k) - f(x^\star) \leqslant \frac{L\|x_0 - x^\star\|^2}{2k} \sim O(\frac{1}{k}) \tag{11}$$

Remarks:

- Such a rate, i.e. $f(x_k) - f(x^\star) \leqslant Ck^q$ where $q < 0$ and $0 < C < \infty$, is called *sublinear*.
- The assumption of convexity here is crucial, recall Lecture on Convexity !

Appendix C - Complexity estimates

Context: find an ϵ-accuracy optimal solution

For a chosen tolerance ϵ (for instance, $\epsilon = 10^{-6}$), we want to find x_k such that $f(x_k) - f(x^\star) \leqslant \epsilon$.

How many k iterations are needed to reach this tolerance?

- If the number of iterations $k = O\left(\frac{1}{\epsilon^2}\right)$, this means that

$$k = \text{constant}\frac{1}{\epsilon^2}.$$

  We then note $\epsilon = O\left(\frac{1}{\sqrt{k}}\right)$
- If the number of iterations $k = O\left(\frac{1}{\epsilon}\right)$, this means that $k = \text{constant}\frac{1}{\epsilon}$. We then note $\epsilon = O\left(\frac{1}{k}\right)$
- If the number of iterations $k = O\left(\frac{1}{\sqrt{\epsilon}}\right)$, this means that $k = \text{constant}\frac{1}{\sqrt{\epsilon}}$. We then note $\epsilon = O\left(\frac{1}{k^2}\right)$
- If the number of iterations $k = O\left(\log\left(\frac{1}{\epsilon}\right)\right)$, this means that $k = \text{constant}\log\left(\frac{1}{\epsilon}\right)$. We then note $\epsilon = O\left(c^k\right)$

Goodbye, So Soon

THANKS FOR THE ATTENTION

- v.leplat@innopolis.ru
- sites.google.com/view/valentinleplat/