Numerical optimization

Mines Nancy – Fall 2024

session 5 – about convergence

Lecturer: Christophe Zhang (INRIA, IECL)

Course material:

- arche.univ-lorraine.fr/course/view.php?id=74098
- github.com/jflamant/mines-nancy-fall24-optimization

Reminder: descent methods

Context: unconstrained optimization

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

Principle of descent methods: solve iteratively using

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \alpha_k \mathbf{d}_k$$

where α_k is the stepsize and $\mathbf{d}_k$ is a descent direction ($\mathbf{d}_k^\top \nabla f(\mathbf{x}^{(k)}) < 0$).

Two important examples

- Gradient descent

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

- Newton's method

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \alpha_k \left[\nabla^2 f(\mathbf{x}^{(k)})\right]^{-1} \nabla f(\mathbf{x}^{(k)})$$

Outline

Characterizing regularity

Context: unconstrained optimization

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

Assumptions

- the function f is (at least) continuously differentiable
- there exists a (global) minimizer $\mathbf{x}^\star$; we note $f(\mathbf{x}^\star)$ the minimum

We'll consider several additional properties of the objective f:

- convexity
- smoothness
- strong convexity

Main references: Recht and Wright (2022), Boyd and Vanderberghe (2004)

Convex functions [reminder]

Theorem (first order)

Let $f : \mathbb{R}^N \to \mathbb{R}$ be differentiable. These statements are equivalent:

1. *f is convex on $\mathbb{R}^N$*
2. *for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^N$, $f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x})$*
3. *for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^N$, $(\nabla f(\mathbf{y}) - \nabla f(\mathbf{x}))^\top (\mathbf{y} - \mathbf{x}) \geq 0$*

The equivalence is preserved for strict convexity, with $\mathbf{x} \neq \mathbf{y}$ and strict inequalities.

Theorem (second order)

Let $f : \mathbb{R}^N \to \mathbb{R}$ be twice differentiable. We have equivalence between

1. *f is convex*
2. *for all $\mathbf{x} \in \mathbb{R}^N$, $\nabla^2 f(\mathbf{x}) \succeq 0$, i.e. the Hessian is positive semidefinite.*

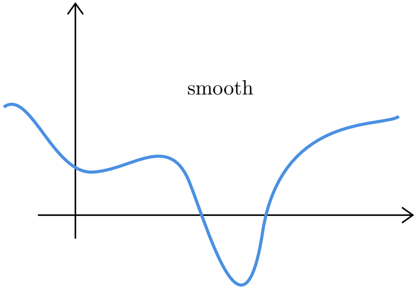
Smoothness

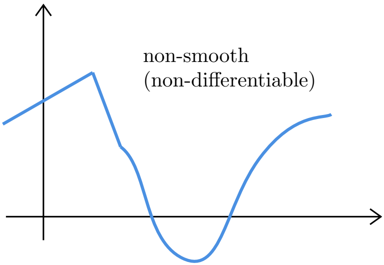
Definition (L-smooth functions)

A continuously differentiable function $f : \mathbb{R}^N \to \mathbb{R}$ is said to be L-smooth if

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_2 \leq L\|\mathbf{x} - \mathbf{y}\|_2 \quad \text{for all } \mathbf{x}, \mathbf{y} \in \mathbb{R}^N,$$

i.e., its gradient ∇f is L-Lipschitz continuous.





Properties of smooth functions

Property 1 (Quadratic upper bound)

Let f be continuously differentiable and L-smooth on $\mathbb{R}^N$. Then

$$f(\mathbf{y}) \leq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) + \frac{L}{2}\|\mathbf{y} - \mathbf{x}\|_2^2 \quad \text{for all } \mathbf{x}, \mathbf{y} \in \mathbb{R}^N$$

Property 2 (Bounds on eigenvalues of the Hessian)

Let f be twice continuously differentiable on $\mathbb{R}^N$.

- If f is L-smooth, one has $\nabla^2 f(\mathbf{x}) \preceq L\mathbf{I}_N$ for all $\mathbf{x} \in \mathbb{R}^N$.
- Conversely, if $-L\mathbf{I}_N \preceq \nabla^2 f(\mathbf{x}) \preceq L\mathbf{I}_N$, then f is L-smooth.

remark: notation $\nabla^2 f(\mathbf{x}) \preceq L\mathbf{I}_N$ means that the matrix $L\mathbf{I}_N - \nabla^2 f(\mathbf{x})$ is positive semidefinite, or in other terms, the eigenvalues of the Hessian are upper-bounded by L.

proofs: in class!

Strong convexity

Definition (m-strongly convex functions)

Let f be a differentiable function and let $m > 0$. Then f is said to be m-strongly convex if

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) + \frac{m}{2}\|\mathbf{y} - \mathbf{x}\|_2^2 \quad \text{for all } \mathbf{x}, \mathbf{y} \in \mathbb{R}^N$$

remarks

- the definition can be extended to non-differentiable functions
- if f is m-strongly convex, then f is also strictly convex
- the function f is m-strongly convex iff $g(\mathbf{x}) = f(\mathbf{x}) - \frac{m}{2}\|\mathbf{x}\|_2^2$ is convex

Implications of strong convexity and convexity

Property 1 (Hessian characterization of strongly convex functions)

Suppose that f is twice continuously differentiable on $\mathbb{R}^N$. The function f is m-strongly convex if and only if $\nabla^2 f(\mathbf{x}) \succeq m\mathbf{I}_N$ for all $\mathbf{x}$.

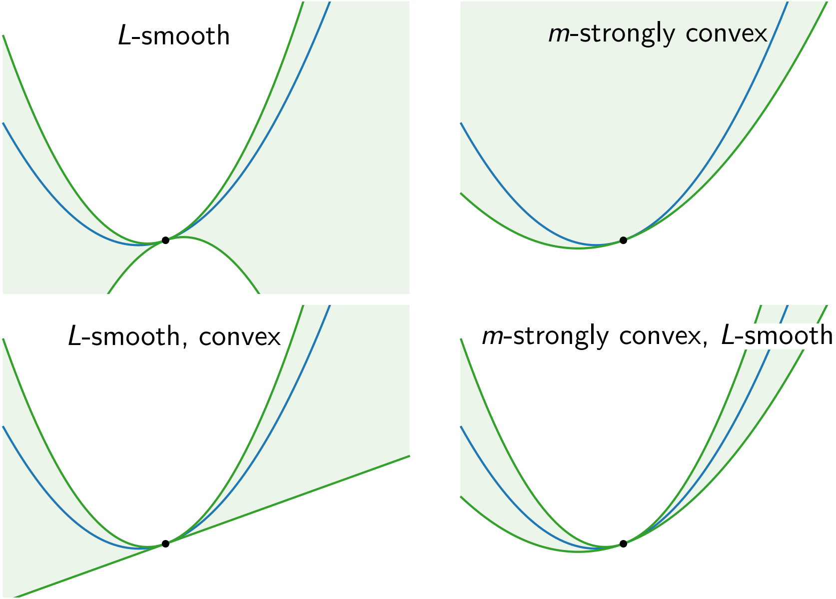
This result shows that Hessian eigenvalues are lower-bounded by m.

Property 2 (Convexity and L-smooth functions)

Let f be twice continuously differentiable on $\mathbb{R}^N$. Suppose that f is convex. Then f is L-smooth if and only if $\mathbf{0} \preceq \nabla^2 f(\mathbf{x}) \preceq L\mathbf{I}_N$.

proofs: in class!

Summary: smoothness and strong convexity



Assessing convergence

The kind of convergence guarantees strongly depend on the regularity properties (convexity, smoothness, strong convexity) of the objective f.

Recall that $\mathbf{x}^\star$ is a minimizer of f.

Convergence in objective function values

In this case, we bound the distance to the minimum

$$f(\mathbf{x}^{(k)}) - f(\mathbf{x}^\star)$$

this condition is usually weaker and requires less assumptions about f.

Convergence in iterates

Here, we bound the distance between the current iterate and a minimizer $\mathbf{x}^\star$

$$\|\mathbf{x}^{(k)} - \mathbf{x}^\star\|_2$$

This usually requires strong convexity of f. In some cases, both type of convergence can be related.

Outline

Setting: constant stepsize gradient descent

Context: Solve the unconstrained optimization problem

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

through gradient descent with constant step size $\alpha > 0$

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \alpha\nabla f(\mathbf{x}^{(k)}), \quad k = 0, 1, 2, \dots$$

Motivations

- simplest case, yet already illuminating
- common strategy for large scale applications (e.g. machine learning)
- results can be extended to backtracking / optimal step size without too much trouble.

Smooth gradient descent: first results

Assumptions: f is continuously differentiable and f is L-smooth

Let us exploit Property 1 of smooth functions. Let $\mathbf{y} = \mathbf{x}^{(k+1)}$ and $x = \mathbf{x}^{(k)}$. We get the inequality

$$f(\mathbf{x}^{(k+1)}) \leq f(\mathbf{x}^{(k)}) + \nabla f(\mathbf{x}^{(k)})^\top (\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}) + \frac{L}{2} \|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\|_2^2$$

Using the gradient descent update, this simplifies as

$$f(\mathbf{x}^{(k+1)}) \leq f(\mathbf{x}^{(k)}) + \alpha \left(\alpha \frac{L}{2} - 1 \right) \|\nabla f(\mathbf{x}^{(k)})\|_2^2$$

The RHS is minimized for $\alpha = 1/L$. For this choice of stepsize, one gets

$$f(\mathbf{x}^{(k+1)}) = f(\mathbf{x}^{(k)} - (1/L)\nabla f(\mathbf{x}^{(k)})) \leq f(\mathbf{x}^{(k)}) - \frac{1}{2L} \|\nabla f(\mathbf{x}^{(k)})\|_2^2$$

in particular $f(\mathbf{x}^{(k+1)}) \leq f(\mathbf{x}^{(k)})$ for this choice of stepsize.

Smooth gradient descent: general case (i)

Assumptions: f is continuously differentiable and f is L-smooth
A minimizer $\mathbf{x}^\star$ exists such that $f(\mathbf{x}) \geq f(\mathbf{x}^\star)$ for all $\mathbf{x}$

Now, summing the previous inequalities from $k = 0$ to $k + 1 = K$ one gets

$$f(\mathbf{x}^{(K)}) \leq f(\mathbf{x}^{(0)}) - \frac{1}{2L} \sum_{k=0}^{K-1} \|\nabla f(\mathbf{x}^{(k)})\|_2^2$$

Since $f(\mathbf{x}^{(K)}) \geq f(\mathbf{x}^\star)$,

$$\sum_{k=0}^{K-1} \|\nabla f(\mathbf{x}^{(k)})\|_2^2 \leq 2L \left[f(\mathbf{x}^{(0)}) - f(\mathbf{x}^\star) \right]$$

and thus $\lim_{K\to\infty} \|\nabla f(\mathbf{x}^{(K)})\|_2^2 = 0$

Hence, gradient descent converges to a stationary point of f.

Smooth gradient descent: general case (ii)

Assumptions: f is continuously differentiable and f is L-smooth
A minimizer $\mathbf{x}^\star$ exists such that $f(\mathbf{x}) \geq f(\mathbf{x}^\star)$ for all $\mathbf{x}$

In addition, we get that

$$\min_{0 \leq k \leq K-1} \|\nabla f(\mathbf{x}^{(k)})\|_2 \leq \sqrt{\frac{2L[f(\mathbf{x}^{(0)}) - f(\mathbf{x}^{(K)})]}{K}} \leq \sqrt{\frac{2L[f(\mathbf{x}^{(0)}) - f(\mathbf{x}^\star)]}{K}}$$

After K steps of gradient descent, we can find a point $\mathbf{x}$ such that

$$\|\nabla f(\mathbf{x})\|_2 \leq \sqrt{\frac{2L[f(\mathbf{x}^{(0)}) - f(\mathbf{x}^\star)]}{K}}$$

Comments

- rate is very slow, in $K^{-1/2}$
- we cannot say much more without additional assumptions on f.

Smooth gradient descent: convex case

Assumptions: f is continuously differentiable and f is L-smooth
A minimizer $\mathbf{x}^\star$ exists such that $f(\mathbf{x}) \geq f(\mathbf{x}^\star)$ for all $\mathbf{x}$
f is convex

Theorem

Under the above assumptions, the gradient descent method with constant stepsize $\alpha_k = 1/L$ generates a sequence $\{\mathbf{x}^{(k)}\}$ such that, after K iterations,

$$f(\mathbf{x}^{(K)}) - f(\mathbf{x}^\star) \leq \frac{L}{2K} \|\mathbf{x}^{(0)} - \mathbf{x}^\star\|_2^2$$

Proof: in class.

Comments

- rate has improved, in K^{-1}
- simple proof for quadratic functions (see next slide)
- convergence in cost, but we can only show boundedness of iterates, i.e., : $\|\mathbf{x}^{(k)} - \mathbf{x}^\star\| \leq \|\mathbf{x}^{(0)} - \mathbf{x}^\star\|$

Illustration for convex quadratic functions

Let $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top \mathbf{Q}\mathbf{x} - \mathbf{p}^\top \mathbf{x}$ with $\mathbf{Q} \geq 0$, hence f is convex.

Recall $\nabla f(\mathbf{x}) = \mathbf{Q}\mathbf{x} - \mathbf{p}$. Moreover f is L-smooth with $L = \|\mathbf{Q}\|$.

Since $\mathbf{Q} \geq 0$ a minimizer always exists, given by $\mathbf{x}^\star = \mathbf{Q}^\dagger \mathbf{p}$ where $\mathbf{Q}^\dagger$ is the pseudo inverse of $\mathbf{Q}$ ($\mathbf{Q}^\dagger = \mathbf{Q}^{-1}$ when $\mathbf{Q} \succ 0$)

Iterates read

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \frac{1}{L}(\mathbf{Q}\mathbf{x}^{(k)} - \mathbf{p}) = \mathbf{x}^{(k)} - \frac{1}{L}\mathbf{Q}(\mathbf{x}^{(k)} - \mathbf{x}^\star)$$

Therefore

$$\mathbf{x}^{(k+1)} - \mathbf{x}^\star = \left[\mathbf{I}_N - \frac{1}{L}\mathbf{Q}\right](\mathbf{x}^{(k)} - \mathbf{x}^\star) = \left[\mathbf{I}_N - \frac{1}{L}\mathbf{Q}\right]^{k+1}(\mathbf{x}^{(0)} - \mathbf{x}^\star)$$

Since $\mathbf{Q} \geq 0$, eigenvalues of $\mathbf{I}_N - \frac{1}{L}\mathbf{Q}$ are in $[0, 1]$. Therefore $\|\mathbf{I}_N - \frac{1}{L}\mathbf{Q}\|_2 \leq 1$ and thus we have boundedness of the iterates

$$\|\mathbf{x}^{(k+1)} - \mathbf{x}^\star\|_2 \leq \|\mathbf{x}^{(0)} - \mathbf{x}^\star\|$$

Smooth gradient descent: strongly convex case

Assumptions: f is continuously differentiable and f is L-smooth
A minimizer $\mathbf{x}^\star$ exists such that $f(\mathbf{x}) \geq f(\mathbf{x}^\star)$ for all $\mathbf{x}$
f is m-strongly convex

Theorem

Under the above assumptions, the gradient descent method with $\alpha_k = 1/L$ generates a sequence $\{\mathbf{x}^{(k)}\}$ such that, after K iterations,

$$f(\mathbf{x}^{(K)}) - f(\mathbf{x}^\star) \leq \left(1 - \frac{m}{L}\right)^K \left(f(\mathbf{x}^{(0)}) - f(\mathbf{x}^\star)\right)$$

$$\|\mathbf{x}^{(K)} - \mathbf{x}^\star\|_2^2 \leq \left(1 - \frac{m}{L}\right)^K \|\mathbf{x}^{(0)} - \mathbf{x}^\star\|_2^2$$

Comments

- convergence in objective and iterates;
- exponential rate (also known as *linear convergence*)
- rate depends on the ratio m/L - yet not the best in this case

Easy proof for strongly convex quadratic functions

Let $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top \mathbf{Q}\mathbf{x} - \mathbf{p}^\top \mathbf{x}$ with $\mathbf{Q} \succ 0$. Note $m = \lambda_{\min}(\mathbf{Q})$ and $L = \lambda_{\max}(\mathbf{Q})$. Hence f is m–strongly convex and L–smooth

Then

$$\begin{aligned}
\mathbf{x}^{(k+1)} - \mathbf{x}^* &= \mathbf{x}^{(k)} - \frac{1}{L}(\mathbf{Q}\mathbf{x}^{(k)} - \mathbf{p}) - \mathbf{x}^* \\
&= \mathbf{x}^{(k)} - \mathbf{x}^* - \frac{1}{L}\mathbf{Q}(\mathbf{x}^{(k)} - \mathbf{x}^*) \\
&= (\mathbf{I}_N - \frac{1}{L}\mathbf{Q})(\mathbf{x}^{(k)} - \mathbf{x}^*)
\end{aligned}$$

Since $\mathbf{Q} \succ 0$, the eigenvalues of $\mathbf{I}_N - \frac{1}{L}\mathbf{Q}$ are in $[0, 1 - (m/L)]$. Therefore, by recursion

$$\|\mathbf{x}^{(K)} - \mathbf{x}^*\|_2^2 \le \left(1 - \frac{m}{L}\right)^K \|\mathbf{x}^{(0)} - \mathbf{x}^*\|_2^2.$$

similar proof for objective values

Smooth gradient descent: strongly convex case

When f is L-smooth and m-strongly convex, we can further refine converge rates through additional characterizations.

Lemma (Coercivity of the gradient)

Let f be L-smooth and m-strongly convex on $\mathbb{R}^N$. Then for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^N$, one has

$$(\nabla f(\mathbf{x}) - \nabla f(\mathbf{y}))^\top (\mathbf{x} - \mathbf{y}) \geq \frac{mL}{m+L} \|\mathbf{x} - \mathbf{y}\|^2 + \frac{1}{L+m} \|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|^2$$

Proof: see e.g., Bubeck (2015)

Smooth gradient descent: strongly convex case [refined]

Assumptions: f is continuously differentiable and f is L-smooth
A minimizer $\mathbf{x}^\star$ exists such that $f(\mathbf{x}) \geq f(\mathbf{x}^\star)$ for all $\mathbf{x}$
f is m-strongly convex

Theorem

Under the above assumptions, the gradient descent method with $\alpha_k = 2/(m+L)$ generates a sequence $\{\mathbf{x}^{(k)}\}$ such that, after K iterations,

$$f(\mathbf{x}^{(K)}) - f(\mathbf{x}^\star) \leq \frac{L}{2}\left(\frac{L-m}{L+m}\right)^{2K} \|\mathbf{x}^{(0)} - \mathbf{x}^\star\|_2^2$$

$$\|\mathbf{x}^{(K)} - \mathbf{x}^\star\|_2^2 \leq \left(\frac{L-m}{L+m}\right)^{2K} \|\mathbf{x}^{(0)} - \mathbf{x}^\star\|_2^2$$

- improved rate: still linear, but better constant
- stepsize incorporate our knowledge of strong convexity

Summary of results

Solve the unconstrained optimization problem

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

through gradient descent with constant step size $\alpha > 0$

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \alpha\nabla f(\mathbf{x}^{(k)}), \quad k = 0, 1, 2, \ldots$$

assumption on f	stepsize	rate	conv iterates
smooth	$\alpha = \frac{1}{L}$	$\|\nabla f(x^{(K)})\| = \mathcal{O}(\sqrt{K})$	-
convex, smooth	$\alpha = \frac{1}{L}$	$f(\mathbf{x}^{(K)}) - f(\mathbf{x}^\star) = \mathcal{O}(K^{-1})$	bounded
str. convex, smooth	$\alpha = \frac{1}{L}$	$f(\mathbf{x}^{(K)}) - f(\mathbf{x}^\star) = \mathcal{O}(c^K)$	yes
str. convex, smooth	$\alpha = \frac{2}{L+m}$	$f(\mathbf{x}^{(K)}) - f(\mathbf{x}^\star) = \mathcal{O}(d^{2K})$	yes

Assessing the trade-off precision - complexity

It is useful to express convergence results as *complexity* bounds, i.e., given a precision ε, provide a bound on the number of iterations K.

Example Consider the smooth convex case. We have from the theorem:

$$f(\mathbf{x}^{(K)}) - f(\mathbf{x}^\star) \leq \frac{L}{2K}\|\mathbf{x}^{(0)} - \mathbf{x}^\star\|_2^2$$

Let $\varepsilon > 0$ be the desired precision after K iterations. Then, if one wants $f(\mathbf{x}^{(K)}) - f(\mathbf{x}^\star) \leq \varepsilon$, this can be satisfied if that

$$\frac{L}{2K}\|\mathbf{x}^{(0)} - \mathbf{x}^\star\|_2^2 \leq \varepsilon$$

and thus whenever

$$K \geq \frac{L}{2\varepsilon}\|\mathbf{x}^{(0)} - \mathbf{x}^\star\|_2^2$$

Assessing the trade-off precision - complexity

Let $\tau^{(k)} = f(\mathbf{x}^{(k)}) - f(\mathbf{x}^\star)$ and note $c = 1 - m/L$.

Exercise

For f smooth and strongly convex, show that the number of iterations of gradient descent with $\alpha = 1/L$ scales with ε as

$$K \geq c_1 \log \frac{1}{\varepsilon} + c_2$$

where $c_1 \geq 0, c_2 \in \mathbb{R}$ are constants to be determined.

About other stepsize strategies

Most of the results for constant step size can be adapted to other stepsize strategies.

For instance, let f be L-smooth and m-strongly convex.

Optimal step size same result as $\alpha = \frac{1}{L}$, i.e.,

$$f(\mathbf{x}^{(K)}) - f(\mathbf{x}^\star) \leq \left(1 - \frac{m}{L}\right)^K \left(f(\mathbf{x}^{(0)}) - f(\mathbf{x}^\star)\right)$$

Backtracking line search similar result, with a different constant

$$f(\mathbf{x}^{(K)}) - f(\mathbf{x}^\star) \leq c^K \left(f(\mathbf{x}^{(0)}) - f(\mathbf{x}^\star)\right)$$

where $c = 1 - \min(2ms, 2\eta sm/L) < 1$, $\quad (s, \eta)$ backtracking parameters

See §9.3.1 Boyd and Vanderberghe (2004) for details

Outline

1. Main mathematical tools
2. Convergence results for gradient descent
3. Convergence results for Newton's method

Setting: Newton's method with unit stepsize

Context: Solve the unconstrained optimization problem

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

through Newton's method with unit step size

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \left[\nabla^2 f(\mathbf{x}^{(k)})\right]^{-1} \nabla f(\mathbf{x}^{(k)}), \quad k = 0, 1, 2, \ldots$$

Basic assumptions

- f is twice differentiable
- a minimizer $\mathbf{x}^\star$ exists and satisfies the sufficient second-order optimality conditions (see Sessio 2), in particular

$$\nabla f(\mathbf{x}^\star) \quad \text{and} \quad \nabla^2 f(\mathbf{x}^\star) \succ 0$$

Convergence result for Newton's method

Additional assumptions the Hessian $\nabla^2 f$ is Lipschitz-continuous in a neighborhood of $\mathbf{x}^\star$.

Theorem (see e.g. Nocedal and Wright (2006))

Under these assumptions, Newton's method with unit step size satisfies

1. *for $\mathbf{x}^{(0)}$ close enough to $\mathbf{x}^\star$, the sequence $\{\mathbf{x}^{(k)}\}$ converges to $\mathbf{x}^\star$;*
2. *the rate of convergence of $\{\mathbf{x}^{(k)}\}$ is quadratic*
3. *the sequence of gradient norms $\{\|\nabla f(\mathbf{x}^{(k)})\|_2\}$ converges quadratically to zero*

Comments

- Convergence guarantees are local, i.e., apply near the minimizer $\mathbf{x}^\star$
- Convergence is quadratic, i.e.,

$$\|\mathbf{x}^{(k+1)} - \mathbf{x}^\star\|_2 \leq c\|\mathbf{x}^{(k)} - \mathbf{x}^\star\|_2^2$$

(Compare with the linear rate of gradient descent)

Proof [from Nocedal and Wright (2006)]

Start by evaluating the distance and exploit the fact that $\nabla f(\mathbf{x})^\star = 0$

$$\begin{aligned}\mathbf{x}^{(k+1)} - \mathbf{x}^\star &= \mathbf{x}^{(k)} - \mathbf{x}^\star - \left[\nabla^2 f(\mathbf{x}^{(k)})\right]^{-1} \nabla f(\mathbf{x}^{(k)}) \\ &= \left[\nabla^2 f(\mathbf{x}^{(k)})\right]^{-1} \left[\nabla^2 f(\mathbf{x}^{(k)})(\mathbf{x}^{(k)} - \mathbf{x}^\star) - (\nabla f(\mathbf{x}^{(k)}) - \nabla f(\mathbf{x}^\star))\right]\end{aligned}$$

Recall Taylor's theorem (Session 2) applied to gradients:

$$\nabla f(\mathbf{x}^{(k)}) - \nabla f(\mathbf{x}^\star) = \int_0^1 \nabla^2 f(\mathbf{x}^{(k)} + t(\mathbf{x}^\star - \mathbf{x}^{(k)}))(\mathbf{x}^{(k)} - \mathbf{x}^\star)\mathrm{d}t$$

We are now ready to bound the second paranthesis in the RHS above

Proof [from Nocedal and Wright (2006)]

Then

$$
\begin{aligned}
&\left\|\nabla^2 f(\mathbf{x}^{(k)})(\mathbf{x}^{(k)} - \mathbf{x}^\star) - (\nabla f(\mathbf{x}^{(k)}) - \nabla f(\mathbf{x}^\star))\right\|_2 \\
&= \left\| \int_0^1 \left[\nabla^2 f(\mathbf{x}^{(k)}) - \nabla^2 f(\mathbf{x}^{(k)} + t(\mathbf{x}^\star - \mathbf{x}^{(k)}))\right](\mathbf{x}^{(k)} - \mathbf{x}^\star)\mathrm{d}t \right\|_2 \\
&\leq \|\mathbf{x}^{(k)} - \mathbf{x}^\star\|_2 \int_0^1 \left\|\nabla^2 f(\mathbf{x}^{(k)}) - \nabla^2 f(\mathbf{x}^{(k)} + t(\mathbf{x}^\star - \mathbf{x}^{(k)}))\right\| \mathrm{d}t \\
&\leq \|\mathbf{x}^{(k)} - \mathbf{x}^\star\|_2^2 \int_0^1 Lt\mathrm{d}t \quad \text{(Lipschitz-continuity of the Hessian near } \mathbf{x}^\star\text{)} \\
&= \frac{1}{2} L \|\mathbf{x}^{(k)} - \mathbf{x}^\star\|_2^2
\end{aligned}
$$

Moreover, $\nabla^2 f(\mathbf{x}^\star)$ is non-singular and continuous. Thus there is a radius $r > 0$ such that $\|\nabla^2 f(\mathbf{x}^{(k)})^{-1}\| \leq 2\|\nabla^2 f(\mathbf{x}^\star)^{-1}\|$ for every $\mathbf{x}^{(k)}$ s.t. $\|\mathbf{x}^{(k)} - \mathbf{x}^\star\| \leq r$.

Proof [from Nocedal and Wright (2006)]

By substitution, we get

$$\|\mathbf{x}^{(k+1)} - \mathbf{x}^\star\|_2 \leq \frac{L}{2}\|\nabla^2 f(\mathbf{x}^{(k)})^{-1}\|\|\mathbf{x}^{(k)} - \mathbf{x}^\star\|_2^2 \leq \underbrace{L\|\nabla^2 f(\mathbf{x}^\star)^{-1}\|}_{=L'}\|\mathbf{x}^{(k)} - \mathbf{x}^\star\|_2^2$$

If we choose $\mathbf{x}^{(0)}$ such that $\|\mathbf{x}^{(0)} - \mathbf{x}^\star\| \leq \min(r, 1/(2L'))$ we obtain quadratic convergence to $\mathbf{x}^\star$ as

$$\begin{aligned}
\|\mathbf{x}^{(k+1)} - \mathbf{x}^\star\|_2 &\leq L'\|\mathbf{x}^{(k)} - \mathbf{x}^\star\|_2^2 \\
&\leq L' \cdot (L')^2\|\mathbf{x}^{(k-1)} - \mathbf{x}^\star\|_2^4 \\
&\leq L'(L')^2 \cdots (L')^{2^{k+1}}\|\mathbf{x}^{(0)} - \mathbf{x}^\star\|_2^{2^{k+1}} \\
&= (L')^{2^{k+1}-1}\|\mathbf{x}^{(0)} - \mathbf{x}^\star\|_2^{2^{k+1}}
\end{aligned}$$

Proof [from Nocedal and Wright (2006)]

Regarding the gradients, note $\mathbf{p}^{(k)} = -\left[\nabla^2 f(\mathbf{x}^{(k)})\right]^{-1} \nabla f(\mathbf{x}^{(k)})$ and let us observe that

$$\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)} = \mathbf{p}^{(k)}$$
$$\text{and } \nabla f(\mathbf{x}^{(k)}) + \nabla^2 f(\mathbf{x}^{(k)})\mathbf{p}^{(k)} = 0$$

Then,

$$\begin{aligned}
\|\nabla f(\mathbf{x}^{(k+1)})\|_2 &= \|\nabla f(\mathbf{x}^{(k+1)}) - \nabla f(\mathbf{x}^{(k)}) - \nabla^2 f(\mathbf{x}^{(k)})\mathbf{p}^{(k)}\|_2 \\
&= \left\| \int_0^1 \nabla^2 f(\mathbf{x}^{(k)} + t\mathbf{p}^{(k)})\mathbf{p}^{(k)} \mathrm{d}t - \nabla^2 f(\mathbf{x}^{(k)})\mathbf{p}^{(k)} \right\|_2 \\
&\leq \frac{1}{2} L \|\mathbf{p}^{(k)}\|_2^2 \\
&\leq L' \|\nabla f(\mathbf{x}^{(k)})\|_2^2
\end{aligned}$$

which shows that gradient norms converge quadratically.

Assessing precision

Exercise

For Newton's method, show that the number of iterations K that guarantees $\tau^{(K)} = \|\mathbf{x}^{(K)} - \mathbf{x}^\star\|_2 \leq \varepsilon$ is such that

$$K \geq c_1 \log\log\frac{1}{\varepsilon} + c_2$$

where c_1, c_2 are constants to be determined.

Hint: use the fact that the condition that $\mathbf{x}^{(0)}$ must be sufficiently close from $\mathbf{x}^\star$ can be formulated as $L'\tau^{(0)} < 1$.

Final comments on convergence

- convergence results can be stated in terms of objective values or in terms of convergence of iterates (sometimes, both)
- convergence rates correspond to worst-case scenarios: they do not tell about the exact rate of algorithms for a given problem
- this means that gradient descent, under the right assumptions, converges at least with a linear rate
- this means that Newton's method, under the right assumptions, converges at least with a quadratic rate
- Statements on number of iterations such as

$$K \geq c_1 \log \frac{1}{\varepsilon} + c_2 \text{ or } \qquad K \geq c_1 \log \log \frac{1}{\varepsilon} + c_2$$

are "at most" statements

⚠ constants can strongly affect the rate!