

第 5 章 4-5 节牛顿法、拟牛顿法

SMaLL

¹ 中国石油大学（华东）

SMaLL 课题组

small.sem.upc.edu.cn

liangxijunsd@163.com

2023

纲要

1. 牛顿法

2. 非线性优化中的一阶方法

2.1 梯度法和牛顿法：有什么不同？

3. 拟牛顿法

牛顿法求解非线性方程

牛顿法：可用于一元函数求根

- ▶ 设 $\phi(t) : R \rightarrow R$. 考虑等式

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

- ▶ 牛顿法基于线性近似的思想

$$\phi(t + \Delta t) = \phi(t) + \phi'(t)\Delta t + o(|\Delta t|).$$

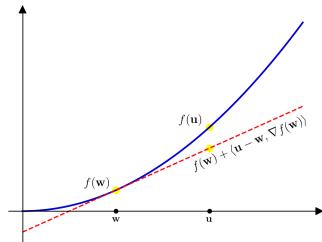
⇒ 因此方程 $\phi(t + \Delta t) = 0$ 可以由以下线性方程近似:

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

这个等式的解是最优位移 $\Delta t^* = t^* - t$
的近似 → 算法形式:



$$t_{k+1} = t_k - \frac{\phi(t_k)}{\phi'(t_k)}$$



牛顿法求解非线性方程

求解非线性方程组,

$$F(x) = 0,$$

其中 $x \in R^n$ 且 $F(x) : R^n \rightarrow R^n$.

► Δx : 线性方程组的一种解法:

$$F(x) + \nabla F(x)^T \Delta x = 0,$$

(它称为牛顿方程), 其中 $\nabla F(x) = [\nabla f_1(x) \cdots \nabla f_n(x)] \in R^{n \times n}$,

$$F(x) = \begin{bmatrix} f_1(x) \\ f_2(x) \\ \vdots \\ f_n(x) \end{bmatrix}.$$

► **备注:** 牛顿方程还可写为 $F(x) + JF(x)\Delta x = 0$, 其中 $JF(x) = \nabla F(x)^T$ 是雅可比行列式. 如果 $JF(x)$ 是非退化的,

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

牛顿法求解无约束极小化问题

无约束极小化问题 $\min_x f(x)$ \rightarrow 可以转化为非线性方程求根问题

$$f'(x) = 0. \quad (1)$$

(虽然这种转化与原问题并不完全等价, 但它在非退化情况下是有效的.) 应用牛顿法

\rightarrow 得到如下牛顿方程:

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

无约束问题的牛顿法表示为:

$$x_{k+1} = x_k - [f''(x_k)]^{-1} f'(x_k). \quad (2)$$

利用二次逼近的思想求牛顿法

利用二次逼近的思想，关于点 x_k :

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

假设 $f''(x_k) \succ 0$. 我们可以选择二次函数 $\phi(x)$ 的最小值点作为下个迭代点 x_{k+1} . 这意味着

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

\Rightarrow 再次得到牛顿迭代公式 (2): $x_{k+1} = x_k - [f''(x_k)]^{-1}f'(x_k)$.

牛顿法的优缺点

- ▶ 优点: 牛顿法在严格局部极小值附近的收敛速度非常快.
- ▶ 缺点:
 1. 如果 $f''(x_k)$ 是奇异阵, 那么该方法将无法使用.
 2. 牛顿迭代过程可能不收敛.

示例

例 1 (示例)

[牛顿迭代过程可能不收敛] 用牛顿法求下列一元函数的零点:

$$\phi(t) = \frac{t}{\sqrt{1+t^2}}.$$

注意到, 有唯一的零点 $t^* = 0$. 同时有 $\phi'(t) = \frac{1}{(1+t^2)^{3/2}}$. 牛顿迭代过程如下:

$$t_{k+1} = t_k - \frac{\phi(t_k)}{\phi'(t_k)} = t_k - \frac{t_k}{\sqrt{1+t_k^2}} \cdot [1+t_k^2]^{3/2} = -t_k^3.$$

因此,

- ▶ 若 $|t_0| < 1$, 那么该方法收敛, 且收敛速度非常快.
- ▶ 若 $t_0 = \pm 1$ 是这个方法的 震荡点
- ▶ 若 $|t_0| > 1$, 该方法将 不收敛

函数 $\phi(t) = \frac{t}{\sqrt{1+t^2}}$. 是由 $f(t) = \sqrt{1+t^2}$. 派生函数

阻尼牛顿法

为了避免不收敛的问题，可以应用阻尼牛顿法：

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

其中 $h_k > 0$ 是步长参数.

- ▶ 在方法的初始阶段，我们可以使用与梯度下降法相同的步长策略.
- ▶ 在最后阶段，可以选择 $h_k = 1$.

牛顿法的局部收敛性

接下来研究牛顿方法的局部收敛性。考虑问题

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

满足以下假设:

1. $f \in C_M^{2,2}(R^n)$, 且 Hessian 矩阵 Lipschitz 连续, Lipschitz 常数为 $M > 0$.
2. 具有正定的 Hessian 矩阵 f , f 在 x^* 处存在局部极小点:

$$f''(x^*) \succeq lI_n, l > 0. \quad (3)$$

3. 初始点 x_0 离 x^* 足够近.

牛顿方法的局部收敛性：准备工作

推论 2 (推论)

[Nesterov. Intro. Lec. on Convex Opt. pp.24, Coro.1.2.2]

设 $f \in C_M^{2,2}(R^n)$ 且 $\|y - x\| = r$, 因此

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

牛顿方法的局部收敛性：分析 I

考虑迭代过程: $x_{k+1} = x_k - [f'(x_k)]^{-1}f(x_k)$. 然后使用与梯度法相同的推理, 我们得到:

$$\begin{aligned}x_{k+1} - x^* &= x_k - x^* - [f'(x_k)]^{-1}f(x_k) \\&= x_k - x^* - [f'(x_k)]^{-1} \int_0^1 f'(x^* + \tau(x_k - x^*))(x_k - x^*) d\tau \\&= [f'(x_k)]^{-1} G_k(x_k - x^*),\end{aligned}$$

其中 $G_k = \int_0^1 [f'(x_k) - f'(x^* + \tau(x_k - x^*))] d\tau$.

记 $r_k = \|x_k - x^*\|$. 则

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

牛顿方法的局部收敛性：分析 II

由定理 2, 和 (3), 得到

$$f''(x_k) \succeq f''(x^*) - Mr_k I_n \succeq (l - Mr_k) I_n.$$

因此, 若 $r_k < \frac{l}{M}$, 则 $f''(x_k)$ 是正定的且

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

对于足够小的 r_k ($r_k < \frac{2l}{3M}$), 有

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

这种类型的收敛速度叫做二次收敛

$$\frac{M(r_k)^2}{2(l - Mr_k)} < r_k \iff Mr_k < 2(l - M \cdot r_k) \iff r_k < \frac{2l}{3M}.$$

牛顿法的局部收敛性

定理 3 (定理)

若函数 $f(x)$ 满足我们的假设. 且初始点 x_0 足够接近 x^* :

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

那么对于所有的 k , $\|x_k - x^*\| < \bar{r}$ 且牛顿法二次收敛

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

- ▶ 与梯度下降法的收敛速度进行比较 → 牛顿法要快得多.
- ▶ 牛顿法迭代点列的区域几乎与梯度下降迭代法点列的收敛区域相同
- ▶ → 可在算法的初始阶段使用梯度下降法的迭代格式 → 接近局部极小解 → 在算法的最后阶段用牛顿法来完成迭代

收敛速度和复杂性界限

收敛速度和算法复杂度的界互为反函数.

1. **次线性速率**. 次线性速率常用**迭代次数的幂函数来描述**. 比如

$r_k \leq \frac{c}{\sqrt{k}}$. 相应问题类的复杂度上界是 $(\frac{c}{\epsilon})^2$.

► 次线性速率较慢, 求得的解中每计算一个新的有效数字所需要的计算量与之前所有有效数字的计算量之和相当。

► 另外, 常数 c 在相应的复杂度估计中起着重要的作用。

当 $\epsilon = 10^{-n}$, 必要的复杂性界限 $k_n \geq \frac{c^2}{\epsilon^2} = c^2 \cdot 10^{2n}$. 对于

$\epsilon = 10^0, 10^{-1}, 10^{-2}, \dots, 10^{-n+1}$, 必要迭代的总量为:

$$c^2(10^0 + 10^2 + 10^4 + \dots + 10^{2n-2}) = c^2 \cdot \frac{100(10^{2n} - 1)}{99} \approx c^2 \cdot 10^{2n}.$$

收敛速度和复杂性界限

2. **线性收敛速率**线性收敛速率是 **迭代次数的指数函数**. 例如,

$$r_k \leq c(1 - q)^k.$$

相应的复杂性上界是 $\frac{1}{q}(\ln c + \ln \frac{1}{\epsilon})$.

- ▶ 这个速度很快, 每计算解一个新的有效数字仅需要恒定的计算量
- ▶ 而且复杂度上界对常数 c 的依赖性较弱.

对于 $\epsilon = 10^{-n}$, 必要的复杂性界限 $O(\ln \frac{1}{\epsilon}) = O(n \ln 10)$. 当 ϵ 从 10^{-n} 下降到 10^{-n-1} , 复杂性界限增加了 常数 $\cdot \ln 10$.

收敛速度和复杂性界限

收敛速度和算法复杂度的界互为反函数.

3. **二次速率**二次速率是 **迭代次数的双指数函数** 比如,

$$r_{k+1} \leq cr_k^2.$$

相应的复杂度估计是精度的二次对数: $\ln \ln \frac{1}{\epsilon}$.

- ▶ 这个速度非常快, 算法每迭代一次都会使解中正确的有效数字的数量翻倍。
- ▶ 常数 c 主要对算法的起始点起作用 ($cr_k < 1$).

纲要

1. 牛顿法

2. 非线性优化中的一阶方法

2.1 梯度法和牛顿法：有什么不同？

3. 拟牛顿法

梯度法和牛顿法：有什么不同？

在最简单的最小化问题中找到局部最小值的两种局部方法。

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

其中 $f \in C_L^{2,2}(R^n)$.

梯度下降法:

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

牛顿法:

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

问题：它们的局部收敛差异的原因是什么？

梯度法具有线性速率

而牛顿法二次收敛.

分析

- ▶ 在梯度方法中，搜索方向是负梯度方向，
- ▶ 在牛顿方法中，搜索方向是负梯度方向左乘 Hessian 矩阵的逆矩阵。
- ▶ 接下来我们从另一个角度来推导这些搜索方向。

梯度法分析：线性近似

对于 $\bar{x} \in R^n$. 考虑函数 $f(x)$ 的近似:

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

其中 $h > 0$. $x_1^* \stackrel{\text{def}}{=} \operatorname{argmin} \phi_1(x)$:

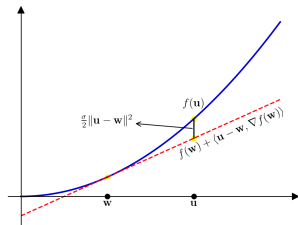
$$\phi_1'(x_1^*) = f'(\bar{x}) + \frac{1}{h}(x_1^* - \bar{x}) = 0.$$

得到, $x_1^* = \bar{x} - hf'(\bar{x})$. ← 这正是梯度下降法的迭代公式

若 $h \in (0, \frac{1}{L}]$, 则函数 $\phi_1(x)$ 是关于 $f(x)$ 全局上近似:

$$f(x) \leq \phi_1(x), \quad \forall x \in R^n,$$

(见引理 1.2.3). 这也是梯度下降法全局收敛的原因。



牛顿法分析：近似

考虑函数 $f(x)$ 的二次近似:

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

→ 该二次函数的最小值点是

$$x_2^* = \bar{x} - [f''(\bar{x})]^{-1} f'(\bar{x}),$$

这正是牛顿法的迭代公式.

牛顿法分析：近似

考虑函数 $f(x)$ 的二次近似:

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

→ 该二次函数的最小值点是

$$x_2^* = \bar{x} - [f''(\bar{x})]^{-1} f'(\bar{x}),$$

这正是牛顿法的迭代公式.

另一种解释 I

- ▶ 非线性函数 $f(x)$ 矩阵的梯度和 Hessian 是关于标准欧几里得内积在 R^n 定义的:

$$\langle x, y \rangle = \sum_{i=1}^n x_i y_i, \quad x, y \in R^n, \quad \|x\| = \langle x, x \rangle^{1/2}.$$

实际上, 梯度的定义是

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

- ▶ 现在介绍一种新的内积. 假设矩阵 A 为 $n \times n$ 阶正定矩阵. 对于 $x, y \in R^n$ 记

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

另一种解释 II

- ▶ 函数 $\|x\|_A$ 是 R^n 上的一个范数. 这个新范数与欧式范数等价:

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

其中 $\lambda_n(A)$ 和 $\lambda_1(A)$ 是矩阵 A 的最小特征值和最大特征值.

使用新的内积计算的梯度和 Hessian 矩阵发生了改变:

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

$\Rightarrow f'_A(x) \triangleq A^{-1}f'(x)$ 是新的梯度, $f''_A \triangleq A^{-1}f''(x)$ 是新的 Hessian 矩阵.

牛顿法的下降方向可以视为 $A = f''(x)$ 定义的度量下的梯度方向.

- ▶ 相对于 $A = f''(x)$ 计算的 $f(x)$ 在 x 处的 Hessian 矩阵为单位矩阵 I_n .

示例：二次函数牛顿法的一步收敛

例 4 (示例)

考虑二次函数

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

其中 $A = A^T \succ 0$. 因此 $f'(x) = Ax + a$, $f''(x) = A$ 令

$$f'(x^*) = Ax^* + a = 0$$

得 $x^* = -A^{-1}a$. 计算牛顿方向 $x \in R^n$:

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

\Rightarrow 对于任意的 $x \in R^n$, $x - d_N(x) = -A^{-1}a = x^*$. \Rightarrow 对于二次函数, 牛顿法迭代一步即可收敛.

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

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

纲要

1. 牛顿法

2. 非线性优化中的一阶方法

2.1 梯度法和牛顿法：有什么不同？

3. 拟牛顿法

拟牛顿法

可以尝试使用函数 $f(x)$ 的其他近似方法, 这些方法比 $\phi_1(x)$ 近似程度更好, 而且比 $\phi_2(x)$ 的计算代价更少.

设 G 为一个正定的 $n \times n$ 阶矩阵. 有

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

为了计算其最小值点, 令

$$\phi'_G(x) = f'(\bar{x}) + G(x_G^* - \bar{x}) = 0,$$

我们得到

$$x_G^* = \bar{x} - G^{-1}f'(\bar{x}). \quad (4)$$

若序列满足

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

(或 $\{H_k\} : H_k \equiv G_k^{-1} \rightarrow [f''(x^*)]^{-1}$, 称为变尺度方法或拟牛顿方法. 在这些方法中, 生成序列的过程中只涉及梯度 $\{G_k\}$ 或 $\{H_k\}$).

更新规则 (4): $x_G^* = \bar{x} - G^{-1}f'(\bar{x})$, 在优化算法中非常常见.

变尺度法

变尺度法

0. 设 $x_0 \in R^n$. 令 $H_0 = I_n$. 计算 $f(x_0)$ 和 $f'(x_0)$.
1. 第 k 次迭代 ($k \geq 0$).
 - (a) 设 $p_k = H_k f'(x_k)$.
 - (b) $x_{k+1} = x_k - h_k p_k$
 - (c) 计算 $f(x_{k+1})$ and $f'(x_{k+1})$.
 - (d) 更新矩阵 $H_k : H_k \rightarrow H_{k+1}$.

- ▶ 不同的变尺度法区别: 步骤的实施 1(d)
- ▶ 步骤 1(d): 更新 $H_k \leftarrow$ 梯度 + 函数值

拟牛顿公式: 如何构造 H_{k+1} 以近似 $f''(x_{k+1})^{-1}$?

假设 H_k 以近似 $f''(x_k)^{-1}$

- Motivation 1. 函数 $f(x)$ 在 $x = x_{k+1}$ 处的二阶近似:

$$f(x) \approx f(x_{k+1}) + \langle f'(x_{k+1}), x - x_{k+1} \rangle + \frac{1}{2}(x - x_{k+1})^T f''(x_{k+1})(x - x_{k+1})$$

$$\rightarrow f(x) \approx f(x_{k+1}) + f'(x_{k+1})(x - x_{k+1})$$

$$\text{令 } x = x_k, f'(x_{k+1}) \approx H_{k+1} \rightarrow H_{k+1}(f'(x_{k+1}) - f'(x_k)) = x_{k+1} - x_k$$

- Motivation 2. 二次函数的性质:

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

$$\Rightarrow \text{对于任意 } x, y \in R^n, f'(x) - f'(y) = A(x - y).$$

拟牛顿公式

选择 H_{k+1} 满足:

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

拟牛顿公式

例 5 (示例)

记

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

一些经典的拟牛顿法:

1. Rank-one correlation 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).

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

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

从计算角度来看, BFGS 被认为是最稳定的方案.

拟牛顿法另一解释

拟牛顿法与梯度下降法有相似迭代格式：

$$x_{k+1} = x_k + \alpha_k d_k \quad (5)$$

$$d_k = -D_k \nabla f(x_k). \quad (6)$$

其中 D_k 是正定矩阵。可以不断校正矩阵 D_k ，使方向 d_k 趋近于牛顿方向。

矩阵 D_k 的更新公式

在最流行的一类拟牛顿算法中，矩阵 D_{k+1} 是由 δ_k 和 γ_k 通过如下公式计算得到的

$$D_{k+1} = D_k + \frac{\delta_k \delta_k^T}{\delta_k^T \gamma_k} - \frac{D_k \gamma_k \gamma_k^T D_k}{\gamma_k^T D_k \gamma_k} + \xi_k \tau_k v_k v_k^T \quad (7)$$

其中

$$v_k = \frac{\delta_k}{\delta_k^T \gamma_k} - \frac{D_k \gamma_k}{\tau_k}$$
$$\tau_k = \gamma_k^T D_k \gamma_k$$

标量 ξ_k 满足：对于任意的 k ， $0 \leq \xi_k \leq 1$, D_0 是任意正定矩阵。

如果对于所有的 k ， $\xi_k = 0$ ，我们得到 Davidon-Fletcher-Powell scheme (DFP) 方法。

如果对于所有的 k ， $\xi_k = 1$ ，我们得到

Broyden-Fletcher-Goldfarb-Shanno scheme (BFGS) 方法。

定理 6 (命题)

如果 D_k 是正定的且步长 α_k 满足:

$$\nabla f(x_k)' d_k < \nabla f(x_{k+1})' d_k \quad (8)$$

那么由 (7) 式给定的 D_k 是正定的。

注意: 如果 x_k 不是驻点, 那么有 $\nabla f(x_k)' d_k < 0$, 因此为了满足条件 (8), 执行线搜索得到满足以下条件的点 x_{k+1} 就足够了。

$$|\nabla f(x_k)' d_k| < |\nabla f(x_{k+1})' d_k|$$

特别的, 如果 α_k 是由最小化线搜索规则确定的, 有 $\nabla f(x_k)' d_k = 0$ 且 (8) 式成立。

定理 7 (命题)

设 x_k, d_k 和 D_k 是拟牛顿算法 (5) – (7) 式生成的序列, 将其应用于正定二次函数的极小化

$$f(x) = \frac{1}{2} x' Q x - b' x$$

其中 α_k 由下面的公式决定

$$f(x_k + \alpha_k d_k) = \min_{\alpha} f(x_k + \alpha d_k)$$

假设向量 x_0, \dots, x_{n-1} 是迭代产生的最优解, 则有

1. 向量 d_0, \dots, d_{n-1} 是 Q 共轭的
2. 有 $D_n = Q^{-1}$

► **超线性局部收敛速度**

在严格最小值附近，有超线性收敛速度：对于任意 $x_0 \in R^n$ 存在 N 对于所有的 $k \geq N$ 有

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

(证明过程很长)

► **需要存贮 $n \times n$ 矩阵**

拟牛顿法需要存储和更新对称的 $n \times n$ 矩阵 \rightarrow 需要 $O(n^2)$ 辅助算术运算 (变尺度方法的主要缺点) \rightarrow 共轭梯度法

作业 I

1. (选作) 基于牛顿法设计求 $a^{1/3}$ 的快速迭代算法, 其中 $a > 0$.
2. 编程题: 第 1 部分第 1 节中, 有练习题: 自行选取二分类数据集, 编程实现梯度下降法求解 Logistic 回归模型, 并对数据集进行分类, 计算分类准确率。请在此基础上, 分别使用 (阻尼) 牛顿法和 BFGS 方法确定每一次迭代的步长。
3. 基于以下方法编程求解 $\min_x f(x)$ (1) 梯度下降法; (2) (阻尼) 牛顿法; (3) BFGS 方法 (选作)

3.1 $f(x) = \frac{1}{2}x^T A x - b^T x,$

其中 $A = \text{diag}(20, 10, 2, 1), \quad b = (1, 1, 1, 1)^T.$

初始点 $x_0 = (0, 0, 0, 0)^T$; 最佳解决方案 $x^* = (0.05, 0.1, 0.5, 1)^T$; 最优值 $f^* = 0.8250$.

3.2 $f = \sum_{i=1}^{N/2} \left[(x_{2i} - x_{2i-1}^2)^2 + (1 - x_{2i-1})^2 \right]$

初始点 $x_0 = (-1.2, 1, -1.2, 1, \dots, -1.2, 1)^T$; 最优解 $x^* = (1, 1, \dots, 1)^T$; 最优值 $f^* = 0$.

作业 II

3.3 (选作) [扩展 Dixon 函数]

$$f(x) = \sum_{i=1}^{n/10} \left[(1 - x_{10i-9})^2 + (1 - x_{10i})^2 + \sum_{j=10i-9}^{10i-1} (x_j^2 - x_{j+1})^2 \right]$$

初始点 $x_0 = (-2, -2, \dots, -2)^T$; 最佳解决方案 $x^* = (1, 1, \dots, 1)^T$;
最优值 $f^* = 0$.