Lecture 16: Introduction to Nonlinear Optimization

Zizhuo Wang

Institute of Data and Decision Analytics (iDDA)
Chinese University of Hong Kong, Shenzhen

Nov 2, 2018

Agenda

- Nonlinear optimization (about 4 weeks)
- Integer optimization (about 2 weeks)

Introduction to Nonlinear Optimization

So far we have discussed linear optimization problems. However, in practice, there are many interesting optimization problems that do not take a linear form.

In general, we can write a nonlinear optimization problem as:

$$\begin{aligned} \text{minimize}_{\mathbf{x}} \quad & f(\mathbf{x}) \\ \text{s.t.} \quad & \mathbf{x} \in F \end{aligned}$$

We call F the feasible region and $\mathbf{x} \in F$ feasible solutions.

In the following, we study such nonlinear optimization problems.

- The properties of such problems.
- How to find the optimal solution?
- Without otherwise specified, we always assume we are solving a minimization problem.

Global and Local Optimizers

We call $\bar{\mathbf{x}}$ a *global optimizer (minimizer)* of the optimization problem:

$$\begin{aligned} \text{minimize}_{\mathbf{x}} \quad & f(\mathbf{x}) \\ \text{s.t.} \quad & \mathbf{x} \in F \end{aligned}$$

if for all $\mathbf{x} \in F$, $f(\bar{\mathbf{x}}) \leq f(\mathbf{x})$.

- The solution we obtained for LP is a global optimizer
- We always want to find global optimizers. However, sometimes this is not easy, we may have to settle on *local optimizers*

Local Minimizers

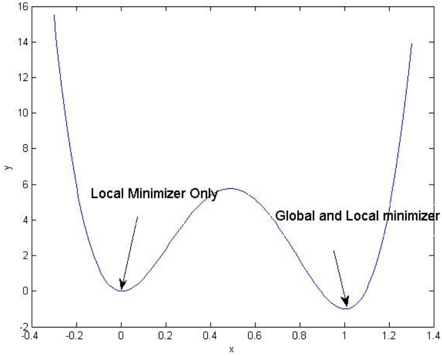
We call $\bar{\mathbf{x}}$ a *local optimizer (minimizer)* of:

$$\begin{aligned} \text{minimize}_{\mathbf{x}} \quad & f(\mathbf{x}) \\ \text{s.t.} \quad & \mathbf{x} \in F \end{aligned}$$

if there exists a neighborhood $N(\bar{\mathbf{x}})$ of $\bar{\mathbf{x}}$ (a small ball around $\bar{\mathbf{x}}$) such that for all $\mathbf{x} \in N(\bar{\mathbf{x}}) \bigcap F$, $f(\bar{\mathbf{x}}) \leq f(\mathbf{x})$.

- Global minimizer is always local minimizer, however, the reverse is not true

Example: $f(x) = 100x^2(1-x)^2 - x$



Review: Gradient, Hessian Matrix and Taylor Expansion

Let $f : \mathbb{R}^n \to \mathbb{R}$.

- Assume $f(\mathbf{x}) = f(x_1, x_2, ..., x_n)$ is continuously differentiable. Then we denote the gradient of f by (an $n \times 1$ vector)

$$\nabla f(\mathbf{x}) = \left(\frac{\partial f}{\partial x_1}; \frac{\partial f}{\partial x_2}; ...; \frac{\partial f}{\partial x_n}\right)$$

  By Taylor expansion, we have

$$f(\mathbf{x} + \alpha \mathbf{d}) = f(\mathbf{x}) + \alpha \nabla f(\mathbf{x})^T \mathbf{d} + o(\alpha)$$

- Assume f is second-order differentiable. Then we denote the Hessian matrix of f by (an $n \times n$ matrix)

$$\nabla^2 f(\mathbf{x}) = \left(\frac{\partial^2 f}{\partial x_i \partial x_j}\right)_{i,j}$$

  By Taylor expansion, we have

$$f(\mathbf{x} + \alpha \mathbf{d}) = f(\mathbf{x}) + \alpha \nabla f(\mathbf{x})^T \mathbf{d} + \frac{1}{2}\alpha^2 \mathbf{d}^T \nabla^2 f(\mathbf{x}) \mathbf{d} + o(\alpha^2)$$

Example

Suppose

$$f(x_1, x_2, x_3) = x_1^2 + x_1 x_2 + x_1 e^{x_3} + x_2 \log x_3$$

Then

$$\nabla f(\mathbf{x}) = \left(2x_1 + x_2 + e^{x_3}; x_1 + \log x_3; x_1 e^{x_3} + \frac{x_2}{x_3} \right)$$

$$\nabla^2 f(\mathbf{x}) = \begin{bmatrix} 2 & 1 & e^{x_3} \\ 1 & 0 & \frac{1}{x_3} \\ e^{x_3} & \frac{1}{x_3} & x_1 e^{x_3} - \frac{x_2}{x_3^2} \end{bmatrix}$$

Optimality Conditions

In the following, we first study what conditions an optimal solution has to satisfy for nonlinear optimization problems

- Optimality conditions
- We will start with local optimal solutions

Optimality Conditions: Unconstrained Problems

Let's start from the easiest case in which $F = \mathbb{R}^n$ (unconstrained problems).

What are the optimality conditions for local minimizers for unconstrained problems?

- Claim: We must have

$$\nabla f(\mathbf{x}) = 0$$

Reason: If $\nabla f(\mathbf{x}) \neq 0$, then we can find a vector $\mathbf{d}$ such that $\nabla f(\mathbf{x})^T \mathbf{d} < 0$. Therefore by Taylor expansion

$$f(\mathbf{x} + \alpha \mathbf{d}) = f(\mathbf{x}) + \alpha \nabla f(\mathbf{x})^T \mathbf{d} + o(\alpha)$$

By choosing α small enough, we can find a point $\mathbf{x}' = \mathbf{x} + \alpha \mathbf{d}$ in the neighborhood of $\mathbf{x}$ such that $f(\mathbf{x}') < f(\mathbf{x})$.

First-Order Necessary Condition (FONC)

Theorem (First-Order Necessary Condition)

If $\mathbf{x}^*$ is a local minimizer of $f(\cdot)$ for the unconstrained problem, then we must have $\nabla f(\mathbf{x}^*) = 0$.

Remark

First-order necessary condition provides all the candidates for local minimizers.

Example: $f(\mathbf{x}) = x_1^2 - x_1 x_2 + x_2^2 - 3x_2$.

The FONC is

$$2x_1 - x_2 = 0, \quad -x_1 + 2x_2 = 3$$

There is a unique solution ($x_1 = 1$, $x_2 = 2$), which turns out to be the global minimizer for f.

Another Example: Least Squares Problem

Assume a variable y is affected by n factors $x_1, ..., x_n$. We know that they approximately have a linear relationship:

$$y \approx \beta_1 x_1 + \cdots + \beta_n x_n$$

Now we want to find out this relationship (parameters βs).

- We have m observations ($m > n$):

$$\{\mathbf{x}_i, y_i\} = \{(x_{i1}, ..., x_{in}), y_i\}, i = 1, ..., m$$

Ideally, we want to find $\boldsymbol{\beta} = (\beta_1, ..., \beta_n)$ such that $y_i = \mathbf{x}_i^T \boldsymbol{\beta}$. Or equivalently, $\mathbf{y} = X\boldsymbol{\beta}$ where X is a matrix whose ij-th entry is x_{ij}

- However, this may not be feasible
- Usually the observations do not follow $y_i = \mathbf{x}_i^T \boldsymbol{\beta}$ exactly. There are noises in the observations.

Least Squares Problem Continued

Instead, we try to minimize the sum of the squared errors

$$\text{minimize}_{\boldsymbol{\beta}} \quad \sum_{i=1}^{m} \left(y_i - \sum_{j=1}^{n} \beta_j x_{ij} \right)^2$$

The matrix form of this problem is

$$\text{minimize}_{\boldsymbol{\beta}} \quad ||X\boldsymbol{\beta} - \mathbf{y}||_2^2 = \boldsymbol{\beta}^T X^T X \boldsymbol{\beta} - 2\boldsymbol{\beta}^T X^T \mathbf{y} + \mathbf{y}^T \mathbf{y}$$

where $||\mathbf{w}||_2^2 = \mathbf{w}^T \mathbf{w} = w_1^2 + \cdots + w_n^2$.

Facts:

- If $f(\mathbf{x}) = \mathbf{x}^T M \mathbf{x}$ (M is symmetric), then $\nabla f(\mathbf{x}) = 2M\mathbf{x}$
- If $f(\mathbf{x}) = \mathbf{c}^T \mathbf{x}$, then $\nabla f(\mathbf{x}) = \mathbf{c}$

Therefore, the FONC for the least squares problem is

$$X^T X \boldsymbol{\beta} = X^T \mathbf{y}$$

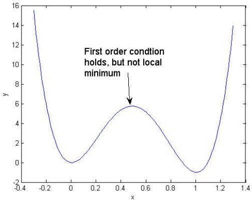
Solving this equation gives candidates for local minimizer.

FONC is Not Sufficient

In the example $f(x) = 100x^2(1-x)^2 - x$. The FONC is

$$f'(x) = 400x^3 - 600x^2 + 200x - 1 = 0$$

with solutions $x_1 = 0.01032$, $x_2 = 0.47997$ and $x_3 = 1.00972$.



We see that FONC is not sufficient

- In fact, each local maximum also satisfies the FONC.
- Or it could be neither a local minimum nor maximum (x^3)

Second-Order Necessary Condition

Consider the Taylor expansion again but to the 2nd order (assuming f is second-order differentiable):

$$f(\mathbf{x}+\alpha\mathbf{d}) = f(\mathbf{x}) + \alpha\nabla f(\mathbf{x})^T\mathbf{d} + \frac{1}{2}\alpha^2\mathbf{d}^T\nabla^2 f(\mathbf{x})\mathbf{d} + o(\alpha^2)$$

When the first-order necessary condition holds, we have

$$f(\mathbf{x}+\alpha\mathbf{d}) = f(\mathbf{x}) + \frac{1}{2}\alpha^2\mathbf{d}^T\nabla^2 f(\mathbf{x})\mathbf{d} + o(\alpha^2)$$

In order for $\mathbf{x}$ to be a local minimizer, we also need $\mathbf{d}^T\nabla^2 f(\mathbf{x})\mathbf{d}$ to be nonnegative for any $\mathbf{d}$.

Second-Order Necessary Condition (SONC)

Theorem (Second-Order Necessary Condition)

If $\mathbf{x}^*$ *is a local minimizer of* $f(\cdot)$ *for an unconstrained problem, then we must have*

1. $\nabla f(\mathbf{x}^*) = 0$;
2. *For all* $\mathbf{d}$, $\mathbf{d}^T \nabla^2 f(\mathbf{x}^*) \mathbf{d} \geq 0$.

Definition

We call a (symmetric) matrix A positive semi-definite (PSD) if and only if for all $\mathbf{x}$, $\mathbf{x}^T A \mathbf{x} \geq 0$.

Remark

Therefore, the second-order necessary condition requires the Hessian matrix at $\mathbf{x}^*$ *is PSD. In the one-dimensional case, this is equivalent to that the second derivative at* $\mathbf{x}^*$ *is nonnegative.*

Positive Semidefinite Matrices

Here are some useful facts about PSD matrices:

- We usually only talk about PSD for symmetric matrix. If a matrix A is not symmetric, we use $\frac{1}{2}(A + A^T)$ to define the PSD properties (because $\mathbf{x}^T A\mathbf{x} = \frac{1}{2}\mathbf{x}^T(A + A^T)\mathbf{x}$)
- A symmetric matrix is PSD if and only if all the eigenvalues are nonnegative.
- A symmetric matrix is PSD if and only if all the principal submatrices have nonnegative determinants
- For any matrix A, $A^T A$ is a (symmetric) PSD matrix

If A is PSD, we call $-A$ a negative semi-definite matrix.

Example Continued

In the example $f(x) = 100x^2(1-x)^2 - x$, the second-order condition is

$$6x^2 - 6x + 1 \geq 0$$

Only $x_1 = 0.01032$ and $x_3 = 1.00972$ satisfy the condition. But $x_2 = 0.47997$ does not (thus x_2 is not a local minimizer)

In the example of least squares problem, we have the following fact:

- If $f(\mathbf{x}) = \mathbf{x}^T M\mathbf{x}$ (M is symmetric), then $\nabla^2 f(\mathbf{x}) = 2M$

Therefore, the Hessian matrix in that problem is $2X^T X$, which is always a PSD matrix. Therefore, the SONC always holds.

SONC is Not Sufficient

However, even both the first- and second-order necessary conditions hold, it still can't guarantee a local minimum.

Consider $f(x) = x^3$ at 0.

- $f'(0) = f''(0) = 0$, thus FONC and SONC hold.
- 0 is not a local minimum

By modifying the SONC, we can get a sufficient condition.

Second-Order Sufficient Condition (SOSC)

Theorem

Let f be second-order continuously differentiable. If $\mathbf{x}^*$ satisfies:

1. $\nabla f(\mathbf{x}^*) = 0$;
2. *For all* $\mathbf{d} \neq 0$, $\mathbf{d}^T \nabla^2 f(\mathbf{x}^*) \mathbf{d} > 0$.

Then $\mathbf{x}^*$ is a local minimum of f for the unconstrained problem.

Definition

We call a (symmetric) matrix A positive definite (PD) if and only if for all $\mathbf{x} \neq 0$, $\mathbf{x}^T A \mathbf{x} > 0$.

- PD matrix must be PSD (thus PD is a stronger notion)
- A symmetric matrix is PD if and only if all its eigenvalues are positive
- A symmetric matrix is PD if and only if the determinants of all leading principal submatrices are positive
- If A is PD, then we call $-A$ a negative definite matrix.

Proof

The proof is again by Taylor expansion.

When $\nabla^2 f(\mathbf{x}^*)$ is positive definite, we have $\mathbf{d}^T \nabla^2 f(\mathbf{x}^*)\mathbf{d} > c||\mathbf{d}||^2$ where $c > 0$ is the smallest eigenvalue of $\nabla^2 f(\mathbf{x}^*)$.

Thus for any $\mathbf{x}^*$ that satisfies the second-order sufficient condition, for any $\mathbf{d}$, and small enough α

$$f(\mathbf{x}^* + \alpha\mathbf{d}) = f(\mathbf{x}^*) + \frac{1}{2}\alpha^2\mathbf{d}^T\nabla^2 f(\mathbf{x}^*)\mathbf{d} + o(\alpha^2) > f(\mathbf{x}^*)$$

which implies that $\mathbf{x}^*$ is a local minimizer.