

Lecture 3.

First-Order Methods in Nonlinear Optimization. Constrained Optimization.

- Gradient Method and Newton Method:
What is different?
- Idea of Variable Metric.
- Variable Metric Methods.
- Conjugate Gradient Methods.
- Constrained Minimization.
- Penalty Functions.
- Penalty Function Methods.
- Barrier Functions.
- Barrier Function Methods.

Gradient method and Newton method

Problem:

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

Problem class:

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

Gradient method:

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

Local linear rate of convergence.

Newton method:

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

Local quadratic rate of convergence.

Formal difference:

$$h_k f'(x_k) \quad \leftrightarrow \quad [f''(x_k)]^{-1} f'(x_k).$$

Essential difference

Gradient method:

Uses the *linear* approximation:

$$\phi_1(x) = f(\bar{x}) + \langle f'(\bar{x}), x - \bar{x} \rangle + \frac{1}{h} \|x - \bar{x}\|^2.$$

Then $\phi_1'(x) = 0$ means that

$$f'(\bar{x}) + \frac{1}{h}(x - \bar{x}) = 0 \quad \Rightarrow \quad x = \bar{x} - hf'(\bar{x}).$$

Note: if $0 < h \leq \frac{1}{L}$, then

$$f(x) \leq \phi_1(x)$$

(see Lemma 2.3).

Newton method:

Uses the *quadratic* approximation:

$$\phi_2(x) = f(\bar{x}) + \langle f'(\bar{x}), x - \bar{x} \rangle + \frac{1}{2} \langle f''(\bar{x})(x - \bar{x}), x - \bar{x} \rangle.$$

Then $\phi_2'(x) = 0$ means that

$$f'(\bar{x}) + f''(\bar{x})(x - \bar{x}) = 0 \quad \Rightarrow \quad x = \bar{x} - [f''(\bar{x})]^{-1} f'(\bar{x}).$$

Question:

Can we use some simple approximations which are better than the linear one?

Let G be a positive-definite $n \times n$ -matrix. Denote

$$\phi_G(x) = f(\bar{x}) + \langle f'(\bar{x}), x - \bar{x} \rangle + \frac{1}{2} \langle G(x - \bar{x}), x - \bar{x} \rangle.$$

Then $\phi'_G(x) = 0$ means that

$$f'(\bar{x}) + G(x - \bar{x}) = 0 \quad \Rightarrow \quad x = \bar{x} - G^{-1}f'(\bar{x}).$$

The first-order methods which form a sequence

$$\{G_k\} : G_k \rightarrow f''(x^*)$$

(or $\{H_k\} : H_k \equiv G_k^{-1} \rightarrow [f''(x^*)]^{-1}$) are called *variable metric* methods. (Sometimes the name *Quasi-Newton* methods is used.)

Note: For generating the sequences $\{G_k\}$ or $\{H_k\}$ we can use only the gradients.

Varying the Metric

Standard inner product: $x, y \in R^n$

$$\langle x, y \rangle = \sum_{i=1}^n x^{(i)}y^{(i)}, \quad \|x\| = \left[\sum_{i=1}^n (x^{(i)})^2 \right]^{1/2}.$$

$\|x\|$ is the *standard* Euclidean metric in R^n .

Gradient:

$$f(x+h) = f(x) + \langle f'(x), h \rangle + o(\|h\|),$$

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

New inner product

Let A be a symmetric positive-definite $n \times n$ matrix. For $x, y \in R^n$ denote

$$\langle x, y \rangle_A = \langle Ax, y \rangle, \quad \|x\|_A = \langle Ax, x \rangle^{1/2}.$$

$\|x\|_A$ is a new metric on R^n defined by A .

Note: Topologically this metric is equivalent to $\|\cdot\|$:

$$\lambda_1(A)^{1/2}\|x\| \leq \|x\|_A \leq \lambda_n(A)^{1/2}\|x\|.$$

Note: the gradient and the Hessian are changing!

$$\begin{aligned} f(x+h) &= f(x) + \langle f'(x), h \rangle + \frac{1}{2} \langle f''(x)h, h \rangle \\ &= f(x) + \langle A^{-1}f'(x), h \rangle_A + \frac{1}{2} \langle A^{-1}f''(x)h, h \rangle_A. \end{aligned}$$

Thus,

$$f'_A(x) = A^{-1}f'(x)$$

is the new gradient and

$$f''_A(x) = A^{-1}f''(x)$$

is the new Hessian (with respect to the metric defined by A).

Important: 1. The Newton direction is just the gradient computed with respect to the metric defined by $A = f''(x)$.

2. The Hessian of $f(x)$ at x computed with respect to $A = f''(x)$ is the *unit* matrix.

Example: quadratic function

$$f(x) = \alpha + \langle a, x \rangle + \frac{1}{2} \langle Ax, x \rangle,$$

where $A = A^T \succ 0$.

Note that

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

$$f'(x^*) = Ax^* + a = 0, \quad \Rightarrow \quad x^* = -A^{-1}a.$$

Newton direction:

$$d_N(x) = [f''(x)]^{-1} f'(x) = A^{-1}(Ax + a) = x + A^{-1}a.$$

Therefore for any $x \in R^n$ we have:

$$x - d_N(x) = -A^{-1}a = x^*.$$

Thus, the Newton method converges for a quadratic function in one step.

Note also that

$$f(x) = \alpha + \langle A^{-1}a, x \rangle_A + \frac{1}{2} \|x\|_A^2,$$

$$f'_A(x) = A^{-1} f'(x) = d_N(x),$$

$$f''_A(x) = A^{-1} f''(x) = I_n.$$

Variable Metric Methods

Problem:

$$\min_{x \in \mathbb{R}^n} f(x), \quad f(x) \in C^{1,1}(\mathbb{R}^n).$$

General scheme:

0. Choose $x_0 \in \mathbb{R}^n$. Set $H_0 = I_n$.
Compute $f(x_0)$ and $f'(x_0)$.
1. k th iteration ($k \geq 0$).
 - a). Set $p_k = H_k f'(x_k)$.
 - b). Find $x_{k+1} = x_k - h_k p_k$
(see Lecture 2 for the step-size rules).
 - c). Compute $f(x_{k+1})$ and $f'(x_{k+1})$.
 - d). Update the matrix $H_k : H_k \rightarrow H_{k+1}$.

Note: variable metric methods differ one from another in implementation of Step 1.d only.

Updating the metric

Note: At Step 1.c we obtain the new information:

$$f(x_{k+1}), \quad f'(x_{k+1}).$$

How we can use it for improving H_k ?

Basic model: quadratic function

$$f(x) = \alpha + \langle a, x \rangle + \frac{1}{2} \langle Ax, x \rangle,$$

$$f'(x) = Ax + a.$$

Therefore, for any $x, y \in R^n$ we have

$$f'(x) - f'(y) = A(x - y).$$

Quasi-Newton equation:

Choose H_{k+1} satisfying

$$H_{k+1}(f'(x_{k+1}) - f'(x_k)) = x_{k+1} - x_k.$$

Examples

Denote $\Delta H_k = H_{k+1} - H_k$,

$$\gamma_k = f'(x_{k+1}) - f'(x_k), \quad \delta_k = x_{k+1} - x_k.$$

1. Rank-one correction scheme:

$$\Delta H_k = \frac{(\delta_k - H_k \gamma_k)(\delta_k - H_k \gamma_k)^T}{\langle \delta_k - H_k \gamma_k, \gamma_k \rangle}.$$

2. Davidon-Fletcher-Powell scheme (DFP):

$$\Delta H_k = \frac{\delta_k \delta_k^T}{\langle \gamma_k, \delta_k \rangle} - \frac{H_k \gamma_k \gamma_k^T H_k}{\langle H_k \gamma_k, \gamma_k \rangle}.$$

3. Broyden-Fletcher-Goldfarb-Shanno scheme (BFGS):

$$\Delta H_k = \frac{H_k \gamma_k \delta_k^T + \delta_k \gamma_k^T H_k}{\langle H_k \gamma_k, \gamma_k \rangle} - \beta_k \frac{H_k \gamma_k \gamma_k^T H_k}{\langle H_k \gamma_k, \gamma_k \rangle},$$

where $\beta_k = 1 + \langle \gamma_k, \delta_k \rangle / \langle H_k \gamma_k, \gamma_k \rangle$.

And hundreds of others.

Remark. BFGS is considered as the most stable scheme.

What can be proved?

1. Finite termination. If $f(x)$ is quadratic, then no more than n iterations is necessary to find x^* .

2. Local superlinear convergence:

If $f \in C^{2,2}(R^n)$ and $f''(x^*) \succ 0$, then for any $x_0 \in R^n$ and k large enough we have

$$\|x_{k+1} - x^*\| \leq \text{const} \cdot \|x_k - x^*\| \cdot \|x_{k-n} - x^*\|$$

(the proofs are very long and technical.

Disadvantages:

1. It is necessary to store $n \times n$ -matrix.

2. The global convergence results are very weak (worse than that of the gradient method).

Conjugate gradients

Basic problem:

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

$$f(x) = \alpha + \langle a, x \rangle + \frac{1}{2} \langle Ax, x \rangle,$$

$$A = A^T \succ 0.$$

Then $x^* = -A^{-1}a$ and

$$\begin{aligned} f(x) &= \alpha + \langle a, x \rangle + \frac{1}{2} \langle Ax, x \rangle \\ &= \alpha - \langle Ax^*, x \rangle + \frac{1}{2} \langle Ax, x \rangle \\ &= \alpha - \frac{1}{2} \langle Ax^*, x \rangle + \frac{1}{2} \langle A(x - x^*), x - x^* \rangle. \end{aligned}$$

Thus, $f^* = \alpha - \frac{1}{2} \langle Ax^*, x^* \rangle$.

Therefore, without loss of generality we assume

$$f^* = 0, \quad x^* = 0.$$

From now on, our problem is

$$\min_{x \in \mathbb{R}^n} f(x), \quad f(x) = \frac{1}{2} \langle Ax, x \rangle, \quad A = A^T \succ 0.$$

Let we are given by starting point x_0 . Consider the linear subspaces

$$\mathcal{L}_k = \text{Lin} \{Ax_0, \dots, A^k x_0\}, \quad k \geq 1.$$

Conjugate gradient method: (CG)

$$\begin{aligned} x_k &= \arg \min \{f(x) : x \in x_0 + \mathcal{L}_k\}, \\ k &= 1, 2, \dots \end{aligned}$$

Lemma 3.1: $\mathcal{L}_k = \text{Lin} \{f'(x_0), \dots, f'(x_{k-1})\}$

Proof: 1. For $k = 1$ we have $f'(x_0) = Ax_0$.

2. Let the statement be true for some $k \geq 1$. Note:

$$x_k = x_0 + \sum_{i=1}^k \lambda_i A^i x_0$$

with some $\lambda_i \in R$. Therefore

$$\begin{aligned} f'(x_k) &= Ax_0 + \sum_{i=1}^k \lambda_i A^{i+1} x_0 \\ &= y + \lambda_k A^{k+1} x_0 \end{aligned}$$

with some $y \in \mathcal{L}_k$. Thus,

$$\begin{aligned} \mathcal{L}_{k+1} &= \text{Lin} \{\mathcal{L}_k, A^{k+1} x_0\} = \text{Lin} \{\mathcal{L}_k, f'(x_k)\} \\ &= \text{Lin} \{f'(x_0), \dots, f'(x_k)\}. \end{aligned}$$

□

Lemma 3.2 For any $k \neq i$ we have

$$\langle f'(x_k), f'(x_i) \rangle = 0.$$

Proof: Let $k > i$. For $\lambda \in R^k$, consider the function

$$\begin{aligned}\phi(\lambda) &= \phi(\lambda_1, \dots, \lambda_k) \\ &= f\left(x_0 + \sum_{j=1}^k \lambda_j f'(x_{j-1})\right).\end{aligned}$$

And let λ^* be defined by $x_k = x_0 + \sum_{j=1}^k \lambda_j^* f'(x_{j-1})$.

Then $\phi'(\lambda^*) = 0$. Therefore

$$0 = \frac{\partial \phi(\lambda^*)}{\partial \lambda_j} = \langle f'(x_k), f'(x_j) \rangle.$$

□

Corollary 3.1 The sequence generated by CG-method is finite.

(Since no more than n orthogonal directions do exist.)

Corollary 3.2 For any $p \in \mathcal{L}_k$ we have

$$\langle f'(x_k), p \rangle = 0.$$

Denote $\delta_i = x_{i+1} - x_i$. It is clear that

$$\mathcal{L}_k = \text{Lin} \{\delta_0, \dots, \delta_{k-1}\}.$$

Lemma 3.3 *For any $k \neq i$ we have*

$$\langle A\delta_k, \delta_i \rangle = 0.$$

Proof: Let $k > i$. Then

$$\begin{aligned} \langle A\delta_k, \delta_i \rangle &= \langle A(x_{k+1} - x_k), x_{i+1} - x_i \rangle \\ &= \langle f'(x_{k+1}) - f'(x_k), x_{i+1} - x_i \rangle \\ &= 0 \end{aligned}$$

since $\delta_i = x_{i+1} - x_i \in \mathcal{L}_{i+1} \subseteq \mathcal{L}_k$. □

Such directions are called *conjugate* with respect to A .

From point to point

Since $\mathcal{L}_k = \text{Lin} \{ \delta_0, \dots, \delta_{k-1} \}$, we can represent x_{k+1} as

$$x_{k+1} = x_k - h_k f'(x_k) + \sum_{j=0}^{k-1} \lambda_j \delta_j.$$

That is

$$\delta_k = -h_k f'(x_k) + \sum_{j=0}^{k-1} \lambda_j \delta_j.$$

Multiplying this equation by A and δ_i , $0 \leq i \leq k-1$, we obtain:

$$\begin{aligned} 0 &= \langle A\delta_k, \delta_i \rangle = -h_k \langle Af'(x_k), \delta_i \rangle + \sum_{j=0}^{k-1} \lambda_j \langle A\delta_j, \delta_i \rangle \\ &= -h_k \langle Af'(x_k), \delta_i \rangle + \lambda_i \langle A\delta_i, \delta_i \rangle \\ &= -h_k \langle f'(x_k), f'(x_{i+1}) - f'(x_i) \rangle + \lambda_i \langle A\delta_i, \delta_i \rangle. \end{aligned}$$

Therefore, $\lambda_i = 0$ for $i < k-1$ and

$$\lambda_{k-1} = \frac{h_k \|f'(x_k)\|^2}{\langle A\delta_{k-1}, \delta_{k-1} \rangle} = \frac{h_k \|f'(x_k)\|^2}{\langle f'(x_k) - f'(x_{k-1}), \delta_{k-1} \rangle}.$$

Thus, $x_{k+1} = x_k - h_k p_k$, where

$$\begin{aligned} p_k &= f'(x_k) - \frac{\|f'(x_k)\|^2 \cdot \delta_{k-1}}{\langle f'(x_k) - f'(x_{k-1}), \delta_{k-1} \rangle} \\ &= f'(x_k) - \frac{\|f'(x_k)\|^2 \cdot p_{k-1}}{\langle f'(x_k) - f'(x_{k-1}), p_{k-1} \rangle}. \end{aligned}$$

General Algorithmic Scheme

0. Choose $x_0 \in R^n$.

Compute $f(x_0)$ and $f'(x_0)$.

Set $p_0 = f'(x_0)$.

1. k th iteration ($k \geq 0$).

a). Find $x_{k+1} = x_k - h_k p_k$
(using an exact line-search procedure).

b). Compute $f(x_{k+1})$ and $f'(x_{k+1})$.

c). Compute coefficient β_k .

d). Set $p_{k+1} = f'(x_{k+1}) - \beta_k p_k$.

Rules for β_k :

$$1. \beta_k = \frac{\|f'(x_{k+1})\|^2}{\langle f'(x_{k+1}) - f'(x_k), p_k \rangle},$$

$$2. \beta_k = \frac{\|f'(x_{k+1})\|^2}{\|f'(x_k)\|^2}, \quad (\text{Fletcher-Rieves}),$$

$$3. \beta_k = \frac{\langle f'(x_{k+1}), f'(x_{k+1}) - f'(x_k) \rangle}{\|f'(x_k)\|^2}, \quad (\text{Polak-Ribbiere}).$$

Restart:

Terminate the process after n iterations. Use the last point as a starting one for the new cycle.

What can be proved?

1. Finite termination for quadratic functions.
2. Global convergence (relaxation + restart).
3. n -step local quadratic convergence:

$$\|x_{n+1} - x^*\| \leq \text{const} \cdot \|x_0 - x^*\|^2.$$

Advantages: Low memory requirements.

Disadvantages:

The global convergence can be even worse than that of the gradient method.

Constrained Minimization

Problem:

$$\begin{aligned} f_0(x) &\rightarrow \min, \\ f_i(x) &\leq 0, \quad i = 1, \dots, m, \end{aligned}$$

where $f_i(x)$ are smooth functions. For example,

$$f_i(x) \in C_L^{1,1}(\mathbb{R}^n).$$

General framework:

1. Unconstrained minimization problems are simpler than the constrained ones. (?)
2. Let us try to approximate solution of a constrained problem by a sequence of solutions of auxiliary unconstrained problems.

Sequential Unconstrained Minimization:

- Penalty function methods.
- Barrier function methods.

Penalty Functions

Definition. A continuous function $\Phi(x)$ is called a *penalty function* for a closed set Q if

- $\Phi(x) = 0$ for any $x \in Q$.
- $\Phi(x) > 0$ for any $x \notin Q$.

Property: If $\Phi_1(x)$ is a penalty function for Q_1 and $\Phi_2(x)$ is a penalty function for Q_2 , then $\Phi_1(x) + \Phi_2(x)$ is a penalty function for $Q_1 \cap Q_2$.

Examples.

Denote $(a)_+ = \max\{a, 0\}$. Let

$$Q = \{x \in R^n : f_i(x) \leq 0, i = 1, \dots, m\}.$$

1. Quadratic penalty:

$$\Phi(x) = \sum_{i=1}^m (f_i(x))_+^2.$$

2. Nonsmooth penalty:

$$\Phi(x) = \sum_{i=1}^m (f_i(x))_+,$$

and many others.

Penalty Function Method

0. Choose $x_0 \in R^n$.

Choose a sequence of penalty coefficients:

$$0 < t_k < t_{k+1}, \quad t_k \rightarrow \infty.$$

1. k th iteration ($k \geq 0$).

Find a point

$$x_{k+1} = \arg \min_{x \in R^n} \{f_0(x) + t_k \Phi(x)\}$$

using x_k as a starting point.

Note: x_{k+1} must be a global minimum of the auxiliary function.

Convergence

Denote

$$\Psi_k(x) = f_0(x) + t_k\Phi(x),$$

$$\Psi_k^* = \min_{x \in R^n} \Psi_k(x),$$

(that is a global minimum).

Assumption: there exist $\bar{t} > 0$ such that the set

$$S = \{x \in R^n : f_0(x) + \bar{t}\Phi(x) \leq f^*\}$$

is bounded.

Theorem 3.1

$$\lim_{k \rightarrow \infty} f(x_k) = f^*, \quad \lim_{k \rightarrow \infty} \Phi(x_k) = 0.$$

Proof: Note that $\Psi_k^* \leq \Psi_k(x^*) = f^*$.

Further, for any $x \in R^n$ we have:

$$\Psi_{k+1}(x) \geq \Psi_k(x).$$

Therefore $\Psi_{k+1}^* \geq \Psi_k^*$. Thus, $\exists \lim_{k \rightarrow \infty} \Psi_k^* \equiv \Psi^* \leq f^*$.

If $t_k > \bar{t}$, then

$$f_0(x_k) + \bar{t}\Phi(x_k) \leq f_0(x_k) + t_k\Phi(x_k) \leq f^*.$$

Therefore, the sequence $\{x_k\}$ has limit points.

For any limit point x_* we have $\Phi(x_*) = 0$. Therefore $x_* \in Q$ and

$$\Psi^* \geq f_0(x_*) \geq f^*. \quad \square$$

Questions:

1. What penalty function should we use?
2. What should be the rules for choosing penalty coefficients?
3. What should be the accuracy for solving auxiliary problems?

For all these questions we have almost no theoretical answers.

They are readdressed by the general NLP theory towards the computational practice.

Barrier Functions

Definition. A continuous function $F(x)$ is called a *barrier function* for a closed set Q with nonempty interior if

$$F(x) \rightarrow \infty \quad \text{when} \quad x \rightarrow \partial Q.$$

Property. If $F_1(x)$ is a barrier for Q_1 and $F_2(x)$ is a barrier for Q_2 , then $F_1(x) + F_2(x)$ is a barrier function for $Q_1 \cap Q_2$.

Slater condition:

$$\exists \bar{x} : f_i(\bar{x}) < 0, \quad i = 1, \dots, m.$$

Examples. Let

$$Q = \{x \in R^n : f_i(x) \leq 0, \quad i = 1, \dots, m\}.$$

1. Power-function barrier:

$$F(x) = \sum_{i=1}^m \frac{1}{(-f_i(x))^p}, \quad p \geq 1.$$

2. Logarithmic barrier:

$$F(x) = - \sum_{i=1}^m \ln(-f_i(x)).$$

3. Exponential barrier:

$$F(x) = \sum_{i=1}^m \exp\left(\frac{1}{-f_i(x)}\right).$$

And many others.

Barrier Function Method

0. Choose $x_0 \in \text{int } Q$.

Choose a sequence of penalty coefficients:

$$0 < t_k < t_{k+1}, \quad t_k \rightarrow \infty.$$

1. k th iteration ($k \geq 0$).

Find a point

$$x_{k+1} = \arg \min_{x \in Q} \left\{ f_0(x) + \frac{1}{t_k} F(x) \right\}$$

using x_k as a starting point.

Note: x_{k+1} must be a global minimum of the auxiliary function.

Convergence

Denote

$$\Psi_k(x) = f_0(x) + \frac{1}{t_k}F(x),$$

$$\Psi_k^* = \min_{x \in Q} \Psi_k(x)$$

(that is a global minimum).

Assumption:

$$F(x) \geq F^* \quad \forall x \in Q.$$

Theorem 3.2: $\lim_{k \rightarrow \infty} \Psi_k^* = f^*.$

Proof. Let $\bar{x} \in \text{int } Q$. Then

$$\limsup_{k \rightarrow \infty} \Psi_k^* \leq \lim_{k \rightarrow \infty} \left[f_0(\bar{x}) + \frac{1}{t_k}F(\bar{x}) \right] = f_0(\bar{x}).$$

Therefore, $\limsup_{k \rightarrow \infty} \Psi_k^* \leq f^*.$

Further,

$$\begin{aligned} \Psi_k^* &= \min_{x \in Q} \left\{ f_0(x) + \frac{1}{t_k}F(x) \right\} \\ &\geq \min_{x \in Q} \left\{ f_0(x) + \frac{1}{t_k}F^* \right\} \\ &= f^* + \frac{1}{t_k}F^*. \end{aligned}$$

Thus, $\lim_{k \rightarrow \infty} \Psi_k^* = f^*.$

□

Questions:

1. How to find the starting point?
2. Choice of barrier function.
3. Rules for updating penalty coefficients.
4. Accuracy of solutions of the auxiliary problems.
5. Efficiency estimates.

Note:

All these questions get *precise* answers in the framework of Convex Optimization.

We will see that in Lectures 10-12.

What is else in NLP?

1. Optimality conditions for Constrained Optimization.
2. Duality theory (Lecture 8).
3. Methods based on Lagrange function (Lecture 9).
4. Sequential Quadratic Optimization (Lecture 6).

Some of these topics will be discussed later in the framework of Convex Optimization.

Bibliography

1. A.Fiacco and G.McCormik.

Nonlinear Programming: sequential unconstrained optimization technique.

J.Willey & Sons, New York, 1968.

2. D. Luenberger.

Linear and Nonlinear Programming.

Addison-Wesley, 1984.