Chapter I. General Optimization

Lecture 2: Local methods in Unconstrained Optimization

Yurii Nesterov
EPLouvain (UCL)

Outline

- Relaxation and Approximation
- Necessary Optimality Conditions
- Sufficient Optimality Conditions
- Class of differentiable functions
- Class of twice differentiable functions
- Gradient method
- Rate of convergence
- Newton method

Relaxation and Approximation

Rules of this field

Goals: Find a local minimum.

Problem Class: Differentiable functions.

Oracle: $1-2$ order black box.

Desired properties: Convergence to a local minimum. Fast convergence.

NB:

We work with general nonlinear functions. Therefore we must be very careful.

Conclusion:

We have no choice except applying the idea of *Relaxation*.

Relaxation

We call a sequence $\{a_k\}_{k=0}^{\infty}$ a *relaxation sequence* if

$$a_{k+1} \leq a_k, \quad k = 0, 1, \ldots \quad .$$

Unconstrained Minimization: $\min_{x \in \mathbb{R}^n} f(x).$

We try to construct a sequence $\{x_k\}_{k=0}^{\infty}$ such that

$$f(x_{k+1}) \leq f(x_k), \quad k = 0, 1, \ldots \quad .$$

Immediate Benefits:

- If $f(x)$ is bounded below on $\mathbb{R}^n$ then the sequence $\{f(x_k)\}_{k=0}^{\infty}$ converges.
- In any case we improve the initial function value.

Note: Relaxation is always based on *Approximation*.

Approximation

We replace initial complicated object by a simpler one.

First-order approximation:

Let $f(x)$ be differentiable at $\bar{x} \in \mathbb{R}^n$. Then

$$f(y) = \qquad \underbrace{f(\bar{x}) + \langle f'(\bar{x}), y - \bar{x} \rangle}_{\text{linear approximation of } f \text{ at } \bar{x}} \qquad + \quad o(\| y - \bar{x} \|).$$

Here and in the sequel we denote by $o(r)$ some function of $r \geq 0$ such that

$$\lim_{r \to +0} \frac{1}{r} o(r) = 0, \quad o(0) = 0.$$

Vector $f'(x) = \left(\frac{\partial f(x)}{\partial x_1}, \dots, \frac{\partial f(x)}{\partial x_n} \right)^T$ is called the *gradient* of function f at point x.

Properties of the gradient

1. Denote by $\mathcal{L}(f, r)$ the sublevel set of $f(x)$:

$$\mathcal{L}(f, r) = \{x \in \mathbb{R}^n \mid \; f(x) \le r\}.$$

Consider the set of directions *tangent* to $\mathcal{L}(f, r)$ at $\bar{x}$, $f(\bar{x}) = r$:

$$S(f, \bar{x}) = \{s \in \mathbb{R}^n \mid \; s = \lim_{\substack{y_k \to \bar{x}, \\ f(y_k) = r}} \tfrac{y_k - \bar{x}}{\|y_k - \bar{x}\|}\}.$$

Lemma 1

If $s \in S(f, \bar{x})$, then $\langle f'(\bar{x}), s\rangle = 0$.

Proof: Since $f(y_k) = f(\bar{x})$, we have:

$$f(y_k) = f(\bar{x}) + \langle f'(\bar{x}), y_k - \bar{x}\rangle + o(\| \; y_k - \bar{x} \; \|) = f(\bar{x}).$$

Therefore $\langle f'(\bar{x}), y_k - \bar{x}\rangle + o(\| \; y_k - \bar{x} \; \|) = 0$.

Dividing that by $\| \; y_k - \bar{x} \; \|$ and taking the limit, we obtain the result. □

2. Let s be a direction in $\mathbb{R}^n$, $\| s \|= 1$. Consider the local decrease of $f(x)$ along s:

$$\Delta(s) = \lim_{\alpha \to +0} \frac{1}{\alpha}[f(\bar{x} + \alpha s) - f(\bar{x})].$$

Note that $f(\bar{x} + \alpha s) - f(\bar{x}) = \alpha \langle f'(\bar{x}), s \rangle + o(\alpha)$. Therefore

$$\Delta(s) = \langle f'(\bar{x}), s \rangle.$$

Further, using Cauchy-Schwartz inequality:

$$- \| x \| \cdot \| y \| \leq \langle x, y \rangle \leq \| x \| \cdot \| y \|,$$

we obtain: $\Delta(s) = \langle f'(\bar{x}), s \rangle \geq - \| f'(\bar{x}) \|$.

Let us take $\bar{s} = -f'(\bar{x}) / \| f'(\bar{x} \|$. Then

$$\Delta(\bar{s}) = -\langle f'(\bar{x}), f'(\bar{x}) \rangle / \| f'(\bar{x}) \| = - \| f'(\bar{x}) \|.$$

Conclusion: $-f'(\bar{x})$ (the *antigradient*) is the direction is of fastest local decrease of $f(x)$ at $\bar{x}$.

First-order optimality conditions

Let x^* be a local minimum of function $f(x)$:

$$\exists r > 0 : \quad \forall y \in B_n(x^*, r) \quad \Rightarrow \quad f(y) \geq f(x^*),$$

where $B_n(x, r) = \{y \in \mathbb{R}^n : \| y - x \| \leq r\}$.

Then for any y from $B_n(x^*, r)$ we have:

$$f(y) = f(x^*) + \langle f'(x^*), y - x^* \rangle + o(\| y - x^* \|) \geq f(x^*).$$

Thus, $\langle f'(x^*), s \rangle \geq 0, \quad \forall s, \ \| s \| = 1.$

However, this implies that

$$\langle f'(x^*), s \rangle = 0, \quad \forall s, \ \| s \| = 1,$$

(consider the directions s and $-s$).

Further, considering directions $s_i = e_i$, where e_i is the ith coordinate vector of $\mathbb{R}^n$, we come to the following

Conclusion: $\boxed{f'(x^*) = 0}$

Note: This condition is only a *necessary* characteristics of local minimum.

Second-order approximation

Let $f(x)$ be twice differentiable at $\bar{x}$. Then

$$f(y) = \underbrace{f(\bar{x}) + \langle f'(\bar{x}), y-\bar{x}\rangle + \frac{1}{2}\langle f''(\bar{x})(y-\bar{x}), y-\bar{x}\rangle}_{\text{quadratic approximation of } f \text{ at } \bar{x}} + o(\| y-\bar{x} \|^2).$$

The $(n \times n)$-matrix $f''(x)$: $\quad (f''(x))_{i,j} = \frac{\partial^2 f(x)}{\partial x_i \partial x_j}$ is called the *Hessian* of function f at x.

Note that the Hessian is a symmetric matrix: $f''(x) = [f''(x)]^T$.

Important: $f'(y) = f'(\bar{x}) + f''(\bar{x})(y-\bar{x}) + \mathbf{o}(\| y-\bar{x} \|)$.

(This is a vector equation)

Second-order optimality conditions

Let x^* be a local minimum of function $f(x)$:

$$\exists r > 0 : \quad \forall y \in B_n(x^*, r) \quad \Rightarrow \quad f(y) \geq f(x^*).$$

Since $f'(x^*) = 0$, for any y from $B_n(x^*, r)$ we have:

$$f(y) = f(x^*) + \tfrac{1}{2}\langle f''(x^*)(y - x^*), y - x^*\rangle + o(\| y - x^* \|^2) \geq f(x^*).$$

Thus, $\langle f''(x^*)s, s\rangle \geq 0, \quad \forall s \in \mathbb{R}^n$.

In other words, the Hessian $f''(x^*)$ is *positive semidefinite* (notation $f''(x^*) \succeq 0$).

Conclusion: $\boxed{f'(x^*) = 0, \quad f''(x^*) \succeq 0.}$

Note: This condition is again a *necessary* characteristics of a local minimum.

Reminder from Linear Algebra

1. Matrix A is called *positive semidefinite* if
$$\langle As, s\rangle \geq 0, \quad \forall s \in \mathbb{R}^n, \text{ (notation } A \succeq 0).$$
2. Matrix A is called *positive definite* if
$$\langle As, s\rangle > 0, \quad \forall s \in \mathbb{R}^n,\ s \neq 0, \text{ (notation } A \succ 0).$$
3. We write $A \succeq B$ if $A - B \succeq 0$.
4. If matrix A is *symmetric*: $A = A^T \ \Leftrightarrow\ a_{i,j} = a_{j,i}$, then all its eigenvalues $\{\lambda_i(A)\}_{i=1}^n$ are real.
5. $A \succeq 0$ iff $\lambda_i(A) \geq 0$, $i = 1 \dots n$.
6. $A \succ 0$ iff $\lambda_i(A) > 0$, $i = 1 \dots n$.
7. We assume that all eigenvalues are enumerated in the *increasing* order: $\lambda_i(A) \leq \lambda_{i+1}(A)$.
8. $\lambda_1(A) I_n \preceq A \preceq \lambda_n(A) I_n$, where I_n is the identity matrix in $\mathbb{R}^n$.
9. If $A \succ 0$ then the inverse matrix A^{-1} exists and $[\lambda_n(A)]^{-1} I_n \preceq A^{-1} \preceq [\lambda_1(A)]^{-1} I_n$.
10. $\| A \| = \max\limits_{\|x\|=1} \| Ax \| = \max\limits_{1 \leq i \leq n} | \lambda_i(A) |$.
11. $\| Ax \| \leq \| A \| \cdot \| x \|$.

Second-order sufficient conditions

Theorem 2

Let function $f(x)$ be twice differentiable on $\mathbb{R}^n$ and x^* satisfy the following conditions:

$$f'(x^*) = 0, \quad f''(x^*) \succ 0.$$

Then x^* is a strict local optimum of $f(x)$.

(strict $\equiv$ isolated)

Proof: Note that in a small neighborhood of point x^* function $f(x)$ can be represented as follows:

$$f(y) = f(x^*) + \tfrac{1}{2}\langle f''(x^*)(y - x^*), y - x^* \rangle + o(\| y - x^* \|^2).$$

Since $\frac{1}{r}o(r) \to 0$, there exists some $\bar{r} > 0$ such that

$$| \, o(r^2) \, | \leq \tfrac{r^2}{4}\lambda_1(f''(x^*))$$

for all $r \in [0, \bar{r}]$. Therefore for any $y \in B_n(x^*, \bar{r})$ we have:

$$\begin{aligned} f(y) \geq \ & f(x^*) + \tfrac{1}{2}\lambda_1(f''(x^*)) \| y - x^* \|^2 + o(\| y - x^* \|^2) \\ \geq \ & f(x^*) + \tfrac{1}{4}\lambda_1(f''(x^*)) \| y - x^* \|^2 > f(x^*). \end{aligned}$$

□

Class of differentiable functions

Let $f(x)$ be differentiable on $\mathbb{R}^n$ and its gradient is Lipshitz continuous:

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

Notation: $f \in C_L^{1,1}(\mathbb{R}^n)$.

Note:

1. Notation $f \in C_L^{k,p}(Q)$, where Q is a subset of $\mathbb{R}^n$, means:
   - f is k times continuously differentiable on Q.
   - Its pth derivative is Lipshitz continuous on Q with the constant L.
2. We always have $p \leq k$.
3. If $q \geq k$ then $C_L^{q,p}(Q) \subseteq C_L^{k,p}(Q)$.

   Example: $C_L^{2,1}(Q) \subseteq C_L^{1,1}(Q)$.
4. Notation $f \in C^k(Q)$ means that f is k times continuously differentiable on Q.

Lemma 3

Function $f(x)$ belongs to $C_L^{2,1}(\mathbb{R}^n)$ iff $\quad \| f''(x) \| \leq L, \quad \forall x \in \mathbb{R}^n$.

Proof: Indeed, for any $x, y \in \mathbb{R}^n$ we have:

$$\begin{aligned} f'(y) &= f'(x) + \int_0^1 f''(x+\tau(y-x))(y-x)d\tau \\ &= f'(x) + \left(\int_0^1 f''(x+\tau(y-x))d\tau \right) \cdot (y-x). \end{aligned}$$

Therefore, if the condition of lemma is satisfied, then

$$\begin{aligned} &\| f'(y) - f'(x) \| = \| \left(\int_0^1 f''(x+\tau(y-x))d\tau \right) \cdot (y-x) \| \\ &\leq \| \int_0^1 f''(x+\tau(y-x))d\tau \| \cdot \| y-x \| \\ &\leq \int_0^1 \| f''(x+\tau(y-x)) \| d\tau \cdot \| y-x \| \leq L \| y-x \| . \end{aligned}$$

On the other hand, if $f \in C_L^{2,1}(\mathbb{R}^n)$, then for any $s \in \mathbb{R}^n$, $\alpha > 0$, we have:

$$\| \left(\int_0^\alpha f''(x+\tau s)d\tau \right) \cdot s \| = \| f'(x+\alpha s) - f'(x) \| \leq \alpha L \| s \|.$$

Dividing this by α and taking the limit, we obtain the result.

Examples

1. Quadratic function

$$f(x) = \alpha + \langle a, x\rangle + \tfrac{1}{2}\langle Ax, x\rangle, \quad A = A^T.$$

We have:

$$f'(x) = a + Ax, \quad f''(x) = A.$$

Therefore $f(x) \in C_L^{1,1}(\mathbb{R}^n)$ with $L = \| A \|$.

2. $f(x) = \sqrt{1+x^2}$, $x \in \mathbb{R}$. We have:

$$f'(x) = \frac{x}{\sqrt{1+x^2}}, \quad f''(x) = \frac{1}{(1+x^2)^{3/2}} \leq 1.$$

Therefore $f(x) \in C_1^{1,1}(\mathbb{R})$.

Lemma 4

Let $f \in C_L^{1,1}(\mathbb{R}^n)$. Then for any x, y from $\mathbb{R}^n$ we have:

$$| f(y) - f(x) - \langle f'(x), y - x\rangle | \le \tfrac{L}{2} \parallel y - x \parallel^2 . \qquad (1)$$

Proof:

$$\begin{aligned} f(y) = & \quad f(x) + \int\limits_0^1 \langle f'(x + \tau(y-x)), y - x\rangle d\tau \\ = & \quad f(x) + \langle f'(x), y - x\rangle + \int\limits_0^1 \langle f'(x+\tau(y-x)) - f'(x), y - x\rangle d\tau. \end{aligned}$$

Therefore

$$\begin{aligned} & | f(y) - f(x) - \langle f'(x), y - x\rangle | \\ & = \left| \int\limits_0^1 \langle f'(x + \tau(y-x)) - f'(x), y - x\rangle d\tau \right| \\ & \le \int\limits_0^1 | \langle f'(x+\tau(y-x)) - f'(x), y - x\rangle | \, d\tau \\ & \le \int\limits_0^1 \parallel f'(x + \tau(y-x)) - f'(x) \parallel \cdot \parallel y - x \parallel d\tau \\ & \le \int\limits_0^1 \tau L \parallel y - x \parallel^2 d\tau = \tfrac{L}{2} \parallel y - x \parallel^2 . \qquad \square \end{aligned}$$

Class of twice differentiable functions

Let $f(x)$ be twice differentiable on $\mathbb{R}^n$ and its Hessian is Lipshitz continuous: $\quad \forall x, y \in \mathbb{R}^n : \quad \| f''(x) - f''(y) \| \leq M \| x - y \|$.
(Notation: $f \in C_M^{2,2}(\mathbb{R}^n)$.)

Lemma 5. Let $f \in C_L^{2,2}(\mathbb{R}^n)$. Then for any x, y from $\mathbb{R}^n$ we have:

$$\| f'(y) - f'(x) - f''(x)(y-x) \| \leq \tfrac{M}{2} \| y - x \|^2 . \tag{2}$$

Proof: $$f'(y) = f'(x) + \int_0^1 f''(x + \tau(y-x))(y-x)d\tau$$

$$= f'(x) + f''(x)(y-x) + \int_0^1 (f''(x+\tau(y-x)) - f''(x))(y-x)d\tau.$$

Therefore
$$\begin{aligned} & \| f'(y) - f'(x) - f''(x)(y-x) \| \\ & = \| \int_0^1 (f''(x+\tau(y-x)) - f''(x))(y-x)d\tau \| \\ & \leq \int_0^1 \| (f''(x+\tau(y-x)) - f''(x))(y-x) \| \, d\tau \\ & \leq \int_0^1 \| f''(x+\tau(y-x)) - f''(x) \| \cdot \| y - x \| \, d\tau \\ & \leq \int_0^1 \tau M \| y - x \|^2 \, d\tau = \tfrac{M}{2} \| y - x \|^2 . \qquad \square \end{aligned}$$

Lemma 6

Let $f \in C_M^{2,2}(\mathbb{R}^n)$ and $\parallel y - x \parallel = r$. Then

$$f''(x) - MrI_n \preceq f''(y) \preceq f''(x) + MrI_n.$$

Proof: Denote $G = f''(y) - f''(x)$. Since $f \in C_M^{2,2}(\mathbb{R}^n)$, we have:

$$\parallel G \parallel \leq Mr.$$

This means that

$$\mid \lambda_i(G) \mid \leq Mr, \quad i = 1, \ldots, n.$$

Consequently,

$$-MrI_n \preceq G \equiv f''(y) - f''(x) \preceq MrI_n.$$

□

Gradient method

Scheme: Choose $x_0 \in \mathbb{R}^n$.

Iterate $x_{k+1} = x_k - h_k f'(x_k), \quad k = 0, 1, \dots .$

Here $h_k > 0$ is the *step size*.

Step size rules:

1. Sequence $\{h_k\}_{k=0}^{\infty}$ is fixed *apriori*:

$$h_k \;=\; h > 0, \quad \text{or} \quad h_k \;=\; \tfrac{h}{\sqrt{k+1}}, \quad \text{etc.}$$

2. *Full relaxation*: $h_k = \arg\min\limits_{h \geq 0} \; f(x_k - hf'(x_k))$.

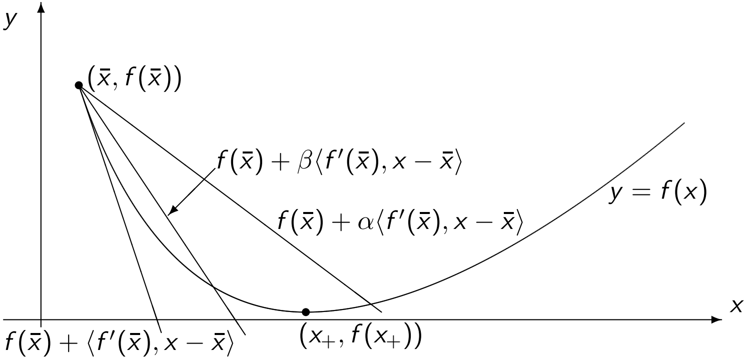
3. *Goldstein-Armijo* rule: Find $x_{k+1} = x_k - hf'(x_k)$ such that

$$\alpha \langle f'(x_k), x_k - x_{k+1} \rangle \leq f(x_k) - f(x_{k+1}), \tag{3}$$

$$\beta \langle f'(x_k), x_k - x_{k+1} \rangle \geq f(x_k) - f(x_{k+1}), \tag{4}$$

where $0 < \alpha < \beta < 1$ are some fixed parameters.

Picture


y
$(\bar{x}, f(\bar{x}))$
$f(\bar{x}) + \beta\langle f'(\bar{x}), x - \bar{x}\rangle$
$y = f(x)$
$f(\bar{x}) + \alpha\langle f'(\bar{x}), x - \bar{x}\rangle$
x
$f(\bar{x}) + \langle f'(\bar{x}), x - \bar{x}\rangle$
$(x_+, f(x_+))$

Note:

1. Rule (1) is very simple. It is often used in Convex Programming methods.

2. Rule (2) is completely theoretical. It is never used in practice.

3. Rule (3) works in the majority of the practical algorithms.

Gradient Method: Global Convergence

Problem: $\min\limits_{x \in \mathbb{R}^n} f(x).$

Assumptions:

1. $f \in C_L^{1,1}(\mathbb{R}^n)$. 2. Function $f(x)$ is bounded below on $\mathbb{R}^n$.

Main inequality: Let $y = x - hf'(x)$. In view of (1), we have:

$$\begin{array}{rl} f(y) & \leq f(x) + \langle f'(x), y - x \rangle + \frac{L}{2} \parallel y - x \parallel^2 \\ & = f(x) - h \parallel f'(x) \parallel^2 + \frac{h^2}{2} L \parallel f'(x) \parallel^2 \\ & = f(x) - h(1 - \frac{h}{2} L) \parallel f'(x) \parallel^2 . \end{array} \tag{5}$$

Optimal step size: $\Delta(h) = -h(1 - \frac{h}{2} L) \to \min\limits_h.$

$$\Delta'(h) = hL - 1 = 0 \quad \Rightarrow \quad h^* = \frac{1}{L}.$$

That is a minimum since $\Delta''(h) = L > 0$.

Optimal decrease: $f(y) \leq f(x) - \frac{1}{2L} \parallel f'(x) \parallel^2.$

Checking the rules ...

Now, let $\quad x_{k+1} = x_k - h_k f'(x_k)$.

1. **Constant step**: If $h_k = \frac{2\alpha}{L}$ with $\alpha \in (0,1)$, then

$$f(x_k) - f(x_{k+1}) \geq \tfrac{2}{L}\alpha(1-\alpha) \parallel f'(x_k) \parallel^2 .$$

Optimal choice: $h_k = \frac{1}{L}$.

2. **Full relaxation**: $f(x_k) - f(x_{k+1}) \geq \frac{1}{2L} \parallel f'(x_k) \parallel^2$.

3. **Goldstein-Armijo rule**: From (4) we have:

$$f(x_k) - f(x_{k+1}) \leq \beta \langle f'(x_k), x_k - x_{k+1} \rangle = \beta h_k \parallel f'(x_k) \parallel^2 .$$

From (5) we obtain: $f(x_k) - f(x_{k+1}) \geq h_k(1 - \frac{h_k}{2}L) \parallel f'(x_k) \parallel^2$.

Therefore $h_k \geq \frac{2}{L}(1-\beta)$.

Further, using (3) we have:

$$f(x_k) - f(x_{k+1}) \geq \alpha \langle f'(x_k), x_k - x_{k+1} \rangle = \alpha h_k \parallel f'(x_k) \parallel^2 .$$

Combining this inequality with the previous one, we conclude that

$$f(x_k) - f(x_{k+1}) \geq \tfrac{2}{L}\alpha(1-\beta) \parallel f'(x_k) \parallel^2 .$$

First convergence results

Thus, in all cases $f(x_k) - f(x_{k+1}) \geq \frac{\omega}{L} \parallel f'(x_k) \parallel^2$, where ω is a constant. Summing up these inequalities in $k = 0, \ldots, N$, we obtain:

$$\frac{\omega}{L} \sum_{k=0}^{N} \parallel f'(x_k) \parallel^2 \leq f(x_0) - f(x_N) \leq f(x_0) - f^*.$$

Conclusion: 1. $\parallel f'(x_k) \parallel \to 0$ as $k \to \infty$.

2. Let $g_N^* = \min_{0 \leq k \leq N} g_k$, where $g_k = \parallel f'(x_k) \parallel$. Then

$$g_N^* \leq \frac{1}{\sqrt{N+1}} \left[\frac{1}{\omega} L(f(x_0) - f^*) \right]^{1/2}.$$

The right hand side of this inequality describes the *rate of convergence* of the sequence $\{g_N^*\}$ to zero.

Note: 1. We cannot say anything about the rate of convergence of sequences $\{f(x_k)\}$ or $\{x_k\}$.

2. This is the only global result known for this problem class.

Rate of convergence and Complexity Estimate

Complexity Estimate $\equiv$ Inverse function of the Rate of Convergence

Example:

Problem class: 1. unconstrained minimization.
2. $f \in C_L^{1,1}(\mathbb{R}^n)$.
3. $f(x)$ is bounded below.

Oracle: First order oracle.

$\epsilon -$ **solution:** 1. $f(\bar{x}) \leq f(x_0)$,
2. $\| f'(\bar{x}) \| \leq \epsilon$.

Complexity estimate: We need $\frac{1}{\sqrt{N+1}} \left[\frac{1}{\omega} L(f(x_0) - f^*)\right]^{1/2} \leq \epsilon$.
Hence,

$$N + 1 \geq \frac{L}{\omega \epsilon^2}(f(x_0) - f^*).$$

Thus, $\frac{L}{\omega \epsilon^2}(f(x_0) - f^*)$ is an *upper complexity estimate* for this class.

Note: 1. This estimate does not depend on n.

2. This complexity bound is the *best possible* for the 1st-order methods.

Gradient method: local convergence

Problem: $\min\limits_{x \in \mathbb{R}^n} \; f(x)$ (find a local minimum).

Assumptions:

1. $f \in C_M^{2,2}(\mathbb{R}^n)$.
2. We know some bounds $0 < \ell \leq L < \infty$ for the Hessian at x^*:
$$\ell I_n \preceq f''(x^*) \preceq L I_n. \tag{6}$$
3. Our starting point x_0 is close enough to x^*.

Consider the process: $x_{k+1} = x_k - h_k f'(x_k)$. Note that

$$\begin{aligned} f'(x_k) &= f'(x_k) - f'(x^*) \\ &= \int\limits_0^1 f''(x^* + \tau(x_k - x^*))(x_k - x^*) d\tau = G_k(x_k - x^*), \end{aligned}$$

where $G_k = \int\limits_0^1 f''(x^* + \tau(x_k - x^*)) d\tau$. Therefore

$$x_{k+1} - x^* \;=\; x_k - x^* - h_k G_k(x_k - x^*) \;=\; (I - h_k G_k)(x_k - x^*).$$

Standard technique

If $a_{k+1} = A_k a_k$ and $\| A_k \| \leq 1 - q$ for some $q \in (0,1)$, then

$$\| a_{k+1} \| \leq (1-q) \| a_k \| \leq (1-q)^{k+1} \| a_0 \| \to 0.$$

Thus, we need to estimate $\| I_n - h_k G_k \|$. Denote $r_k = \| x_k - x^* \|$.

In view of Lemma 6, we have:

$$f''(x^*) - \tau M r_k I_n \preceq f''(x^* + \tau(x_k - x^*)) \preceq f''(x^*) + \tau M r_k I_n.$$

Therefore, using our assumption (6), we obtain:

$$(\ell - \tfrac{r_k}{2} M) I_n \preceq G_k \preceq (L + \tfrac{r_k}{2} M) I_n.$$

Hence, $(1 - h_k(L + \frac{r_k}{2} M)) I_n \preceq I_n - h_k G_k \preceq (1 - h_k(\ell - \frac{r_k}{2} M)) I_n$.

Thus, $\| I_n - h_k G_k \| \leq \max \left\{ 1 - h_k(\ell - \frac{r_k}{2} M), h_k(L + \frac{r_k}{2} M) - 1 \right\}$.

This means that if $r_k < \bar{r} \equiv \frac{2\ell}{M}$ then we can choose h_k: $\| I_n - h_k G_k \| < 1$.

In this case $r_{k+1} < r_k$.

Many step size strategies are possible:

1. $h_k = \frac{1}{L}$;
2. Optimal strategy: $\max\left\{1 - h(\ell - \frac{r_k}{2}M), h(L + \frac{r_k}{2}M) - 1\right\} \to \min_h$,

and others.

Optimal strategy:

We assume that $r_0 < \bar{r} \equiv \frac{2\ell}{M}$. Optimal step size h_k^* can be found from the equation:

$$1 - h(\ell - \tfrac{r_k}{2}M) = h(L + \tfrac{r_k}{2}M) - 1.$$

Hence $h_k^* = \frac{2}{L+\ell}$. Under this choice we obtain:

$$r_{k+1} \le \tfrac{(L-\ell)r_k}{L+\ell} + \tfrac{Mr_k^2}{L+\ell}.$$

Let us estimate the rate of convergence. Note that $r_0 < \bar{r}$.

Moreover, if $r_k < \bar{r}$ then $r_{k+1} < \left(\frac{L-\ell}{L+\ell} + \frac{M\bar{r}}{L+\ell}\right) r_k < \bar{r}$.

Therefore all $r_k < \bar{r}$. Denote $q = \frac{2\ell}{L+\ell}$, $a_k = \frac{M}{L+\ell} r_k$ $(< q)$. Then

$$a_{k+1} \le (1-q)a_k + a_k^2 = a_k(1 + (a_k - q)) = a_k \tfrac{1-(a_k-q)^2}{1-(a_k-q)} \le \tfrac{a_k}{1+q-a_k}.$$

Local rate of convergence

Therefore $\frac{1}{a_{k+1}} \geq \frac{1+q}{a_k} - 1$, or

$$\frac{q}{a_{k+1}} - 1 \geq \frac{q(1+q)}{a_k} - q - 1 = (1+q)\left(\frac{q}{a_k} - 1\right).$$

Hence, $\frac{q}{a_k} - 1 \;\geq\; (1+q)^k \left(\frac{q}{a_0} - 1\right)$

$$= \; (1+q)^k \left(\frac{2\ell}{L+\ell} \cdot \frac{L+\ell}{r_0 M} - 1\right) \; = \; (1+q)^k \left(\frac{\bar{r}}{r_0} - 1\right).$$

Thus, $a_k \;\leq\; \frac{q r_0}{r_0 + (1+q)^k (\bar{r} - r_0)} \;\leq\; \frac{q r_0}{\bar{r} - r_0} \left(\frac{1}{1+q}\right)^k.$

Thus, we have proved a theorem.

Theorem 7

Let function $f(x)$ satisfy our assumptions and the starting point x_0 be close enough to a local minimum:

$$r_0 = \parallel x_0 - x^* \parallel < \bar{r} = \frac{2\ell}{M}.$$

Then the gradient method with optimal step size converges linearly:

$$\parallel x_k - x^* \parallel \leq \; \frac{\bar{r} r_0}{\bar{r} - r_0} \left(\frac{L+\ell}{L+3\ell}\right)^k.$$

Remarks

1. The rate of convergence is fast.
2. We managed to prove only a local result.
3. Some unknown parameters are involved in the scheme.
4. In a smaller neighborhood $\{x \mid \| x - x^* \| \leq \frac{\ell}{M}\}$ we can guarantee that

$$f''(x) \succeq 0.$$

In this case we can obtain some stronger results using technique of Lecture 6.

Newton method

Classical scheme. Find a root of the function $\phi(t)$, $t \in \mathbb{R}$:

$$\phi(t^*) = 0.$$

Note that $\phi(t + \Delta t) = \phi(t) + \phi'(t)\Delta t + o(\mid \Delta t \mid)$.

Therefore equation $\phi(t + \Delta t) = 0$ can be approximated by

$$\phi(t) + \phi'(t)\Delta t = 0.$$

Thus, we obtain the following scheme $t_{k+1} = t_k - \frac{\phi(t_k)}{\phi'(t_k)}$.

System of nonlinear equations. Find a solution to the system

$$F(x) = 0,$$

where $x \in \mathbb{R}^n$ and $F(x) : \mathbb{R}^n \to \mathbb{R}^n$.

For that, find the displacement Δx from the following linear system:

$$F(x) + F'(x)\Delta x = 0,$$

and iterate the process: $x_{k+1} = x_k - [F'(x_k)]^{-1}F(x_k)$.

Optimization

Find a solution of the equation: $f'(x) = 0$.

For that, solve the *Newton system* $\quad f'(x) + f''(x)\Delta x = 0$,

and iterate the process: $x_{k+1} = x_k - [f''(x_k)]^{-1} f'(x_k)$.

Another interpretation. Consider the quadratic approximation of $f(x)$ at x_k:

$$\phi(x) = \quad f(x_k) + \langle f'(x_k), x - x_k \rangle + \frac{1}{2} \langle f''(x_k)(x - x_k), x - x_k \rangle.$$

Let us choose x_{k+1} as a point of minimum of $\phi(x)$:

$$\phi'(x_{k+1}) = 0 \quad \Leftrightarrow \quad f'(x_k) + f''(x_k)(x_{k+1} - x_k) = 0.$$

Thus, we obtain: $x_{k+1} = x_k - [f''(x_k)]^{-1} f'(x_k)$.

Main disadvantages

- The Hessian $f''(x_k)$ can be degenerate. Then the method breaks down.
- The method can be divergent.

In order to avoid the second trouble in practice we usually apply a *Damped Newton method*:

$$x_{k+1} = x_k - h_k[f''(x_k)]^{-1}f'(x_k),$$

where $h_k > 0$ is a stepsize parameter.

At the initial stage of the method we can apply the same step size strategies as for the gradient method.

At the final stage it is reasonable to chose $h_k = 1$.

Newton Method: local convergence

Problem: $\min\limits_{x\in\mathbb{R}^n} f(x)$ (find a local minimum).

Assumptions:

1. $f \in C_M^{2,2}(\mathbb{R}^n)$.
2. The Hessian $f''(x^*)$ is nondegenerate: $f''(x^*) \succeq \ell I_n$ for some $\ell > 0$.
3. Our starting point x_0 is close enough to x^*.

Consider the process: $x_{k+1} = x_k - [f''(x_k)]^{-1} f'(x_k)$. Then

$$\begin{aligned} x_{k+1} - x^* &= x_k - x^* - [f''(x_k)]^{-1} f'(x_k) \\ &= x_k - x^* - [f''(x_k)]^{-1} \int\limits_0^1 f''(x^* + \tau(x_k - x^*))(x_k - x^*)d\tau \\ &= [f''(x_k)]^{-1} G_k (x_k - x^*), \end{aligned}$$

where $G_k = \int\limits_0^1 [f''(x_k) - f''(x^* + \tau(x_k - x^*))]d\tau$.

Proof of the local rate of convergence

Denote $r_k = \| x_k - x^* \|$. Then

$$\begin{aligned} \| G_k \| &= \| \int_0^1 [f''(x_k) - f''(x^* + \tau(x_k - x^*))] d\tau \| \\ &\leq \int_0^1 \| f''(x_k) - f''(x^* + \tau(x_k - x^*)) \| d\tau \\ &\leq \int_0^1 M(1-\tau) r_k d\tau = \frac{r_k}{2} M. \end{aligned}$$

In view of Lemma 6, and Assumption 2 we have:

$$f''(x_k) \succeq f''(x^*) - M r_k I_n \succeq (\ell - M r_k) I_n.$$

Therefore $\| [f''(x_k)]^{-1} \| \leq (\ell - M r_k)^{-1}$. Hence, $r_{k+1} \leq \frac{M r_k^2}{2(\ell - M r_k)}$.

This type of convergence rate is called <u>*quadratic*</u>.

Theorem 8

Let function $f(x)$ satisfy our assumptions. Suppose that the initial starting point x_0 is close enough to x^*: $\| x_0 - x^* \| < \bar{r} = \frac{2\ell}{3M}$.

Then $\| x_k - x^* \| < \bar{r}$ for all k and the Newton method converges quadratically: $\| x_{k+1} - x^* \| \leq \frac{M \| x_k - x^* \|^2}{2(\ell - M \| x_k - x^* \|)}$.

Types of Convergence and Complexity

Sublinear rate. Example: $r_k \leq \frac{c}{\sqrt{k}}$. Complexity: $\boxed{\frac{c^2}{\epsilon^2}}$.

- Rather slow.
- Each new right digit of the answer takes the amount of computations comparable with the total previous work.
- Strongly depends on the constant c.

Linear rate. Example: $r_k \leq c(1-q)^k \leq ce^{-qk}$. Complexity: $\boxed{\frac{1}{q} \ln \frac{c}{\epsilon}}$.

- Fast.
- Each new right digit of the answer takes a constant amount of computations.
- The constant c plays almost no role.

Quadratic rate. Example: $r_{k+1} \leq cr_k^2$. Complexity: $\boxed{\frac{\log_2 \log_2 \frac{1}{\epsilon}}{-\log_2(c\, r_0)}}$.

- Extremely fast.
- Each iteration doubles the number of right digits in the answer.
- The constant c is important only for the starting moment of quadratic convergence ($cr_0 < 1$).