Lecture 2.

Local methods in Nonlinear Minimization

– 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.

Recall the 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.

Note:

We work with general nonlinear functions. Therefore we must be extremely 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: $\min\limits_{x \in R^n} f(x)$.

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:

1. If $f(x)$ is bounded below on R^n then the sequence $\{f(x_k)\}_{k=0}^{\infty}$ converges.

2. In any case we improve the initial function value.

Note: Relaxation is always based on *Approximation*.

Approximation.

We replace an initial complicated object by a simpler one.

1st order approximation:

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

$$f(y) = \boxed{f(\bar{x}) + \langle f'(\bar{x}), y - \bar{x} \rangle} + o(\| y - \bar{x} \|)$$

(linear approximation of f at $\bar{x}$).

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

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

The vector

$$f'(x) = \left(\frac{\partial f(x)}{\partial x_1}, \ldots, \frac{\partial f(x)}{\partial x_n} \right)$$

is called the *gradient* of function f at x.

Properties of the gradient

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

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

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

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

Lemma 2.1

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

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. $\square$

2. Let s be a direction in 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-Shwartz 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 the direction of the fastest local decrease of $f(x)$ at $\bar{x}$.

1st 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

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

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 R^n, we come to the following

Conclusion:

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

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

2nd order approximation.

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

$$f(y) = f(\bar{x}) + \langle f'(\bar{x}), y - \bar{x}\rangle + \tfrac{1}{2}\langle f''(\bar{x})(y - \bar{x}), y - \bar{x}\rangle$$

(quadratic approximation of f 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:

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

(that is a vector equation)

2nd 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:

$$\begin{aligned} f(y) = \; & f(x^*) + \langle f''(x^*)(y - x^*), y - x^* \rangle \\ & + o(\| y - x^* \|^2) \geq f(x^*). \end{aligned}$$

Thus,

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

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

Conclusion:

$$\boxed{f'(x^*) = 0, \quad f''(x^*) \geq 0}$$

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

Reminder from Linear Algebra

1. A matrix A is called *positive semidefinite* if
$$\langle As, s\rangle \geq 0, \quad \forall s \in R^n, \quad \text{(notation } A \geq 0).$$
2. A matrix A is called *positive definite* if
$$\langle As, s\rangle > 0, \quad \forall s \in R^n,\ s \neq 0, \quad \text{(notation } A > 0).$$
3. We write $A \geq B$ if $A - B \geq 0$.

4. If matrix A is *symmetric*:
$$A = A^T \quad \Leftrightarrow \quad a_{i,j} = a_{j,i},$$
then all its eigenvalues $\{\lambda_i(A)\}_{i=1}^n$ are real.

5. $A \geq 0$ iff $\lambda_i(A) \geq 0$, $i = 1 \ldots n$.

6. $A > 0$ iff $\lambda_i(A) > 0$, $i = 1 \ldots n$.

7. We assume that the eigenvalues are enumerated in *increasing* order: $\lambda_i(A) \leq \lambda_{i+1}(A)$.

8. $\lambda_1(A) I_n \leq A \leq \lambda_n(A) I_n$, where I_n is the identity matrix in R^n.

9. If $A > 0$ then the inverse matrix A^{-1} exists and
$$[\lambda_n(A)]^{-1} I_n \leq A^{-1} \leq [\lambda_1(A)]^{-1} I_n.$$

10. $\| A \| = \max\limits_{\|x\|=1} \| Ax \| = \max\limits_{1 \leq i \leq n} | \lambda_i(A) |$.

11. $\| Ax \| \leq \| A \| \cdot \| x \|$.

2nd order sufficient conditions.

Theorem 2.1 *Let function $f(x)$ be twice differentiable on R^n and x^* satisfy the following conditions:*

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

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

(strict $\equiv$ isolated)

Proof. Note that in a small neighborhood of the point x^* the 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 a value $\bar{r}$ such that

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

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 R^n and its gradient is Lipshitz continuous:

$$\forall x, y \in R^n : \quad \| f'(x) - f'(y) \| \leq L \| x - y \| . \quad (2.1)$$

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

Note:

1. Notation $f \in C_L^{k,p}(Q)$, where Q is a subset of 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 2.2 *Function $f(x)$ belongs to $C_L^{2,1}(R^n)$ iff*

$$\| f''(x) \| \leq L, \quad \forall x \in R^n. \tag{2.2}$$

Proof. Indeed, for any $x, y \in R^n$ 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 condition (2.2) 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 $f \in C_L^{2,1}(R^n)$ for any $s \in R^n$ and $\alpha > 0$, we have:

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

Dividing this inequality by α and taking the limit, we obtain (2.2). $\square$

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

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

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

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

Lemma 2.3 *Let $f \in C_L^{1,1}(R^n)$. Then for any x, y from R^n we have:*

$$| f(y) - f(x) - \langle f'(x), y - x \rangle | \leq \frac{L}{2} \| y - x \|^2 . \quad (2.3)$$

Proof.

$$\begin{aligned} f(y) = & f(x) + \int_0^1 \langle f'(x + \tau(y - x)), y - x \rangle d\tau \\ = & f(x) + \langle f'(x), y - x \rangle \\ & + \int_0^1 \langle f'(x + \tau(y - x)) - f'(x), y - x \rangle d\tau. \end{aligned}$$

Therefore

$$\begin{aligned} & | f(y) - f(x) - \langle f'(x), y - x \rangle | \\ = & | \int_0^1 \langle f'(x + \tau(y - x)) - f'(x), y - x \rangle d\tau | \\ \leq & \int_0^1 | \langle f'(x + \tau(y - x)) - f'(x), y - x \rangle | d\tau \\ \leq & \int_0^1 \| f'(x + \tau(y - x)) - f'(x) \| \cdot \| y - x \| d\tau \\ \leq & \int_0^1 \tau L \| y - x \|^2 d\tau = \frac{L}{2} \| y - x \|^2 . \end{aligned}$$

□

Class of twice differentiable functions.

Let $f(x)$ be twice differentiable on R^n and its Hessian is Lipshitz continuous:

$$\forall x, y \in R^n : \quad \| f''(x) - f''(y) \| \leq M \| x - y \| . \quad (2.4)$$

(Notation: $f \in C^{2,2}_M(R^n)$.)

Lemma 2.4 *Let $f \in C^{2,2}_L(R^n)$. Then for any x, y from R^n we have:*

$$\| f'(y) - f'(x) - f''(x)(y-x) \| \leq \frac{M}{2} \| y - x \|^2 . \quad (2.5)$$

Proof.

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

Therefore

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

□

Lemma 2.5 *Let $f \in C_M^{2,2}(R^n)$ and $\| y-x \| = r$. Then*

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

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

$$\| G \| \leq Mr.$$

This means that

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

Consequently,

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

□

Gradient method.

Scheme: Choose $x_0 \in R^n$.

Iterate

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

Here h_k is the *step size*.

Step size rules:

1. The sequence $\{h_k\}_{k=0}^{\infty}$ is given *apriori*:

$$h_k = h > 0,$$

$$h_k = \frac{h}{\sqrt{k+1}}.$$

2. *Full relaxation*:

$$h_k = \arg\min_{h \geq 0} \ f(x_k - h f'(x_k)).$$

3. *Goldstein-Armijo* rule:

Find $x_{k+1} = x_k - h f'(x_k)$ such that

$$\alpha \langle f'(x_k), x_k - x_{k+1} \rangle \leq f(x_k) - f(x_{k+1}), \qquad (2.6)$$

$$\beta \langle f'(x_k), x_k - x_{k+1} \rangle \geq f(x_k) - f(x_{k+1}), \qquad (2.7)$$

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

Picture

Note:

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

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

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

Gradient Method: Global Convergence.

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

Assumptions:

1. $f \in C^{1,1}_L(R^n)$.

2. Function $f(x)$ is bounded below on R^n.

Main inequality:

Let $y = x - hf'(x)$. In view of (2.3), we have:

$$\begin{aligned} f(y) &\leq f(x) + \langle f'(x), y - x \rangle + \tfrac{L}{2} \| y - x \|^2 \\ &= f(x) - h \| f'(x) \|^2 + \tfrac{h^2}{2} L \| f'(x) \|^2 \qquad (2.8) \\ &= f(x) - h(1 - \tfrac{h}{2} L) \| f'(x) \|^2 . \end{aligned}$$

Optimal step size:

$$\Delta(h) = -h(1 - \frac{h}{2} L) \to \min_h .$$

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

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

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

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

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

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

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 (2.7) 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 (2.8) 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 (2.6) 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

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

Thus, in all cases

$$f(x_k) - f(x_{k+1}) \geq \frac{\omega}{L} \parallel f'(x_k) \parallel^2,$$

where ω is some constant.

Summarizing 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 $g_N^* = \min\limits_{0 \leq k \leq N} g_k$, 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. That is the only global result known for that class.

Rate of convergence and Complexity Estimate.

Complexity Estimate $\equiv$ an inverse function of the Rate of Convegence.

Example.

Problem class: 1. unconstrained minimization.
2.$f \in C_L^{1,1}(R^n)$.
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:

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

Thus,

$$\frac{L}{\omega\epsilon^2}(f(x_0) - f^*)$$

is an *upper complexity estimate* for that class.

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

2. The corresponding lower complexity bound is not known.

Gradient method: local convergence.

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

Assumptions:

1. $f \in C^{2,2}_M(R^n)$.

2. We know some bounds $0 < l \leq L < \infty$ for the Hessian at x^*:
$$lI_n \leq f''(x^*) \leq LI_n. \tag{2.9}$$

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

Therefore
$$\begin{aligned} x_{k+1} - x^* &= x_k - x^* - h_k G_k(x_k - x^*) \\ &= (I - h_k G_k)(x_k - x^*). \end{aligned}$$

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 2.5, we have:

$$\begin{aligned} f''(x^*) - \tau M r_k I_n &\leq f''(x^* + \tau(x_k - x^*)) \\ &\leq f''(x^*) + \tau M r_k I_n. \end{aligned}$$

Therefore, using our assumption (2.9), we obtain:

$$(l - \frac{r_k}{2} M) I_n \leq G_k \leq (L + \frac{r_k}{2} M) I_n.$$

Hence,

$$\begin{aligned} (1 - h_k(L + \tfrac{r_k}{2} M)) I_n &\leq I_n - h_k G_k \\ &\leq (1 - h_k(l - \tfrac{r_k}{2} M)) I_n. \end{aligned}$$

Thus,

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

This means that if $r_k < \bar{r} \equiv \frac{2l}{M}$ then we can chose 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:

$$\max\left\{1 - h(l - \tfrac{r_k}{2}M), h(L + \tfrac{r_k}{2}M) - 1\right\} \to \min_h;$$

and others.

Optimal strategy.

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

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

Hence

$$h_k^* = \tfrac{2}{L+l}. \tag{2.10}$$

Under this choice we obtain:

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

Let us estimate the rate of convergence. Note that $r_0 < \bar{r}$. Moreover, if $r_k < \bar{r}$ then

$$r_{k+1} < \left(\tfrac{L-l}{L+l} + \tfrac{M\bar{r}}{L+l}\right) r_k < \bar{r}.$$

Therefore all $r_k < \bar{r}$. Denote

$$q = \tfrac{2l}{L+l}, \quad a_k = \tfrac{M}{L+l} r_k \ (< q).$$

Then

$$\begin{aligned} a_{k+1} &\leq (1-q)a_k + a_k^2 = a_k(1 + (a_k - q)) \\ &= a_k \tfrac{1-(a_k-q)^2}{1-(a_k-q)} \leq \tfrac{a_k}{1+q-a_k}. \end{aligned}$$

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,

$$\begin{aligned}\frac{q}{a_k} - 1 &\geq (1+q)^k \left(\frac{q}{a_0} - 1\right) = (1+q)^k \left(\frac{2l}{L+l} \cdot \frac{L+l}{r_0 M} - 1\right) \\ &= (1+q)^k \left(\frac{\bar{r}}{r_0} - 1\right).\end{aligned}$$

Thus,

$$a_k \leq \frac{qr_0}{r_0 + (1+q)^k(\bar{r} - r_0)} \leq \frac{qr_0}{\bar{r} - r_0}\left(\frac{1}{1+q}\right)^k \leq \frac{qr_0(1-q)^k}{\bar{r} - r_0}.$$

Thus, we have proved a theorem.

Theorem 2.2 *Let function $f(x)$ satisfy our assumptions and the starting point x_0 be close enough to a local minimum:*

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

Then the gradient method with the optimal step size (2.10) converges linearly:

$$\| x_k - x^* \| \leq \frac{\bar{r} r_0}{\bar{r} - r_0}\left(\frac{L-l}{L+l}\right)^k.$$

Remarks.

1. The rate of convergence is fast.

2. We managed to prove only a local result.

3. Too many unknown parameters are involved in the scheme.

4. In a smaller neighborhood $\{x \mid \| x - x^* \| \leq \frac{l}{M}\}$ we can guarantee that $f''(x) \geq 0$. In this case we can obtain some stronger results using the technique of Lecture 6.

Newton method.

Classical scheme.

Find a root of the function $\phi(t)$, $t \in R$:

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

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 of the system

$$F(x) = 0,$$

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

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*

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

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:

$$\begin{aligned}\phi(x) = \; & f(x_k) + \langle f'(x_k), x - x_k\rangle \\ & + \tfrac{1}{2}\langle f''(x_k)(x - x_k), x - x_k\rangle.\end{aligned}$$

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

Assumptions:

1. $f \in C_M^{2,2}(R^n)$.

2. The Hessian $f''(x^*)$ is nondegenerate:

$$f''(x^*) \geq lI_n \tag{2.11}$$

for some $l > 0$.

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

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

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

$$= [f''(x_k)]^{-1} G_k (x_k - x^*),$$

where

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

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 2.5, and (2.11), we have:

$$f''(x_k) \geq f''(x^*) - M r_k I_n \geq (l - M r_k) I_n.$$

Therefore $\| [f''(x_k)]^{-1} \| \leq (l - M r_k)^{-1}$.

Hence,

$$r_{k+1} \leq \frac{M r_k^2}{2(l - M r_k)}.$$

The rate of convergence of this type is called *quadratic*.

Thus, we have proved a theorem.

Theorem 2.3 *Let function $f(x)$ satisfy our assumptions. Suppose that the initial starting point x_0 is close enough to x^*:*

$$\| x_0 - x^* \| < \bar{r} = \frac{2l}{3M}.$$

Then $\| x_k - x^* \| < \bar{r}$ for all k and the Newton method converges quadratically:

$$\| x_{k+1} - x^* \| \leq \frac{M \| x_k - x^* \|^2}{2(l - M \| x_k - x^* \|)}.$$

Types of convergence and Complexity

Sublinear rate.

Example: $r_k \le \frac{c}{\sqrt{k}}$. Complexity: $\frac{c^2}{\epsilon^2}$.

- 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: $r_k \le c(1-q)^k$. Complexity: $\frac{1}{q}(\ln c + \ln \frac{1}{\epsilon})$.

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

Quadratic rate.

Example: $r_{k+1} \le c r_k^2$. Complexity: $\ln \ln \frac{1}{\epsilon}$.

- 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 ($cr_k < 1$).