Numerical Optimisation 2022
Line Search Methods

François Lauze
francois@di.ku.dk

Department of Computer Science
University of Copenhagen

February 21, 2022

Optimisation

- Search directions and search steps
- Wolfe Conditions
- Obtaining search steps
- The mother theorem of many convergence theorems in continuous optimisation: Zoutendijk's theorem.
- Line search Newton

Steps, descent direction

- Iteration step: $x_{k+1} = x_k + \alpha_k \mathbf{p}_k$
- Wish (in general) $f_{k+1} := f(x_{k+1}) < f_k := f(x_k)$.
- Choose $\mathbf{p}_k$ to be a *descent direction*.

$$\mathbf{p}_k^\top \nabla f_k < 0$$

- Taylor's theorem

$$f(x_k + \alpha \mathbf{p}_k) = f(x_k) + \alpha \nabla f(\mathbf{x}_k)^\top \mathbf{p}_k + o(\alpha)$$

- Thus for α small enough, $f(x_k + \alpha \mathbf{p}_k) < f(x_k)$

Descent directions

- In general
$$\mathbf{p}_k = -B_k^{-1}\nabla f_k$$
with B_k is in general a symmetric, non singular matrix.

- Steepest descent: $B_k = \mathrm{id}$, $\mathbf{p}_k = -\nabla f_k$.
This is at first order the fastest way to decrease f. But it requires in general *very small* steps.

- Newton's method: $B_k = \nabla^2 f_k$. When $\nabla^2 f_k$ keeps being positive definite, one of the fastest method!
- When B_k chosen positive definite , $\mathbf{p}_k$ is guaranteed to be a descent direction.

Bad example

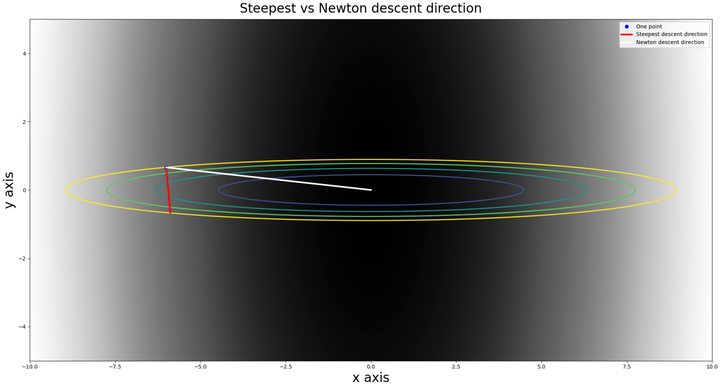
$f(x) = x^2$.

- Descent $x_{k+1} = x_k - 2\alpha x_k$ since $f'(x) = 2x$.
- $\alpha_k = e^{-k}$ is OK?
- Unless $x_1 = 0$, it will never converge to the 0 solution!

$$x_\infty \approx 0.164 x_1$$

(Dixit Mathematica...)

Example with Ellipsoid function



To descend enough!

- Descent direction given: How far should I go?

$$\alpha_k = \mathrm{argmin}_\alpha \varphi(\alpha) = f(x_k + \alpha \mathbf{p}_k)$$

- Optimisation along the (half-)line $x_k + \alpha \mathbf{p}_k$ (the name of the game!)

- Full optimisation can be prohibitive / unnecessary:
- trade-off
  - large enough improvement: $\alpha_k \gg 0$
  - Not to many evaluations of φ and φ'.

Wolfe Conditions

Conditions that define acceptable α_ks

- First Wolfe condition (or Armijo's condition): For some $c_1 \in (0,1)$,

$$f(x_k + \alpha \mathbf{p}_k) \leq f(x_k) + c_1 \alpha \nabla f_k^\top \mathbf{p}_k$$

- but not enough to guarantee $\alpha_k \gg 0$ as it will always work when α small enough (Taylor!)
- Curvature condition: For $c_2 \in (c_1, 1)$,

$$\varphi'(\alpha_k) = \nabla f(x_k + \alpha_k \mathbf{p}_k)^\top \mathbf{p}_k \geq c_2 \nabla f_k^\top \mathbf{p}_k$$

- Remember, this slope should be negative because we want to descend, but at convergence it should be 0! No descent direction should exist!
- But if α is large enough, the slope could be positive. Strong Wolfe conditions as a remedy!

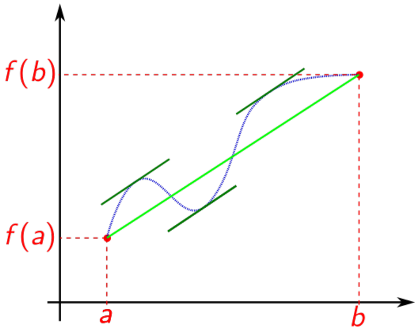
Non vacuity of Wolfe conditions

Lemma

Let $f : \mathbb{R}^n \to \mathbb{R}$ of class $\mathcal{C}^1$ and $\mathbf{p}_k$ a descent direction at x_k. Assume that f is bounded below along the ray $x_k + \alpha\mathbf{p}_k$. Then for c_1 and c_2 such that $0 < c_1 < c_2 < 1$, There are non empty intervals of step lengths satisfying the Wolfe condition (and the strong one.)

Assumption on the ray: if not satisfied, f would go toward $-\infty$ along the ray, no minimiser would exist in that case!

The mean value Theorem



There exists at least a $x \in [a, b]$ with

$$f'(x) = \frac{f(b) - f(a)}{b - a}$$

Sketch of proof of the lemma

- Line $l(\alpha) = f(\mathbf{x}_k) + c_a \alpha \nabla f_k^\top \mathbf{p}_k$ unbounded below: it cuts the graph of φ as φ starts below the line but is *bounded*.
- Let $\alpha' > 0$ the smallest of all the αs lying at the intersection (there can be one or more!)

$$f(x_k + \alpha' \mathbf{p}_k) = f(x_k) + c_1 \alpha' \nabla f_k^\top \mathbf{p}_k$$

- Mean value theorem: there exists $\alpha'' \in (0, \alpha')$,

$$\frac{\varphi(\alpha') - \varphi(0)}{\alpha'} = \varphi'(\alpha'') \iff \frac{f(x_k + \alpha' \mathbf{p}_k) - f(x_k)}{\alpha'} = \nabla f(x_k + \alpha'' \mathbf{p}_k)$$

- Then, as $c_1 < c_2$ and $\nabla f_k^\top \mathbf{p}_k < 0$

$$\nabla f(x_k + \alpha'' \mathbf{p}_k)^\top \mathbf{p}_k = c_1 \nabla f_k^\top \mathbf{p}_k > c_2 \nabla f_k^\top \mathbf{p}_k$$

- The rest is left to the reader!

Backtracking Line Search algorithm

- the books proposes a simple line search strategy: Start with an $\bar{\alpha}$ large enough and decrease it by a factor $\rho \in (0, 1)$ as long as Armijo's condition is not satisfied.
- Some variations record the last α accepted and use a fixed increase of it as starting step to a next iteration to ensure that such an α does not vanish.
- Might be difficult with some maths about the convergence (but not too) but works often well in practice.

Convergence of Line Search - Zoutendijk's Theorem

- Descent direction $\mathbf{p}_k$: angle with steepest descent direction:

$$\cos\theta_k = \frac{-\nabla f_k^\top \mathbf{p}_k}{\|\nabla f_k\| \|\mathbf{p}_k\|}$$

Theorem (Zoutendijk)

Consider iterations with $\mathbf{p}_k$ descent direction sand α_k satisfying Wolfe conditions for a function f. Suppose that f is bounded below and that it is $\mathcal{C}^1$ on an open set containing $\mathcal{N} = \{x, f(x) \le f(x_0)\}$ where x_0 is the starting point of the iterative procedure. Assume also that ∇f is Lipschitz-continuous on $\mathcal{N}$:

$$\exists L > 0, \|\nabla f(x) - \nabla f(y)\| \le L\|x - y\|, \quad \forall x, y \in \mathcal{N}.$$

Then

$$\sum_{k=0}^{\infty} \cos^2\theta_k \|\nabla f_k\|^2 < \infty.$$

What does it mean?

- A Recall: when a series $\sum_{n\geq 0} a_n$ converges, its *general term* $a_n \to 0$.
- Here it means that the sequence $\cos^2\theta_k \|\nabla f_k\|^2 \to 0$.
- if $\theta_k \geq \delta > 0$ for all $k \geq 0$, $\|\nabla f_k\| \to 0$.
- For steepest descent, $\theta_k \equiv 1$.

- Question: Can it fail to converge to a critical point in that case?

- Zoutendijk's theorem has served as model for a lot of convergence results!

Sketch of Proof - Cauchy-Schwarz inequality

General inequality between inner product of vectors and their norms

- If x and y are vector of an Euclidean space or Hilbert space E (e.g $\mathbb{R}^n$), with inner product $\langle x, y \rangle_E$ and norm $\|x\|_E$, then

$$|\langle x, y \rangle_E| \leq \|x\|_E \|y\|_E$$

- With classical norm and inner products on $\mathbb{R}^n$,

$$|x^\top y| \leq \|x\| \|y\|$$

- Generalises too much more "exotic" situations!

Sketch of proof – I

- From Wolfe condition 2,

$$(\nabla f_{k+1} - \nabla f_k))^\top \mathbf{p}_k \geq (c_2 - 1)\nabla f_k^\top \mathbf{p}_k \quad (> 0).$$

- Lipschitz condition and Cauchy-Schwarz

$$\begin{aligned}(\nabla f_{k+1} - \nabla f_k)^\top \mathbf{p}_k &\leq \|\nabla f_{k+1} - \nabla f_k\| \|\mathbf{p}_k\| \\ &\leq L\|x_{x+1} - x_k\| \|\mathbf{p}_k\| \\ &\leq \alpha_k L \|\mathbf{p}_k\|^2\end{aligned}$$

  since $x_{k+1} - x_k = \alpha_k \mathbf{p}_k$.
- Combination gives

$$(c_1 - 1)\nabla f_k^\top \mathbf{p}_k \leq (\nabla f_{k+1} - \nabla f_k)^\top \mathbf{p}_k \leq \alpha_k L \|\mathbf{p}_k\|^2$$

- Thus

$$\alpha_k \geq \frac{c_2 - 1}{L} \frac{\nabla f_k^\top \mathbf{p}_k}{\|\mathbf{p}_k\|^2}$$

Sketch of proof – II

- Recall WC 1. $f_{k+1} - f_k \leq c_1 \alpha_k \nabla f_k^\top \mathbf{p}_k$:
- By clever combination (beware signs!): decrease bound

$$f_{k+1} - f_k \leq -\underbrace{c_1 \frac{1-c_2}{L}}_{:=c} \underbrace{\frac{\left(\nabla f_k^\top \mathbf{p}_k\right)^2}{\|\mathbf{p}_k\|^2}}_{\cos^2 \theta_k \|\nabla f_k\|^2}$$

- Then

$$f_{k+1} - f_0 = f_{k+1} - f_k + f_k - f_{k-1} + \cdots - f_0 \leq -c \sum_{j=0}^{k} \cos^2 \theta_j \|\nabla f_j\|^2$$

- Boundedness $f(x) \geq A$, thus $f_k \geq A$ for all $k \geq 0$:

$$\sum_{j=0}^{k} \cos^2 \theta_j \|\nabla f_j\|^2 \leq \frac{f_0 - A}{c} \quad (> 0)$$

- Since it true for all ks,

$$\sum_{j=0}^{\infty} \cos^2 \theta_j \|\nabla f_j\|^2 \leq \frac{f_0 - A}{c}$$

- This is a classical theorem: a series with only positive terms and bounded above converges.

A few words on poor convergence rate for steepest gradient: Always kind of work, but terribly slow! Linear convergence rate.

But you will have to think about it in the theory part!

Newton descent

- Descent direction taken to be $\mathbf{p}_k = -\nabla^2 f_k^{-1} \nabla f_k$.
- Can be tricky when the Hessian $\nabla^2 f_k$is not SPD. Can fail to obtain a descent direcxtion.
- In the neighbourhood of a minimum, this holds. In that case Newton's method is **very fast**!

Newton Descent Theorem

Theorem (Newton descent)

Assume f twice differentiable and its Hessian $\nabla^2 f$ is Lipschitz-continuous around a solution x^*, with a constant L, for which $\nabla^2 f(x^*)$ is positive definite. Consider the line search iteration with $\alpha_k \equiv 1$ with Newton descent direction $\mathbf{p}_k$. Then

- *if x_0 close enough of x^*, the sequence of iterates $(x_k)_k$ converges to x^*,*
- *The convergence rate is quadratic, and*
- *the sequence of iterates $\|\nabla f_k\|$ converges quadratically to 0.*

Sketch of Proof.

- Uses a version of the *Fundamental Theorem of Calculus*

$$F(y) - F(x) = \int_0^1 DF(x + t(y - x))(y - x)\, dt$$

(integration of DF along the line segment between x and y).

- Use it with $F = \nabla f$, $x = x_k$, $y = x^*$:

$$\nabla f_k - \nabla f_* = \int_0^1 \nabla^2 f(x_k + t(x^* - x_k))(x_k - x^*)\, dt$$

- Write $x_{k+1} = x_k + \mathbf{p}_k = x_k - \nabla^2 f_k^{-1} \nabla f_k$ and subtract x^* on both sides, using $\nabla f_* = 0$

$$x_k + \mathbf{p}_k - x^* = \nabla^2 f_k^{-1} \left[\nabla^2 f_k (x_k - x^*) - (\nabla f_k - \nabla f_*) \right]$$

- I factored via the *invertible Hessian* $\nabla^2 f_k$ from the theorem's assumptions.
- from the line segment integration in previous slide

$$\nabla f_k - \nabla f_* = \int_0^1 \nabla^2 f(x_k + t(x^* - x_k))(x_k - x^*)\, dt$$

So that

$$\begin{aligned} \nabla^2 f_k (x_k - x^*) - (\nabla f_k - \nabla f_*) &= \int_0^1 \nabla^2 f_k (x_k - x^*) dt \\ &\quad - \int_0^1 \nabla^2 f(x_k + t(x^* - x_k))(x_k - x^*) dt \end{aligned}$$

- Factor and get $\nabla^2 f_k(x_k - x^*) - (\nabla f_k - \nabla f_*) =$

$$\int_0^1 \left[\nabla^2 f_k(x_k - x^*) - \nabla^2 f(x_k + t(x^* - x_k))(x_k - x^*)dt\right]$$

- Use the fact that $\| \int_a^b f(x)\, dx\| \leq \int_a^b \|f(x)\|\, dx$ Then

$$\|\nabla^2 f_k(x_k - x^*) - (\nabla f_k - \nabla f_*)\| \leq$$

$$\int_0^1 \underbrace{\|\nabla^2 f_k(x_k - x^*) - \nabla^2 f(x_k + t(x^* - x_k))\|}_{\text{Lipschitz assumption on } \nabla^2 f} \underbrace{\|x_k - x^*\|}_{\text{constant}} dt$$

$$\leq \|x_k - x^*\|^2 \int_0^1 Lt\, dt = \frac{1}{2} L \|x_k - x^*\|^2$$

I used $\|x_k - x^* - t(x^* - x_k)\| = t\|x_k - x^*\|$.

- Because $\nabla^2 f(x^*)$ is nonsingular, there is a ball of radius r around x^* for which $\nabla^2 f(x)$ is non singular too, and for which

$$\|\nabla^2 f(x)^{-1}\| \leq 2\|\nabla^2 f(x_*)^{-1}\|.$$

(continuity of $x \mapsto \|\nabla^2 f(x)^{-1}\|$ around x^*)

- Then for all x_k in the ball or radius r around x^*,

$$\|x_{k+1}-x^*\| = \|x_k+\mathbf{p}_k-x^*\| \leq L\|\nabla^2 f_*^{-1}\|\|x_k-x^*\|^2 = \tilde{L}\|x_k-x^*\|^2.$$

(with $\tilde{L} = L\|\nabla^2 f_*^{-1}\|$)

- This means exactly that $(x_k)_k$ converges quadratically to x^* as soon as x_k is close enough of x^*.
- Similar calculations for the gradient iterates.