最优化方法
Optimization Theory

龚舒凯
中国人民大学
Renmin University of China
School of Applied Economics/ School of Statistics
shukai_gong@ruc.edu.cn

目录

1 引论

1.1 最优化问题分类

设目标函数为 f，约束函数为 c_i

1. 根据变量取值
   (a) 连续最优化问题，$\mathbf{x} \in \mathbb{R}$
   (b) 离散最优化问题，$\mathbf{x} \in \mathbb{Z}$
2. 根据光滑性
   (a) 光滑最优化问题：f, c_i 均连续可微
   (b) 非光滑最优化问题：f 或 c_i 非光滑
3. 根据线性性质
   (a) 线性规划问题：f, c_i 都是关于决策变量的线性函数
   (b) 二次规划问题：f 是关于决策变量的二次函数，c_i 是决策变量的线性函数
   (c) 非线性最优化问题：f 或 c_i 中有一个函数关于决策变量是非线性的

1.2 凸集与凸函数

凸集

设集合 $C \subset \mathbb{R}^n$，若 $\forall x, y \in C$，成立

$$\theta x + (1-\theta)y \in C$$

则 C 为凸集。

凸集的几何意义：如果$x, y \in C$，则x, y的连线上的所有点都属于C。

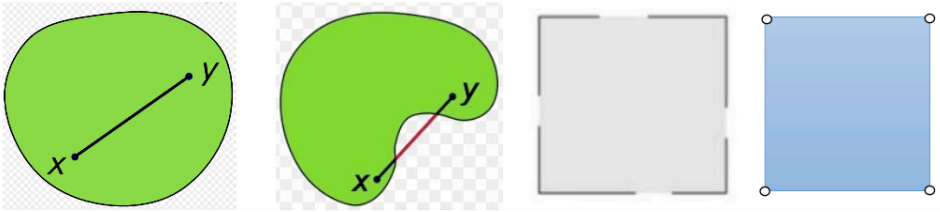


图 1: 凸集的几何意义

常见的凸集包括：$\varnothing$，$\mathbb{R}^n$，超平面 $H = \{\mathbf{x} \in \mathbb{R}^n | \mathbf{G}^\top \mathbf{x} = b\}$，半空间 $H^+ = \{\mathbf{x} \in \mathbb{R}^n | \mathbf{G}^\top \mathbf{x} \geq b\}, H^- = \{\mathbf{x} \in \mathbb{R}^n | \mathbf{G}^\top \mathbf{x} \leq b\}$

凸集计算的封闭性

凸集的加法、数乘、交运算仍然是凸集。

上图 (Epigraph)

设集合 $C \subset \mathbb{R}^n \neq \varnothing$，定义在 C 上的函数 f 的上图定义为

$$\text{epi}(f) = \{(\mathbf{x}, t) | \mathbf{x} \in C, t \in R, f(\mathbf{x}) \leq t\} \subset \mathbb{R}^{n+1}$$

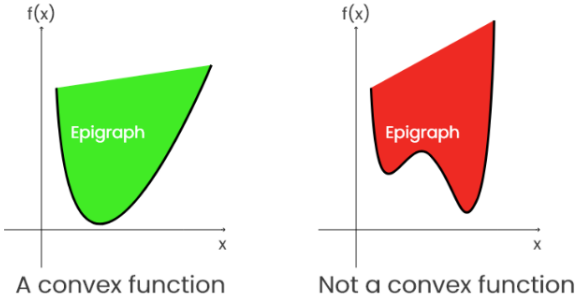


图 2: 上图的几何意义：一个函数的上半区域

凸 (Convex) 函数

对于定义在 $D \subset \mathbb{R}^n$ 的函数 f，若 $\forall \mathbf{x}, \mathbf{y} \in D, \lambda \in [0, 1]$，都有

$$f(\lambda \mathbf{x} + (1 - \lambda)\mathbf{y}) \leq \lambda f(\mathbf{x}) + (1 - \lambda) f(\mathbf{y})$$

则 f 是凸函数。

凸函数的性质

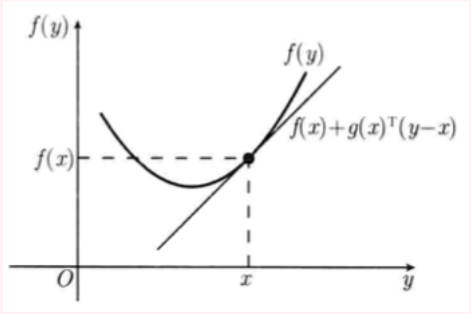
1. 对于定义在 $D \subset \mathbb{R}^n$ 的凸函数 f 以及 $k \geq 0$，kf 也是 D 上的凸函数。
2. 对于定义在 $D \subset \mathbb{R}^n$ 的凸函数 f_1, f_2 以及 $\lambda \geq 0, \mu \geq 0$，$\lambda f_1 + \mu f_2$ 也是 D 上的凸函数。
3. 对于定义在 $D \subset \mathbb{R}^n$ 的凸函数 f 以及 $\beta \in \mathbb{R}$，水平集 $S(f, \beta) = \{\mathbf{x} | \mathbf{x} \in D, f(\mathbf{x}) \leq \beta\}$ 是凸集

凸函数判定定理

1. **一阶判定条件**：对于定义在 $D \subset \mathbb{R}^n$ 的可微函数 f，

$$f\text{在}D\text{上为凸函数} \iff \forall \mathbf{x}, \mathbf{y} \in D, f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x})$$

$$f\text{在}D\text{上为严格凸函数} \iff \forall \mathbf{x} \neq \mathbf{y} \in D, f(\mathbf{y}) > f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x})$$



2. **二阶判定条件**：对于定义在开凸集 $D \subset \mathbb{R}^n$ 上的二阶可微函数 f，设其，

$$f\text{在}D\text{上为凸函数} \iff \forall \mathbf{x} \in D, \mathbf{H}_f(\mathbf{x})\text{半正定}$$

$$f\text{在}D\text{上为严格凸函数} \Leftarrow \forall \mathbf{x} \in D, \mathbf{H}_f(\mathbf{x})\text{正定}$$

其中 $\mathbf{H}_f(\mathbf{x})$ 是 $f(\mathbf{x})$ 的 Hessian 矩阵：$\mathbf{H}_f(\mathbf{x}) = \nabla^2 f(\mathbf{x})$。

3. 设集合 $C \subset \mathbb{R}^n$ 为非空凸集，则

$$\text{函数}f : C \mapsto \mathbb{R}\text{是凸函数} \iff \text{epi}(f)\text{是凸集}$$

[注]：二阶判定条件中，严格凸函数的判定条件**是一个必要条件，充分性不成立**，例如 $f(x) = x^4$

2 无约束优化基础

2.1 基础概念

- **全局最优解**：若 $\forall \mathbf{x} \in \mathbb{R}^n$，$f(x) \geq f(x^*)$，则 x^* 是问题的全局最优解。
- **局部最优解**：对 $\forall \mathbf{x}^* \in \mathbb{R}^n$，存在 $\epsilon > 0$，使得 $\forall \mathbf{x} \in \mathbb{R}^n$，当 $\|\mathbf{x} - \mathbf{x}^*\| < \epsilon$，$f(x) \geq f(x^*)$，则 x^* 是问题的局部最优解。
- **一阶方向导数**：若 $f \in C^1$，则 $\mathbf{x}$ 处沿方向 $\mathbf{d} \neq \mathbf{0}$ 的一阶方向导数可以表示为 $\dfrac{\partial f(\mathbf{x})}{\partial \mathbf{d}} = \dfrac{\nabla f(\mathbf{x})^\top \mathbf{d}}{\|\mathbf{d}\|}$
- **二阶方向导数**：若 $f \in C^2$，则 $\mathbf{x}$ 处沿方向 $\mathbf{d} \neq \mathbf{0}$ 的一阶方向导数可以表示为 $\dfrac{\partial f^2(\mathbf{x})}{\partial \mathbf{d}^2} = \dfrac{\mathbf{d}^\top \nabla^2 f(\mathbf{x}) \mathbf{d}}{\|\mathbf{d}\|}$

2.2 最优性条件

最优性条件

- **一阶必要条件**：设 $f \in C^1$，$\nabla f(\mathbf{x}^*) = 0 \Leftarrow \mathbf{x}^*$ 为 $f(x)$ 的一个局部极小点。
- **一阶必要条件 2**：设 $f \in C^1$，$\forall \mathbf{d} \in \mathbb{R}^n, \dfrac{\partial f(\mathbf{x}^*)}{\partial \mathbf{d}} = 0 \Leftarrow \mathbf{x}^*$ 为 $f(x)$ 的一个局部极小点。
- **二阶必要条件**：设 $f \in C^2$，$\nabla^2 f(\mathbf{x}^*)$ 半正定 $\Leftarrow \mathbf{x}^*$ 为 $f(x)$ 的一个局部极小点。
- **二阶充分条件**：设 $f \in C^2$，$\forall \mathbf{d} \in \mathbb{R}^n$，$\dfrac{\partial f(\mathbf{x}^*)}{\partial \mathbf{d}} = 0, \dfrac{\partial f^2(\mathbf{x}^*)}{\partial \mathbf{d}^2} > 0 \Rightarrow \mathbf{x}^*$ 为 $f(x)$ 的一个局部极小点。

2.3 最优化方法准则和常用性质

最优化方向

为了求极小值点，我们希望迭代过程满足 $f(\mathbf{x}_{k+1}) = f(\mathbf{x}_k + \alpha_k \mathbf{d}_k) < f(\mathbf{x}_k)$。对不等式左侧作 Taylor 展开得

$$f(\mathbf{x}_k + \alpha_k \mathbf{d}_k) = f(\mathbf{x}_k) + \alpha_k \nabla f(\mathbf{x}_k)^\top \mathbf{d}_k + o(\|\alpha_k \mathbf{d}_k\|) < f(\mathbf{x}_k)$$

这天然要求 $\nabla f(\mathbf{x}_k)^\top \mathbf{d}_k < 0$，即梯度方向与迭代方向成” 钝角”。

终止准则

有以下终止准则可采取

1. $\forall \epsilon > 0, \|\nabla f(\mathbf{x}_k)\| \leq \epsilon$
2. $\forall \epsilon > 0, \|\mathbf{x}_k - \mathbf{x}_{k+1}\| \leq \epsilon$ 或 $f(\mathbf{x}_k) - f(\mathbf{x}_{k+1}) \leq \epsilon$

[注]：两种终止准则存在的缺陷为

- **准则 1:** 对于极小点附近较为陡峭的函数，迭代难以停止

- **准则 2:** 尽管迭代点/迭代点函数值变化足够小，但不能保证其距离最优点距离足够小，即不能保证 $\|\mathbf{x}_k - \mathbf{x}^*\| \leq \epsilon$ 或 $f(\mathbf{x}_k) - f(\mathbf{x}^*) \leq \epsilon$

二次终止性

对于任意正定二次函数 $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top \mathbf{G}\mathbf{x} + \mathbf{b}^\top \mathbf{x} + c$，其中 $\mathbf{G}$ 为对称正定矩阵，若从任意初始点出发，收敛算法经过有限次迭代可以求得最小点，则称收敛算法具有二次终止性。

使用正定二次函数的原因是：一般函数极小点附近**可以用正定二次函数来近似**，故能否有效求得正定二次函数的极小点是检验一个算法好坏的标准之一。

收敛性

若算法产生的点列 $\{\mathbf{x}_k\}$ 在某种范数 $\|\cdot\|$ 意义下满足 $\lim\limits_{k\to\infty} \|\mathbf{x}_k - \mathbf{x}^*\| = 0$ 称这个算法是收敛的。

- **全局收敛:** 从任意初始点出发，$\{\mathbf{x}_k\}$ 都能收敛到 $\mathbf{x}^*$
- **局部收敛:** 仅当初始点与 $\mathbf{x}^*$ 充分接近时，$\{\mathbf{x}_k\}$ 都能收敛到 $\mathbf{x}^*$

收敛速度

- 若 $\lim\limits_{k\to\infty} \dfrac{\|\mathbf{x}_{k+1} - \mathbf{x}^*\|}{\|\mathbf{x}_k - \mathbf{x}^*\|} = a$，那么收敛算法
  - 当 $0 < a < 1$ 时：**线性收敛**
  - 当 $a = 0$ 时：**超线性收敛**
- 若 $\lim\limits_{k\to\infty} \dfrac{\|\mathbf{x}_{k+1} - \mathbf{x}^*\|}{\|\mathbf{x}_k - \mathbf{x}^*\|^2} = a, \forall a \in \mathbb{R}$，那么收敛算法二**阶收敛**

3 线搜索方法与信赖域方法

最优化方法分为两种结构：线搜索方法、信赖域方法

- **线搜索方法**：先确定方向 $\mathbf{d}_k$，再确定步长 α_k
- **信赖域方法**：先确定步长最大范围 Δ_k，再同时确定方向 $\mathbf{d}_k$ 和步长 α_k

在 $\mathbf{x}_k$ 位置，假定通过某种算法已经得到下降方向 $\mathbf{d}_k$，求步长 α_k 的问题称为线搜索问题。所以我们关心两个关键点：

1. 确定步长的准则
2. 如何求出满足准则的步长

处于便捷的考虑，以下记 $\mathbf{g}_k = \nabla f(\mathbf{x}_k)$，$\mathbf{G}_k = \nabla^2 f(\mathbf{x}_k)$.

3.1 精确线搜索准则

精确线搜索准则

在 $\mathbf{x}_k$ 位置，当迭代方向 $\mathbf{d}_k$ 已知时，令

$$\alpha_k = \min_{\alpha} f(\mathbf{x}_k + \alpha \mathbf{d}_k)$$

不妨记 $\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha \mathbf{d}_k$，则根据链式法则可求得一阶条件

$$\mathbf{g}_{k+1}^\top \mathbf{d}_k = 0$$

也就是说，下一迭代点 $\mathbf{x}_{k+1}$ 处的梯度必须与迭代方向正交。

[注]：然而，要从这一条件解出 α_k 是非常困难的。

正定二次函数的精确线搜索步长

对于一般的正定二次函数 $\phi(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top \mathbf{G}\mathbf{x} - \mathbf{b}^\top \mathbf{x}$，如果按照如下的方式进行迭代：

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{d}_k$$

则精确线搜索的步长 $\alpha_k = -\dfrac{\mathbf{g}_k^\top \mathbf{d}_k}{\mathbf{d}_k^\top \mathbf{G}\mathbf{d}_k}$。

证明. 注意到

$$\begin{aligned}\phi(\mathbf{x}_k + \alpha_k \mathbf{d}_k) &= \frac{1}{2}(\mathbf{x}_k + \alpha_k \mathbf{d}_k)^\top \mathbf{G}(\mathbf{x}_k + \alpha_k \mathbf{d}_k) - \mathbf{b}^\top(\mathbf{x}_k + \alpha_k \mathbf{d}_k) \\ &= \frac{1}{2}\mathbf{x}_k^\top \mathbf{G}\mathbf{x}_k + \alpha_k \mathbf{d}_k^\top \mathbf{G}\mathbf{x}_k + \frac{1}{2}\alpha_k^2 \mathbf{d}_k^\top \mathbf{G}\mathbf{d}_k - \mathbf{b}^\top(\mathbf{x}_k + \alpha_k \mathbf{d}_k)\end{aligned}$$

关于 α_k 求导得

$$\frac{\partial \phi(\mathbf{x}_k+\alpha_k\mathbf{d}_k)}{\partial \alpha_k} = \mathbf{d}_k^\top \mathbf{G}\mathbf{x}_k + \alpha_k \mathbf{d}_k^\top \mathbf{G}\mathbf{d}_k - \mathbf{b}^\top \mathbf{d}_k = 0$$
$$\Rightarrow \alpha_k = \frac{\mathbf{b}^\top \mathbf{d}_k - \mathbf{x}_k^\top \mathbf{G}\mathbf{d}_k}{\mathbf{d}_k^\top \mathbf{G}\mathbf{d}_k} = -\frac{(\mathbf{G}\mathbf{x}_k - \mathbf{b})^\top \mathbf{d}_k}{\mathbf{d}_k^\top \mathbf{G}\mathbf{d}_k} = -\frac{\mathbf{g}_k^\top \mathbf{d}_k}{\mathbf{d}_k^\top \mathbf{G}\mathbf{d}_k}$$

□

3.2 非精确线搜索准则

在 $\mathbf{x}_k$ 位置，我们放宽条件，只需要选取步长 α_k 使得函数值充分下降: $f(\mathbf{x}_k+\alpha_k\mathbf{d}_k) < f(\mathbf{x}_k)$ 即可。由于我们关心步长 α_k 的选取，不妨记 $\phi(\alpha) = f(\mathbf{x}_k+\alpha_k\mathbf{d}_k)$，那么可将其放至二维坐标系可视化：如图3所示，满足准则的 $\alpha \in [0,\beta_1],[\beta_2,\beta_3]$

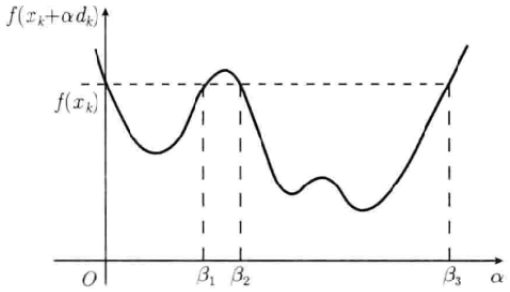


图 3: $\phi(\alpha) = f(\mathbf{x}_k+\alpha_k\mathbf{d}_k)$ 的几何意义

Goldstein 准则

选取 $\alpha > 0$，使得

$$f(\mathbf{x}_k+\alpha\mathbf{d}_k) \geq f(\mathbf{x}_k) + (1-\rho)\mathbf{g}_k^\top \mathbf{d}_k\alpha$$
$$f(\mathbf{x}_k+\alpha\mathbf{d}_k) \leq f(\mathbf{x}_k) + \rho\mathbf{g}_k^\top \mathbf{d}_k\alpha$$

其中 $\rho \in (0,0.5)$.

直观上而言，由于 $\nabla f(\mathbf{x}_k)^\top \mathbf{d}_k < 0$，Goldstein 准则用两条关于 α 斜率为负的直线缩小 α 的范围到 $[\beta_7,\beta_4],[\beta_5,\beta_6]$。

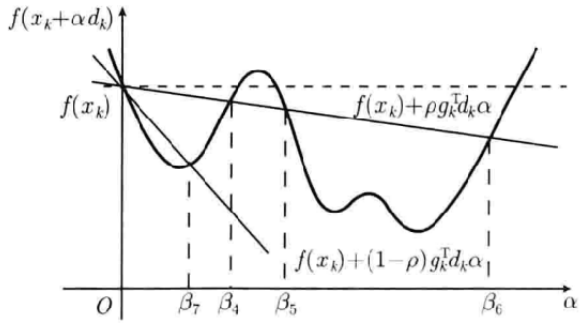


图 4: Goldstein 准则

Wolfe 准则

选取 $\alpha > 0$，使得

$$f(\mathbf{x}_k + \alpha \mathbf{d}_k) \leq f(\mathbf{x}_k) + \rho \mathbf{g}_k^\top \mathbf{d}_k \alpha$$
$$g(\mathbf{x}_k + \alpha \mathbf{d}_k)^\top \mathbf{d}_k > \sigma \mathbf{g}_k^\top \mathbf{d}_k$$

其中 $0 < \rho < \sigma < 1$。

直观上而言，除了用一条关于 α 斜率为负的直线缩小 α 范围外，还要求更新后 $\mathbf{x}_k + \alpha \mathbf{d}_k$ 处梯度必须大于原 $\mathbf{x}_k$ 处梯度的 σ 倍。

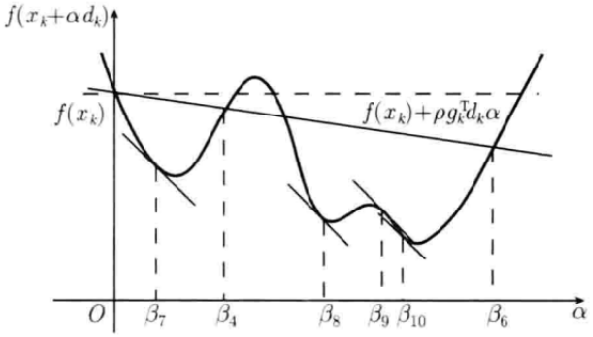


图 5: Wolfe 准则

强 Wolfe 准则

选取 $\alpha > 0$，使得

$$f(\mathbf{x}_k + \alpha \mathbf{d}_k) \leq f(\mathbf{x}_k) + \rho \mathbf{g}_k^\top \mathbf{d}_k \alpha$$
$$|g(\mathbf{x}_k + \alpha \mathbf{d}_k)^\top \mathbf{d}_k| < -\sigma \mathbf{g}_k^\top \mathbf{d}_k$$

其中 $0 < \rho < \sigma < 1$.

相比 Wolfe 准则，强 Wolfe 准则要求更新后 $\mathbf{x}_k + \alpha \mathbf{d}_k$ 处梯度的必须介于原 $\mathbf{x}_k$ 处梯度的 $-\sigma$ 倍到 σ 倍之间。

非精确性线性搜索步长 α_k 的存在性定理

设

- $f(\mathbf{x}_k + \alpha \mathbf{d}_k)$ 在 $\alpha > 0$ 时有下界
- $\mathbf{g}_k^\top \mathbf{d}_k < 0$

则必存在 $\alpha_k > 0$，在 $\mathbf{x}_k + \alpha_k \mathbf{d}_k$ 处满足 Goldstein/Wolfe 准则。

证明. (作业题) 由于 $\phi(\alpha) = f(\mathbf{x}_k + \alpha \mathbf{d}_k)$ 在 $\alpha > 0$ 有下界，而直线 $l(\alpha) = f(\mathbf{x}_k) + \rho \mathbf{g}_k^\top \mathbf{d}_k \alpha$ 在 $\alpha > 0$ 无下界，

因此 $\phi(\alpha)$ 与 $l(\alpha)$ 在 $\alpha>0$ 处必有一交点。不妨记这一交点为 $\alpha'>0$，则

$$f(\mathbf{x}_k+\alpha'\mathbf{d}_k)=f(\mathbf{x}_k)+\alpha'\rho\mathbf{g}_k^\top\mathbf{d}_k$$

注意到 $\alpha=0$ 也是 $\phi(\alpha)$ 和 $l(\alpha)$ 的一个交点。根据中值定理，必存在 $\alpha''\in(0,\alpha')$，使得

$$f(\mathbf{x}_k+\alpha'\mathbf{d}_k)-f(\mathbf{x}_k)=\alpha' g(\mathbf{x}_k+\alpha''\mathbf{d}_k)^\top\mathbf{d}_k$$

对比上两式，容易发现

$$g(\mathbf{x}_k+\alpha''\mathbf{d}_k)^\top\mathbf{d}_k=\rho\mathbf{g}_k^\top\mathbf{d}_k>\sigma\mathbf{g}_k^\top\mathbf{d}_k$$

且在 $\alpha''\in(0,\alpha')$ 上，由于 α' 是 $\phi(\alpha)$ 与 $l(\alpha)$ 的第一个交点，因此

$$f(\mathbf{x}_k+\alpha''\mathbf{d}_k)\le f(\mathbf{x}_k)+\alpha''\rho\mathbf{g}_k^\top\mathbf{d}_k$$

这说明存在 $\alpha''>0$ 满足 Wolfe 准则。 □

非精确性线性搜索方法的收敛性定理

设在水平集 $\{\mathbf{x}|f(\mathbf{x})\le f(\mathbf{x}_0)\}$ 上，$f(\mathbf{x})$ 有下界，梯度 $g(\mathbf{x})$ 一致连续，且给定任意线搜索方法（精确/非精确），第 k 次迭代方向 $\mathbf{d}_k$ 与 $-\nabla f(\mathbf{x}_k)$ 的夹角一致有界：

$$0\le\theta_k\le\frac{\pi}{2}-\mu,\qquad\exists\mu>0,\forall k$$

若 Wolfe 准则（或 Goldstein 准则、强 Wolfe 准则、精确线搜索条件）对 $\forall k$ 都成立，则要么 $\exists N$, s.t. $\mathbf{g}_N=0$，要么 $\mathbf{g}_k\to 0,k\to\infty$

换言之，只要**负梯度和迭代方向**的夹角为锐角，那么线搜索方法必定收敛。

证明. **(用 Wolfe 准则，定理 2.8)** 设 $\forall k,\mathbf{g}_k\neq\mathbf{0}$，如果 $\mathbf{g}_k\not\to 0,k\to\infty$。不妨设对 $\mathbf{x}_{k+1}=\mathbf{x}_k+\alpha\mathbf{d}_k$，Wolfe 准则成立，即

$$\begin{aligned}&f(\mathbf{x}_{k+1})\le f(\mathbf{x}_k)+\rho\mathbf{g}_k^\top\mathbf{d}_k\alpha_k\\ \Rightarrow\ &\frac{1}{\rho}[f(\mathbf{x}_k)-f(\mathbf{x}_{k+1})]\ge-\mathbf{g}_k^\top\mathbf{d}_k\alpha_k\triangleq-\mathbf{g}_k\mathbf{s}_k=\|\mathbf{g}_k\|\|\mathbf{s}_k\|\cos\theta_k\\ &\ge\|\mathbf{g}_k\|\|\mathbf{s}_k\|\cos(\frac{\pi}{2}-\mu)=\|\mathbf{g}_k\|\|\mathbf{s}_k\|\sin\mu\ge 0\end{aligned}$$

由于 $\{f(\mathbf{x}_k)\}$ 单调递减有下界，因此 $f(\mathbf{x}_k)-f(\mathbf{x}_{k+1})\to 0,k\to\infty$。若 $\mathbf{g}_k\not\to 0$，那么只能 $\|\mathbf{s}_k\|\to 0$。由 $g(\mathbf{x})$ 的一致连续性

$$\mathbf{g}_{k+1}^\top\mathbf{s}_k=g(\mathbf{x}_k+\mathbf{s}_k)^\top\mathbf{s}_k=\mathbf{g}_k^\top\mathbf{s}_k+o(\|\mathbf{s}_k\|)\Rightarrow\frac{\mathbf{g}_{k+1}^\top\mathbf{s}_k}{\mathbf{g}_k^\top\mathbf{s}_k}\to 1,k\to\infty$$

然而 Wolfe 准则要求 $\dfrac{\mathbf{g}_{k+1}^\top\mathbf{s}_k}{\mathbf{g}_k^\top\mathbf{s}_k}<\sigma<1$，矛盾！假设不成立，原命题得证。

(用精确线搜索准则，习题 2-12) 设 $\forall k,\mathbf{g}_k\neq\mathbf{0}$，如果 $\mathbf{g}_k\not\to 0,k\to\infty$。则存在子列 $\{k_j\}$ 以及正常数 $\delta>0$ 使得

$$\|\mathbf{g}_{k_j}\|\ge\delta$$

在精确线搜索准则下，Zoutendijk 条件成立，即

$$\sum_{k\geq 0}\|\mathbf{g}_k\|^2\cos^2\theta_k<\infty$$

然而由题意

$$\cos\theta_k\geq\cos(\frac{\pi}{2}-\mu)=\sin\mu$$

那么

$$\sum_{k\geq 0}\|\mathbf{g}_k\|^2\cos^2\theta_k\geq\sum_{j\geq 0}\|\mathbf{g}_{k_j}\|^2\cos^2\theta_{k_j}\geq\sum_{j\geq 0}\delta^2\sin^2\mu=\infty$$

这与 Zoutendijk 条件矛盾！假设不成立，原命题得证。 □

3.3 线搜索求步长

0.618 法

1. 确定初始区间 $[a,b]$，其包含 $\phi(\alpha)=f(\mathbf{x}_k+\alpha\mathbf{d}_k)$ 的极小点。令 $[a_0,b_0]=[a,b]$。
2. 在第 i 次迭代，若 $b_i-a_i<\epsilon$，则 $a^*=\dfrac{a_i+b_i}{2}$，停止线搜索
3. 否则继续迭代：
   (a) 取两个点 a_i^l,a_i^r 满足：$a_i^r-a_i=b_i-b_i^l=\tau(b_i-a_i)$，这里 $\tau=\dfrac{\sqrt{5}-1}{2}$
   (b) 如果 $\phi(a_i^l)<\phi(a_i^r)$，则 $a^*\in[a_i,a_i^r]$，取 $[a_{i+1},b_{i+1}]=[a_i,a_i^r]$
   (c) 如果 $\phi(a_i^l)\geq\phi(a_i^r)$，则 $a^*\in[b_i^l,b_i]$，取 $[a_{i+1},b_{i+1}]=[b_i^l,b_i]$

步骤 1 中**确定初始区间的方法称为进退法**：我们要的初始区间 $[a,b]$ 必须满足：$\phi(\alpha)$ 在 $\alpha\in[a,b]$ 上是单峰函数（$[a,a^*]$ 上 $\downarrow$，$[a^*,b]$ 上 $\uparrow$）。

直观上说：给定初始点 $\alpha_0>0$ 和初始步□ $\gamma_0>0$，向前一步 $a_0+\gamma_0$，如 $\phi(\alpha_0+\gamma_0)<\phi(\alpha_0)$，则加大 γ_0，继续向前，直到新一点的函数值较前一点的函数值增大了，否则从 α_0 起以 $-\gamma_0$ 为步□向相反方向搜索，其余过程相同，

至于为什么取 $\tau=\dfrac{\sqrt{5}-1}{2}$，我们希望第 i 次迭代时区间端点能够对齐，即 $a_i^l=a_{i+1}^r$ 或 $a_i^r=a_{i+1}^l$。以下图为例，要 a_0^l 和 a_1^l 对齐，有

$$a_0^l-a_0=a_1^r-a_1\Rightarrow(1-\tau)(b_0-a_0)=\tau\cdot\tau(b_0-a_0)\Rightarrow 1-\tau=\tau^2$$

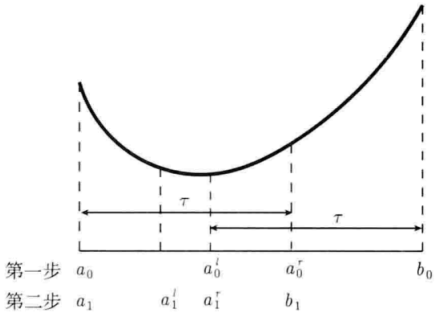


图 6: 0.618 法

多项式插值法

1. 构造近似 $\phi(\alpha)$ 的多项式函数：已知 $\phi(\alpha)$ 在 $m+1$ 个点不同处的函数值 $\phi(\alpha_i), i=0,\cdots,m$，我们要求近似 $\leq m$ 多项式 $p(\alpha)$ 满足

$$p(\alpha_i)=\phi(\alpha_i), \qquad i=0,\cdots,m$$

2. 求出 $p(\alpha)$ 的极小点（多项式函数容易求极小点），检验是否满足非精确线搜索准则（Goldstein, Wolfe, ...)。如果不满足，根据新的信息构造新的多项式函数，重复上一步骤。

3.4 信赖域方法

信赖域方法

1. 给定初始迭代点 $\mathbf{x}_0$，初始信赖域半径 $\Delta_k>0$，终止准则参数 $\epsilon>0$。
2. 在第 k 次迭代时，若满足终止准则，输出 $\mathbf{x}_k$，停止迭代。
3. 否则求解信赖域子问题：令 $q_k(\mathbf{d})=f(\mathbf{x}_k)+\mathbf{g}_k^\top\mathbf{d}+\frac{1}{2}\mathbf{d}^\top\mathbf{G}_k\mathbf{d}$ 为 $f(\mathbf{x}_k+\mathbf{d})$ 在 $\mathbf{x}_k$ 附近的 Taylor 展开，求解

$$\mathbf{d}_k=\min_{\mathbf{d}} q_k(\mathbf{d}) \qquad \text{s.t.}\|\mathbf{d}\|\leq\Delta_k$$

4. 计算比值 $\rho_k=\dfrac{f(\mathbf{x}_k)-f(\mathbf{x}_k+\mathbf{d}_k)}{q_k(\mathbf{0})-q_k(\mathbf{d}_k)}$ 来表现二次函数 $q_k(\mathbf{d}_k)$ 在 $\mathbf{x}_k+\mathbf{d}_k$ 附近近似目标函数 $f(\mathbf{x}_k)$ 的好坏程度：
   (a) 若 $\rho_k>0.75$，说明近似好，下一步增大 $\Delta_{k+1}:=2\Delta_k$
   (b) 若 $\rho_k<0.25$，说明近似差，下一步缩小 $\Delta_{k+1}=\frac{1}{4}\Delta_k$
   (c) 否则维持信赖域半径，$\Delta_{k+1}:=\Delta_k$
5. 若 $\rho_k\leq 0$，则迭代点不变：$\mathbf{x}_{k+1}:=\mathbf{x}_k$，缩小 Δ_k 重新求解信赖域问题，否则 $\mathbf{x}_{k+1}:=\mathbf{x}_k+\mathbf{d}_k$，返回第 2 步。

[注]：由于 $q_k(\mathbf{d})$ 是关于 $\mathbf{d}$ 的二次函数（碗形），从 $\mathbf{x}_k$ 到 $\mathbf{x}_k+\mathbf{d}_k$ 必然使得 q_k 函数值减小，$q_k(\mathbf{0})-q_k(\mathbf{d}_k)>0$。

如果 $f(\mathbf{x}_k) - f(\mathbf{x}_k + \mathbf{d}_k) < 0$，说明从 $\mathbf{x}_k$ 起走 $\mathbf{d}_k$ 的距离反而使得函数值“上升”了，这是因为我们将信赖域取得太大，让走 $\mathbf{d}_k$ 的距离“走过头了”极小点。因此，在这一情况下我们不应该进一步迭代 $\mathbf{x}_k$，而是缩小 Δ_k 以后重新求解信赖域问题。

关于信赖域子问题 $\mathbf{d}_k = \min_{\mathbf{d}} q_k(\mathbf{d}) \quad \text{s.t.} \|\mathbf{d}\| \le \Delta_k$ 的求解，可采用 **Dogleg 算法**：

Dogleg 算法

1. 给出 $\Delta_k > 0$
2. 计算 Newton 方向 (最优方向)$\mathbf{d}_k^N = -\mathbf{G}_k^{-1}\mathbf{g}_k$，若 $\|\mathbf{d}_k^N\| \le \Delta_k$，则 $\mathbf{d}_k = \mathbf{d}_k^N$，停止迭代。
3. 计算负梯度方向 (最速下降方向)$\mathbf{d}_k^{SD} = \dfrac{-\mathbf{g}_k^\top \mathbf{g}_k}{\mathbf{g}_k^\top \mathbf{G}_k \mathbf{g}_k}\mathbf{g}_k$，若 $\|d_k^{SD}\| \ge \Delta_k$，则 $\mathbf{d}_k = \Delta_k \cdot \dfrac{\mathbf{g}_k}{\|\mathbf{g}_k\|}$，停止迭代。
4. Doglet 点为 $\mathbf{d}_k = (1-\beta)\mathbf{d}_k^{SD} + \beta\mathbf{d}_k^N$，其中 β 应使得 $\|\mathbf{d}_k\| = \Delta_k$

如图7所示，Dogleg 点恰好夹在最速下降方向和 Newton 方向之间

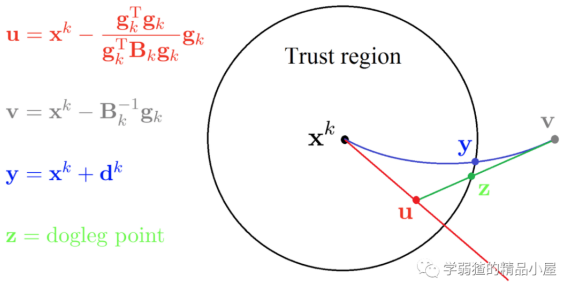


图 7: Dogleg 点

4 无约束优化

4.1 最速下降法

假设当前迭代点为 $\mathbf{x}_k$，我们要求 $\mathbf{x}_k$ 处 $f(\mathbf{x})$ 下降最快的方向，考虑其 Taylor 展开

$$f(\mathbf{x}_k+\mathbf{d}_k)=f(\mathbf{x}_k)+\mathbf{g}_k^\top\mathbf{d}_k+o(\|\mathbf{d}_k\|^2)$$

要让其下降最多，只需取负梯度方向 $\mathbf{d}_k=-\mathbf{g}_k$。又根据精确线搜索条件 $\mathbf{g}_{k+1}^\top\mathbf{d}_k=0$，因此 $\mathbf{d}_{k+1}^\top\mathbf{d}_k=0$，**即相邻两步的迭代方向正交**。由此，可得到最速下降算法：

最速下降法

1. 给定初始迭代点 $\mathbf{x}_0$，终止准则参数 $\epsilon>0$。
2. 在第 k 次迭代时，若满足终止准则，输出 $\mathbf{x}_k$，停止迭代。
3. $\mathbf{d}_k=-\nabla f(\mathbf{x}_k)$
4. 精确线搜索求 α_k（0.618 法）
5. $\mathbf{x}_{k+1}=\mathbf{x}_k+\alpha_k\mathbf{d}_k$，转第 2 步。

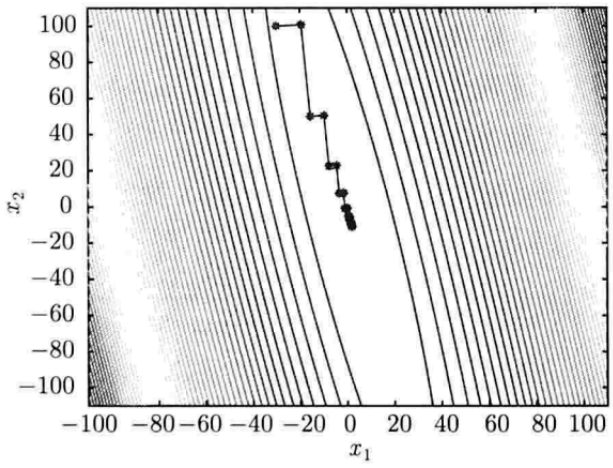


图 8: 最速下降法

最速下降法的收敛性

最速下降法是全局收敛的。

证明. 注意到 $\mathbf{d}_k$ 与 $-\mathbf{g}_k$ 同方向 (夹角为 0)，根据非精确线搜索的收敛性定理，最速下降法是全局收敛的。 □

最速下降法的收敛速度

速下降法是线性收敛的。具体收敛速度与 $\mathbf{G}_k$ 的最大/最小特征值有关。

证明. 首先定义正定对称矩阵 $\mathbf{G}\in\mathbb{R}^{n\times n}$ 度量下的范数与内积

$$(\mathbf{u}^\top\mathbf{v})_{\mathbf{G}}=\mathbf{u}^\top\mathbf{G}\mathbf{v},\|\mathbf{u}\|_{\mathbf{G}}=\mathbf{u}^\top\mathbf{G}\mathbf{u}$$

为了数学性质的优美,我们仅考虑正定二次函数 $f(\mathbf{x})=\frac{1}{2}\mathbf{x}^\top\mathbf{G}\mathbf{x}+\mathbf{b}^\top\mathbf{x}+a$ 最速下降方法的收敛速度,其梯度显然为 $\mathbf{g}_k=\nabla f(\mathbf{x})=\mathbf{G}\mathbf{x}+\mathbf{b}$。由于 $\mathbf{G}$ 对称正定,函数的极小点即为 $\mathbf{x}^*=-\mathbf{G}^{-1}\mathbf{b}$,极小值 $f(\mathbf{x}^*)=-\frac{1}{2}\mathbf{b}^\top\mathbf{G}^{-1}\mathbf{b}+a$。接下来求解最速下降法的步长。根据精确线搜索准则:

$$\begin{aligned}\alpha_k&=\min_\alpha f(\mathbf{x}_k+\alpha\mathbf{d}_k)=\min_\alpha f(\mathbf{x}_k-\alpha\mathbf{g}_k)\\&=\min\left(\frac{1}{2}(\mathbf{x}_k-\alpha\mathbf{g}_k)^\top\mathbf{G}(\mathbf{x}_k-\alpha\mathbf{g}_k)+\mathbf{b}^\top(\mathbf{x}_k-\alpha\mathbf{g}_k)+a\right)\\&=\min_\alpha\left(\frac{1}{2}\mathbf{x}_k^\top\mathbf{G}\mathbf{x}_k+\mathbf{b}^\top\mathbf{x}_k+a-\alpha\mathbf{g}_k^\top\mathbf{G}\mathbf{x}_k+\frac{1}{2}\alpha^2\mathbf{g}_k^\top\mathbf{G}\mathbf{g}_k-\alpha\mathbf{b}^\top\mathbf{g}_k\right)\\&=\min_\alpha\left(f(\mathbf{x}_k)-\alpha\mathbf{g}_k^\top(\mathbf{g}_k-\mathbf{b})+\frac{1}{2}\alpha^2\mathbf{g}_k^\top\mathbf{G}\mathbf{g}_k-\alpha\mathbf{b}^\top\mathbf{g}_k\right)\\&=\min_\alpha\left(f(\mathbf{x}_k)-\alpha\mathbf{g}_k^\top\mathbf{g}_k+\frac{1}{2}\alpha^2\mathbf{g}_k^\top\mathbf{G}\mathbf{g}_k\right)\end{aligned}$$

对目标式关于 α 求偏导容易得知 $\alpha_k=\dfrac{\mathbf{g}_k^\top\mathbf{g}_k}{\mathbf{g}_k^\top\mathbf{G}\mathbf{g}_k}$，进而 $\mathbf{x}_{k+1}=\mathbf{x}_k-\dfrac{\mathbf{g}_k^\top\mathbf{g}_k}{\mathbf{g}_k^\top\mathbf{G}\mathbf{g}_k}\mathbf{g}_k$，注意到在 $\mathbf{G}$ 范数下，

$$\frac{1}{2}\|\mathbf{x}_k-\mathbf{x}^*\|_{\mathbf{G}}=f(\mathbf{x}_k)-f(\mathbf{x}^*)$$

可以证明得到

$$\begin{aligned}\frac{\|\mathbf{x}_{k+1}-\mathbf{x}^*\|_{\mathbf{G}}}{\|\mathbf{x}_k-\mathbf{x}^*\|_{\mathbf{G}}}=\frac{f(\mathbf{x}_{k+1})-f(\mathbf{x}^*)}{f(\mathbf{x}_k)-f(\mathbf{x}^*)}&=\frac{f(\mathbf{x}_k)-\frac{1}{2}\frac{(\mathbf{g}_k^\top\mathbf{g}_k)^2}{\mathbf{g}_k^\top\mathbf{G}\mathbf{g}_k}-f(\mathbf{x}^*)}{f(\mathbf{x}_k)-f(\mathbf{x}^*)}\\&=1-\frac{\frac{1}{2}\frac{(\mathbf{g}_k^\top\mathbf{g}_k)^2}{\mathbf{g}_k^\top\mathbf{G}\mathbf{g}_k}}{\frac{1}{2}\mathbf{x}_k^\top\mathbf{G}\mathbf{x}_k+\mathbf{b}^\top\mathbf{x}_k+a-\frac{1}{2}\mathbf{b}^\top\mathbf{G}^{-1}\mathbf{b}-a}\\&=1-\frac{\frac{(\mathbf{g}_k^\top\mathbf{g}_k)^2}{\mathbf{g}_k^\top\mathbf{G}\mathbf{g}_k}}{(\mathbf{G}\mathbf{x}_k+\mathbf{b})^\top\mathbf{G}^{-1}(\mathbf{G}\mathbf{x}_k+\mathbf{b})}\\&=1-\frac{(\mathbf{g}_k^\top\mathbf{g}_k)^2}{(\mathbf{g}_k^\top\mathbf{G}\mathbf{g}_k)(\mathbf{g}_k^\top\mathbf{G}^{-1}\mathbf{g}_k)}\underbrace{\leq}_{\text{Kantorovich 不等式}}\left(\frac{\lambda_{\max}-\lambda_{\min}}{\lambda_{\max}+\lambda_{\min}}\right)^2\end{aligned}$$

其中 $\lambda_{\max},\lambda_{\min}$ 为 $\mathbf{G}$ 的最大、最小特征值。从上述不等式容易看出，最速下降方法是**线性收敛**的，但是具体收敛速度的多少与 $\mathbf{G}$ 的最大、最小特征值相关。 □

那么，直观上，$\mathbf{G}$ 的特征值 λ 和收敛速度为什么会有关联呢？**事实上，λ 是函数 $f(\mathbf{x})$ 在 $\mathbf{d}$ 上的二阶方向导数**，衡量了 $f(\mathbf{x})$ 等高线的“弯曲程度”。对实对称阵 $\mathbf{G}=\mathbf{Q}^\top\mathbf{\Lambda}\mathbf{Q}$ 作特征值分解，其中 $\mathbf{Q}$ 为正交矩阵。由于在特定方向 $\mathbf{d}$ 上，函数 $f(\mathbf{x})$ 的二阶方向导数为 $\mathbf{d}^\top\mathbf{G}\mathbf{d}$，当 $\mathbf{d}$ 为对应特征值为 λ_i 的 $\mathbf{G}$ 的特征向量时 (这里规定 $\|\mathbf{d}\|=1$)，二阶方向导数为

$$\mathbf{d}^\top\mathbf{G}\mathbf{d}=\mathbf{d}^\top\lambda_i\mathbf{d}=\lambda_i$$

换言之函数 $f(\mathbf{x})$ **在特征向量的方向上的二阶导数恰好为这个特征向量所对应的特征值!**

- 对于非特征向量的方向 $\mathbf{d}$，其二阶方向导数是所有特征值的加权平均，且与 $\mathbf{d}$ 夹角越小的特征向量有更大的权重。

因此，最大特征值确定最大方向二阶导数，最小特征值确定最小方向二阶导数。从等高线上理解更为直观，如图9所示，$\lambda_{\max}$ 对应的特征向量的方向等高线更密 (更陡峭)，$\lambda_{\min}$ 对应的特征向量的方向等高线更稀疏 (更平坦)。

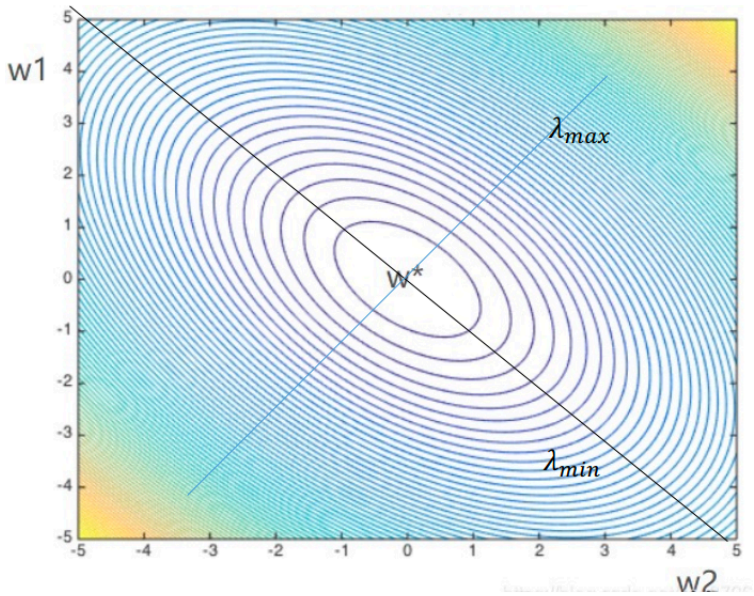


图 9: 二次函数等高线与 Hessian 矩阵特征值的关系

当 $\lambda_{\max} \approx \lambda_{\min}$ 时，等高线趋于圆形，收敛速度趋于超线性；当 $\lambda_{\max} \gg \lambda_{\min}$，等高线是很扁的椭圆，收敛速度逐渐变慢 (在 $\mathbf{d}_{\lambda_{\max}}$ 方向下降的快，但在 $\mathbf{d}_{\lambda_{\min}}$ 几乎没法下降)。

另一种描述函数等高线的指标为条件数：

条件数

定义矩阵 $\mathbf{G}$ 的条件数为

$$\kappa(\mathbf{G}) = \frac{\lambda_{\max}}{\lambda_{\min}}$$

为矩阵的条件数。如果矩阵 $\mathbf{G}$ 的条件数很大，则称其为病态的，反之为良态的。

容易发现最速下降法的收敛速度可以用条件数来描述，

$$\frac{\|\mathbf{x}_{k+1} - \mathbf{x}^*\|_{\mathbf{G}}}{\|\mathbf{x}_k - \mathbf{x}^*\|_{\mathbf{G}}} \leq \left(\frac{\lambda_{\max} - \lambda_{\min}}{\lambda_{\max} + \lambda_{\min}}\right)^2 = \left(\frac{\kappa(\mathbf{G}) - 1}{\kappa(\mathbf{G}) + 1}\right)^2$$

条件数越大，μ 越接近 1，收敛速度越慢。

4.2 随机梯度下降 (考试不考)

随机梯度下降（stochastic gradient descent, SGD）是在每一个迭代步的梯度上加入随机性的负梯度方法。SGD 考虑最小化如下目标函数：

$$f(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{x})$$

其中 $f_i(\mathbf{x})$ 表示第 i 个样本的目标函数 (比如 Loss Function)。当样本个数很多，且 $f_i(\mathbf{x})$ 计算复杂时，依次计算所有样本的 $f_i(\mathbf{x})$ 会很耗时。

为了降低计算量，随机梯度方法每一个迭代步只选择一个（或者一部分）样本的子函数，将它们梯度的负值作为迭代方向：(一个 iteration)

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha \frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}} \nabla f_i(\mathbf{x})$$

依次选中每一个样本或者每一个样本子集，直到所有的数据都用过一遍 (一个 epoch)。为了避免出现死循环，可以每一个 epoch 后对样本进行随机重排。

随机梯度下降

1. 给定 $\mathbf{x}_0$，步长 α，终止准则 $\epsilon > 0$，$k = 0$.
2. 在第 k 次迭代，若终止准则满足，则输出 $\mathbf{x}_k$ 迭代停止。
3. 训练样本随机重排 (一个 epoch)
4. 用样本子集 $\mathcal{B}$ 更新 $\mathbf{x}_k$：$\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha \frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}} \nabla f_i(\mathbf{x})$.
5. $k+1 \leftarrow k$，转步骤 2。

关于批量大小的选取：

1. $|\mathcal{B}| = 1$：随机梯度下降，一次只选一个样本计算梯度、更新参数.
   - 计算开销 $O(1)$
   - 梯度方差极大，可能无法收敛。
   - 步长需要在迭代过程中衰减。
2. $|\mathcal{B}| = n$：批量梯度下降，一次选全部样本计算梯度、更新参数.
   - 计算开销 $O(n)$
   - 梯度方差小，但可能陷入局部最小值。
   - 步长不需要进行衰减。
3. $|\mathcal{B}| = B$：小批量梯度下降，一次选 B 个样本计算梯度、更新参数.
   - 计算开销 $O(B)$。B 一般取 $50 \sim 256$。
   - 梯度方差适中且收敛更稳定。
   - 步长需要进行衰减。

以下方法为梯度下降法的变种

动量法 (Momentum)

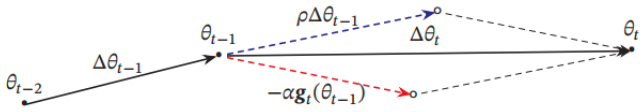
动量法在进行更新时，不仅参考当前点处的梯度方向，还要参考之前累积的梯度方向。每步迭代的方向为两者矢量和的方向：

$$\mathbf{v}_{k+1} = \gamma \mathbf{v}_k + \alpha \mathbf{g}_k$$
$$\mathbf{x}_{k+1} = \mathbf{x}_k - \mathbf{v}_{k+1}$$

动量项的作用在于参考前面累积的梯度方向。

- 若第 $k+1$ 步的梯度方向与前面累积的梯度方向相近，那么就放大第 $k+1$ 步的步长。
- 若第 $k+1$ 步的梯度方向变化很大，那么之前累积的梯度方向便对当前梯度方向起到了个修正的作用。

$\mathbf{v}_0 = 0$，动量超参数 γ 一般取值为 0.9。



(a) 动量法

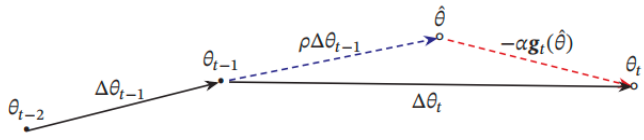
图 10: 动量法

Nesterov 动量法

动量法在到达最小值时通常会具有很高的动量，这会导致算法有可能错过最优解。Nesterov 方法对动量法进行了修正，不再计算当前位置的梯度方向，而是先试着沿之前累积的方向走到下一位置，然后再计算此处的梯度方向。每步迭代的过程为：

$$\mathbf{v}_{k+1} = \gamma \mathbf{v}_k + \alpha g(\mathbf{x}_k - \gamma \mathbf{v}_k)$$
$$\mathbf{x}_{k+1} = \mathbf{x}_k - \mathbf{v}_{k+1}$$

- 如果预测到有可能跨过最优解，那么预测值处的梯度可以对之前的梯度进行修正，从而避免错过最优解。



(b) Nesterov 加速梯度

图 11: Nesterov 动量法

AdaGrad

之前介绍的算法，每一步迭代时，参数中所有元素都采用相同的步长 (也就是 $\alpha\times$ 梯度)。当不同维度的梯度值有很大差异时，可能无法找到统一的步长适用于所有的元素。Adagrad 是一种自动调整步长的算法，它可以根据参数调整步长，**使累积梯度大的参数更新步长更小，累积梯度小的参数更新步长更大**。

$$\mathbf{s}_{k+1} = \mathbf{s}_k + \mathbf{g}_k \odot \mathbf{g}_k,\ \mathbf{s}_0 = \mathbf{0}$$
$$\mathbf{x}_{k+1} = \mathbf{x}_k - \frac{\alpha}{\sqrt{\mathbf{s}_{k+1} + \varepsilon}} \odot \mathbf{g}_k$$

默认取 $\alpha = 0.01, \varepsilon = O(10^{-8})$

然而，在迭代过程中，随着梯度累加，参数中每个元素的步长都在不断衰减 (或不变)，算法收敛速度会越来越慢，以至于在达到最优解之前，参数无法得到有效更新。

RMSProp

RMSProp 相比 AdaGrad，从单纯累加梯度平方改为**对梯度平方进行指数加权移动平均**

$$\mathbf{s}_{k+1} = \gamma \mathbf{s}_k + (1-\gamma)\mathbf{g}_k \odot \mathbf{g}_k,\ \mathbf{s}_0 = \mathbf{0}$$
$$\mathbf{x}_{k+1} = \mathbf{x}_k - \frac{\alpha}{\sqrt{\mathbf{s}_{k+1} + \varepsilon}} \odot \mathbf{g}_k$$

超参数一般取 $\gamma = 0.9$，这样一来 $\mathbf{s}_{k+1}$ 在迭代过程中一般可看作只加和了过去 $\dfrac{1}{1-\gamma} = 10$ 项的梯度平方。仍然默认取 $\alpha = 0.01, \varepsilon = O(10^{-8})$

Adam

Adam 算法**不但使用动量作为参数更新方向，而且可以自适应调整学习率**。一方面，Adam 算法计算梯度 $\mathbf{g}_k$ 的指数加权平均：

$$\mathbf{v}_{k+1} = \beta_1 \mathbf{v}_k + (1-\beta_1)\mathbf{g}_k$$

另一方面又对梯度 $\mathbf{g}_k$ 按元素平方后进行指数加权平均

$$\mathbf{s}_{k+1} = \beta_2 \mathbf{s}_k + (1-\beta_2)\mathbf{g}_k \odot \mathbf{g}_k$$

其中 $\mathbf{v}_0 = \mathbf{0}, \mathbf{s}_0 = \mathbf{0}$。最后，Adam 算法更新参数时，使用了动量和学习率修正：

$$\hat{\mathbf{v}}_{k+1} = \frac{\mathbf{v}_{k+1}}{1-\beta_1^{k+1}},\ \hat{\mathbf{s}}_{k+1} = \frac{\mathbf{s}_{k+1}}{1-\beta_2^{k+1}}$$

从而 Adam 算法的更新公式为

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \frac{\alpha}{\sqrt{\hat{\mathbf{s}}_{k+1} + \varepsilon}} \odot \nabla \hat{\mathbf{v}}_{k+1}$$

其中 $0 \le \beta_1 < 1$，$0 \le \beta_2 < 1$。通常取 $\beta_1 = 0.9, \beta_2 = 0.999$,

之所以作如上的动量和学习率修正，是因为

$$\begin{aligned}\mathbf{v}_{k+1} &= \beta_1\mathbf{v}_k + (1-\beta_1)\nabla f(\mathbf{x}_k)\\ &= \beta_1[\beta_1\mathbf{v}_k + (1-\beta_1)\nabla f(\mathbf{x}_k)] + (1-\beta_1)\nabla f(\mathbf{x}_k)\\ &= \cdots = (1-\beta_1)\sum_{i=0}^{k}\beta_1^{k-i}\nabla f(\mathbf{x}_k)\\ &= (1-\beta_1^{k+1})\nabla f(\mathbf{x}_k)\end{aligned}$$

在迭代初期，过去梯度的权重相加之和比较小，比如 $\mathbf{v}_1 = (1-\beta_1)\nabla f(\mathbf{x}_0) = 0.1\nabla f(\mathbf{x}_0)$。为解决这个问题，Adam 算法将 $\mathbf{v}_{k+1}$ 除以 $1-\beta_{k+1}$ ，从而使过去梯度的权重相加之和始终等于 1。

4.3 坐标下降

坐标下降

坐标下降方法（coordinate decent）是一个非梯度优化算法，每次沿着一个坐标方向进行线搜索，算法如下：

1. 给定 $\mathbf{x}_0, k = 0$.
2. 在第 k 次迭代时，若满足终止准则，则输出 $\mathbf{x}_k$，迭代停止。
3. 否则对坐标方向 i，线搜索计算步长 $\alpha_i = \min\limits_{\alpha} f(x_k^{(i)} - \alpha\dfrac{\partial f(\mathbf{x}_k)}{\partial x^{(i)}})$，$x_{k+1}^{(i)} = x_k^{(i)} - \alpha_i\dfrac{\partial f(\mathbf{x}_k)}{\partial x^{(i)}}$.
4. $k+1 \leftarrow k$，转步骤 2。

4.4 BB 法 (考试不考)

BB 算法是负梯度方法 $\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k\mathbf{g}_k$，但是求 α_k 的思想源于拟 Newton 方法。

BB 法

- 令 $\mathbf{H}_k = \alpha_k\mathbf{I}$，从而迭代变为 $\mathbf{x}_{k+1} = \mathbf{x}_k - \mathbf{H}_k\mathbf{g}_k$。
- 理想的 $\mathbf{H}_k = \alpha_k\mathbf{I}$ 应满足拟 Newton 条件 $\mathbf{H}_{k+1}\mathbf{y}_k = \mathbf{s}_k$，但由于其自由度较低，实际中很难严格满足，于是退而求其次

  $$\alpha_k = \arg\min_{\alpha}\|\mathbf{s}_{k-1} - \alpha\mathbf{y}_{k-1}\|_2^2$$

  可以解得两种步长 $\alpha_k^{\mathrm{BB1}} = \dfrac{\mathbf{s}_{k-1}^\top\mathbf{y}_{k-1}}{\mathbf{y}_{k-1}^\top\mathbf{y}_{k-1}}$，$\alpha_k^{\mathrm{BB2}} = \dfrac{\mathbf{s}_{k-1}^\top\mathbf{s}_{k-1}}{\mathbf{s}_{k-1}^\top\mathbf{y}_{k-1}}$。

因为

$$\mathbf{x}_k = \mathbf{x}_{k-1} - \alpha_{k-1}\mathbf{g}_{k-1} \Rightarrow \mathbf{s}_{k-1} = -\alpha_{k-1}\mathbf{g}_{k-1}$$

且

$$\mathbf{H}_k\mathbf{y}_{k-1} = \mathbf{s}_{k-1} \Rightarrow \mathbf{G}_k^{-1}\mathbf{y}_{k-1} = -\alpha_{k-1}\mathbf{g}_{k-1} \Rightarrow \mathbf{y}_{k-1} = -\alpha_{k-1}\mathbf{G}_k\mathbf{g}_{k-1}$$

因此两种步长可以表示为

$$\alpha_k^{\mathrm{BB1}} = \frac{\mathbf{g}_{k-1}^\top\mathbf{G}\mathbf{g}_{k-1}}{\mathbf{g}_{k-1}^\top\mathbf{G}^2\mathbf{g}_{k-1}},\ \alpha_k^{\mathrm{BB2}} = \frac{\mathbf{g}_{k-1}^\top\mathbf{g}_{k-1}}{\mathbf{g}_{k-1}^\top\mathbf{G}\mathbf{g}_{k-1}}$$

注意到最速下降法和最小梯度法 (没有推导，此处直接给出) 的精确线搜索步长

$$\alpha_k^{\mathrm{SD}} = \frac{\mathbf{g}_k^\top \mathbf{g}_k}{\mathbf{g}_k^\top \mathbf{G}\mathbf{g}_k},\ \alpha_k^{\mathrm{MG}} = \frac{\mathbf{g}_k^\top \mathbf{G}\mathbf{g}_k}{\mathbf{g}_k^\top \mathbf{G}^2\mathbf{g}_k}$$

可以发现 BB2 步长就是最速下降法的前一步步长，BB1 步长就是最小梯度法的前一步步长。

4.5 Newton 法

4.5.1 基本 Newton 法与阻尼 Newton 法

最速下降和随机梯度下降都是一种一阶优化方法（即只用到 $\mathbf{g}_k$），而 Newton 法为二阶优化方法（用到 Hessian 矩阵 $\mathbf{G}_k$）

具体而言，在每个迭代点 $\mathbf{x}_k$ 附近，我们都用二次函数来近似 $f(\mathbf{x}_k)$，然后向着二次函数的极值点方向迭代。假设 $f(\mathbf{x})$ 有连续二阶偏导数，当前迭代点为 $\mathbf{x}_k$，注意到 $f(\mathbf{x})$ 在 $\mathbf{x}_k$ 处的 Taylor 展开式为

$$f(\mathbf{x}_k + \mathbf{d}) = f(\mathbf{x}_k) + \mathbf{g}_k^\top \mathbf{d} + \frac{1}{2}\mathbf{d}^\top \mathbf{G}_k \mathbf{d} + o(\|\mathbf{d}\|^2)$$

容易解得

$$\begin{aligned}\mathbf{d}^* &= \arg\min_{\mathbf{d}} f(\mathbf{x}_k) + \mathbf{g}_k^\top \mathbf{d} + \frac{1}{2}\mathbf{d}^\top \mathbf{G}_k \mathbf{d} \\ &= -\mathbf{G}_k^{-1}\mathbf{g}_k\end{aligned}$$

只要 $\mathbf{G}_k$ 正定，$\mathbf{g}_k^\top \mathbf{d}_k = -\mathbf{g}_k^\top \mathbf{G}_k \mathbf{g}_k < 0$，$\mathbf{d}_k$ 必然是下降方向。以该方向为迭代方向的方法称为 Newton 法。

基本 Newton 法

基本 Newton 法迭代公式为

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \mathbf{G}_k^{-1}\mathbf{g}_k$$

基本 Newton 法的收敛性

基本 Newton 法不具有全局收敛性，收敛结果取决于初始点的选择。

1. 初始点接近极小点：收敛到极小点，收敛速度很快
2. 初始点远离极小点：可能收敛到鞍点、极大点
3. 在迭代过程中可能出现 $\mathbf{G}_k$ 奇异的情况，使得迭代无法继续

基本 Newton 法具有**局部收敛性**：设 $f(\mathbf{x}) \in C^2$，若

1. $f(\mathbf{x})$ 的 Hessian 矩阵 $\mathbf{G}(\mathbf{x})$ 满足 Lipschitz 条件：存在 $\beta > 0, \forall \mathbf{x}, \mathbf{y}$，有 $\|\mathbf{G}(\mathbf{x}) - \mathbf{G}(\mathbf{y})\| \le L\|\mathbf{x} - \mathbf{y}\|$。
2. 初始点 $\mathbf{x}_0$ 充分接近 $f(\mathbf{x})$ 的局部极小点 $\mathbf{x}^\star$，即 $\exists \delta > 0, \|\mathbf{x}_0 - \mathbf{x}^\star\| < \delta$
3. $\mathbf{G}(\mathbf{x}^\star)$ 正定

则基本 Newton 法具有**二阶收敛速度**，且梯度序列 $\mathbf{g}_k$ 二次收敛到 0。

证明. 根据 Newton 方法迭代方向以及 $\mathbf{g}^* = \mathbf{0}$ 可知

$$\mathbf{x}_{k+1} - \mathbf{x}^* = \mathbf{x}_k - \mathbf{x}^* - \mathbf{G}_k^{-1}\mathbf{g}_k = \mathbf{G}_k^{-1}(\mathbf{G}_k(\mathbf{x}_k - \mathbf{x}^*) - (\mathbf{g}_k - \mathbf{g}^*))$$

注意到

$$\mathbf{g}_k - \mathbf{g}^* = \int_{\mathbf{x}^*}^{\mathbf{x}_k} \mathbf{G}(\mathbf{x})\mathrm{d}\mathbf{x}$$

再令 $\mathbf{x} = \mathbf{x}_k + t(\mathbf{x}^* - \mathbf{x}_k)$，则

$$\mathbf{g}_k - \mathbf{g}^* = \int_{\mathbf{x}^*}^{\mathbf{x}_k} \mathbf{G}(\mathbf{x})\mathrm{d}\mathbf{x} = \int_0^1 \mathbf{G}(\mathbf{x}_k + t(\mathbf{x}^* - \mathbf{x}_k))(\mathbf{x}_k - \mathbf{x}^*)\mathrm{d}t$$

于是

$$\begin{aligned}
\|\mathbf{G}_k(\mathbf{x}_k - \mathbf{x}^*) - (\mathbf{g}_k - \mathbf{g}^*)\| &= \left\| \int_0^1 \left[\mathbf{G}_k - \mathbf{G}(\mathbf{x}_k + t(\mathbf{x}^* - \mathbf{x}_k))\right](\mathbf{x}_k - \mathbf{x}^*)\mathrm{d}t \right\| \\
&\le \int_0^1 \|\mathbf{G}_k - \mathbf{G}(\mathbf{x}_k + t(\mathbf{x}^* - \mathbf{x}_k))\|\|(\mathbf{x}_k - \mathbf{x}^*)\|\mathrm{d}t \\
(\text{Lipschitz}) &\le L\|\mathbf{x}_k - \mathbf{x}^*\|^2 \int_0^1 t\mathrm{d}t \\
&= \frac{L}{2}\|\mathbf{x}_k - \mathbf{x}^*\|^2
\end{aligned}$$

注意到 $\exists r > 0$，由 $\mathbf{G}(\mathbf{x})$ 的 Lipschitz 连续性可以保证当 $\|\mathbf{x} - \mathbf{x}^*\| \le r$ 时，有 $\|\mathbf{G}(\mathbf{x})^{-1}\| \le 2\|\mathbf{G}(\mathbf{x}^*)^{-1}\|$，当我们取初始点 $\|\mathbf{x}_0 - \mathbf{x}^*\| \le \min\{\delta, r, \dfrac{1}{L\|\mathbf{G}(\mathbf{x}^*)^{-1}\|}\}$ 时，由于当 $\|\mathbf{x}_0 - \mathbf{x}^*\| < \dfrac{1}{L\|\mathbf{G}(\mathbf{x}^*)^{-1}\|}$ 时，可以保证

$$\frac{\|\mathbf{x}_{k+1} - \mathbf{x}^*\|}{\|\mathbf{x}_k - \mathbf{x}^*\|} \le L\|\mathbf{G}(\mathbf{x}^*)^{-1}\| < 1$$

也就是保证了迭代的点列 $\{\mathbf{x}_k\}$ 的范围越缩越小，总处于我们框定的邻域内。再由

$$\|\mathbf{x}_{k+1} - \mathbf{x}^*\| \le \|\mathbf{G}_k^{-1}\|\frac{L}{2}\|\mathbf{x}_k - \mathbf{x}^*\|^2 \le L\|\mathbf{G}(\mathbf{x}^*)^{-1}\|\|\mathbf{x}_k - \mathbf{x}^*\|^2$$

得知 $\{\mathbf{x}_k\}$ 二次收敛到 $\mathbf{x}^*$.

其次

$$\begin{aligned}
\|\mathbf{g}_{k+1}\| = \|\mathbf{g}_{k+1} - \mathbf{g}_k - \mathbf{G}_k\mathbf{d}_k\| &= \left\| \int_0^1 G(\mathbf{x}_k + t\mathbf{d}_k)\mathbf{d}_k\mathrm{d}t - \mathbf{G}_k\mathbf{d}_k \right\| \\
&\le \int_0^1 \|(G(\mathbf{x}_k + t\mathbf{d}_k) - \mathbf{G}_k)\|\|\mathbf{d}_k\|\mathrm{d}t \\
&\le \frac{1}{2}L\|\mathbf{d}_k\|^2 = \frac{1}{2}L\|\mathbf{G}_k^{-1}\mathbf{g}_k\|^2 \le \frac{1}{2}L\|\mathbf{G}_k^{-1}\|^2\|\mathbf{g}_k\|^2
\end{aligned}$$

这说明梯度序列二阶收敛到 0。 □

基本 Newton 法的优点：

- 当初始点接近极小点 $\mathbf{x}^*$ 时，方法以二阶速度收敛（非常快）
- 方法具有二次终止性，对于二次函数只需迭代一步即可找到最优解

缺点：

- 当初始点没有充分接近极小点 $\mathbf{x}$ 时，Hessian 矩阵 $\mathbf{G}_k$ 可能不正定或者奇异，使得 $\mathbf{x}_k$ 无法收敛到 $\mathbf{x}^*$，甚至迭代无法进行。
- 即使 $\mathbf{G}_k$ 正定，$f(\mathbf{x}_k)$ 也不能保证单调下降。因为固定了 $\alpha_k = 1$，即使方向是下降方向，但可能走过头了。
- 每步迭代需要计算实对称矩阵 $\mathbf{G}_k$，要计算 $\dfrac{n(n+1)}{2}$ 个偏导数。
- 每步迭代需要进行矩阵求逆运算，$O(n^3)$ 计算复杂度。

阻尼 Newton 法

阻尼 Newton 法的迭代公式为

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k \mathbf{G}_k^{-1} \mathbf{g}_k$$

阻尼 Newton 法的收敛性

设 $f(\mathbf{x}) \in C^2$，若 $\forall \mathbf{x}_0 \in \mathbb{R}^n$，存在 $\beta > 0$，使得 $f(\mathbf{x})$ 在水平集 $L(\mathbf{x}_0) = \{\mathbf{x} | f(\mathbf{x}) \le f(\mathbf{x}_0)\}$ 上满足

$$\mathbf{u}^\top \mathbf{G}(\mathbf{x}) \mathbf{u} \ge \beta \|\mathbf{u}\|^2, \ \mathbf{u} \in \mathbb{R}^n, \mathbf{x} \in L(\mathbf{x}_0)$$

则采用精确线搜索/Goldstein 准则/Wolfe 准则的阻尼 Newton 方法具有**全局收敛性**，满足下列两者之一：

- $\{\mathbf{x}_k\}$ 为有穷点列，即存在 N，使得 $\mathbf{g}_N = \mathbf{0}$
- $\{\mathbf{x}_k\}$ 为无穷点列，$\{\mathbf{x}_k\}$ 收敛到 f 的唯一极小点 $\mathbf{x}^*$

直观上，$\mathbf{u}^\top \mathbf{G}(\mathbf{x}) \mathbf{u} \ge \beta \|\mathbf{u}\|^2$ 的条件表明 $\mathbf{G}$ 是一个“更加正定”的正定矩阵。

证明. (课件上说见教材) 只需证明当 $\{\mathbf{x}_k\}$ 为无穷点列时，$\{\mathbf{x}_k\}$ 收敛到 f 的唯一极小点 $\mathbf{x}^*$。要证明此，**需要先说明水平集 $L(\mathbf{x}_0)$ 是有界闭凸集**。

1. $L(\mathbf{x}_0)$ 为凸集：$\forall \mathbf{x}_1, \mathbf{x}_2 \in L(\mathbf{x}_0)$，由于 $f(\mathbf{x}_1) \le f(\mathbf{x}_0), f(\mathbf{x}_2) \le f(\mathbf{x}_0)$，故 $\forall \lambda \in [0,1]$ 以及 $\mathbf{x} = \lambda \mathbf{x}_1 + (1-\lambda)\mathbf{x}_2$，有
$$f(\mathbf{x}) = f(\lambda \mathbf{x}_1 + (1-\lambda)\mathbf{x}_2) \le \lambda f(\mathbf{x}_1) + (1-\lambda) f(\mathbf{x}_2) \le f(\mathbf{x}_0)$$
所以 $\mathbf{x} \in L(\mathbf{x}_0)$，即 $L(\mathbf{x}_0)$ 为凸集。
2. $L(\mathbf{x}_0)$ 为闭集：设存在序列 $\{\mathbf{x}_n\} \in L(\mathbf{x}_0), \mathbf{x}_n \to \mathbf{x}^*$，则由 $f(\mathbf{x}_n) \le f(\mathbf{x}_0)$ 以及 f 的连续性可知 $f(\mathbf{x}^*) \le f(\mathbf{x}_0)$，即 $\mathbf{x}^* \in L(\mathbf{x}_0)$，所以 $L(\mathbf{x}_0)$ 为闭集。
3. $L(\mathbf{x}_0)$ 为有界集：$\forall \mathbf{y} \in L(\mathbf{x}_0), \mathbf{y} \ne \mathbf{x}_0$，根据 Taylor 公式有
$$\begin{aligned} f(\mathbf{y}) &= f(\mathbf{x}_0) + \mathbf{g}_0^\top (\mathbf{y} - \mathbf{x}_0) + \frac{1}{2} (\mathbf{y} - \mathbf{x}_0)^\top \mathbf{G}_0 (\mathbf{y} - \mathbf{x}_0) + o(\|\mathbf{y} - \mathbf{x}_0\|^2) \\ &\ge f(\mathbf{x}_0) + \mathbf{g}_0^\top (\mathbf{y} - \mathbf{x}_0) + \frac{1}{2} \beta \|\mathbf{y} - \mathbf{x}_0\|^2 \\ &\ge f(\mathbf{x}_0) - \|\mathbf{g}_0\| \|\mathbf{y} - \mathbf{x}_0\| + \frac{\beta}{2} \|\mathbf{y} - \mathbf{x}_0\|^2 \end{aligned}$$
因为 $f(\mathbf{y}) \le f(\mathbf{x}_0)$，因此解上述不等式可得 $\|\mathbf{y} - \mathbf{x}_0\| \le 2 \dfrac{\|\mathbf{g}_0\|}{\beta}$，所以 $L(\mathbf{x}_0)$ 为有界集。

由于 $L(\mathbf{x}_0)$ 为凸集，根据凸函数判定定理，以及 $\mathbf{u}^\top \mathbf{G}(\mathbf{x})\mathbf{u}$，知定义在 $L(\mathbf{x}_0)$ 上的 f 为**严格凸函数**，其稳定点为唯一的极小点。

由于 $L(\mathbf{x}_0)$ 为有界闭集，f 定义在 $L(\mathbf{x}_0)$ 上，以及 $G(\mathbf{x})$ 连续 (因为 $f \in C^2$)，$\forall \mathbf{x} \in L(\mathbf{x}_0)$，$\exists \gamma > 0$，使得

$$\|G(\mathbf{x})\| \leq \gamma \Rightarrow \|\mathbf{g}_k\| = \|\mathbf{G}_k \mathbf{d}_k\| \leq \gamma \|\mathbf{d}_k\|$$

接下来我们希望利用非精确线搜索的收敛性定理来证明 $\{\mathbf{x}_k\}$ 收敛到 f 的唯一极小点 $\mathbf{x}^*$：$\mathbf{d}_k$ 和负梯度 $-\mathbf{g}_k$ 的夹角满足

$$\begin{aligned}
&\frac{\pi}{2} - \theta_k \geq \sin\left(\frac{\pi}{2} - \theta_k\right) = \cos\theta_k = -\frac{\mathbf{g}_k^\top \mathbf{d}_k}{\|\mathbf{g}_k\|\|\mathbf{d}_k\|} = \frac{\mathbf{d}_k^\top \mathbf{G}_k \mathbf{d}_k}{\|\mathbf{g}_k\|\|\mathbf{d}_k\|} \geq \frac{\beta}{\gamma} \\
\Rightarrow &\theta_k \leq \frac{\pi}{2} - \frac{\beta}{\gamma}
\end{aligned}$$

根据非精确线搜索的收敛性定理，要么存在 N，使得 $\mathbf{g}_N = \mathbf{0}$，要么 $\mathbf{g}_k \to 0, k \to \infty$。后者说明 $\{\mathbf{x}_k\}$ 收敛到 f 的唯一极小点 $\mathbf{x}^*$。 □

阻尼 Newton 法的优点：在基本 Newton 法上引入线搜索，即 $\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{d}_k$，能够保证

- 当 $\mathbf{G}_k$ 正定时，$f(\mathbf{x}_k)$ 单调下降
- 即使 $\mathbf{x}_k$ 离 $\mathbf{x}^*$ 稍远，产生的点列 $\{\mathbf{x}_k\}$ 仍可能收敛到 $\mathbf{x}^*$，**改善基本 Newton 法的局部收敛性问题**。

缺点：

- 阻尼 Newton 法还是没有解决 $\mathbf{G}_k$ 奇异/非正定的问题。

4.5.2 混合法

基本/阻尼 Newton 方法始终出现几个问题：

- $\mathbf{G}_k$ 不正定
- $\mathbf{G}_k$ 不可逆
- 迭代方向 $\mathbf{d}_k$ 几乎与 $-\mathbf{g}_k$ 正交（我们希望 $-\mathbf{g}_k^\top \mathbf{d}_k$ 越负越好）

混合法的基本思想就是，当一种方法无法继续迭代时，采用另一种方法可以使迭代进行下去，具体而言：

- $\mathbf{G}_k$ 不正定但可逆时，注意到 $\mathbf{g}_k^\top (\mathbf{G}_k^{-1} \mathbf{g}_k) < 0$，因此取反向 $\mathbf{d}_k = \mathbf{G}_k^{-1} \mathbf{g}_k$
- $\mathbf{G}_k$ 不可逆或迭代方向 $\mathbf{d}_k$ 几乎与 $-\mathbf{g}_k$ 正交时，直接取负梯度方向 $\mathbf{d}_k = -\mathbf{g}_k$

混合法

1. 给定 $\mathbf{x}_0$，$\epsilon_1 > 0, \epsilon_2 > 0$，$k = 0$.
2. 在第 k 次迭代，若终止准则满足，则输出 $\mathbf{x}_k$ 迭代停止。
3. 若 $\mathbf{G}_k$ 非奇异，则 $\mathbf{d}_k \leftarrow -\mathbf{G}_k^{-1} \mathbf{g}_k$，否则转到第 6 步
4. 若 $\mathbf{g}_k^\top \mathbf{d}_k > \varepsilon \|\mathbf{g}_k\| \|\mathbf{d}_k\|$（$\mathbf{G}_k$ 不正定），则 $\mathbf{d}_k \leftarrow -\mathbf{d}_k$，转到第 7 步

5. 若 $|\mathbf{g}_k^\top \mathbf{d}_k| \le \epsilon_2 \|\mathbf{g}_k\| \|\mathbf{d}_k\|$（$\mathbf{d}_k$ 几乎与 $-\mathbf{g}_k$ 正交），转到第 6 步，否则转到第 7 步
6. $\mathbf{d}_k = -\mathbf{g}_k$
7. 精确线搜索求 $\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{d}_k$ 的步长 α_k，$k+1 \leftarrow k$，转到第 2 步

4.5.3 LM 法

Levenberg-Marquardt 法是处理 $\mathbf{G}_k$ 奇异/非正定情况的一个简单有效的方法。对于非正定矩阵 $\mathbf{G}_k$，我们可以选一个足够大的 v，保证 $\mathbf{G}_k + v\mathbf{I}$ 是正定矩阵。

LM 法

LM 法的迭代方向为

$$\mathbf{d}_k = -(\mathbf{G}_k + v_k \mathbf{I})^{-1} \mathbf{g}_k$$

随后通过精确线搜索求 $\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{d}_k$ 的步长 α_k。

- 当 v_k 很小时，LM 方向接近 Newton 方向。
- 当 v_k 很大时，LM 方向接近负梯度方向。

具体而言，v_k 如何取值呢？以下定理指出，**LM 法法等效于信赖域框架下的 Newton 方法，从而 v_k 的取值可以参考信赖域半径 Δ_k 进行修正**。

LMF 法是信赖域框架下的 Newton 法

$\mathbf{d}_k$ 是信赖域子问题

$$\min_d q_k(\mathbf{d}) = \min_d \frac{1}{2}\mathbf{d}^\top \mathbf{G}_k \mathbf{d} + \mathbf{g}_k^\top \mathbf{d},\ \text{s.t.} \|\mathbf{d}\| \le \|\Delta_k\|$$

的解 $\iff$ 存在 $v_k \ge 0$ 使得 $(\mathbf{G}_k^\top + v_k \mathbf{I})\mathbf{d}_k = -\mathbf{g}_k$，其中 $v_k(\Delta_k - \|\mathbf{d}_k\|) = 0$

证明. 证明过程见4.8.2中“LMF 法是信赖域框架下的 Gauss-Newton 法”。 □

根据上述定理，

$$\|\mathbf{G}_k + v_k \mathbf{I}\| \|\mathbf{d}_k\| = \|\mathbf{g}_k\|$$

当 v_k 变大时，$\|\mathbf{d}_k\|$ 变小，从而 Δ_k 变小，反之亦然。因此，v_k 的调整方向应与信赖域的半径 Δ_k 的调整方向相反。类似于之前信赖域方法所述，定义比值

$$\gamma_k = \frac{f(\mathbf{x}_k) - f(\mathbf{x}_k + \mathbf{d}_k)}{q_k(\mathbf{0}) - q_k(\mathbf{d}_k)}$$

来表现二次函数 $q_k(\mathbf{d}_k)$ 在 $\mathbf{x}_k + \mathbf{d}_k$ 附近近似目标函数 $f(\mathbf{x}_k)$ 的好坏程度。

- γ_k 越小，说明二次函数 $q_k(\mathbf{d}_k)$ 近似效果差，**应缩小信赖域半径 Δ_k，即增大 v_k**
- γ_k 越大，说明二次函数 $q_k(\mathbf{d}_k)$ 近似效果好，**应扩大信赖域半径 Δ_k，即减小 v_k**

由此可以给出 LM 方法的算法流程

LMF算法流程

1. 给出$x_0 \in \mathbb{R}^n$，$v_0 > 0$，$\epsilon > 0$，$k \leftarrow 0$；
2. 若终止准则满足，则输出有关信息，停止迭代；
3. 求解方程$(G_k + v_k I)d_k = -g_k$，得到d_k；
4. 计算$\gamma_k = \frac{\Delta f_k}{\Delta q_k}$；
5. 若$\gamma_k < 0.25$，则$v_{k+1} \leftarrow 4v_k$；若$\gamma_k > 0.75$，则$v_{k+1} \leftarrow \frac{v_k}{2}$，否则$v_{k+1} \leftarrow v_k$；
6. 若$\gamma_k \leq 0$，则$x_{k+1} \leftarrow x_k$；否则$x_{k+1} \leftarrow x_k + d_k$；
7. 若$G_{k+1} + v_{k+1}I$负定，则$v_{k+1} \leftarrow 2v_{k+1}$；
8. $k \leftarrow k + 1$，转到第2步。

4.6 拟 Newton 法

无论用什么改进方法，Newton 方法的最大问题是

- 计算量大（Hessian 矩阵、求逆）
- 不稳定性（Hessian 矩阵不正定/奇异）

我们想找到一种**不需要计算二阶梯度，收敛速度又比较快的方法**。一条基本思路是：利用 $\mathbf{x}_k, \mathbf{x}_{k+1}$ 和 $\mathbf{g}_k, \mathbf{g}_{k+1}$ 来构造矩阵 $\mathbf{B}_{k+1}$ 来近似 $\mathbf{G}_{k+1}$（或构造 $\mathbf{H}_{k+1}$ 近似 $\mathbf{G}_{k+1}^{-1}$，且近似矩阵 $\mathbf{B}_{k+1}$(或 $\mathbf{H}_{k+1}$) 要满足

- 只需要 $f(\mathbf{x})$ 的一阶梯度信息
- $\mathbf{B}_k$(或 $\mathbf{H}_k$) 正定，保证 $\mathbf{d}_k$ 为下降方向
- 方法收敛速度快

拟 Newton 条件

令 $\mathbf{s}_k = \mathbf{x}_{k+1} - \mathbf{x}_k, \mathbf{y}_k = \mathbf{g}_{k+1} - \mathbf{g}_k$，则近似矩阵 $\mathbf{B}_k$ 或 $\mathbf{H}_k$ 应满足

$$\mathbf{B}_{k+1}\mathbf{s}_k = \mathbf{y}_k, \mathbf{H}_{k+1}\mathbf{y}_k = \mathbf{s}_k$$

证明. 近似的度量可以从 Taylor 展开考虑

$$g(\mathbf{x}) = \mathbf{g}_{k+1} + \mathbf{G}_{k+1}(\mathbf{x} - \mathbf{x}_{k+1}) + o(\|\mathbf{x} - \mathbf{x}_{k+1}\|^2)$$

取 $\mathbf{x} = \mathbf{x}_k$，当 $\mathbf{x}_{k+1}$ 与 $\mathbf{x}_k$ 很接近时，成立

$$\begin{aligned} \mathbf{g}_k - \mathbf{g}_{k+1} &= \mathbf{G}_{k+1}(\mathbf{x}_k - \mathbf{x}_{k+1}) \\ \Rightarrow \mathbf{y}_k &= \mathbf{G}_{k+1}\mathbf{s}_k \end{aligned}$$

从这个性质入手，我们只需用 $\mathbf{B}_{k+1}$ 替代 $\mathbf{G}_{k+1}$，或用 $\mathbf{H}_{k+1}$ 替代 $\mathbf{G}_{k+1}^{-1}$ 即可。 □

拟 Newton 法

1. 给定初始迭代点 $\mathbf{x}_0$，对称正定矩阵 $\mathbf{H}_0$，$\epsilon > 0$，$k = 0$.
2. 若终止准则满足，则输出有关信息，迭代停止.

3. 计算 $\mathbf{d}_k = -\mathbf{H}_k\mathbf{g}_k$.
4. 沿 $\mathbf{d}_k$ 进行线搜索求出步长 α_k，作迭代 $\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k\mathbf{d}_k$.
5. **修正** $\mathbf{H}_k$ 得到 $\mathbf{H}_{k+1}$，使得 $\mathbf{H}_{k+1}\mathbf{y}_k = \mathbf{s}_k$(即满足拟 Newton 条件)。$k \leftarrow k+1$，转到第 2 步.

重点在于**如何修正** $\mathbf{H}_k$ **得到** $\mathbf{H}_{k+1}$。在确定了修正量 $\Delta\mathbf{B}_k$(或 $\Delta\mathbf{H}_k$) 后，我们才能更新近似矩阵

$$\mathbf{B}_{k+1} = \mathbf{B}_k + \Delta\mathbf{B}_k, \mathbf{H}_{k+1} = \mathbf{H}_k + \Delta\mathbf{H}_k$$

4.6.1 Symmetric Rank 1 (SR1) 方法

SR1 方法

Symmetric Rank 1 方法中，增量矩阵 $\Delta\mathbf{B}_k$ 或者 $\Delta\mathbf{H}_k$ 为**对称且秩为 1 的矩阵**。具体而言，SR1 以如下方法更新 $\mathbf{B}_k$ 或 $\mathbf{H}_k$:

$$\mathbf{B}_{k+1}^{\mathrm{SR1}} = \mathbf{B}_k + \frac{(\mathbf{y}_k - \mathbf{B}_k\mathbf{s}_k)(\mathbf{y}_k - \mathbf{B}_k\mathbf{s}_k)^\top}{(\mathbf{y}_k - \mathbf{B}_k\mathbf{s}_k)^\top\mathbf{s}_k}$$
$$\mathbf{H}_{k+1}^{\mathrm{SR1}} = \mathbf{H}_k + \frac{(\mathbf{s}_k - \mathbf{H}_k\mathbf{y}_k)(\mathbf{s}_k - \mathbf{H}_k\mathbf{y}_k)^\top}{(\mathbf{s}_k - \mathbf{H}_k\mathbf{y}_k)^\top\mathbf{y}_k}$$

($\mathbf{B}_k$ 和 $\mathbf{H}_k$ 的迭代式是完全对称的)

证明. SR1 方法用 $\sigma \in \mathbb{R}, \mathbf{v} \in \mathbb{R}^n$ 来近似 $\Delta\mathbf{B}_k$:

$$\mathbf{B}_{k+1} = \mathbf{B}_k + \sigma\mathbf{v}\mathbf{v}^\top$$

从而 $\Delta\mathbf{B}_k$ 是秩为 1 的矩阵。我们希望通过选择合适的 σ 和 $\mathbf{v}$ 来使得 $\mathbf{B}_{k+1}$ 满足拟 Newton 条件，即

$$\mathbf{y}_k = \mathbf{B}_{k+1}\mathbf{s}_k = \mathbf{B}_k\mathbf{s}_k + \sigma\mathbf{v}\mathbf{v}^\top\mathbf{s}_k = \mathbf{B}_k\mathbf{s}_k + (\sigma\mathbf{v}^\top\mathbf{s}_k)\mathbf{v}$$

注意到 $\sigma\mathbf{v}^\top\mathbf{s}_k$ 是标量，因此 $\mathbf{y}_k - \mathbf{B}_k\mathbf{s}_k$ 一定与 $\mathbf{v}$ 共线。可记 $\mathbf{v} = \delta(\mathbf{y}_k - \mathbf{B}_k\mathbf{s}_k)$，再代回上式

$$\mathbf{y}_k - \mathbf{B}_k\mathbf{s}_k = \delta^2\sigma(\mathbf{y}_k - \mathbf{B}_k\mathbf{s}_k)^\top\mathbf{s}_k(\mathbf{y}_k - \mathbf{B}_k\mathbf{s}_k) \Rightarrow \delta^2\sigma(\mathbf{y}_k - \mathbf{B}_k\mathbf{s}_k)^\top\mathbf{s}_k = 1$$

再将 $\mathbf{v} = \delta(\mathbf{y}_k - \mathbf{B}_k\mathbf{s}_k)$ 代回 $\mathbf{B}_k$ 的迭代式

$$\begin{aligned}\mathbf{B}_{k+1}^{\mathrm{SR1}} &= \mathbf{B}_k + \delta^2\sigma(\mathbf{y}_k - \mathbf{B}_k\mathbf{s}_k)^\top(\mathbf{y}_k - \mathbf{B}_k\mathbf{s}_k) \\ &= \mathbf{B}_k + \frac{(\mathbf{y}_k - \mathbf{B}_k\mathbf{s}_k)(\mathbf{y}_k - \mathbf{B}_k\mathbf{s}_k)^\top}{(\mathbf{y}_k - \mathbf{B}_k\mathbf{s}_k)^\top\mathbf{s}_k}\end{aligned}$$

对称地，$\mathbf{H}_{k+1}$ 也有迭代式

$$\mathbf{H}_{k+1}^{\mathrm{SR1}} = \mathbf{H}_k + \frac{(\mathbf{s}_k - \mathbf{H}_k\mathbf{y}_k)(\mathbf{s}_k - \mathbf{H}_k\mathbf{y}_k)^\top}{(\mathbf{s}_k - \mathbf{H}_k\mathbf{y}_k)^\top\mathbf{y}_k}$$

□

SR1 方法的收敛性

SR1 方法具有二次终止性，即假设 $f(\mathbf{x})=\frac{1}{2}\mathbf{x}^\top\mathbf{G}\mathbf{x}+\mathbf{b}^\top\mathbf{x}$(其中 $\mathbf{G}$ 对称正定)，若对于 $\forall\mathbf{x}_0\in\mathbb{R}^n$ 和任意对称正定矩阵 $\mathbf{B}_0$，只要 $\forall k,(\mathbf{y}_k-\mathbf{B}_k\mathbf{s}_k)^\top\mathbf{s}_k\neq 0$，则按照 SR1 迭代式，最多经过有限次迭代即可求得 $f(\mathbf{x})$ 的最小点。

若确实迭代了 n 步且 $\mathbf{s}_0,\cdots,\mathbf{s}_{n-1}$ 线性无关，则 $\mathbf{B}_n=\mathbf{G}$。

[注]：以上结论换为 $\mathbf{H}_k$ 的迭代式同样成立。

实际运行 SR1 算法时，可能出现可能出现三种情况：

1. $(\mathbf{y}_k-\mathbf{B}_k\mathbf{s}_k)^\top\mathbf{s}_k\neq 0$，SR1 公式可以用.
2. $\mathbf{y}_k=\mathbf{B}_k\mathbf{s}_k$，根据拟 Newton 条件还是可以通过 $\mathbf{B}_{k+1}=\mathbf{B}_k$ 更新.
3. $\mathbf{y}_k\neq\mathbf{B}_k\mathbf{s}_k$ 但 $(\mathbf{y}_k-\mathbf{B}_k\mathbf{s}_k)^\top\mathbf{s}_k=0$，此时 SR1 公式不可用了.

故而在实际使用 SR1 算法时，为避免 3，**一种解决方法是**在计算 SR1 前再做一小步判断：设置一个很小的超参数 $r=10^{-8}$：

- 如果 $|(\mathbf{y}_k-\mathbf{B}_k\mathbf{s}_k)^\top\mathbf{s}_k|\geq r\|\mathbf{s}_k\|\|\mathbf{y}_k-\mathbf{B}_k\mathbf{s}_k\|$，则使用 SR1 公式.
- 否则用 $\mathbf{B}_{k+1}\leftarrow\mathbf{B}_k$ 更新.

SR1 方法的缺点：

1. 迭代过程中 $\mathbf{B}_{k+1}$ 不一定正定。$(\mathbf{y}_k-\mathbf{B}_k\mathbf{s}_k)^\top\mathbf{s}_k>0$ 才能推出 $\mathbf{B}_{k+1}$ 正定。
2. $(\mathbf{y}_k-\mathbf{B}_k\mathbf{s}_k)^\top\mathbf{s}_k=0$ 时，迭代会出问题。本质是因为在 SR1 方法中 $\Delta\mathbf{B}_k$ 是由两个向量的外积构成的，这种修正的自由度非常有限。遇到某些特殊情况时，SR1 更新可能无法充分调整矩阵。

4.6.2 BFGS 和 DFP 方法

BFGS 和 DFP 方法都是在 rank=1 的 SR1 方法上拓展，提出的 rank=2 的拟 Newton 方法。

BFGS 方法

BFGS 方法中，增量矩阵 $\Delta\mathbf{B}_k$ 或者 $\Delta\mathbf{H}_k$ 为**对称且秩为 2 的矩阵**。具体而言，BFGS 以如下方法更新 $\mathbf{B}_k$ 或 $\mathbf{H}_k$：

$$\mathbf{B}_{k+1}^{\text{BFGS}}=\mathbf{B}_k+\frac{\mathbf{y}_k\mathbf{y}_k^\top}{\mathbf{y}_k^\top\mathbf{s}_k}-\frac{\mathbf{B}_k\mathbf{s}_k\mathbf{s}_k^\top\mathbf{B}_k^\top}{\mathbf{s}_k^\top\mathbf{B}_k\mathbf{s}_k}$$

$$\mathbf{H}_{k+1}^{\text{BFGS}}=(\mathbf{I}-\frac{\mathbf{s}_k\mathbf{y}_k^\top}{\mathbf{y}_k^\top\mathbf{s}_k})\mathbf{H}_k(\mathbf{I}-\frac{\mathbf{y}_k\mathbf{s}_k^\top}{\mathbf{y}_k^\top\mathbf{s}_k})+\frac{\mathbf{s}_k\mathbf{s}_k^\top}{\mathbf{y}_k^\top\mathbf{s}_k}$$

[注]：有关初始矩阵的选定，可以先初始取 $\mathbf{H}_0=\mathbf{I}$，计算出 $\mathbf{x}_1$ 后，在 BFGS 更新之前，更新 $\mathbf{H}_0$ 为 $\mathbf{H}_0=\frac{\mathbf{y}_k^\top\mathbf{s}_k}{\mathbf{y}_k^\top\mathbf{y}_k}\mathbf{I}$，再通过 BFGS 计算 $\mathbf{H}_1$.

证明. BFGS 方法用 $\beta,\gamma\in\mathbb{R},\mathbf{u},\mathbf{v}\in\mathbb{R}^n$ 来近似 $\Delta\mathbf{B}_k$:

$$\mathbf{B}_{k+1}=\mathbf{B}_k+\beta\mathbf{u}\mathbf{u}^\top+\gamma\mathbf{v}\mathbf{v}^\top$$

我们希望通过选择合适的 β,γ 和 $\mathbf{u},\mathbf{v}$ 来使得 $\mathbf{B}_{k+1}$ 满足拟 Newton 条件，即

$$\mathbf{y}_k=\mathbf{B}_{k+1}\mathbf{s}_k=\mathbf{B}_k\mathbf{s}_k+\beta(\mathbf{u}^\top\mathbf{s}_k)\mathbf{u}+\gamma(\mathbf{v}^\top\mathbf{s}_k)\mathbf{v}$$

这里 $\mathbf{u},\mathbf{v}$ 的选择不唯一了，一种简单的取法是

$$\beta(\mathbf{u}^\top\mathbf{s}_k)=1,\mathbf{u}=\mathbf{y}_k$$
$$\gamma(\mathbf{v}^\top\mathbf{s}_k)=-1,\mathbf{v}=\mathbf{B}_k\mathbf{s}_k$$

把这些代回 $\mathbf{B}_{k+1}$ 的更新公式

$$\mathbf{B}_{k+1}^{\text{BFGS}}=\mathbf{B}_k+\frac{\mathbf{y}_k\mathbf{y}_k^\top}{\mathbf{y}_k^\top\mathbf{s}_k}-\frac{\mathbf{B}_k\mathbf{s}_k\mathbf{s}_k^\top\mathbf{B}_k^\top}{\mathbf{s}_k^\top\mathbf{B}_k\mathbf{s}_k}$$

对称地，如果 $\mathbf{B}_k,\mathbf{B}_{k+1}$ 可逆，依据 Shermann-Morrison-Woodbury 公式可得出 $\mathbf{H}_{k+1}^{\text{BFGS}}$ 公式

$$\mathbf{H}_{k+1}^{\text{BFGS}}=(\mathbf{I}-\frac{\mathbf{s}_k\mathbf{y}_k^\top}{\mathbf{y}_k^\top\mathbf{s}_k})\mathbf{H}_k(\mathbf{I}-\frac{\mathbf{y}_k\mathbf{s}_k^\top}{\mathbf{y}_k^\top\mathbf{s}_k})+\frac{\mathbf{s}_k\mathbf{s}_k^\top}{\mathbf{y}_k^\top\mathbf{s}_k}$$

□

DFP 方法

DFP 方法中，增量矩阵 $\Delta\mathbf{B}_k$ 或者 $\Delta\mathbf{H}_k$ 为**对称且秩为 2 的矩阵**:

$$\mathbf{H}_{k+1}^{\text{DFP}}=\mathbf{H}_k+\frac{\mathbf{s}_k\mathbf{s}_k^\top}{\mathbf{s}_k^\top\mathbf{y}_k}-\frac{\mathbf{H}_k\mathbf{y}_k\mathbf{y}_k^\top\mathbf{H}_k^\top}{\mathbf{y}_k^\top\mathbf{H}_k\mathbf{y}_k}$$
$$\mathbf{B}_{k+1}^{\text{DFP}}=(\mathbf{I}-\frac{\mathbf{y}_k\mathbf{s}_k^\top}{\mathbf{s}_k^\top\mathbf{y}_k})\mathbf{B}_k(\mathbf{I}-\frac{\mathbf{s}_k\mathbf{y}_k^\top}{\mathbf{s}_k^\top\mathbf{y}_k})+\frac{\mathbf{y}_k\mathbf{y}_k^\top}{\mathbf{s}_k^\top\mathbf{y}_k}$$

证明. DFP 用 $\beta,\gamma\in\mathbb{R},\mathbf{u},\mathbf{v}\in\mathbb{R}^n$ 来近似 $\Delta\mathbf{H}_k$:

$$\mathbf{H}_{k+1}=\mathbf{H}_k+\beta\mathbf{u}\mathbf{u}^\top+\gamma\mathbf{v}\mathbf{v}^\top$$

得到 $\mathbf{H}_{k+1}^{\text{DFP}}$ 公式

$$\mathbf{H}_{k+1}^{\text{DFP}}=\mathbf{H}_k+\frac{\mathbf{s}_k\mathbf{s}_k^\top}{\mathbf{s}_k^\top\mathbf{y}_k}-\frac{\mathbf{H}_k\mathbf{y}_k\mathbf{y}_k^\top\mathbf{H}_k^\top}{\mathbf{y}_k^\top\mathbf{H}_k\mathbf{y}_k}$$

对称地，依据 Shermann-Morrison-Woodbury 公式可得出 $\mathbf{B}_{k+1}^{\text{DFP}}$ 公式

$$\mathbf{B}_{k+1}^{\text{DFP}}=(\mathbf{I}-\frac{\mathbf{y}_k\mathbf{s}_k^\top}{\mathbf{s}_k^\top\mathbf{y}_k})\mathbf{B}_k(\mathbf{I}-\frac{\mathbf{s}_k\mathbf{y}_k^\top}{\mathbf{s}_k^\top\mathbf{y}_k})+\frac{\mathbf{y}_k\mathbf{y}_k^\top}{\mathbf{s}_k^\top\mathbf{y}_k}$$

□

容易看出，BFGS 和 DFP 是互为对偶的方法。但是 BFGS 和 DFP 代表的是两种解。

从优化意义上得出 BFGS 和 DFP 迭代公式

$\mathbf{B}_{k+1}^{\mathrm{DFP}}$ 是以下矩阵优化问题的解

$$\min_{\mathbf{B}} \|\mathbf{W}^{-\top}(\mathbf{B}-\mathbf{B}_k)\mathbf{W}^{-1}\|_F,\ \text{s.t.}\begin{cases}\mathbf{B}=\mathbf{B}^\top\\ \mathbf{B}\mathbf{s}_k=\mathbf{y}_k\\ \mathbf{W}^\top\mathbf{W}=\mathbf{B},\mathbf{W}\in\mathbb{R}^n\end{cases}$$

即：在所有满足拟 Newton 条件的对称矩阵中，寻找在加权 F 范数意义下与 $\mathbf{B}_k$ 距离最近的矩阵。

[注]：将 $\mathbf{B}$ 换为 $\mathbf{H}$，将 $\mathbf{s}_k,\mathbf{y}_k$ 互换，就能从矩阵优化角度理解 $\mathbf{H}_{k+1}^{\mathrm{BFGS}}$ 公式。

下面介绍 BFGS 和 DFP 方法一些性质

对称正定性

假设 $\mathbf{H}_k$ 对称正定，如果 $\mathbf{s}_k^\top\mathbf{y}_k>0$，则构造出的 $\mathbf{H}_{k+1}^{\mathrm{DFP}}$ 或 $\mathbf{H}_{k+1}^{\mathrm{BFGS}}$ 也对称正定。
如果 $\mathbf{B}_k$ 对称正定，如果 $\mathbf{y}_k^\top\mathbf{s}_k>0$，则构造出的 $\mathbf{B}_{k+1}^{\mathrm{DFP}}$ 或 $\mathbf{B}_{k+1}^{\mathrm{BFGS}}$ 也对称正定。

证明. 就以 $\mathbf{H}_{k+1}^{\mathrm{DFP}}$ 为例：$\forall\mathbf{x}\in\mathbb{R}^n$，

$$\begin{aligned}\mathbf{x}^\top\mathbf{H}_{k+1}^{\mathrm{DFP}}\mathbf{x}&=\mathbf{x}^\top\mathbf{H}_k\mathbf{x}+\frac{(\mathbf{x}^\top\mathbf{s}_k)^2}{\mathbf{s}_k^\top\mathbf{y}_k}-\frac{(\mathbf{x}^\top\mathbf{H}_k\mathbf{y}_k)^2}{\mathbf{y}_k^\top\mathbf{H}_k\mathbf{y}_k}\\&=\frac{(\mathbf{x}^\top\mathbf{H}_k\mathbf{x})(\mathbf{y}_k^\top\mathbf{H}_k\mathbf{y}_k)-(\mathbf{x}^\top\mathbf{H}_k\mathbf{y}_k)^2}{\mathbf{y}_k^\top\mathbf{H}_k\mathbf{y}_k}+\frac{(\mathbf{x}^\top\mathbf{s}_k)^2}{\mathbf{s}_k^\top\mathbf{y}_k}\\(\text{Cauchy-Schwartz})&\geq\frac{(\mathbf{x}^\top\mathbf{s}_k)^2}{\mathbf{s}_k^\top\mathbf{y}_k}>0\end{aligned}$$

□

上述定理要求 $\mathbf{s}_k^\top\mathbf{y}_k>0$，那么如何保证这一点？**关于此，有以下定理**

保证 $\mathbf{s}_k^\top\mathbf{y}_k>0$ 的定理

对于使用精确线搜索或非精确线搜索 Wolfe 准则的 DFP 方法和 BFGS 方法，有 $\mathbf{s}_k^\top\mathbf{y}_k>0$。

证明. 对于精确线搜索方法 $\alpha_k=\arg\min_\alpha f(\mathbf{x}_k+\alpha\mathbf{d}_k)$，其能导出 $\mathbf{g}_{k+1}^\top\mathbf{d}_k=0$。另外，$\mathbf{s}_k=\mathbf{x}_{k+1}-\mathbf{x}_k=\alpha_k\mathbf{d}_k$，那么

$$\mathbf{s}_k^\top\mathbf{y}_k=\alpha_k\mathbf{d}_k^\top(\mathbf{g}_{k+1}-\mathbf{g}_k)=-\alpha_k(-\mathbf{H}_k\mathbf{g}_k)^\top\mathbf{g}_k=\alpha_k\mathbf{g}_k^\top\mathbf{H}_k\mathbf{g}_k>0$$

对于 Wolfe 准则的非精确线搜索，

$$\mathbf{s}_k^\top\mathbf{y}_k=\alpha_k(\mathbf{d}_k^\top\mathbf{g}_{k+1}-\mathbf{d}_k^\top\mathbf{g}_k)\geq\alpha_k(\sigma-1)\mathbf{d}_k^\top\mathbf{g}_k>0$$

□

需要注意的是，以上定理对强 Wolfe 准则也成立，**但 Goldstein 准则不能保证 $\mathbf{s}_k^\top\mathbf{y}_k>0$，所以拟 Newton 方法中，一般我们不使用 Goldstein 准则。**

BFGS 和 DFP 方法的收敛性

假设初始矩阵 $\mathbf{B}_0$ 对称正定，目标函数 $f(\mathbf{x})$ 二阶连续可微，其水平集 $L = \{\mathbf{x}|f(\mathbf{x}) \leq f(\mathbf{x}_0)\}$ 凸，则 $\exists m, M \in \mathbb{R}^+$ 使得 $\forall \mathbf{z} \in \mathbb{R}^n$，$\mathbf{x} \in L$，且满足

$$m\|\mathbf{z}\|^2 \leq \mathbf{z}^\top \mathbf{G}(\mathbf{x})\mathbf{z} \leq M\|\mathbf{z}\|^2$$

则使用精确线搜索或者非精确线搜索 Wolfe 准则的 BFGS 方法和 DFP 方法具有**全局收敛性**。

BFGS 和 DFP 方法的收敛速度

如果 $\mathbf{G}(\mathbf{x})$ 在极小点 $\mathbf{x}^*$ 上具有 Lipschitz 连续性，则 BFGS 和 DFP 方法以**超线性速度**收敛到 $\mathbf{x}^*$

以 BFGS 格式为代表的拟 Newton 类算法由于仅仅使用了 $\mathbf{G}(\mathbf{x})$ 的近似，因此很难达到二阶收敛速度，最多只能达到超线性收敛速度。但是，由于拟 Newton 方法对近似矩阵的更新代价 ($O(n^2)$) 可能远小于 Newton 方法计算 $\mathbf{G}(\mathbf{x})$ 的代价（$O(n^3)$），因此它在大规模问题中的开销可能远小于牛顿算法，更为实用。

证明. (证明见 https://bicmr.pku.edu.cn/ wenzw/optbook/lect/12-lect-QN.pdf，考试可能不作要求) □

BFGS 和 DFP 法的算法复杂度

迭代过程可以描述为

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{B}_k^{-1}\mathbf{g}_k$$
$$\mathbf{B}_{k+1} \leftarrow \mathbf{B}_k + \Delta\mathbf{B}_k$$

或

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{H}_k\mathbf{g}_k$$
$$\mathbf{H}_{k+1} \leftarrow \mathbf{H}_k + \Delta\mathbf{H}_k$$

如果直接采用 $\mathbf{H}_k$，则迭代的计算复杂度是 $O(n^2)$(只涉及 $\mathbb{R}^{n\times n}$ 和 $\mathbb{R}^n$ 的矩阵乘法).
如果采用 $\mathbf{B}_k$，用 Gauss 消元法求逆，则迭代的计算复杂度为 $O(n^3)$(求逆的运算为 $O(n^3)$ 的).
如果用对 $\mathbf{B}_k$ 作 LDL 分解: $\mathbf{B}_k = \mathbf{L}_k\mathbf{D}_k\mathbf{L}_k^\top$，储存下三角阵 $\mathbf{L}_k$ 和对角阵 $\mathbf{D}_k$，每次迭代时更新 $\mathbf{L}_k$ 和 $\mathbf{D}_k$ 即可，迭代的计算复杂度为 $O(n^2)$.

[注]：逆运算的复杂度为 $O(n^3)$ 的解释如下：高斯消元法首先将矩阵转换为上三角形式。在这个过程中，对于每一列，我们需要将其下面的所有元素消为 0。对于第 k 列，需要处理 $n-k$ 行，每行需要执行 $n-k$ 次乘法和减法操作。总操作数为

$$\sum_{k=1}^{n-1}(n-k)^2 \sim \sum_{k=1}^{n-1}k^2 \sim \frac{n(n-1)(2n-1)}{6} \sim O(n^3)$$

[注]：另一个用 $\mathbf{B}_k$ 而非 $\mathbf{H}_k$ 的好处在于，每次更新完 $\mathbf{D}_k$ 之后，如果其对角元素全为正，则能直观判断 $\mathbf{B}_k$ 正定；如果有负值，则修改为一个很小的正值即可。但采用 $\mathbf{H}_k$ 就无法如此直观的判断。

4.6.3 L-BFGS (考试不考)

尽管 BFGS 等近似 Hessian 矩阵 $\mathbf{G}$ 的方法降低了直接对 Hessian 矩阵求逆的复杂度（$O(n^3) \to O(n^2)$，但 $\mathbf{B}_k$ 或 $\mathbf{H}_k$ 仍然需要 $O(n^2)$ 规模的存储空间，当 n 很大时仍然是困难的。对此，研究者提出 Limited-memory BFGS (L-BFGS)，**只保存最近 m 次迭代的信息来构造 Hessian 矩阵（或其逆矩阵）的近似矩阵，从而大大减少数据的存储空间**。具体原理如下：令 $\rho_k = \dfrac{1}{\mathbf{y}_k^T \mathbf{s}_k}$，其中 $\mathbf{s}_k = \mathbf{x}_{k+1} - \mathbf{x}_k$, $\mathbf{y}_k = \mathbf{g}_{k+1} - \mathbf{g}_k$，则 $\mathbf{H}_{k+1}$ 的 BFGS 公式可写为：

$$\mathbf{H}_{k+1} = \left(\mathbf{I} - \rho_k \mathbf{s}_k \mathbf{y}_k^T\right) \mathbf{H}_k \left(\mathbf{I} - \rho_k \mathbf{y}_k \mathbf{s}_k^T\right) + \rho_k \mathbf{s}_k \mathbf{s}_k^T.$$

令 $\mathbf{V}_k = \mathbf{I} - \rho_k \mathbf{y}_k \mathbf{s}_k^T$，则

$$\mathbf{H}_{k+1} = \mathbf{V}_k^T \mathbf{H}_k \mathbf{V}_k + \rho_k \mathbf{s}_k \mathbf{s}_k^T$$

用这个递推公式，不难发现可以只用 $\{\mathbf{s}_i, \mathbf{y}_i\}_{i=1}^k$ 将 $\mathbf{H}_{k+1}$ 表示为

$$\begin{aligned}
\mathbf{H}_{k+1} = & \left(\mathbf{V}_k^\top \mathbf{V}_{k-1}^\top \cdots \mathbf{V}_1^\top \mathbf{V}_0^\top\right) \mathbf{H}_0 \left(\mathbf{V}_0 \mathbf{V}_1 \cdots \mathbf{V}_{k-1} \mathbf{V}_k\right) \\
& + \left(\mathbf{V}_k^\top \mathbf{V}_{k-1}^\top \cdots \mathbf{V}_2^\top \mathbf{V}_1^\top\right) \left(\rho_0 \mathbf{s}_0 \mathbf{s}_0^\top\right) \left(\mathbf{V}_1 \mathbf{V}_2 \cdots \mathbf{V}_{k-1} \mathbf{V}_k\right) \\
& + \left(\mathbf{V}_k^\top \mathbf{V}_{k-1}^\top \cdots \mathbf{V}_3^\top \mathbf{V}_2^\top\right) \left(\rho_1 \mathbf{s}_1 \mathbf{s}_1^\top\right) \left(\mathbf{V}_2 \mathbf{V}_3 \cdots \mathbf{V}_{k-1} \mathbf{V}_k\right) \\
& + \cdots \\
& + \left(\mathbf{V}_k^\top \mathbf{V}_{k-1}^\top\right) \left(\rho_{k-2} \mathbf{s}_{k-2} \mathbf{s}_{k-2}^\top\right) \left(\mathbf{V}_{k-1} \mathbf{V}_k\right) \\
& + \mathbf{V}_k^\top \left(\rho_{k-1} \mathbf{s}_{k-1} \mathbf{s}_{k-1}^\top\right) \mathbf{V}_k \\
& + \rho_k \mathbf{s}_k \mathbf{s}_k^\top
\end{aligned}$$

但是这个储存量还是太大了，我们只希望保留最近的 m 个 $\{\mathbf{s}_i, \mathbf{y}_i\}$。因此，考虑舍弃那些最早生成的向量。例如，计算 $\mathbf{H}_{m+1}$ 时，保存的信息为 $\{\mathbf{s}_i, \mathbf{y}_i\}_{i=1}^m$，舍弃了 $\{\mathbf{s}_0, \mathbf{y}_0\}$；计算 $\mathbf{H}_{m+2}$ 时，保存的信息为 $\{\mathbf{s}_i, \mathbf{y}_i\}_{i=2}^m$，舍弃了 $\{\mathbf{s}_i, \mathbf{y}_i\}_{i=0}^1$。以此类推。不妨记 $\hat{m} = \min\{k, m-1\}$(这意味着当 $k \le m-1$，时，我们用满 k 个最近的信息，当 $k > m$ 时，我们只用最近的 $m-1$ 个信息)

$$\begin{aligned}
\mathbf{H}_{k+1} = & \left(\mathbf{V}_k^\top \mathbf{V}_{k-1}^\top \cdots \mathbf{V}_{k-\hat{m}+1}^\top \mathbf{V}_{k-\hat{m}}^\top\right) \mathbf{H}_{k+1}^0 \left(\mathbf{V}_{k-\hat{m}} \mathbf{V}_{k-\hat{m}+1} \cdots \mathbf{V}_{k-1} \mathbf{V}_k\right) \\
& + \left(\mathbf{V}_k^\top \mathbf{V}_{k-1}^\top \cdots \mathbf{V}_{k-\hat{m}+2}^\top \mathbf{V}_{k-\hat{m}+1}^\top\right) \left(\rho_0 \mathbf{s}_0 \mathbf{s}_0^\top\right) \left(\mathbf{V}_{k-\hat{m}+1} \mathbf{V}_{k-\hat{m}+2} \cdots \mathbf{V}_{k-1} \mathbf{V}_k\right) \\
& + \left(\mathbf{V}_k^\top \mathbf{V}_{k-1}^\top \cdots \mathbf{V}_{k-\hat{m}+3}^\top \mathbf{V}_{k-\hat{m}+2}^\top\right) \left(\rho_1 \mathbf{s}_1 \mathbf{s}_1^\top\right) \left(\mathbf{V}_{k-\hat{m}+2} \mathbf{V}_{k-\hat{m}+3} \cdots \mathbf{V}_{k-1} \mathbf{V}_k\right) \\
& + \cdots \\
& + \left(\mathbf{V}_k^\top \mathbf{V}_{k-1}^\top\right) \left(\rho_{k-2} \mathbf{s}_{k-2} \mathbf{s}_{k-2}^\top\right) \left(\mathbf{V}_{k-1} \mathbf{V}_k\right) \\
& + \mathbf{V}_k^\top \left(\rho_{k-1} \mathbf{s}_{k-1} \mathbf{s}_{k-1}^\top\right) \mathbf{V}_k \\
& + \rho_k \mathbf{s}_k \mathbf{s}_k^\top.
\end{aligned}$$

4.6.4 Broyden 族方法

Broyden 族方法

将 BFGS 和 DFP 公式进行加权求和，就得到了 Broyden 族公式：

$$\mathbf{B}_{k+1}^{\text{Broyden}} = (1 - \phi_k)\mathbf{B}_{k+1}^{\text{BFGS}} + \phi_k \mathbf{B}_{k+1}^{\text{DFP}}$$

[注]: 显然，BFGS($\phi_k = 0$)、DFP($\phi_k = 1$) 都属于 Broyden 族。

[注]: SR1 也属于 Broyden 族，此时 $\phi_k = \dfrac{\mathbf{s}_k^\top \mathbf{y}_k}{\mathbf{s}_k^\top \mathbf{y}_k - \mathbf{s}_k^\top \mathbf{B}_k \mathbf{s}_k} \notin [0,1]$。

[注]: 对 Broyden 族公式作变形

$$\mathbf{B}_{k+1}^{\text{Broyden}} = \mathbf{B}_{k+1}^{\text{BFGS}} + \phi_k(\mathbf{B}_{k+1}^{\text{DFP}} - \mathbf{B}_{k+1}^{\text{BFGS}}) = \mathbf{B}_{k+1}^{\text{BFGS}} + \phi_k \mathbf{v}_k \mathbf{v}_k^\top$$

其中 $\mathbf{v}_k = (\mathbf{s}_k^\top \mathbf{B}_k \mathbf{s}_k)^{\frac{1}{2}} \left(\dfrac{\mathbf{y}_k}{\mathbf{y}_k^\top \mathbf{s}_k} - \dfrac{\mathbf{B}_k \mathbf{s}_k}{\mathbf{s}_k^\top \mathbf{B}_k \mathbf{s}_k} \right)$，这说明 Broyden 族公式与 BFGS 方法也就相差一个秩为 1 的矩阵。

Broyden 族方法的对称正定性

假设 $\mathbf{B}_{k+1}^{\text{BFGS}}$ 为对称正定矩阵，当 $\phi_k \geq 0$ 时得到的 Broyden 族公式得到的 $\mathbf{B}_{k+1}$ 也为对称正定矩阵。

证明. 由 $\mathbf{B}_{k+1}^{\text{Broyden}} = \mathbf{B}_{k+1}^{\text{BFGS}} + \phi_k \mathbf{v}_k \mathbf{v}_k^\top$ 易知 $\mathbf{B}_{k+1}^{\text{Broyden}}$ 也对称正定。 □

Broyden 族方法的二次终止性

设 $f(\mathbf{x}) = \dfrac{1}{2}\mathbf{x}^\top \mathbf{G}\mathbf{x} + \mathbf{b}^\top \mathbf{x}$，其中 $\mathbf{G}$ 为对称正定阵。对任意初始点 $\mathbf{x}_0$ 和任意对称正定阵 $\mathbf{B}_0 \in \mathbb{R}^{n\times n}$，$\alpha_k$ 为精确线搜索得到的步长，则 Broyden 族的拟 Newton 方法最多经过 n 次迭代，可求得二次函数 $f(\mathbf{x})$ 的极小点。此外，如果确实迭代了 n 步，那么 $\mathbf{B}_n = \mathbf{G}$。

证明. 证明见 4.7 共轭梯度法中的介绍 □

4.7 共轭梯度法

4.7.1 变度量意义的最速下降法

之前介绍的最速下降法是 $\|\cdot\|_2$ 范数度量意义下的最速下降，这仅局限于当前迭代点的局部范围内是下降最快的，整体上看，迭代效果并不是很好 (回忆: 函数的等高线是非常扁的椭圆的情况)。自然想到，是否度量变了，最速下降的方向也会不同? 效果更好呢? 如图 12所示，对于等高线为左图的二次型 $\dfrac{1}{2}\mathbf{x}^\top \mathbf{G}\mathbf{x}$，如果对向量 $\mathbf{x}$ 作线性变换 $\mathbf{x} \mapsto \mathbf{W}\mathbf{x}$，能否将等高线变为圆的，从而方便优化?

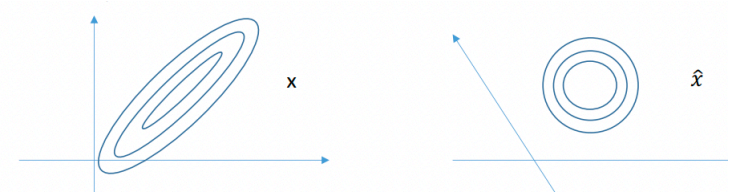


图 12: 变度量意义的最速下降法

设 $\hat{\mathbf{x}} = \mathbf{W}\mathbf{x}$，则变换后梯度 $g(\hat{\mathbf{x}}) = (\mathbf{W}^{-1})^\top g(\mathbf{x})$。如果在变换后的空间，等高线变为标准的圆，那么

$$\frac{1}{2}\hat{\mathbf{x}}^\top \hat{\mathbf{x}} = \frac{1}{2}\mathbf{x}^\top \mathbf{W}^\top \mathbf{W}\mathbf{x}$$

那么这要求 $\mathbf{W}$ 满足 $\mathbf{W}^\top \mathbf{W} = \mathbf{G}$。从优化的角度看，原空间的牛顿方向

$$\mathbf{d}_k = -\mathbf{G}_k^{-1}\mathbf{g}_k = -\mathbf{W}^{-1}(\mathbf{W}^\top)^{-1}\mathbf{g}_k$$

从而

$$\hat{\mathbf{d}}_k = \mathbf{W}\mathbf{d}_k = -(\mathbf{W}^\top)^{-1}\mathbf{g}_k = g(\hat{\mathbf{x}}_k)$$

即**原始空间里的 Newton 方向等价于变换空间的负梯度方向**。

- 原始空间的最速下降法不考虑函数的形状，在周围一个很小的“圆”里找到最快下降的方向
- Newton 法计算出了曲率，可以认知周围的函数形状，在周围一个很小的“椭圆”里找到最快下降的方向。
- 拟 Newton 法也要计算曲率，只是曲率是一个近似的曲率。

4.7.2 共轭梯度法的基本概念

如图所示，假如我们在 3 维空间内有一个正定二次函数 $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top\mathbf{G}\mathbf{x} + \mathbf{b}^\top\mathbf{x}$，等高线为椭圆面。从任意初始点 $\mathbf{x}_0$ 出发，沿着其短轴方向 $\mathbf{d}_0$，长轴方向 $\mathbf{d}_1$，纵轴方向 $\mathbf{d}_2$ 作精确线搜索，只需要 3 步就能走到最优点 $\mathbf{x}^*$。这三个方向 $\mathbf{d}_0, \mathbf{d}_1, \mathbf{d}_2$ 两两正交，被称为**共轭方向**。

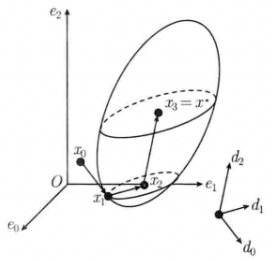


图 13: 共轭方向

共轭方向的重要性

对于 $\mathbb{R}^n$ 上一般正定的二次函数而言，依次沿着共轭方向迭代就可以在最多 n 步内得到函数的极小点。

证明. 回顾正定二次函数精确线搜索的步长

$$\alpha_k = \frac{-\mathbf{g}_k^\top\mathbf{d}_k}{\mathbf{d}_k^\top\mathbf{G}\mathbf{d}_k} = \frac{\mathbf{d}_k(\mathbf{G}\mathbf{x}_k - \mathbf{b})}{\mathbf{d}_k^\top\mathbf{G}\mathbf{d}_k}$$

由于共轭方向线性无关，因此这组方向构成 $\mathbb{R}^n$ 的一组基，又 $\mathbf{x}^* - \mathbf{x}_0 = \beta_0\mathbf{d}_0 + \cdots + \beta_{n-1}\mathbf{d}_{n-1}$，等式两端同时左乘 $\mathbf{d}_k^\top\mathbf{G}$ 可得

$$\beta_k = \frac{\mathbf{d}_k^\top\mathbf{G}(\mathbf{x}^* - \mathbf{x}_0)}{\mathbf{d}_k^\top\mathbf{G}\mathbf{d}_k}$$

根据迭代过程，$\mathbf{x}_k - \mathbf{x}_0 = \alpha_0\mathbf{d}_0 + \cdots + \alpha_{k-1}\mathbf{d}_{k-1}$，等式两端同时左乘 $\mathbf{d}_k^\top\mathbf{G}$ 可得

$$\mathbf{d}_k^\top\mathbf{G}(\mathbf{x}_k - \mathbf{x}_0) = 0$$

于是

$$\begin{aligned}\beta_k &= \frac{\mathbf{d}_k^\top\mathbf{G}(\mathbf{x}^* - \mathbf{x}_0)}{\mathbf{d}_k^\top\mathbf{G}\mathbf{d}_k} = \frac{\mathbf{d}_k^\top\mathbf{G}(\mathbf{x}^* - \mathbf{x}_k + \mathbf{x}_k - \mathbf{x}_0)}{\mathbf{d}_k^\top\mathbf{G}\mathbf{d}_k} = \frac{\mathbf{d}_k^\top\mathbf{G}(\mathbf{x}^* - \mathbf{x}_k)}{\mathbf{d}_k^\top\mathbf{G}\mathbf{d}_k} \\ &= \frac{-\mathbf{d}_k^\top(\mathbf{G}\mathbf{x}_k - \mathbf{b})}{\mathbf{d}_k^\top\mathbf{G}\mathbf{d}_k} = \alpha_k\end{aligned}$$

这说明了经过精确线搜索的每一步迭代后，$\mathbf{x}_0$ 就能到达最优点 $\mathbf{x}^*$。得证。 □

那么，这样的共轭方向怎么求？**我们可以将原空间仿射变换到一个新的 $\mathbb{R}^n$ 空间，使得在新空间中，原本的 n 个共轭方向恰好平行于 n 根坐标轴**。为实现此，假设映射 f 将新空间的点 $\tilde{\mathbf{x}}$ 映射到原空间 $\mathbf{x}$

$$f:\mathbb{R}^n\mapsto\mathbb{R}^n,\ \mathbf{x}=\mathbf{W}\tilde{\mathbf{x}}$$

其中 $\mathbf{W}=[\mathbf{d}_0,\cdots,\mathbf{d}_{n-1}]$ 于是

$$g(\tilde{\mathbf{x}})=f(\mathbf{W}\tilde{\mathbf{x}})=\frac{1}{2}\tilde{\mathbf{x}}^\top(\mathbf{W}^\top\mathbf{G}\mathbf{W})\tilde{\mathbf{x}}+(\mathbf{b}^\top\mathbf{W})\tilde{\mathbf{x}}$$

在新空间中，变换后的正定二次函数各轴与坐标轴平行，那么 $\mathbf{W}^\top\mathbf{G}\mathbf{W}$ 必然是一个对角阵。非对角处为 0，即 $\mathbf{d}_i^\top\mathbf{G}\mathbf{d}_j=0,\ \forall i\neq j$。由此，引出如下关于共轭方向的定义。

G 共轭

设 $\mathbf{G}$ 是对阵正定矩阵，若非零向量组 $\{\mathbf{d}_0,\cdots,\mathbf{d}_{n-1}\}$ 满足 $\mathbf{d}_i^\top\mathbf{G}\mathbf{d}_j=0,\ i\neq j$，则称这个非零向量组是矩阵 $\mathbf{G}$ 的共轭方向，简称为 $\mathbf{G}$ 共轭。

- 正交方向是共轭方向的特例，$\mathbf{G}=\mathbf{I}$
- 显然，共轭向量组中的向量必然线性无关。(若不然，设 $\mathbf{d}_i=\sum\limits_{j\neq i}\alpha_j\mathbf{d}_j$，则 $\mathbf{d}_i^\top\mathbf{G}\mathbf{d}_i=\sum\limits_{j\neq i}\alpha_j\mathbf{d}_j^\top\mathbf{G}\mathbf{d}_i=0$，与共轭性矛盾)

子空间扩展定理

设正定二次函数 $f(\mathbf{x})=\frac{1}{2}\mathbf{x}^\top\mathbf{G}\mathbf{x}+\mathbf{b}^\top\mathbf{x}$，$\mathbf{d}_0,\cdots,\mathbf{d}_{n-1}$ 是 $\mathbf{G}$ 的共轭方向。由任意的 $\mathbf{x}_0$ 出发，依次沿着 $\mathbf{x}_k+\alpha\mathbf{d}_k$ 作精确线搜索得 α_k，则

$$\mathbf{g}_k^\top\mathbf{d}_j=0,\ j=0,\cdots,k-1$$

且 $\mathbf{x}_k$ 是 $f(\mathbf{x})$ 在集合 $X_k=\{\mathbf{x}|\mathbf{x}=\mathbf{x}_0+\mathrm{span}\{\mathbf{d}_0,\cdots,\mathbf{d}_k\}\}=\{\mathbf{x}|\mathbf{x}=\mathbf{x}_0+\sum\limits_{j=0}^{k-1}\beta_j\mathbf{d}_j\}$ 上的极小点。

- **这个定理说明，$\mathbf{g}_k$ 与之前所有的共轭更新方向都正交！(精确线搜索只说 $\mathbf{g}_k$ 与 $\mathbf{d}_{k-1}$ 正交)**

证明. 首先证明

$$\mathbf{g}_k^\top\mathbf{d}_j=0,\ j=0,\cdots,k-1$$

当 $j=k-1$ 时，由精确线搜索的推论易知成立。对于 $j=0,\cdots,k-2$，注意到 $\mathbf{g}_i=\mathbf{G}\mathbf{x}_i+\mathbf{b}$，

$$\begin{aligned}\mathbf{g}_k=&\mathbf{g}_{j+1}+\sum_{i=j+1}^{k-1}(\mathbf{g}_{i+1}-\mathbf{g}_i)=\mathbf{g}_{j+1}+\mathbf{G}\sum_{i=j+1}^{k-1}(\mathbf{x}_{i+1}-\mathbf{x}_i)\\=&\mathbf{g}_{j+1}+\mathbf{G}\sum_{i=j+1}^{k-1}\alpha_i\mathbf{d}_i.\end{aligned}$$

由精确线搜索的结果及 $\mathbf{d}_0,\mathbf{d}_1,\cdots,\mathbf{d}_{k-1}$ 的两两共轭性有

$$\mathbf{g}_k^\top\mathbf{d}_j=\mathbf{g}_{j+1}^\top\mathbf{d}_j+\sum_{i=j+1}^{k-1}\alpha_i\mathbf{d}_i^\top\mathbf{G}\mathbf{d}_j=0.$$

对于定理的第二个结论，只要证明对任给 $\mathbf{x}\in X_k$, $f(\mathbf{x})\geq f(\mathbf{x}_k)$ 即可。注意到

$$\mathbf{x}_k=\mathbf{x}_0+\sum_{j=0}^{k-1}\alpha_j\mathbf{d}_j,\quad \mathbf{x}=\mathbf{x}_0+\sum_{j=0}^{k-1}\beta_j\mathbf{d}_j,\quad \forall\mathbf{x}\in X_k,$$

由 $\mathbf{G}$ 的正定性和 $\mathbf{g}_k^\top\mathbf{d}_j=0,\ j=0,\cdots,k-1$，我们对 $f(\mathbf{x})$ 在 $\mathbf{x}_k$ 附近展开

$$\begin{aligned}f(\mathbf{x})&=f(\mathbf{x}_k)+(\mathbf{x}-\mathbf{x}_k)^\top\mathbf{g}_k+\frac{1}{2}(\mathbf{x}-\mathbf{x}_k)^\top\mathbf{G}(\mathbf{x}-\mathbf{x}_k)\\&\geq f(\mathbf{x}_k)+(\mathbf{x}-\mathbf{x}_k)^\top\mathbf{g}_k\\&=f(\mathbf{x}_k)+\sum_{j=0}^{k-1}(\beta_j\mathbf{d}_j-\alpha_j\mathbf{d}_j)^\top\mathbf{g}_k\\&=f(\mathbf{x}_k)+\sum_{j=0}^{k-1}(\beta_j-\alpha_j)\mathbf{d}_j^\top\mathbf{g}_k\\&=f(\mathbf{x}_k)\end{aligned}$$

由此知定理结论成立。 □

4.7.3 线性共轭梯度法

由于求解正定二次函数的极小化问题 $\min\frac{1}{2}\mathbf{x}^\top\mathbf{G}\mathbf{x}-\mathbf{b}^\top\mathbf{x}$，等价于求解线性方程组 $\mathbf{G}\mathbf{x}=\mathbf{b}$，因此把求解正定二次函数极小化问题的共轭梯度法称为**线性共轭梯度法**。

线性共轭梯度法

线性共轭梯度法中，按照如下方法构造共轭方向：规定 $\mathbf{d}_0=-\mathbf{g}_0$，

$$\mathbf{d}_k=-\mathbf{g}_k+\xi_{k-1}\mathbf{d}_{k-1}=-\mathbf{g}_k+\frac{\mathbf{g}_k^\top\mathbf{g}_k}{\mathbf{g}_{k-1}^\top\mathbf{g}_{k-1}}\mathbf{d}_{k-1}$$

这样生成的新方向 $\mathbf{d}_k$ 仅用到上一个方向 $\mathbf{d}_{k-1}$，而无需知道之前所有的方向，并且能够自动保证 $\mathbf{d}_k$ 与之前所有方向共轭。这样我们不需要储存 $\mathbf{d}_i$ $(i=1,\cdots,k-2)$，是一种更高效的构造。

优化算法迭代步骤如下

$$\mathbf{x}_{k+1}=\mathbf{x}_k+\alpha_k\mathbf{d}_k=\mathbf{x}_k+\frac{\mathbf{g}_k^\top\mathbf{g}_k}{\mathbf{d}_k^\top\mathbf{G}\mathbf{d}_k}\mathbf{d}_k$$

证明. 首先我们给出 $\mathbf{d}_k$ 的构造方法：由 $\mathbf{d}_k$ 与 $\mathbf{d}_0,\cdots,\mathbf{d}_{k-1}$ 线性无关，且 $\mathbf{g}_k\perp\mathbf{d}_0,\cdots,\mathbf{d}_{n-1}$，所以 $\mathbf{d}_k$ 可以表示为 $\mathbf{g}_k,\mathbf{d}_0,\cdots,\mathbf{d}_{n-1}$ 的线性组合：

$$\mathbf{d}_k=-\mathbf{g}_k+\sum_{i=0}^{k-1}\xi_i\mathbf{d}_i$$

这里特意让 $\mathbf{g}_k$ 前系数为 -1。由共轭方向的定义，我们需要确定系数使得 $\mathbf{d}_k^\top\mathbf{G}\mathbf{d}_j=0,\ j=0,\cdots,k-1$，代入上式可以反解出

$$\xi_j=\frac{\mathbf{g}_k^\top\mathbf{G}\mathbf{d}_j}{\mathbf{d}_j^\top\mathbf{G}\mathbf{d}_j},j=0,\cdots,k-1$$

可以证明：$\xi_j = 0, j = 0, \cdots, k-2$，从而 $\mathbf{d}_k = -\mathbf{g}_k + \xi_{k-1}\mathbf{d}_{k-1}$。证明如下：由 $\mathbf{x}_{j+1} - \mathbf{x}_j = \alpha_j \mathbf{d}_j$ 和 $\mathbf{g}_j = \mathbf{G}\mathbf{x}_j + b$ 知，对 ξ_j **的分子乘以** α_j 得

$$\begin{aligned}\alpha_j \mathbf{g}_k^\top \mathbf{G}\mathbf{d}_j &= \mathbf{g}_k^\top \mathbf{G}(\mathbf{x}_{j+1} - \mathbf{x}_j) \\ &= \mathbf{g}_k^\top (\mathbf{g}_{j+1} - \mathbf{g}_j).\end{aligned}$$

由

$$\mathbf{d}_k = -\mathbf{g}_k + \sum_{i=0}^{k-1} \xi_i \mathbf{d}_i \Rightarrow \mathbf{g}_j = -\mathbf{d}_j + \sum_{i=0}^{j-1} \xi_i \mathbf{d}_i$$

再由子空间扩展定理（这里假设了 $\mathbf{G}$ 是正定对称的）知

$$\mathbf{g}_k^\top \mathbf{g}_j = \mathbf{g}_k^\top (-\mathbf{d}_j + \sum_{i=0}^{j-1} \xi_i \mathbf{d}_i) = 0, \quad j = 0, \cdots, k-1,$$

从而

$$\mathbf{g}_k^\top (\mathbf{g}_{j+1} - \mathbf{g}_j) = \begin{cases} 0, & j = 0, \cdots, k-2, \\ \mathbf{g}_k^\top \mathbf{g}_k, & j = k-1. \end{cases}$$

这就证明了 ξ_j 的分子为 0，即 $\xi_j = 0\, j = 0, \cdots, k-2$，从而

$$\mathbf{d}_k = -\mathbf{g}_k + \xi_{k-1}\mathbf{d}_{k-1}.$$

再来看 ξ_{k-1} 的表示。对 ξ_{k-1} 的分母乘以 α_{k-1} 为

$$\begin{aligned}\alpha_{k-1}\mathbf{d}_{k-1}^\top \mathbf{G}\mathbf{d}_{k-1} &= \mathbf{d}_{k-1}^\top \mathbf{G}(\mathbf{x}_k - \mathbf{x}_{k-1}) = \mathbf{d}_{k-1}^\top (\mathbf{g}_k - \mathbf{g}_{k-1}) = -\mathbf{d}_{k-1}^\top \mathbf{g}_{k-1} \\ &= (\mathbf{g}_{k-1} - \xi_{k-2}\mathbf{d}_{k-2})^\top \mathbf{g}_{k-1} \\ &= \mathbf{g}_{k-1}^\top \mathbf{g}_{k-1}, \quad k \geq 2.\end{aligned}$$

对 $k = 1$，回顾 $\mathbf{d}_0 = -\mathbf{g}_0$，有

$$\alpha_0 \mathbf{d}_0^\top \mathbf{G}\mathbf{d}_0 = \mathbf{d}_0^\top (\mathbf{g}_1 - \mathbf{g}_0) = \mathbf{g}_0^\top \mathbf{g}_0.$$

因此，ξ_{k-1} 可表示为

$$\xi_{k-1} = \frac{\mathbf{g}_k^\top \mathbf{g}_k}{\mathbf{g}_{k-1}^\top \mathbf{g}_{k-1}}.$$

从而 $\mathbf{d}_k = -\mathbf{g}_k + \dfrac{\mathbf{g}_k^\top \mathbf{g}_k}{\mathbf{g}_{k-1}^\top \mathbf{g}_{k-1}}\mathbf{d}_{k-1}$。

我们知道正定二次函数在精确线搜索下的步长为 $\alpha_k = -\dfrac{\mathbf{g}_k^\top \mathbf{d}_k}{\mathbf{d}_k^\top \mathbf{G}\mathbf{d}_k}$，故应把共轭方向代入上所求得的步长公式中。

由子空间拓展定理，$\mathbf{g}_k^\top \mathbf{d}_i = 0,\ i = 0, 1, \cdots, k-1$。那么对 $\mathbf{d}_k = -\mathbf{g}_k + \dfrac{\mathbf{g}_k^\top \mathbf{g}_k}{\mathbf{g}_{k-1}^\top \mathbf{g}_{k-1}}\mathbf{d}_{k-1}$ 两边同时左乘 $\mathbf{g}_k^\top$，得

$$\mathbf{g}_k^\top \mathbf{d}_k = -\mathbf{g}_k^\top \mathbf{g}_k + \frac{\mathbf{g}_k^\top \mathbf{g}_k}{\mathbf{g}_{k-1}^\top \mathbf{g}_{k-1}} \mathbf{g}_k^\top \mathbf{d}_{k-1} = -\mathbf{g}_k^\top \mathbf{g}_k$$

那么步长可以化简为 $\alpha_k = \dfrac{\mathbf{g}_k^\top \mathbf{g}_k}{\mathbf{d}_k^\top \mathbf{G}\mathbf{d}_k}$ □

线性共轭梯度法的性质

考虑正定二次函数 $\frac{1}{2}\mathbf{x}^\top\mathbf{G}\mathbf{x}+\mathbf{b}^\top\mathbf{x}$，对任意初始点 $\mathbf{x}_0$，**取 $\mathbf{d}_0=-\mathbf{g}_0$(必须这么取)**，则

1. 采用精确线搜索的共轭梯度法具有二次终止性。

假定共轭梯度方法经过 k 步迭代未达到极小点，则下列关系成立：

1. 共轭性：$\mathbf{d}_k^\top\mathbf{G}\mathbf{d}_i=0,\ i=0,\cdots,k-1$
2. 正交性：$\mathbf{g}_k^\top\mathbf{g}_i=0,\ i=0,\cdots,k-1$
3. 下降性：$\mathbf{g}_k^\top\mathbf{d}_k=-\mathbf{g}_k^\top\mathbf{g}_k<0$
4. $\mathrm{span}\{\mathbf{g}_0,\cdots,\mathbf{g}_k\}=\mathrm{span}\{\mathbf{d}_0,\cdots,\mathbf{d}_k\}=\mathrm{span}\{\mathbf{g}_0,\cdots,\mathbf{G}^k\mathbf{g}_k\}$，即 $\{\mathbf{g}_0,\cdots,\mathbf{g}_k\}$ 和 $\{\mathbf{d}_0,\cdots,\mathbf{d}_k\}$ 是 Krylov 空间 $\{\mathbf{g}_0,\cdots,\mathbf{G}^k\mathbf{g}_k\}$ 的一组正交基和一组共轭正交基。（梯度张成的空间和迭代方向张成的空间相同）

[注]: 如果初始共轭梯度方向没有选 $\mathbf{d}_0=-\mathbf{g}_0$，则线性共轭梯度法生成的迭代方向 $\mathbf{d}_k$ 可能不共轭。

证明. 1,2 的证明已经蕴含在线性共轭梯度方向的构造中。3 的证明是显然的

$$\mathbf{g}_k^\top\mathbf{d}_k=-\mathbf{g}_k^\top(\mathbf{g}_k-\frac{\mathbf{g}_k^\top\mathbf{g}_k}{\mathbf{g}_{k-1}^\top\mathbf{g}_{k-1}}\mathbf{d}_{k-1})=-\mathbf{g}_k^\top\mathbf{g}_k<0$$

4 的证明不做要求。 □

4.7.4 非线性共轭梯度法

将“正定二次函数”的线性共轭梯度法推广到“一般函数”，就称之为非线性共轭梯度法。回顾线性梯度法中迭代方向 $\mathbf{d}_k$ 的更新公式：

$$\mathbf{d}_k=-\mathbf{g}_k+\xi_{k-1}\mathbf{d}_{k-1},\ \xi_{k-1}=\frac{\mathbf{g}_k^\top\mathbf{g}_k}{\mathbf{g}_{k-1}^\top\mathbf{g}_{k-1}}$$

非共轭梯度法的更新方向同样由上式给出，但采取的 ξ_{k-1} 不同：

FR 方法

FR 方法的更新迭代方向仍为

$$\mathbf{d}_k=-\mathbf{g}_k+\xi_{k-1}\mathbf{d}_{k-1},\ \xi_{k-1}=\frac{\mathbf{g}_k^\top\mathbf{g}_k}{\mathbf{g}_{k-1}^\top\mathbf{g}_{k-1}}$$

但 α_k 用非精确线搜索确定。

FR 方法的下降性质

对于 FR 方法，若 α_k 由**强 Wolfe 准则**得到，且 $\sigma\in\left(0,\frac{1}{2}\right)$，则 $\mathbf{d}_k$ 满足

$$-\frac{1}{1-\sigma}\le\frac{\mathbf{g}_k^\top\mathbf{d}_k}{\|\mathbf{g}_k\|^2}\le\frac{2\sigma-1}{1-\sigma}$$

从而 $\mathbf{d}_k$ 是下降方向。

证明. 由于当 $\sigma \in (0, \frac{1}{2})$ 时，$\frac{2\sigma-1}{1-\sigma} \in (-1, 0)$，所以只需要证明 $-\frac{1}{1-\sigma} \leq \frac{\mathbf{g}_k^\top \mathbf{d}_k}{\|\mathbf{g}_k\|^2} \leq \frac{2\sigma-1}{1-\sigma}$ 即可证明 $\mathbf{d}_k$ 是下降方向。下面用数学归纳法说明其成立。

1. 当 $k=0$ 时，$\frac{\mathbf{g}_0^\top \mathbf{d}_0}{\|\mathbf{g}_0\|^2} = -1 \in \left(-\frac{1}{1-\sigma}, \frac{2\sigma-1}{1-\sigma}\right)$，易知成立。

2. 设在 k 时成立 $-\frac{1}{1-\sigma} \leq \frac{\mathbf{g}_k^\top \mathbf{d}_k}{\|\mathbf{g}_k\|^2} \leq \frac{2\sigma-1}{1-\sigma}$，则 $k+1$ 时

$$\frac{\mathbf{g}_{k+1}^\top \mathbf{d}_{k+1}}{\|\mathbf{g}_{k+1}\|^2} = \frac{\mathbf{g}_{k+1}^\top(-\mathbf{g}_{k+1} + \frac{\mathbf{g}_{k+1}^\top \mathbf{g}_{k+1}}{\mathbf{g}_k^\top \mathbf{g}_k}\mathbf{d}_k)}{\|\mathbf{g}_{k+1}\|^2} = -1 + \frac{\mathbf{g}_{k+1}^\top \mathbf{d}_k}{\|\mathbf{g}_k\|^2}$$

   根据强 Wolfe 准则，$|\mathbf{g}_{k+1}^\top \mathbf{d}_k| < -\sigma \mathbf{g}_k^\top \mathbf{d}_k$，因此

$$\frac{\mathbf{g}_{k+1}^\top \mathbf{d}_{k+1}}{\|\mathbf{g}_{k+1}\|^2} > -1 + \sigma\frac{\mathbf{g}_k^\top \mathbf{d}_k}{\|\mathbf{g}_k\|^2} > -1 - \frac{\sigma}{1-\sigma} = -\frac{1}{1-\sigma}$$
$$\frac{\mathbf{g}_{k+1}^\top \mathbf{d}_{k+1}}{\|\mathbf{g}_{k+1}\|^2} < -1 - \sigma\frac{\mathbf{g}_k^\top \mathbf{d}_k}{\|\mathbf{g}_k\|^2} < -1 + \frac{\sigma}{1-\sigma} = \frac{2\sigma-1}{1-\sigma}$$

3. 因此，$\forall k \in \mathbb{N}$，$-\frac{1}{1-\sigma} \leq \frac{\mathbf{g}_k^\top \mathbf{d}_k}{\|\mathbf{g}_k\|^2} \leq \frac{2\sigma-1}{1-\sigma}$，即 $\mathbf{d}_k$ 是下降方向。

□

FR 方法的收敛性

设 $f(\mathbf{x})$ 有下界，$\mathbf{g}(\mathbf{x})$ 满足 Lipschitz 条件，则使用精确线搜索或 $\sigma \in (0, \frac{1}{2})$ 的强 Wolfe 准则的 FR 方法具有**全局收敛性**，即

1. 要么 $\exists N$, s.t. $\mathbf{g}_N = 0$
2. 要么 $\liminf_{k\to\infty} \|\mathbf{g}_k\| = 0$

证明. 要证明此，先证明一个引理：

Zoutendijk 条件

设 $f(\mathbf{x})$ 有下界，$\mathbf{g}(\mathbf{x})$ 满足 Lipschitz 条件，α_k 由 Wolfe 准则或精确线搜索准则得到，则对于具有 $\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{d}_k$ 的迭代格式的下降方法，满足 Zoutendijk 条件

$$\sum_{k\geq 0} \frac{(\mathbf{g}_k^\top \mathbf{d}_k)^2}{\|\mathbf{d}_k\|^2} = \sum_{k\geq 0} \|\mathbf{g}_k\|^2 \cos^2\theta_k < \infty$$

其中 $\theta = \langle -\mathbf{g}_k, \mathbf{d}_k \rangle$ 为负梯度方向与迭代方向的夹角。

证明. **(使用 Wolfe 准则)** 根据 Wolfe 准则，

$$\mathbf{g}_{k+1}^\top \mathbf{d}_k > \sigma \mathbf{g}_k^\top \mathbf{d}_k \Rightarrow (\mathbf{g}_{k+1} - \mathbf{g}_k)^\top \mathbf{d}_k > (\sigma - 1)\mathbf{g}_k^\top \mathbf{d}_k$$

又根据 $g(\mathbf{x})$ 的 Lipschitz 连续性：

$$(\mathbf{g}_{k+1}-\mathbf{g}_k)^\top \mathbf{d}_k \le \|\mathbf{g}_{k+1}-\mathbf{g}_k\|\|\mathbf{d}_k\| \le \alpha_k L\|\mathbf{d}_k\|^2$$

因此

$$(\sigma-1)\mathbf{g}_k^\top \mathbf{d}_k < \alpha_k L\|\mathbf{d}_k\|^2 \Rightarrow \alpha_k > \frac{(\sigma-1)\mathbf{g}_k^\top \mathbf{d}_k}{L\|\mathbf{d}_k\|^2}$$

将上述 α_k 的范围带入 Wolfe 准则

$$\begin{aligned} f(\mathbf{x}_k+\alpha_k\mathbf{d}_k) &\le f(\mathbf{x}_k)+\rho\mathbf{g}_k^\top\mathbf{d}_k\alpha_k \\ \Rightarrow f_{k+1} &\le f_k+\rho\mathbf{g}_k^\top\mathbf{d}_k\frac{(\sigma-1)\mathbf{g}_k^\top\mathbf{d}_k}{L\|\mathbf{d}_k\|^2} = f_k+\rho\frac{(\sigma-1)(\mathbf{g}_k^\top\mathbf{d}_k)^2}{L\|\mathbf{d}_k\|^2} \\ &= f_k+\frac{\rho(\sigma-1)}{L}\|\mathbf{g}_k\|^2\cos^2\theta_k \end{aligned}$$

对上式关于 $k=0$ 到 n 作累加得：

$$\begin{aligned} f_k &\le f_0+\frac{\rho(\sigma-1)}{L}\sum_{i=0}^{k-1}\|\mathbf{g}_i\|^2\cos^2\theta_i \\ \Rightarrow \sum_{i=0}^{k-1}\|\mathbf{g}_i\|^2\cos^2\theta_i &\le \frac{L}{\rho(1-\sigma)}(f_0-f_k) \end{aligned}$$

由于 f_k 是下有界的，所以 $f_0-f_k<\infty$，得证。 □

证明. **(使用精确线搜索准则)** 在精确线搜索准则下，$\mathbf{g}_{k+1}\mathbf{d}_k=0$，于是

$$\begin{aligned} \|\mathbf{d}_k^\top\mathbf{g}_k\| &= \|\mathbf{d}_k^\top(\mathbf{g}_{k+1}-\mathbf{g}_k)\| \le \|\mathbf{d}_k\|\|\mathbf{g}_{k+1}-\mathbf{g}_k\| \le \alpha_k L\|\mathbf{d}_k\|^2 \\ \Rightarrow \alpha_k &\ge \frac{\|\mathbf{d}_k^\top\mathbf{g}_k\|}{L\|\mathbf{d}_k\|^2} \end{aligned}$$

再对 $\phi(\alpha)=f(\mathbf{x}_k+\alpha\mathbf{d}_k)$ 在 $\alpha=\alpha_k$ 处作 Taylor 展开，得

$$\begin{aligned} \phi(\alpha) &= \phi(\alpha_k)+\phi'(\alpha_k)(\alpha-\alpha_k)+\frac{1}{2}\phi''(\xi_k)(\alpha-\alpha_k)^2,\ \xi_k\in(\alpha_k,\alpha) \\ \Rightarrow f(\mathbf{x}_k+\alpha\mathbf{d}_k) &= f(\mathbf{x}_{k+1})+\mathbf{g}_{k+1}^\top\mathbf{d}_k(\alpha-\alpha_k)+\frac{1}{2}\mathbf{d}_k^\top G(\xi_k)\mathbf{d}_k(\alpha-\alpha_k)^2 \end{aligned}$$

令 $\alpha=0$，则

$$\begin{aligned} f(\mathbf{x}_k) &= f(\mathbf{x}_{k+1})+0+\frac{1}{2}\mathbf{d}_k^\top G(\xi_k)\mathbf{d}_k\alpha_k^2 \\ \Rightarrow f_k-f_{k+1} &= \frac{1}{2}\mathbf{d}_k^\top G(\xi_k)\mathbf{d}_k\alpha_k^2 \ge \frac{1}{2}M\mathbf{d}_k^\top\mathbf{d}_k\frac{(\mathbf{d}_k^\top\mathbf{g}_k)^2}{L^2\|\mathbf{d}_k\|^4} \\ f_k-f_{k+1} &\ge \frac{M}{2L^2}\frac{(\mathbf{d}_k^\top\mathbf{g}_k)^2}{\|\mathbf{d}_k\|^2} = \frac{M}{2L^2}\|\mathbf{g}_k\|^2\cos^2\theta_k \end{aligned}$$

对 k 从 0 开始求和，得

$$\sum_{i=0}^{k-1}\|\mathbf{g}_i\|^2\cos^2\theta_i \le \frac{2L^2}{M}(f_0-f_k)<\infty$$

Zoutendijk 条件得证。 □

接下来说明 FR 方法的迭代方向是下降方向。假设 $\forall k, \mathbf{g}_k \neq 0$，不妨假设 $\exists \mu > 0, \forall k \geq 0, \|\mathbf{g}_k\| \geq \mu$。根据 Zoutendijk 条件

$$\sum_{k\geq 0} \|\mathbf{g}_k\| \cos\theta_k < \infty$$

只能令 $\cos\theta_k \to 0$。在第 k 次迭代的时候，注意到

$$\begin{aligned}
&\mathbf{g}_k^\top \mathbf{d}_k = \mathbf{g}_k^\top(-\mathbf{g}_k + \xi_{k-1}\mathbf{d}_{k-1}) = -\|\mathbf{g}_k\|^2 \\
\Rightarrow &-\|\mathbf{g}_k\|^2 = -\|\mathbf{g}_k\|\|\mathbf{d}_k\|\cos\theta_k \\
\Rightarrow &\|\mathbf{g}_k\| = \|\mathbf{d}_k\|\cos\theta_k \\
\Rightarrow &\|\mathbf{d}_k\| = \|\mathbf{g}_k\|\sec\theta_k
\end{aligned}$$

又注意到在第 $k+1$ 次迭代的时候，根据 $\mathbf{d}_{k+1} = -\mathbf{g}_{k+1} + \xi_k \mathbf{d}_k$

$$\begin{aligned}
&\|\mathbf{d}_{k+1}\|^2 = \|\mathbf{g}_{k+1}\|^2 + \xi_k^2\|\mathbf{d}_k\|^2 - 2\xi_k \mathbf{g}_{k+1}^\top \mathbf{d}_k = \|\mathbf{g}_{k+1}\|^2 + \xi_k^2\|\mathbf{d}_k\|^2 \\
\Rightarrow &\frac{\|\mathbf{d}_{k+1}\|^2}{\|\mathbf{g}_{k+1}\|^2} = 1 + \xi_k^2 \frac{\|\mathbf{d}_k\|^2}{\|\mathbf{g}_{k+1}\|^2} \\
\Rightarrow &\sec^2\theta_{k+1} - 1 = \xi_k^2 \frac{\|\mathbf{d}_k\|^2}{\|\mathbf{g}_{k+1}\|^2} \\
\Rightarrow &\tan^2\theta_{k+1} = \xi_k^2 \frac{\|\mathbf{d}_k\|^2}{\|\mathbf{g}_{k+1}\|^2} = \frac{\|\mathbf{g}_{k+1}\|^2\|\mathbf{d}_k\|^2}{\|\mathbf{g}_k\|^4}
\end{aligned}$$

结合上面两个式子

$$\frac{\tan^2\theta_{k+1}}{\|\mathbf{g}_{k+1}\|^2} = \frac{\sec^2\theta_k}{\|\mathbf{g}_k\|^2} = \frac{\tan^2\theta_k + 1}{\|\mathbf{g}_k\|^2} \Rightarrow \frac{\tan^2\theta_{k+1}}{\|\mathbf{g}_{k+1}\|^2} - \frac{\tan^2\theta_k}{\|\mathbf{g}_k\|^2} = \frac{1}{\|\mathbf{g}_k\|^2}$$

对上述递推式从 0 到 $k-1$ 加和，得

$$\frac{\tan^2\theta_k}{\|\mathbf{g}_k\|^2} = \sum_{i=0}^{k-1} \frac{1}{\|\mathbf{g}_k\|^2} \leq \frac{k}{\mu^2} \Rightarrow \frac{\mu^2}{k} \leq \|\mathbf{g}_k\|^2 \cot^2\theta_k$$

当 k 充分大时，因为 $\cos\theta_k \to 0$，因此 $\sin\theta_k \to 1 > \frac{1}{2}$，那么 $\cot^2\theta_k = \dfrac{\cos^2\theta_k}{\sin^2\theta_k} < 4\cos\theta_k^2$，于是

$$\frac{\mu^2}{4k} \leq \|\mathbf{g}_k\|^2\cos^2\theta_k \Rightarrow \sum_{k\geq 0}\|\mathbf{g}_k\|^2\cos^2\theta_k \geq \sum_{k\geq 0}\frac{\mu^2}{4k} = \infty$$

矛盾！假设不成立，原命题得证。 □

FR 方法的缺点在于，如果在某一步迭代方向很差 (前进方向和负梯度方向接近正交，$\mathbf{d}_k^\top(-\mathbf{g}_k) \approx 0$)，并且步长很小，那么之后的迭代方向仍然会很差，从而收敛速度变得很慢。这是因为：假设 $\mathbf{d}_k$ 与 $-\mathbf{g}_k$ 的夹角为 θ_k，当迭代方向很差时，$\theta_k \approx \dfrac{\pi}{2}$，根据确定 FR 方法下降性质的定理，

$$\begin{aligned}
-\frac{1}{1-\sigma} \leq \frac{\mathbf{g}_k^\top \mathbf{d}_k}{\|\mathbf{g}_k\|^2} \leq \frac{2\sigma - 1}{1-\sigma} &\Rightarrow c_1\frac{\|\mathbf{g}_k\|}{\|\mathbf{d}_k\|} \leq \frac{-\mathbf{g}_k^\top\mathbf{d}_k}{\|\mathbf{g}_k\|\|\mathbf{d}_k\|} \leq c_2\frac{\|\mathbf{g}_k\|}{\|\mathbf{d}_k\|} \\
&\Rightarrow c_1\frac{\|\mathbf{g}_k\|}{\|\mathbf{d}_k\|} \leq \cos\theta_k \leq c_2\frac{\|\mathbf{g}_k\|}{\|\mathbf{d}_k\|}
\end{aligned}$$

其中 $c_1, c_2 > 0$，根据夹逼准则知 $\cos\theta_k = 0 \iff \|\mathbf{g}_k\| \ll \|\mathbf{d}_k\|$。如果步长 α_k 很小，那么 $\mathbf{g}_k \approx \mathbf{g}_{k-1}$，进而 $\xi_{k-1} \approx 1$，进而 $\mathbf{d}_k \approx \mathbf{d}_{k-1}$，那么新的迭代方向基本没有改善！

PRP 方法

更新迭代方向为

$$\mathbf{d}_k = -\mathbf{g}_k + \xi_{k-1}\mathbf{d}_{k-1},\ \xi_{k-1} = \frac{\mathbf{g}_k^\top(\mathbf{g}_k - \mathbf{g}_{k-1})}{\mathbf{g}_{k-1}^\top \mathbf{g}_{k-1}}$$

[注]: 由于此时 $\mathbf{G}$ 未必正定，子空间拓展定理不可用。

[注]: PRP 方法修正了 FR 方法的弱点，当 $\mathbf{d}_k$ 是很差的迭代方向，且步长很小时，有 $\mathbf{g}_k \approx \mathbf{g}_{k-1}$，进而 $\xi_{k-1} \approx 0$，进而 $\mathbf{d}_k \approx -\mathbf{g}_k$.

PRP 方法的下降性和收敛性

PRP 方法的前进方向 $\mathbf{d}_k$ 满足下降性，且 PRP 方法具有**全局收敛性**。

二次终止与算法重启

函数极小点附近可以近似为正定二次函数。注意到线性共轭梯度方法对正定二次函数具有二次终止性，但是需要在初始点选择负梯度方向。类似地，可以对非线性共轭梯度方法**采用如下重启策略**：

1. n 步重启：(但如果需要迭代的步数很小，无法使用 n 步重启)

$$\mathbf{d}_k = \begin{cases} -\mathbf{g}_k & k = cn,\ c = 0, 1, 2, \cdots \\ -\mathbf{g}_k + \xi_{k-1}\mathbf{d}_{k-1} & k \neq cn,\ c = 0, 1, 2, \cdots \end{cases}$$

2. 梯度重启：当相邻两个梯度向量**远非正交**（说明当前迭代点附近无法被正定二次函数很好的近似），则下一步迭代方向变为负梯度方向。其中远非正交的标准为

$$\frac{\mathbf{g}_k^\top \mathbf{g}_{k-1}}{\|\mathbf{g}_k\|} \geq 0.1$$

4.7.5 Broyden 族方法的搜索方向共轭性

以下定理非常重要，其揭示了为什么 SR1、BFGS、DFP 等 Broyden 族方法是二次终止的：因为可以证明 Broyden 族方法（包括 SR1、BFGS、DFP）都是共轭梯度法，而共轭梯度法通过 n 步迭代必收敛到二次函数的极小点。

Broyden 族方法的搜索方向共轭性

对正定二次函数 $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top \mathbf{G}\mathbf{x} + \mathbf{b}^\top \mathbf{x}$，对于采用精确线搜索的 Broyden 族方法，当经过 k 步迭代未达到极小点时，满足如下关系：

$$\mathbf{H}_k \mathbf{y}_i = \mathbf{s}_i,\ i = 0, \cdots, k-1 \qquad \text{(拟 Newton 条件)}$$
$$\mathbf{d}_k^\top \mathbf{G}\mathbf{d}_i = 0,\ i = 0, \cdots, k-1 \qquad \text{(共轭性)}$$

在证明前，有必要回顾以下几条关系：由于 f 为正定二次型，$\mathbf{g} = \mathbf{G}\mathbf{x} + \mathbf{b}$，那么

$$\mathbf{g}_{k+1} - \mathbf{g}_k = \mathbf{G}(\mathbf{x}_{k+1} - \mathbf{x}_k) \Rightarrow \mathbf{y}_k = \mathbf{G}\mathbf{s}_k$$

根据迭代公式

$$\mathbf{x}_{k+1}-\mathbf{x}_k=\alpha_k\mathbf{d}_k\Rightarrow\mathbf{s}_k=\alpha_k\mathbf{d}_k\Rightarrow\mathbf{y}_k=\alpha_k\mathbf{G}\mathbf{d}_k$$

根据 Newton 方向（但用的 $\mathbf{H}_k$ 近似），$\mathbf{d}_k=-\mathbf{H}_k\mathbf{g}_k$。

证明. 下用数学归纳法证明：

1. 当 $k=1$ 时，由拟 Newton 条件有 $\mathbf{H}_1\mathbf{y}_0=\mathbf{s}_0$，且 $\mathbf{d}_1^\top\mathbf{G}\mathbf{d}_0=-\dfrac{1}{\alpha_0}(\mathbf{H}_1\mathbf{g}_1)^\top\mathbf{y}_0=-\mathbf{g}_1^\top\mathbf{d}_0=0$

2. 假设当 $k=j$ 时成立，即

$$\mathbf{H}_j\mathbf{y}_i=\mathbf{s}_i,\ i=0,\cdots,j-1$$
$$\mathbf{d}_j^\top\mathbf{G}\mathbf{d}_i=0,\ i=0,\cdots,j-1$$

   则 $k=j+1$ 时，

   (a) 当 $i=j$ 时，根据拟 Newton 条件知 $\mathbf{H}_{j+1}\mathbf{y}_j=\mathbf{s}_j$，且

$$\begin{aligned}\mathbf{d}_{j+1}^\top\mathbf{G}\mathbf{d}_j&=-\frac{1}{\alpha_j}(\mathbf{H}_{j+1}\mathbf{g}_{j+1})^\top\mathbf{y}_j=-\frac{1}{\alpha_j}\mathbf{g}_{j+1}^\top\mathbf{H}_{j+1}\mathbf{y}_j\\&=-\frac{1}{\alpha_j}\mathbf{g}_{j+1}^\top\alpha_j\mathbf{d}_j=-\mathbf{g}_{j+1}^\top\mathbf{d}_j=0\end{aligned}$$

   成立。

   (b) 当 $i=0,\cdots,j-1$，根据 Broyden 族公式

$$\mathbf{H}_{j+1}=\mathbf{H}_{j+1}^{\mathrm{DFP}}+\phi\mathbf{v}_j\mathbf{v}_j^\top$$

   从而

$$\mathbf{H}_{j+1}\mathbf{y}_i=\mathbf{H}_j\mathbf{y}_i-\frac{\mathbf{H}_j\mathbf{y}_j}{\mathbf{y}_j^\top\mathbf{H}_j\mathbf{y}_j}\mathbf{y}_j^\top\mathbf{H}_j\mathbf{y}_i+\frac{\mathbf{s}_j}{\mathbf{s}_j^\top\mathbf{y}_j}\mathbf{s}_j^\top\mathbf{y}_i+\phi\mathbf{v}_j\mathbf{v}_j^\top\mathbf{y}_i=0$$

   其中

$$\begin{aligned}\mathbf{y}_j^\top\mathbf{H}_j\mathbf{y}_i&=\mathbf{y}_j^\top\mathbf{s}_i=\mathbf{s}_j^\top\mathbf{G}\mathbf{s}_i=\alpha_j\alpha_i(\mathbf{d}_j^\top\mathbf{G}\mathbf{d}_i)=0\\\mathbf{s}_j^\top\mathbf{y}_i&=\mathbf{s}_j^\top\mathbf{G}\mathbf{s}_i=0\\\mathbf{v}_j^\top\mathbf{y}_i&=(\mathbf{y}_j^\top\mathbf{H}_j\mathbf{y}_i)^{\frac12}(\frac{\mathbf{s}_j}{\mathbf{s}_j^\top\mathbf{y}_i}-\frac{\mathbf{H}_j\mathbf{y}_j}{\mathbf{y}_j^\top\mathbf{H}_i\mathbf{y}_i})^\top\mathbf{y}_i=(\mathbf{y}_j^\top\mathbf{H}_j\mathbf{y}_i)^{\frac12}(\frac{\mathbf{s}_j^\top\mathbf{y}_i}{\mathbf{s}_j^\top\mathbf{y}_i}-\frac{\mathbf{y}_j^\top\mathbf{H}_j\mathbf{y}_i}{\mathbf{y}_j^\top\mathbf{H}_i\mathbf{y}_i})=0\end{aligned}$$

   故 $\mathbf{H}_{j+1}\mathbf{y}_i=\mathbf{H}_j\mathbf{y}_i=\mathbf{s}_i$ 成立，另外

$$\begin{aligned}\mathbf{d}_{j+1}^\top\mathbf{G}\mathbf{d}_i&=-\frac{1}{\alpha_j}(\mathbf{H}_{j+1}\mathbf{g}_{j+1})^\top\mathbf{y}_i=-\frac{1}{\alpha_j}\mathbf{g}_{j+1}^\top\mathbf{s}_i\\&=-\frac{1}{\alpha_j}(\mathbf{g}_{i+1}+(\mathbf{g}_{i+2}-\mathbf{g}_{i+1})+\cdots+(\mathbf{g}_{j+1}-\mathbf{g}_j))^\top\mathbf{s}_i\\&=-\frac{1}{\alpha_j}(\mathbf{g}_{i+1}+\mathbf{y}_{i+1}+\cdots+\mathbf{y}_j)^\top\mathbf{s}_i\\&=-\frac{1}{\alpha_j}(\mathbf{g}_{i+1}+\mathbf{G}\mathbf{s}_{i+1}+\cdots+\mathbf{G}\mathbf{s}_j)^\top\mathbf{s}_i=0\end{aligned}$$

   综上 $k=j+1$ 时定理成立。

3. 由数学归纳法，定理对任意 $k\in\mathbb{N}$ 均成立。

□

4.8 最小二乘法

最小二乘法的动机是“数据拟合问题”：给定一组实验数据 $(\mathbf{t}_i, y_i), i = 1, \cdots, m$，我们希望寻找一个以 $\mathbf{x}$ 为参数的函数 $f_{\mathbf{x}}(\mathbf{t})$，使得 $f_{\mathbf{x}}(\mathbf{t}_i) \approx y_i$。由此，我们可以定义第 i 组数据 $(\mathbf{t}_i, y_i)$ 的剩余量 $r_i(\mathbf{x})$ 为

$$r_i(\mathbf{x}) = y_i - f_{\mathbf{x}}(\mathbf{t}_i)$$

在剩余量平方和的意义下尽可能好的拟合数据，即最小化损失函数 $L(\mathbf{x}) = \frac{1}{2}\sum_{i=1}^{m} r_i^2(\mathbf{x})$。

最小二乘问题

$$\min L(\mathbf{x}) = \frac{1}{2}\sum_{i=1}^{m} r_i^2(\mathbf{x}) = \frac{1}{2}\mathbf{r}(\mathbf{x})^\top \mathbf{r}(\mathbf{x}), \quad \mathbf{x} \in \mathbb{R}^n$$

其中 $\mathbf{r}(\mathbf{x}) = [r_1(\mathbf{x}), \ldots, r_m(\mathbf{x})]^\top$ 称为剩余函数，剩余函数在 $\mathbf{x}$ 点的值 $r_i(\mathbf{x}) = y_i - f_{\mathbf{x}}(\mathbf{t}_i)$ 称为剩余量。

- 若 $r_i(\mathbf{x}), i = 1, \ldots, m$ 均为线性函数，则问题称为线性最小二乘问题。
- 若至少有一个 $r_i(\mathbf{x})$ 为非线性函数，则问题称为非线性最小二乘问题。

Jacobian 矩阵

Jacobian 矩阵定义为

$$J(\mathbf{x}) = [\nabla r_1(\mathbf{x}), \cdots, \nabla r_m(\mathbf{x})]^\top \in \mathbb{R}^{m \times n}$$

由此

- $L(\mathbf{x})$ 的梯度为

$$g(\mathbf{x}) = \sum_{i=1}^{m} r_i(\mathbf{x}) \nabla r_i(\mathbf{x}) = J(\mathbf{x})^\top \mathbf{r}(\mathbf{x})$$

- $L(\mathbf{x})$ 的 Hessian 矩阵为

$$G(\mathbf{x}) = \sum_{i=1}^{m} (\nabla r_i(\mathbf{x}) \nabla r_i(\mathbf{x})^\top + r_i(\mathbf{x}) \nabla^2 r_i(\mathbf{x})) = J(\mathbf{x})^\top J(\mathbf{x}) + S(\mathbf{x})$$

对于 $S(\mathbf{x}) = \sum_{i=1}^{m} r_i(\mathbf{x}) \nabla^2 r_i(\mathbf{x})$，在极小点 $\mathbf{x}^*$ 处，$S(\mathbf{x}^*)$ 的大小 $\|\mathbf{S}^*\|$ 取决于剩余量和问题的非线性性

- 对于线性最小二乘问题（显然有 $\nabla^2 r_i(\mathbf{x}) = 0$）/剩余量为 0，$\|\mathbf{S}^*\| = 0$
- 当剩余量增大/$r_i(\mathbf{x})$ 非线性的增强时，$\|\mathbf{S}^*\|$ 增大

由此，根据 $\|\mathbf{S}^*\|$ 的大小可将优化算法分为小剩余算法（$\|\mathbf{S}^*\| = 0$ 或较小）与大剩余算法（$\|\mathbf{S}^*\|$ 很大）

4.8.1 Gauss-Newton 法

回顾 Newton 法中求解最优迭代方向的 Newton 方程

$$\mathbf{d}_k = \arg\min_{\mathbf{d}} f(\mathbf{x}_k) + \mathbf{g}_k^\top \mathbf{d} + \frac{1}{2}\mathbf{d}^\top \mathbf{G}_k \mathbf{d}$$

$$\Rightarrow \mathbf{G}_k \mathbf{d}_k = -\mathbf{g}_k$$

记 $J(\mathbf{x}_k)=\mathbf{J}_k, S(\mathbf{x}_k)=\mathbf{S}_k, \mathbf{r}(\mathbf{x}_k)=\mathbf{r}_k$，于是对于最小二乘问题，可以轻松得到如下 Newton 方程

$$(\mathbf{J}_k^\top \mathbf{J}_k+\mathbf{S}_k)\mathbf{d}_k=-\mathbf{J}_k^\top \mathbf{r}_k$$

然而，每次迭代时都要计算 $\mathbf{S}_k$(即 m 个矩阵 $\nabla^2 r_i(\mathbf{x})$)，带来沉重的计算负担。**但对于小剩余问题**，我们可以将 $\mathbf{S}_k$ 近似为 0。

Gauss-Newton 方法

在 Newton 法中忽略 $\mathbf{S}_k$ 便得到 Gauss-Newton 方法，即 Gauss-Newton 方向满足

$$\mathbf{J}_k^\top \mathbf{J}_k\mathbf{d}_k=-\mathbf{J}_k^\top \mathbf{r}_k$$

Gauss-Newton 方向的另一种理解如下，我们可以对剩余函数 $\mathbf{r}(\mathbf{x})$ 作一阶 Taylor 近似：$\mathbf{r}(\mathbf{x}_k+\mathbf{d})\approx\mathbf{r}(\mathbf{x}_k)+\mathbf{J}_k^\top\mathbf{d}$，代回最小二乘问题，即得

$$\begin{aligned}\mathbf{d}_k&=\arg\min_{\mathbf{d}}\frac{1}{2}\|\mathbf{r}_k+\mathbf{J}_k\mathbf{d}\|^2\\&=\arg\min_{\mathbf{d}}\frac{1}{2}\|\mathbf{r}_k\|^2+\mathbf{r}_k^\top\mathbf{J}_k\mathbf{d}+\frac{1}{2}\mathbf{d}^\top\mathbf{J}_k^\top\mathbf{J}_k\mathbf{d}\end{aligned}$$

对上式关于 $\mathbf{d}$ 求导并令导数为 0，即得 Gauss-Newton 方向。**在迭代过程中，我们要求 $J(\mathbf{x})$ 是列满秩矩阵，否则如果 $J(\mathbf{x})$ 不是列满秩矩阵，$J(\mathbf{x})^\top J(\mathbf{x})$ 是奇异阵，Gauss-Newton 方向无法求解。**

对比 Newton 法，Gauss-Newton 法在优化 $L(\mathbf{x})$ 时有三大优点

- 无需计算 $\mathbf{S}_k$，计算量小。事实上，在许多实际问题中 $\mathbf{S}_k\ll\mathbf{J}_k^\top\mathbf{J}_k$，因此 Gauss-Newton 法的效果和 Newton 法类似。
- 当 $\mathbf{J}_k$ 列满秩且 $\mathbf{g}_k\neq 0$ 时，Gauss-Newton 法的迭代方向 $\mathbf{d}_k$ 是下降方向，因为

$$\mathbf{d}_k^\top\mathbf{g}_k=\mathbf{d}_k^\top\mathbf{J}_k^\top\mathbf{r}_k=-\mathbf{d}_k^\top\mathbf{J}_k^\top\mathbf{J}_k\mathbf{d}_k=-\|\mathbf{J}_k\mathbf{d}_k\|^2\leq 0$$

Gauss-Newton 法的算法流程如下：当 $\alpha_k=1$ 时称为基本 Gauss-Newton 法，当 α_k 用线搜索确定时称为阻尼 Gauss-Newton 法。

算法5.1：

1. 给定x_0，$\epsilon>0$，$k=0$；
2. 若终止准则满足，则停止迭代；
3. 求解$J_k^TJ_kd_k=-J_k^Tr_k$，得到d_k；
4. 作一维线搜索求α_k；
5. $x_{k+1}=x_k+\alpha_kd_k$，$k\leftarrow k+1$，转到第2步。

基本 Gauss-Newton 法的收敛性

基本 Gauss-Newton 法具有**局部收敛性**。在 $\mathbf{x}^*$ 为最小二乘问题的最优解的情况下，设

1. $r_i(\mathbf{x})\in C^2, i=1,\cdots,m$；
2. $(\mathbf{J}^*)^\top(\mathbf{J}^*)$ 正定；
3. 由基本 Gauss-Newton 法产生的迭代序列 $\{\mathbf{x}_k\}$ 收敛到 $\mathbf{x}^*$；

4. $G(\mathbf{x})$ 与 $J(\mathbf{x})^\top J(\mathbf{x})$ 在 $\mathbf{x}^*$ 的邻域内 Lipschitz 连续；

则成立

$$\|\mathbf{x}_{k+1}-\mathbf{x}^*\| \le \|[(\mathbf{J}^*)^\top(\mathbf{J}^*)]^{-1}\|\|\mathbf{S}^*\|\|\mathbf{x}_k-\mathbf{x}^*\| + O(\|\mathbf{x}_k-\mathbf{x}^*\|^2)$$

从而基本 Gauss-Newton 法的收敛速度为

- 当 $\|[(\mathbf{J}^*)^\top(\mathbf{J}^*)]^{-1}\|\|\mathbf{S}^*\| < 1$ 时，基本 Gauss-Newton 法收敛且有**线性收敛速度**，收敛速度随 $\|\mathbf{S}^*\|$ 的增大而减慢，甚至对剩余量很大或者剩余函数高度非线性的问题不收敛。
- 当 $\|\mathbf{S}^*\| = 0$，即零剩余问题或者线性最小二乘问题时，基本 Gauss-Newton 法就是 Newton 法，收敛速度为**二次收敛速度**。
- 当 $\|\mathbf{S}^*\| \neq 0$ 时，阻尼 Gauss-Newton 法具有线性收敛速度。

证明. 由剩余量 $r_i(\mathbf{x}) \in C^2$ 立刻知损失函数 $L \in C^2$。首先简记 $\mathbf{h}_k = \mathbf{x}_k - \mathbf{x}^*$。对于充分接近 $\mathbf{x}^*$ 的 $\mathbf{x}_k$(即 $\mathbf{x}_k \in B(\mathbf{x}^*, \delta)$)，有

$$g(\mathbf{x}_k+\mathbf{d}) = g(\mathbf{x}_k) + G(\mathbf{x}_k)\mathbf{d} + O(\|\mathbf{d}\|^2)$$

令 $\mathbf{d} = -\mathbf{h}_k$，

$$\begin{aligned} &g(\mathbf{x}^*) = g(\mathbf{x}_k) - G(\mathbf{x}_k)\mathbf{h}_k + O(\|\mathbf{h}_k\|^2) = 0 \\ \Rightarrow &\mathbf{J}_k^\top\mathbf{r}_k - (\mathbf{J}_k^\top\mathbf{J}_k + \mathbf{S}_k)\mathbf{h}_k + O(\|\mathbf{h}_k\|^2) = 0 \end{aligned}$$

由于 $(\mathbf{J}^*)^\top(\mathbf{J}^*)$ 正定且 $\mathbf{x}_k$ 充分接近 $\mathbf{x}^*$，故 $\mathbf{J}_k^\top\mathbf{J}_k$ 正定，等号两侧左乘 $(\mathbf{J}_k^\top\mathbf{J}_k)^{-1}$，得

$$\begin{aligned} &(\mathbf{J}_k^\top\mathbf{J}_k)^{-1}\mathbf{J}_k^\top\mathbf{r}_k - \mathbf{h}_k - (\mathbf{J}_k^\top\mathbf{J}_k)^{-1}\mathbf{S}_k\mathbf{h}_k + O(\|\mathbf{h}_k\|) = 0 \\ \Rightarrow &\mathbf{d}_k + \mathbf{h}_k = -(\mathbf{J}_k^\top\mathbf{J}_k)^{-1}\mathbf{S}_k\mathbf{h}_k + O(\|\mathbf{h}_k\|) \end{aligned}$$

注意到 $\mathbf{d}_k + \mathbf{h}_k = \mathbf{x}_{k+1} - \mathbf{x}_k + \mathbf{x}_k - \mathbf{x}^* = \mathbf{x}_{k+1} - \mathbf{x}^* = \mathbf{h}_{k+1}$，故

$$\begin{aligned} \mathbf{h}_{k+1} &= -(\mathbf{J}_k^\top\mathbf{J}_k)^{-1}\mathbf{S}_k\mathbf{h}_k + O(\|\mathbf{h}_k\|) \\ \Rightarrow \|\mathbf{h}_{k+1}\| &\le \|(\mathbf{J}_k^\top\mathbf{J}_k)^{-1}\mathbf{S}_k\|\|\mathbf{h}_k\| + O(\|\mathbf{h}_k\|^2) \\ &\le \|(\mathbf{J}_k^\top\mathbf{J}_k)^{-1}\mathbf{S}_k - [(\mathbf{J}^*)^\top(\mathbf{J}^*)]^{-1}\mathbf{S}^*\|\|\mathbf{h}_k\| + \|[(\mathbf{J}^*)^\top(\mathbf{J}^*)]^{-1}\mathbf{S}^*\|\|\mathbf{h}_k\| + O(\|\mathbf{h}_k\|^2) \\ &\le \underbrace{\|(\mathbf{J}_k^\top\mathbf{J}_k)^{-1}\mathbf{S}_k - [(\mathbf{J}^*)^\top(\mathbf{J}^*)]^{-1}\mathbf{S}^*\|\|\mathbf{h}_k\|}_{\text{Term of Interest}} + \|[(\mathbf{J}^*)^\top(\mathbf{J}^*)]^{-1}\|\|\mathbf{S}^*\|\|\mathbf{h}_k\| + O(\|\mathbf{h}_k\|^2) \end{aligned}$$

从不等式的最终形式来看，我们只需要将 Term of Interest 放缩去除即可(放缩成 $\|\mathbf{h}_k\|^2$ 的形式，并入 $O(\|\mathbf{h}_k\|^2)$)。注意到感兴趣的项的形式实际上为 $(J(\mathbf{x})^\top J(\mathbf{x}))^{-1}S(\mathbf{x})$ 中取 $\mathbf{x} = \mathbf{x}_k$ 和 $\mathbf{x} = \mathbf{x}^*$ 的差值，由此想到证明 $(J(\mathbf{x})^\top J(\mathbf{x}))^{-1}S(\mathbf{x})$ 的 Lipschitz 连续性。因为 $G(x)$ 与 $J(x)^\top J(x)$ 在 $\mathbf{x}^*$ 的邻域内 Lipschitz 连续，存在 $\beta, \gamma > 0$，使得 $\forall \mathbf{x}, \mathbf{y} \in B(\mathbf{x}^*, \delta)$，成立

$$\begin{aligned} \|G(\mathbf{x}) - G(\mathbf{y})\| &\le \beta\|\mathbf{x}-\mathbf{y}\| \\ \|J(\mathbf{x})^\top J(\mathbf{x}) - J(\mathbf{y})^\top J(\mathbf{y})\| &\le \gamma\|\mathbf{x}-\mathbf{y}\| \end{aligned}$$

那么

$$\begin{aligned}\|S(\mathbf{x})-S(\mathbf{y})\| &= \|G(\mathbf{x})-J(\mathbf{x})^\top J(\mathbf{x})-G(\mathbf{y})+J(\mathbf{y})^\top J(\mathbf{y})\| \\ &\le \|G(\mathbf{x})-G(\mathbf{y})\|+\|J(\mathbf{x})^\top J(\mathbf{x})-J(\mathbf{y})^\top J(\mathbf{y})\| \\ &\le (\beta+\gamma)\|\mathbf{x}-\mathbf{y}\|\end{aligned}$$

这说明了 $S(\mathbf{x})$ 在 $\mathbf{x}^*$ 的邻域内 Lipschitz 连续。另外，由 $(\mathbf{J}^*)^\top(\mathbf{J}^*)$ 正定，对 $\forall \mathbf{x}\in B(\mathbf{x}^*,\delta)$，$\exists \xi>0$，使得 $\|[J(\mathbf{x})^\top J(\mathbf{x})]^{-1}\|\le\xi$，进而

$$\begin{aligned}\|(J(\mathbf{x})^\top J(\mathbf{x}))^{-1}-(J(\mathbf{y})^\top J(\mathbf{y})^{-1})\| &= \left\| \left(J(\mathbf{x})^\top J(\mathbf{x})\right)^{-1}\left(J(\mathbf{x})^\top J(\mathbf{x})-J(\mathbf{y})^\top J(\mathbf{y})\right)\left(J(\mathbf{y})^\top J(\mathbf{y})^{-1}\right)\right\| \\ &\le \|(J(\mathbf{x})^\top J(\mathbf{x}))^{-1}\|\|J(\mathbf{x})^\top J(\mathbf{x})-J(\mathbf{y})^\top J(\mathbf{y})\|\|J(\mathbf{y})^\top J(\mathbf{y})^{-1}\| \\ &\le \xi^2\gamma\|\mathbf{x}-\mathbf{y}\|\end{aligned}$$

因此

$$\begin{aligned}\|(\mathbf{J}_k^\top\mathbf{J}_k)^{-1}\mathbf{S}_k-[(\mathbf{J}^*)^\top(\mathbf{J}^*)]^{-1}\mathbf{S}^*\|\|\mathbf{h}_k\| &\le \|(\mathbf{J}_k^\top\mathbf{J}_k)^{-1}\mathbf{S}_k-(\mathbf{J}_k^\top\mathbf{J}_k)^{-1}\mathbf{S}^*\|+\|(\mathbf{J}_k^\top\mathbf{J}_k)^{-1}\mathbf{S}^*-(\mathbf{J}^*)^\top(\mathbf{J}^*)^{-1}\mathbf{S}^*\|\|\mathbf{h}_k\| \\ &\le (\beta+\gamma)\xi\|\mathbf{h}_k\|^2+\xi^2\gamma\|\mathbf{h}_k\|^2 \\ &\le \left((\beta+\gamma)\xi+\gamma\xi^2\|\mathbf{S}^*\|\right)\|\mathbf{h}_k\|^2=O(\|\mathbf{h}_k\|^2)\end{aligned}$$

这说明 Term of Interest 是 $O(\|\mathbf{h}_k\|^2)$ 的，不等式得证。 □

- 条件 4 变相要求初始点 $\mathbf{x}_0$ 离 $\mathbf{x}^*$ 足够近。Hessian 矩阵和 Jacobian 矩阵内积的 Lipschitz 连续性无法保证时，点列 $\{\mathbf{x}_k\}$ 可能不收敛。这就是为什么基本 Gauss-Newton 法只有局部收敛性的原因。

阻尼 Gauss-Newton 法的全局收敛性

设在有界水平集 $L(\mathbf{x}_0)=\{\mathbf{x}|f(\mathbf{x})\le f(\mathbf{x}_0)\}$ 上，$r_i(\mathbf{x})$ 连续可微 $(i=1,\cdots,m)$，$J(\mathbf{x})$ 列满秩，则对采用 Wolfe 准则的阻尼 Gauss-Newton 法产生的迭代序列 $\{\mathbf{x}_k\}$，有 $\mathbf{g}_k\to 0, k\to\infty$

4.8.2 LMF 法

Gauss-Newton 法在迭代过程中可能出现 $\mathbf{J}_k^\top\mathbf{J}_k$ 接近奇异的情况，从而无法解出下降方向。为此，LMF 法提出在 $\mathbf{J}_k^\top\mathbf{J}_k$ 上加上一个对角阵 $v_k\mathbf{I}(v_k>0)$，使得 $\mathbf{J}_k^\top\mathbf{J}_k+v_k\mathbf{I}$ 正定。**这个方法有个良好的性质：Fletcher 发现，LMF 法是信赖域框架下的 Gauss-Newton 法**。并根据这一点，可以提出合适的 $v_k>0$ 的方法，**即让 v_k 的取值和信赖域的半径 Δ_k 有关**。

LMF 法是信赖域框架下的 Gauss-Newton 法

$\mathbf{d}_k$ 是信赖域子问题

$$\min_d q_k(\mathbf{d})=\min_d \frac{1}{2}\|\mathbf{J}_k\mathbf{d}+\mathbf{r}_k\|^2,\ \text{s.t.}\|\mathbf{d}\|\le\|\Delta_k\|$$

的解 $\iff$ 存在 $v_k\ge 0$ 使得 $(\mathbf{J}_k^\top\mathbf{J}_k+v_k\mathbf{I})\mathbf{d}_k=-\mathbf{J}_k^\top\mathbf{r}_k$，其中 $v_k(\Delta_k-\|\mathbf{d}_k\|)=0$

[注]：目标函数 $q_k(\mathbf{d})$ 即 $\mathbf{r}(\mathbf{x}_k+\mathbf{d})$ 在 $\mathbf{x}_k$ 处的一阶近似。
[注]：$(\mathbf{J}_k^\top\mathbf{J}_k+v_k\mathbf{I})\mathbf{d}_k=-\mathbf{J}_k^\top\mathbf{r}_k$ 被称为 LM 方程。

证明. **必要性：**(⇒) 建立 Lagrange 函数

$$\mathcal{L}(\mathbf{d}, v_k) = q_k(\mathbf{d}) - \frac{1}{2}v_k(\Delta_k^2 - \|\mathbf{d}\|^2)$$

那么

$$\frac{\partial \mathcal{L}}{\partial \mathbf{d}} = \mathbf{J}_k^\top \mathbf{J}_k \mathbf{d} + \mathbf{J}_k^\top \mathbf{r}_k + v_k \mathbf{d} = 0 \Rightarrow (\mathbf{J}_k^\top \mathbf{J}_k + v_k \mathbf{I})\mathbf{d} = -\mathbf{J}_k^\top \mathbf{r}_k$$

充分性：(⇐) 首先，可以发现 LM 迭代方向 $\mathbf{d}_k$ 就是如下方程的全局极小点 (只需要对下式求导就知道)

$$\tilde{q}_k(\mathbf{d}) = \frac{1}{2}\mathbf{d}^\top(\mathbf{J}_k^\top \mathbf{J}_k + v_k \mathbf{I})\mathbf{d} + \mathbf{d}^\top \mathbf{J}_k^\top \mathbf{r}_k + \frac{1}{2}\mathbf{r}_k^\top \mathbf{r}_k$$

其次，对上式稍作变形得知

$$\tilde{q}_k(\mathbf{d}) = q_k(\mathbf{d}) + \frac{1}{2}v_k\|\mathbf{d}\|^2$$

由于 $\mathbf{d}_k$ 是 $q_k(\mathbf{d})$ 的全局极小点，因此 $\tilde{q}_k(\mathbf{d}) \geq \tilde{q}_k(\mathbf{d}_k)$，那么 $\forall \mathbf{d}$

$$\begin{aligned} q_k(\mathbf{d}) - q_k(\mathbf{d}_k) &= \tilde{q}_k(\mathbf{d}) - \tilde{q}_k(\mathbf{d}_k) + \frac{1}{2}v_k\left(\|\mathbf{d}_k\|^2 - \|\mathbf{d}\|^2\right) \\ &\geq \frac{1}{2}v_k\left(\|\mathbf{d}_k\|^2 - \|\mathbf{d}\|^2\right) \end{aligned}$$

当 $v_k = 0$ 时，有 $q_k(\mathbf{d}) \geq q_k(\mathbf{d}_k)$；当 $v_k > 0$ 时，只能 $\Delta_k = \|\mathbf{d}_k\|$，无论如何，都有

$$q_k(\mathbf{d}) \geq q_k(\mathbf{d}_k) + \frac{1}{2}v_k\left(\Delta_k^2 - \|\mathbf{d}\|^2\right)$$

这意味着 $\forall v_k \geq 0$，当 $\|\mathbf{d}_k\| \leq \Delta_k$ 时，$\mathbf{d}_k$ 就是 $q_k(\mathbf{d})$ 的全局极小点，即 $\mathbf{d}_k$ 是定理中所述信赖域子问题的解。 □

根据 LM 方程

$$\|\mathbf{J}_k^\top \mathbf{J}_k + v_k \mathbf{I}\|\|\mathbf{d}_k\| = \|\mathbf{J}_k^\top \mathbf{r}_k\|$$

当 v_k 变大时，$\|\mathbf{d}_k\|$ 变小，从而 Δ_k 变小，反之亦然。因此，v_k 的调整方向应与信赖域的半径 Δ_k 的调整方向相反。类似于之前 LM 方法所述，可以给出 LMF 方法的算法流程

算法5.2(LMF方法)

1. 给出$x_0 \in \mathbb{R}^n$，$v_0 > 0$，$\epsilon > 0$，$k \leftarrow 0$；
2. 若终止条件满足，则输出有关信息，停止迭代；
3. 求解$(J_k^T J_k + v_k)d_k = -J_k^T r_k$，得到$d_k$；
4. 计算γ_k；
5. 若$\gamma_k < 0.25$，则$v_{k+1} \leftarrow 4v_k$；若$\gamma_k \geq 0.75$，则$v_{k+1} \leftarrow \frac{v_k}{2}$，否则$v_{k+1} \leftarrow v_k$，$\gamma_k < 0.25$说明近似较差，下一步要缩小半径，因此增加$v_k$，$\gamma_k \geq 0.75$说明近似较好，下一步可以扩大半径，因此减少$v_k$；
6. 若$\gamma_k \leq 0$，则$x_{k+1} \leftarrow x_k$；否则$x_{k+1} = x_k + d_k$，$k \leftarrow k+1$，转到第2步，$\gamma_k \leq 0$即 $f(x_k) < f(x_k + d_k)$ 当前半径过大导致二次函数近似失败，因此缩小半径重新近似。

4.8.3 大剩余问题 (考试不考)

Gauss-Newton 方法忽略了 Hessian 矩阵中二阶项 $S(\mathbf{x}) = \sum_{i=1}^{m} r_i(\mathbf{x})\nabla^2 r_i(\mathbf{x})$，这在剩余量小或者剩余函数非线性低的问题是一个较好的近似。但是对于剩余量大或者剩余函数非线性程度高的问题，忽略 $S(\mathbf{x})$ 会影响算法的收敛性和收敛速度。但是，直接计算 $S(\mathbf{x})$ 要计算 m 个 Hessian 矩阵 $\nabla^2 r_i(\mathbf{x})$，计算量巨大。为此，类比拟 Newton 法，我们考虑构造只包含 $r_i(\mathbf{x})$ 一阶梯度信息的近似矩阵 $\hat{\mathbf{B}}$，来近似 $S(\mathbf{x})$.

BFGS/DFP 法求解大剩余量问题

令 $\hat{\mathbf{y}}_k = (\mathbf{J}_{k+1} - \mathbf{J}_k)^\top \mathbf{r}_{k+1}$，则 $\mathbf{S}_k$ 的近似矩阵 $\hat{\mathbf{B}}_k$ 满足拟 Newton 条件，即

$$\hat{\mathbf{B}}_{k+1}\mathbf{s}_k = \hat{\mathbf{y}}_k$$

从而可以通过 BFGS 或 DFP 法迭代求解 $\mathbf{S}_k$ 的近似矩阵 $\hat{\mathbf{B}}_k$

证明. 假定在点 $\mathbf{x}_k$ 处已得 $\mathbf{S}_k$ 的近似矩阵 $\hat{\mathbf{B}}_k$，下面推导 $\mathbf{S}_{k+1}$ 的近似矩阵 $\hat{\mathbf{B}}_{k+1}$。先考虑 $\nabla^2 r_i^{(k+1)} = \nabla^2 r_i(\mathbf{x}_{k+1})$ 的近似矩阵 $\tilde{\mathbf{B}}_i^{(k+1)}$，其应满足拟 Newton 条件

$$\tilde{\mathbf{B}}_i^{(k+1)}\mathbf{s}_k = \nabla r_i^{(k+1)} - \nabla r_i^{(k)} = (\mathbf{J}_{k+1} - \mathbf{J}_k)^\top \mathbf{e}_i$$

这里 $\mathbf{e}_i$ 是第 i 个分量为 1，其他分量为 0 的 m 维向量。于是

$$\hat{\mathbf{B}}_{k+1}\mathbf{s}_k = \sum_{i=1}^{m} r_i^{(k+1)}\tilde{\mathbf{B}}_i^{(k+1)}\mathbf{s}_k = \sum_{i=1}^{m} r_i^{(k+1)}(\mathbf{J}_{k+1} - \mathbf{J}_k)^\top \mathbf{e}_i = (\mathbf{J}_{k+1} - \mathbf{J}_k)^\top \mathbf{r}_{k+1} = \hat{\mathbf{y}}_k$$

□

4.8.4 正交距离回归 (考试不考)

在之前的回归问题中，我们假定每个样本的因变量 y_i 有误差，而自变量 $\mathbf{t}_i$ 无误差。但在现实中，可能出现自变量有误差的情况。考虑自变量误差的模型在统计上被称为“变量误差模型”，对应的优化问题如果是线性的则称为完全最小二乘，非线性的则称为正交距离回归。

正交距离回归

(假设这里自变量 $\mathbf{t}_i$ 和自变量的误差 δ_i 都是 1 维的) 自变量 $\mathbf{t}_i$ 的误差记为 δ_i，因变量 y_i 的误差记为 ε_i，有以下关系

$$y_i = f_{\mathbf{x}}(t_i + \delta_i) + \varepsilon_i,\ i = 1, \cdots, m$$

我们的目标是找到模型参数 $\mathbf{x} \in \mathbb{R}^n$ 和 δ_i，使得加权剩余量最小，即

$$\min_{\mathbf{x},\boldsymbol{\delta}} \sum_{i=1}^{m} w_i^2 \left(y_i - f_{\mathbf{x}}(t_i + \delta_i)\right)^2 + d_i^2\delta_i^2 = \min_{\mathbf{x},\boldsymbol{\delta}_i} \sum_{i=1}^{2m} r_i^2(\mathbf{x}, \boldsymbol{\delta})$$

其中

$$r_i(\mathbf{x}, \boldsymbol{\delta}) = \begin{cases} w_i(y_i - f_{\mathbf{x}}(\mathbf{t}_i + \delta_i)) & i = 1, \cdots, m \\ d_{i-m}\delta_{i-m} & i = m+1, \cdots, 2m \end{cases}$$

其中 w_i 和 d_i 是权重，调节自变量和因变量误差的重要性。

如果权重 $w_i, d_i = 1$，则上述优化问题表示的就是数据点 $(\mathbf{t}_i, y_i)$ 与曲线 $f_{\mathbf{x}}(\mathbf{t}_i)$ 的距离最小。因为点到曲线的最短路径正交于曲线在交点的斜率，故该问题称为正交距离回归。

正交距离回归是一个标准的最小二乘问题，一共有 $2m$ 个剩余项，$m+n$ 个未知数 (n 个模型参数 $\mathbf{x}$ 和 m

个误差 $\boldsymbol{\delta}$)。我们先考虑其 Jacobian 矩阵 $J(\mathbf{x},\boldsymbol{\delta})\in\mathbb{R}^{2m\times(n+m)}$。可将其分成 4 个块

$$J(\mathbf{x},\boldsymbol{\delta})=\begin{bmatrix}\mathbf{J}^1_{m\times n} & \mathbf{J}^2_{m\times m}\\ \mathbf{J}^3_{m\times n} & \mathbf{J}^4_{m\times m}\end{bmatrix}$$

在 $\mathbf{J}^2$ 块中，每个元素为

$$\frac{\partial r_i}{\partial \delta_j}=\begin{cases}-w_i\dfrac{\partial f_{\mathbf{x}}(\mathbf{t}_i+\delta_i)}{\partial\delta_j} & i=j\\ 0 & i\neq j\end{cases},\quad i,j=1,\cdots,m$$

所以 $\mathbf{J}^2$ 块是一个对角阵，记为 $\mathbf{V}=\mathbf{J}^2$。在 $\mathbf{J}^3$ 块中，每个元素为

$$\frac{\partial r_i}{\partial x_j}=0,\quad i=m+1,\cdots,2m,\ j=1,\cdots,n$$

所以 $\mathbf{J}^3$ 块是一个零矩阵，记为 $\mathbf{O}_{m\times n}=\mathbf{J}^3$。在 $\mathbf{J}^4$ 块中，每个元素为

$$\frac{\partial r_i}{\partial \delta_j}=\begin{cases}0 & i\neq j\\ d_{i-m} & i=j\end{cases},\quad i=m+1,\cdots,2m,\ j=1,\cdots,m$$

所以 $\mathbf{J}^4$ 块是一个对角阵，记为 $\mathbf{D}=\mathbf{J}^4$。因此，Jacobian 矩阵 $J(\mathbf{x},\boldsymbol{\delta})$ 的形式为

$$J(\mathbf{x},\boldsymbol{\delta})=\begin{bmatrix}\hat{\mathbf{J}} & \mathbf{V}\\ \mathbf{O} & \mathbf{D}\end{bmatrix}$$

其中 $\hat{\mathbf{J}}=\mathbf{J}^1$ 的每个元素为 $\dfrac{\partial[w_i(y_i-f_{\mathbf{x}}(\mathbf{t}_i+\delta_i))]}{\partial x_j}$。我们想用 LM 方法解决正交距离回归问题，因此需要计算

$$\mathbf{J}^\top\mathbf{J}+v\mathbf{I}=\begin{bmatrix}\hat{\mathbf{J}}^\top\hat{\mathbf{J}}+v\mathbf{I} & \hat{\mathbf{J}}^\top\mathbf{V}\\ \mathbf{V}\hat{\mathbf{J}} & \mathbf{V}^2+\mathbf{D}^2+v\mathbf{I}\end{bmatrix}$$

将迭代方向和剩余量分解为两个列向量

$$\mathbf{d}=\begin{bmatrix}\mathbf{d}_x\\ \mathbf{d}_\delta\end{bmatrix},\mathbf{r}=\begin{bmatrix}\mathbf{r}_x\\ \mathbf{r}_\delta\end{bmatrix},\mathbf{d}_x\in\mathbb{R}^n,\mathbf{d}_\delta\in\mathbb{R}^m,\mathbf{r}_x\in\mathbb{R}^m,\mathbf{r}_\delta\in\mathbb{R}^m$$

于是 LM 法的迭代方程为

$$\begin{bmatrix}\hat{\mathbf{J}}^\top\hat{\mathbf{J}}+v\mathbf{I} & \hat{\mathbf{J}}^\top\mathbf{V}\\ \mathbf{V}\hat{\mathbf{J}} & \mathbf{V}^2+\mathbf{D}^2+v\mathbf{I}\end{bmatrix}\begin{bmatrix}\mathbf{d}_x\\ \mathbf{d}_\delta\end{bmatrix}=-\begin{bmatrix}\hat{\mathbf{J}}^\top\mathbf{r}_x\\ \mathbf{V}\mathbf{r}_x+\mathbf{D}\mathbf{r}_\delta\end{bmatrix}$$

由于 $\mathbf{V}^2+\mathbf{D}^2+v\mathbf{I}$ 是对角阵，很容易将 $\mathbf{d}_\delta$ 消去，从而只用求解 n 维向量 $\mathbf{d}_x$。

5 约束优化

5.1 约束优化理论

5.1.1 一阶条件

约束优化问题

对于一般优化问题

$$\min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x}) \quad \text{s.t.} \quad c_i(\mathbf{x})=0,\ i\in\epsilon, c_i(\mathbf{x})\geq 0,\ i\in\tau$$

其中 f 和 c_i 都是 $\mathbb{R}^n$ 上的光滑实值函数，ϵ 和 τ 分别是等式约束和不等式约束的索引集。记可行域为

$$\Omega=\{\mathbf{x}\in\mathbb{R}^n | c_i(\mathbf{x})=0, i\in\epsilon, c_i(\mathbf{x})\geq 0, i\in\tau\}$$

优化问题可改写为 $\min_{\mathbf{x}\in\Omega} f(\mathbf{x})$。

起作用约束

对于点 $\mathbf{x}\in\Omega$，若 $c_i(\mathbf{x})=0$，则称该约束为**起作用约束** (积极约束)。若 $c_i(\mathbf{x})>0$，则称该约束为**非起作用约束** (非积极约束)。在 $\mathbf{x}$ 处所有其作用约束的集合称为**起作用集**，记为 $\mathcal{A}(\mathbf{x})$。

- 等式约束显然都是其作用约束
- 起作用约束限制了 $\mathbf{x}$ 的迭代方向，而不起作用约束未限制 $\mathbf{x}$ 在小范围内的迭代方向 (直观上，往点 $\mathbf{x}$ 的各个方向走很小的一步仍然满足约束)

在引入一般的约束优化问题之前，我们先以三个例子来引入约束优化的**一阶必要条件**，并观察 Lagrange 乘子的符号：

1. **等式约束优化问题**：问题的最优解 $\mathbf{x}^*=(-1,-1)^\top$，$g(\mathbf{x}^*)=-\frac{1}{2}a_1(\mathbf{x}^*)^2$，**此处 Lagrangian 乘子 λ_1 的符号无所谓.**

   $$\min_{x_1,x_2} x_1+x_2,\ \text{s.t. } x_1^2+x_2^2-2=0$$

   可行条件：$c_1(\mathbf{x}+\mathbf{d})\geq 0\Rightarrow a_1(\mathbf{x})^\top\mathbf{d}=0$，下降条件：$f(\mathbf{x}+\mathbf{d})<f(\mathbf{x})\Rightarrow g(\mathbf{x})^\top\mathbf{d}<0$
   如果 $\mathbf{x}^*$ 为最优点，要让上述两条件不能同时成立，那么

   - $g(\mathbf{x}^*)$ 和 $a_1(\mathbf{x}^*)$ 同向，即 $g(\mathbf{x}^*)=\lambda_1 a_1(\mathbf{x}^*)$

2. **一个不等式约束**：问题的最优解 $\mathbf{x}^*=(-1,-1)^\top$，$g(\mathbf{x}^*)=\frac{1}{2}a_1(\mathbf{x}^*)^2$，**但此处 Lagrange 乘子 λ_1 的符号必须大于等于 0.**

   $$\min_{x_1,x_2} x_1+x_2,\ \text{s.t. } 2-x_1^2-x_2^2\geq 0$$

   可行条件：$c_1(\mathbf{x}+\mathbf{d})\geq 0\Rightarrow c_1(\mathbf{x})+a_1(\mathbf{x})^\top\mathbf{d}\geq 0$，下降条件：$f(\mathbf{x}+\mathbf{d})<f(\mathbf{x})\Rightarrow g(\mathbf{x})^\top\mathbf{d}<0$
   如果 $\mathbf{x}^*$ 为最优点，要让上述两条件不能同时成立，那么

   - 如果 $c_1(\mathbf{x}^*)>0$，可行条件不起作用，从而必须 $g(\mathbf{x}^*)=0$

- 如果 $c_1(\mathbf{x}^*)=0$，要么 $g(\mathbf{x}^*)=0$，要么 $g(\mathbf{x}^*)$ 和 $a_1(\mathbf{x}^*)$ 同向，即 $g(\mathbf{x}^*)=\lambda_1 a_1(\mathbf{x}^*), \lambda_1>0$

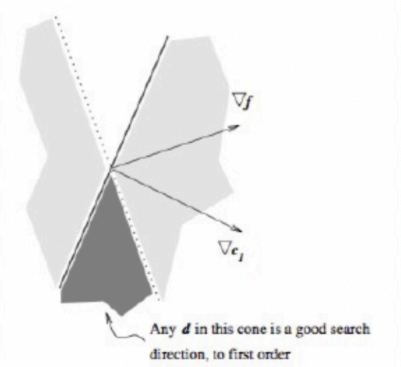


图 14: 只有 $g(\mathbf{x}^*)$ 和 $a_1(\mathbf{x}^*)$ 同向时，半平面 $g(\mathbf{x})^\top\mathbf{d}<0$ 和半平面 $a_1(\mathbf{x})^\top\mathbf{d}\geq 0$ 不会有重叠的部分

用 Lagrangian 函数可以将上述情况总结在一起：记 $\mathcal{L}(\mathbf{x},\lambda_1)=f(\mathbf{x})-\lambda_1 c_1(\mathbf{x})$，如果 $\mathbf{x}^*$ 为最优解，那么必然满足

- $\nabla\mathcal{L}(\mathbf{x}^*,\lambda_1^*)=g(\mathbf{x}^*)-\lambda_1^* a_1(\mathbf{x}^*)=0, \lambda_1^*\geq 0$
- **互补条件**：$\lambda_1^* c_1(\mathbf{x}^*)=0$，Lagrangian 乘子和约束至少有一个为 0。

3. **两个不等式约束**:问题的最优解 $\mathbf{x}^*=(-\sqrt{2},0)^\top, g(\mathbf{x}^*)=(1,1), a_1(\mathbf{x}^*)=(0,2\sqrt{2}), a_2(\mathbf{x}^*)=(1,0), g(\mathbf{x}^*)=\frac{1}{2\sqrt{2}}a_1(\mathbf{x}^*)+a_2(\mathbf{x}^*)$，**此处 Lagrange 乘子 λ_1,λ_2 的符号必须大于等于 0.**

$$\min_{x_1,x_2} x_1+x_2,\ \text{s.t. } 2-x_1^2-x_2^2\geq 0, x_2\geq 0$$

可行条件：$c_i(\mathbf{x}+\mathbf{d})\geq 0 \Rightarrow c_i(\mathbf{x})+a_i(\mathbf{x})^\top\mathbf{d}\geq 0$，下降条件：$f(\mathbf{x}+\mathbf{d})<f(\mathbf{x})\Rightarrow g(\mathbf{x})^\top\mathbf{d}<0$
如果 $\mathbf{x}^*$ 是最优点：

(a) 如果 $c_1(\mathbf{x}^*)>0, c_2(\mathbf{x}^*)>0$，两个可行条件都不起作用，从而必须 $g(\mathbf{x}^*)=0$

(b) 如果 $c_1(\mathbf{x}^*), c_2(\mathbf{x}^*)$ 中有一个为 0，转到“一个不等式约束”的情况

(c) 如果 $c_1(\mathbf{x}^*)=c_2(\mathbf{x}^*)=0$，那么 $a_i(\mathbf{x}^*)^\top\mathbf{d}\geq 0$ 和 $g(\mathbf{x}^*)^\top\mathbf{d}<0$ 不能同时成立，只有当 $g(\mathbf{x}^*)$ 在 $a_1(\mathbf{x}^*), a_2(\mathbf{x}^*)$ 之间才行，换言之

$$g(\mathbf{x}^*)=\lambda_1 a_1(\mathbf{x}^*)+\lambda_2 a_2(\mathbf{x}^*), \lambda_1,\lambda_2\geq 0$$

用 Lagrangian 函数可以将上述情况总结在一起：记 $\mathcal{L}(\mathbf{x},\boldsymbol{\lambda})=f(\mathbf{x})-\lambda_1 c_1(\mathbf{x})-\lambda_2 c_2(\mathbf{x})$，如果 $\mathbf{x}^*$ 为最优解，那么必然满足

- $\nabla\mathcal{L}(\mathbf{x}^*,\boldsymbol{\lambda}^*)=g(\mathbf{x}^*)-\lambda_1^* a_1(\mathbf{x}^*)-\lambda_2^* a_2(\mathbf{x}^*)=0, \lambda_1^*,\lambda_2^*\geq 0$
- **互补条件**：$\lambda_1^* c_1(\mathbf{x}^*)=0, \lambda_2^* c_2(\mathbf{x}^*)=0$

从上述三个例子可以抽象出**一般约束问题的求解方法**：对于一般约束优化问题，列出 Lagrangian 函数

$$\mathcal{L}(\mathbf{x},\boldsymbol{\lambda})=f(\mathbf{x})-\sum_{i\in\epsilon}\lambda_i c_i(\mathbf{x})-\sum_{i\in\tau}\mu_i c_i(\mathbf{x})$$

从上式可以看出，如果 $\mathbf{x}^*$ 为最优解，那么必然满足 $\nabla\mathcal{L}(\mathbf{x}^*,\boldsymbol{\lambda}^*)=\mathbf{g}(\mathbf{x}^*)-\sum_{i\in\epsilon\cup\tau}\lambda_i^* a_i(\mathbf{x}^*)=0$。换言之，在最优解 $\mathbf{x}^*$ 上，**梯度 $g(\mathbf{x}^*)$ 是约束梯度 $a_i(\mathbf{x}^*)$ 的线性组合**。

然而在某些非常特殊的约束条件下，未必**梯度 $g(\mathbf{x}^*)$ 是约束梯度 $a_i(\mathbf{x}^*)$ 的线性组合**。例如

$$\min_{x_1,x_2} x_1 + x_2,\ \text{s.t. } x_1^3 \geq x_2,\ x_2 \geq 0$$

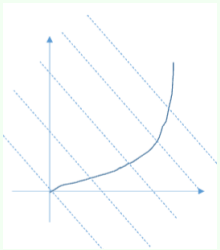


图 15: 函数等高线与约束条件

容易注意到约束下的最优解为 $(0,0)^\top$，但最优解处 $g(\mathbf{x}^*) = (1,1)^\top$，$a_1(\mathbf{x}^*) = (0,-1)^\top$，$a_2(\mathbf{x}^*) = (0,1)^\top$，显然 $g(\mathbf{x}^*)$ 不是约束梯度 $a_i(\mathbf{x}^*)$ 的线性组合。**出现这种情况的原因是 $a_i(\mathbf{x}^*)$ 是互相线性相关的。为了避免这种情况，我们会引入约束规范。**

线性独立约束规范 (LICQ)

给定最优解 $\mathbf{x}^*$ 和其对应的起作用集 $\mathcal{A}(\mathbf{x}^*)$，若 $a_i(\mathbf{x}^*) = \nabla c_i(\mathbf{x}^*),\ i \in \mathcal{A}(\mathbf{x}^*)$ 线性无关，则称约束优化问题满足**线性独立约束规范** (LICQ)。

LICQ 成立可以确保

- $g(\mathbf{x}^*) = \sum_i \lambda_i a_i(\mathbf{x}^*)$ 一定成立
- 所有的 λ_i 唯一

LICQ 是一个比较强的约束规范，有时候会过于严格。还有一些更弱的规范条件 (如 KT 约束规范、正则约束规范等)，此处略去不表。

KKT 条件 (一阶必要条件)

$\mathbf{x}^*, \boldsymbol{\lambda}^*$ 一定满足 **KKT 条件**：

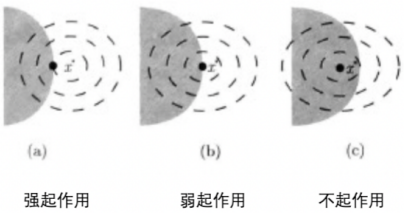
- **可行性条件**：$c_i(\mathbf{x}^*) = 0,\ i \in \epsilon, c_i(\mathbf{x}^*) \geq 0,\ i \in \tau$
- **梯度条件**：$\nabla \mathcal{L}(\mathbf{x}^*, \boldsymbol{\lambda}) = g(\mathbf{x}^*) - \sum_{i \in \epsilon \cup \tau} \lambda_i^* a_i(\mathbf{x}^*) = 0$
- **互补条件**：$\lambda_i^* c_i(\mathbf{x}^*) = 0,\ i \in \epsilon \cup \tau$
- **非负性条件**：$\lambda_i^* \geq 0,\ i \in \tau$

$\Leftarrow \mathbf{x}^*$ 是约束优化问题的最优解，且约束优化问题满足 LICQ

[注]: 满足 KKT 条件的点对 $(\mathbf{x}^*, \boldsymbol{\lambda}^*)$ 称为 **KKT 对**。

- 互补条件说明 $\lambda_i^*, c_i(\mathbf{x}^*)$ 不可能同时为正。
- 如果某个约束条件 i 不起作用，那么 $\lambda_i^* = 0$。

- 如果某个约束条件 i 起作用，那么 λ_i^* 可能等于 0，可能大于 0。
  - $\lambda_i^* = 0$ 时，如果撤掉此约束条件 i，最优解不变，$\Rightarrow$ **弱起作用约束**。
  - $\lambda_i^* > 0$ 时，如果撤掉此约束条件 i，最优解会发生变化，$\Rightarrow$ **强起作用约束**。



- 如果对每一个起作用的不等式约束 i，$\lambda_i^* > 0$，则称之为**严格互补条件**。

实际上，Lagrangian 乘子 λ_i^* 度量了最优目标函数值 $f(\mathbf{x}^*)$ 对于约束 c_i 的敏感程度。根据 KKT 条件，对于不起作用的约束 $c_i(\mathbf{x}^*) \geq 0$，$\lambda_i^* = 0$，说明对 $c_i(\mathbf{x}^*)$ 的扰动不会影响最优解 $\mathbf{x}^*$。而对于起作用的约束 $c_i(\mathbf{x}^*) = 0$，不妨设对其右端进行微小的扰动 $c_i(\mathbf{x}) \geq -\varepsilon\|a_i(\mathbf{x}^*)\|$，以及经扰动后的最优解 $\mathbf{x}^*(\varepsilon)$。**这里假设 $\varepsilon > 0$ 充分小，以至于起作用的约束集不会发生变化**。那么

$$f(\mathbf{x}^*(\varepsilon)) - f(\mathbf{x}^*) \approx (\mathbf{x}^*(\varepsilon) - \mathbf{x}^*)^\top g(\mathbf{x}^*) = \sum_{i \in \mathcal{A}(\mathbf{x}^*)} \lambda_i^* (\mathbf{x}^*(\varepsilon) - \mathbf{x}^*)^\top a_i(\mathbf{x}^*)$$

其中

$$\begin{aligned} -\varepsilon\|a_i(\mathbf{x}^*)\| \approx c_i(\mathbf{x}^*(\varepsilon)) - c_i(\mathbf{x}^*) \approx (\mathbf{x}^*(\varepsilon) - \mathbf{x}^*)^\top a_i(\mathbf{x}^*) \\ 0 = c_j(\mathbf{x}^*(\varepsilon)) - c_j(\mathbf{x}^*) \approx (\mathbf{x}^*(\varepsilon) - \mathbf{x}^*)^\top a_j(\mathbf{x}^*),\ j \in \mathcal{A}(\mathbf{x}^*), j \neq i \end{aligned}$$

于是

$$f(\mathbf{x}^*(\varepsilon)) - f(\mathbf{x}^*) \approx -\varepsilon\lambda_i^*\|a_i(\mathbf{x}^*)\| \Rightarrow \frac{\mathrm{d}f(\mathbf{x}^*(\varepsilon))}{\mathrm{d}\varepsilon} = -\lambda_i^*\|a_i(\mathbf{x}^*)\|$$

容易看出，如果 $\lambda_i\|a_i(\mathbf{x}^*)\|$ 大，则最优值对约束 $c_i(\mathbf{x}) \geq 0$ 敏感。

线性可行方向集

设 $\mathbf{x}$ 为问题的可行点，$\tau(\mathbf{x}) = \{i : c_i(\mathbf{x}) = 0, i \in \tau\}$ 为 $\mathbf{x}$ 处起作用的不等式约束集合，$\mathcal{F}(\mathbf{x}) = \{d : \|d\| = 1, a_i(\mathbf{x})^\top \mathbf{d} = 0, i \in \epsilon, a_i(\mathbf{x})^\top \mathbf{d} \geq 0, i \in \tau(\mathbf{x})\}$ 为 $\mathbf{x}$ 处的**线性可行方向集**。

假设 $\mathbf{x}$ 为可行点，$\mathbf{x} + \mathbf{d}$ 仍为可行点，那么

$$c_i(\mathbf{x} + \mathbf{d}) = c_i(\mathbf{x}) + a_i(\mathbf{x})^\top \mathbf{d}$$

如果 $c_i(\mathbf{x})$ 为等式约束 (一定起作用)，可以推出 $a_i(\mathbf{x})^\top \mathbf{d} = 0$，如果 $c_i(\mathbf{x})$ 为不等式约束 (但在 $\mathbf{x}$ 处起作用)，可以推出 $a_i(\mathbf{x})^\top \mathbf{d} \geq 0$。由此构造出 $\mathcal{F}(\mathbf{x})$. **值得注意的是，这里只用了约束的一阶梯度信息构造可行方向，故称为「线性」可行方向**。如果约束条件很复杂，仅用一阶梯度信息可能构造出错误的可行方向。但在这门课中，处于简便起见，**我们默认线性可行方向等于可行方向 (本质就是 KT 约束规范)**。

一阶充分条件

$\mathbf{x}^*$ 在可行域中，$g(\mathbf{x}^*)^\top \mathbf{d} > 0,\ \forall \mathbf{d} \in \mathcal{F}(\mathbf{x}^*) \Rightarrow \mathbf{x}^*$ 是问题的严格局部最优解。

5.1.2 二阶条件

但对于 $g(\mathbf{x}^*)^\top \mathbf{d} = 0$ 的可行方向，只通过一阶梯度无法判断 $\mathbf{x}^*$ 是否为最优解 (例如鞍点)，需要二阶梯度信息。定义

$$\mathcal{F}_1(\mathbf{x}^*) = \{\mathbf{d} : g(\mathbf{x}^*)^\top \mathbf{d} = 0, \mathbf{d} \in \mathcal{F}(\mathbf{x}^*)\}$$

显然 $\mathcal{F}_1(\mathbf{x}^*) \subset \mathcal{F}(\mathbf{x}^*)$。注意到在 KKT 点 $\mathbf{x}^*$，因为起作用约束的 Lagrangian 乘子 λ_i^* 都是非负的，所以

$$g(\mathbf{x}^*)^\top \mathbf{d} = \sum_{c_i \in \mathcal{A}(\mathbf{x}^*)} \lambda_i^* a_i(\mathbf{x}^*)^\top \mathbf{d} \geq 0,\ \mathbf{d} \in \mathcal{F}(\mathbf{x}^*)$$

即 KKT 点上的所有可行方向要么是上升方向，要么是满足 $g(\mathbf{x}^*)^\top \mathbf{d} = 0$ 的「不变方向」。所以 $\mathcal{F}_1$ 可以写为

$$\begin{aligned}\mathcal{F}_1(\mathbf{x}^*) = \{\mathbf{d} : \mathbf{d} \neq \mathbf{0}, a_i(\mathbf{x}^*)^\top \mathbf{d} \geq 0, \lambda_i^* = 0, i \in \tau(\mathbf{x}^*) \quad &\text{(弱起作用约束)}\\ , a_i(\mathbf{x}^*)^\top \mathbf{d} = 0, \lambda_i^* > 0, i \in \tau(\mathbf{x}^*) \quad &\text{(强起作用约束)}\\ , a_i(\mathbf{x}^*)^\top \mathbf{d} = 0, i \in \epsilon\} \quad &\text{(等式约束)}\end{aligned}$$

这样就可以使得 $g(\mathbf{x}^*)^\top \mathbf{d} = \sum_{c_i \in \mathcal{A}(\mathbf{x}^*)} \lambda_i^* a_i(\mathbf{x}^*)^\top \mathbf{d} = 0$。

二阶必要条件

$\forall \mathbf{d} \in \mathcal{F}_1(\mathbf{x}^*),\ \mathbf{d}^\top \nabla^2 \mathcal{L}(\mathbf{x}^*, \boldsymbol{\lambda}^*)\mathbf{d} \geq 0 \Leftarrow \mathbf{x}^*$ 是问题的局部最优解，$\mathbf{x}^*$ 处 LICQ 条件成立，从而存在 $\boldsymbol{\lambda}^*$，使得 $\mathbf{x}^*, \boldsymbol{\lambda}^*$ 满足 KKT 条件.

二阶充分条件

$\mathbf{x}^*, \boldsymbol{\lambda}^*$ 满足 KKT 条件，且 $\forall \mathbf{d} \in \mathcal{F}_1(\mathbf{x}^*),\ \mathbf{d}^\top \nabla^2 \mathcal{L}(\mathbf{x}^*, \boldsymbol{\lambda}^*)\mathbf{d} > 0 \Rightarrow \mathbf{x}^*$ 是问题的严格局部最优解.

[注]: $\nabla^2 \mathcal{L}(\mathbf{x}^*, \boldsymbol{\lambda}^*) = \nabla^2 f(\mathbf{x}^*) - \sum_{i \in \epsilon \cup \tau} \lambda_i^* \nabla^2 c_i(\mathbf{x}^*)$

[注]: 如果我们能直接判断 $\nabla^2 \mathcal{L}(\mathbf{x}^*, \boldsymbol{\lambda}^*)$ 是正定的，那么都不需要计算出 $\mathcal{F}_1(\mathbf{x}^*)$ 就知道 $\mathbf{x}^*$ 是严格局部最优解。但如果不能判断，那么还需要计算 $\mathcal{F}_1(\mathbf{x}^*)$，只要 $\mathbf{d}^\top \nabla^2 \mathcal{L}(\mathbf{x}^*, \boldsymbol{\lambda}^*)\mathbf{d},\ \forall \mathbf{d} \in \mathcal{F}_1(\mathbf{x}^*)$ 都大于 0，即使 $\nabla^2 \mathcal{L}(\mathbf{x}^*, \boldsymbol{\lambda}^*)$ 本身不是正定的，在约束条件下 $\mathbf{x}^*$ 也是严格局部最优解。

5.2 罚函数方法

5.2.1 外点罚函数方法

一般约束优化问题的罚函数方法

对于一般约束优化问题

$$\min_{\mathbf{x} \in \mathbb{R}^n} f(\mathbf{x}) \quad \text{s.t.} \quad c_i(\mathbf{x}) = 0,\ i \in \epsilon, c_i(\mathbf{x}) \geq 0,\ i \in \tau$$

定义如下罚函数

$$P(\mathbf{x}, \sigma) = f(\mathbf{x}) + \frac{\sigma}{2}\left(\sum_{i \in \epsilon} c_i(\mathbf{x})^2 + \sum_{i \in \tau} \min\{0, c_i(\mathbf{x})\}^2\right)$$

然后做无约束优化。

- 非可行点的惩罚项非 0，故称为**外点罚函数**。
- $P(\mathbf{x},\sigma)$ 的最优点往往不是约束优化的最优点。对非可行点，让 $\sigma\to\infty$，增大惩罚项的比重，则最小化 $P(\mathbf{x},\sigma)$ 会迫使最优解靠近可行域。
- 当 $\mathbf{x}$ 已经在可行域内时 (即 $c_i(\mathbf{x})\geq 0$)，惩罚项 $\min\{0,c_i(\mathbf{x})\}^2$ 就是 0；只有当 $\mathbf{x}$ 不在可行域内时，惩罚项才会起作用，迫使 $\mathbf{x}$ 逼近可行域。

直观上，最小化 $P(\mathbf{x},\sigma)$ 的过程中，一方面在最小化 $f(\mathbf{x})$，另一方面也让约束 $c_i(\mathbf{x})$ 越来越趋近于 0。算法步骤如下所示：

算法 7.1 (外点罚函数方法)

步 1　给定 $\sigma_1>0,\varepsilon_1>0,\varepsilon>0,x_0,k:=1$.

步 2　以 x_{k-1} 为初始点，求 $x(\sigma_k)=\arg\min \mathrm{P_E}(x,\sigma_k)$；求解该无约束最优化问题的算法当 $\|\nabla \mathrm{P_E}(x(\sigma_k),\sigma_k)\|\leqslant\varepsilon_1$ 时停止.

步 3　当 $\|c(x(\sigma_k))\|\leqslant\varepsilon$，迭代停止.

步 4　$x_k:=x(\sigma_k)$，选 $\sigma_{k+1}>\sigma_k,k:=k+1$，转步 2.

关于算法 7.1 的说明如下：

• 算法的步 2 为内层子迭代，我们可用某一合适的求解无约束最优化问题的方法迭代求解 $\min \mathrm{P_E}(x,\sigma_k)$，用 $x(\sigma_{k-1})$ 作为这一子迭代的初始点. 在第 $k-1$ 步迭代得到的在点 $x(\sigma_{k-1})$ 的其他信息，亦可代入第 k 步迭代中.

• 在算法的步 2 中，我们假定 $\mathrm{P_E}(x,\sigma_k)$ 的局部极小点 $x(\sigma_k)$ 是存在的. 若非如此，增大 σ_k 后重新求解.

• 如何选取递增序列 $\{\sigma_k\}$ 的问题直接影响到算法的有效性. 如果我们让该序列增长很快，这会影响无约束子问题的求解，因为每一步无约束最优化问题的解是下一步无约束最优化问题的初始点. 而如果我们让该序列增长缓慢，无约束最优化问题固然可以更好地求解，然而迭代的速度必然会受到影响. 一般地，我们可选该序列为 $\{10^k\}$.

外点罚函数方法的全局收敛性

假设 $f(\mathbf{x})$ 在可行域上有下界，且 $\mathbf{x}_k$ 是 $P(\mathbf{x},\sigma_k)$ 的全局极小点，若序列 $\{\sigma_k\}$ 满足 $\sigma_{k+1}\geq\sigma_k$，则当 $\sigma_k\to\infty$ 时，$\{\mathbf{x}_k\}$ 的任何聚点 $\mathbf{x}^*$ 为约束优化问题的全局最优解。

证明. □

但这一收敛定理在每一步都需要求解 $P(\mathbf{x},\sigma_k)$ 的全局最小点，这是很难做到的，但下面的定理没有这样的要求

外点罚函数方法的收敛性

若在算法步骤 2 中 $\|\nabla_{\mathbf{x}}P(\mathbf{x}_k,\sigma_k)\|\leq\varepsilon_k$，且 $\lim\limits_{k\to\infty}\varepsilon_k=0,\sigma_k\to\infty$，对 $\{\mathbf{x}_k\}$ 的任意极限点 $\mathbf{x}^*,\nabla c_i(\mathbf{x}),i\in\epsilon$ 线性无关，则 $\mathbf{x}^*$ 是约束优化问题的 KKT 点，且

$$\lim_{k\to\infty}(-\sigma_k c_i(\mathbf{x}_k))=\lambda_i^*$$

其中 $\boldsymbol{\lambda}^*$ 是 $\mathbf{x}^*$ 的 Lagrangian 乘子。

证明. 注意到

$$\nabla_{\mathbf{x}} P(\mathbf{x}_k, \sigma_k) = \mathbf{g}_k + \sum_{i\in\epsilon} c_i(\mathbf{x}_k)\sigma_k \nabla c_i(\mathbf{x}_k)$$

当 $k \to \infty$ 时，由于 $\|\nabla_{\mathbf{x}} P(\mathbf{x}_k, \sigma_k)\| \to 0$，等式右边也应趋于 0，因此 $\{\mathbf{x}_k\}$ 的极限点 $\mathbf{x}^*$ 满足

$$g(\mathbf{x}^*) = \sum_{i\in\epsilon} -\sigma_\infty c_i(\mathbf{x}^*)\nabla c_i(\mathbf{x}^*)$$

变形得 $\sum_{i\in\epsilon} c_i(\mathbf{x}^*)\nabla c_i(\mathbf{x}^*) = -\dfrac{1}{\sigma_\infty} g(\mathbf{x}^*) \to 0$。又因为 $\{\nabla c_i(\mathbf{x}), i \in \epsilon\}$ 线性无关，所以 $c_i(\mathbf{x}^*) = 0, i \in \epsilon$。所以 $\mathbf{x}^*$ 为约束优化问题的 KKT 点。

根据 $g(\mathbf{x}^*) = \sum_{i\in\epsilon} -\sigma_\infty c_i(\mathbf{x}^*)\nabla c_i(\mathbf{x}^*)$，又 $\mathbf{x}^*$ 为 KKT 点，对比系数知 Lagrangian 乘子满足

$$\lambda_i^* = -\sigma_\infty c_i(\mathbf{x}^*) = \lim_{k\to\infty} (-\sigma_k c_i(\mathbf{x}_k))$$

□

外点罚函数方法的数值困难问题

当 σ_k 越来越大时，$\nabla^2_{\mathbf{xx}} P(\mathbf{x}, \sigma_k)$ 的病态性越来越严重，极小化 $P(\mathbf{x}, \sigma_k)$ 的数值困难性也越来越大。

证明. 以等式约束优化问题为例，令

$$\mathbf{G}(\mathbf{x}) = \Big[\nabla c_1(\mathbf{x}), \nabla c_2(\mathbf{x}), \cdots, \nabla c_m(\mathbf{x})\Big]$$

可以计算出

$$\nabla^2_{\mathbf{xx}} P(\mathbf{x}, \sigma_k) = \nabla^2_{\mathbf{xx}} f(\mathbf{x}) + \sum_{i\in\epsilon} \sigma_k c_i(\mathbf{x})\nabla^2 c_i(\mathbf{x}) + \sigma_k \mathbf{G}(\mathbf{x})\mathbf{G}(\mathbf{x})^\top$$

根据上面外点罚函数方法的收敛性定理，可以注意到当 $\sigma_k \to \infty$ 时，$\nabla^2_{\mathbf{xx}} P(\mathbf{x}, \sigma_k)$ 展开的前两项近似于 Lagrangian 函数的 Hessian 矩阵，即

$$\nabla^2_{\mathbf{xx}} P(\mathbf{x}, \sigma_k) \approx \nabla^2_{\mathbf{xx}} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}^*) + \sigma_k \mathbf{G}(\mathbf{x})\mathbf{G}(\mathbf{x})^\top$$

后面一项 $\sigma_k \mathbf{G}(\mathbf{x})\mathbf{G}(\mathbf{x})^\top$ 是一个秩为 m 的矩阵，其中 m 个特征值随 $\sigma_k \to \infty$ 而趋于无穷。故 $\nabla^2_{\mathbf{xx}} P(\mathbf{x}, \sigma_k)$ 的条件数会趋于无穷，这就是病态性的来源。
(更直观一些的说，如果退化到二维情况，等高线会随着 $\sigma_k \to \infty$ 而变得越来越扁，这就是病态性的体现。) □

5.2.2 障碍函数法/内点罚函数方法

障碍函数法也是把约束优化问题转化为无约束优化问题。与外点罚函数方法不同

1. 外点罚函数方法是由**可行域外**逼近约束优化问题的最优解的，障碍函数方法是由**可行域内**部逼近约束优化问题的最优解。
2. 外点罚函数方法可以解决**等式和不等式约束**优化问题，而障碍函数方法则适宜解决**不等式约束**优化问题。

不等式约束问题的障碍函数法

考虑不等式约束优化问题

$$\min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x}) \quad \text{s.t.} \quad c_i(\mathbf{x}) \geq 0,\ i \in \tau$$

可以写出倒数障碍函数

$$B_I(\mathbf{x},\mu) = f(\mathbf{x}) + \mu \sum_{i\in\tau} \frac{1}{c_i(\mathbf{x})}$$

或对数障碍函数

$$B_L(\mathbf{x},\mu) = f(\mathbf{x}) - \mu \sum_{i\in\tau} \ln c_i(\mathbf{x})$$

其中 $\mu > 0$ 为障碍因子。

- 当不等式约束 $c_i(\mathbf{x})$ 趋近于 0(也就是逼近可行域边界时)，障碍项 $\mu \sum_{i\in\tau} \frac{1}{c_i(\mathbf{x})}$ 或 $-\mu \sum_{i\in\tau} \ln c_i(\mathbf{x})$ 会趋于无穷，就好像“一堵墙”一样阻止 $\mathbf{x}$ 跳到可行域外。
- 障碍函数的极小点一定为**严格可行点** (即 $c_i(\mathbf{x}) > 0$)，在最小化 $B_I(\mathbf{x},\mu)$ 或 $B_L(\mathbf{x},\mu)$ 的过程中，为了最小化惩罚项，$c_i(\mathbf{x})$ 会变为很大的正值。
- 因为约束优化问题的最优解可能在边界上，为了使障碍函数极小点逼近可行域边界，需要令 $\mu \to 0$，尽可能降低障碍项的值。

如果约束条件中含有等式约束，也可以使用一种结合了**障碍函数和外点罚函数的方法**

一般约束优化问题的障碍函数方法

对于一般约束优化问题

$$\min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x}) \quad \text{s.t.} \quad c_i(\mathbf{x}) = 0,\ i \in \epsilon, c_i(\mathbf{x}) \geq 0,\ i \in \tau$$

定义如下障碍函数

$$B(\mathbf{x},\mu) = f(\mathbf{x}) - \mu \sum_{i\in\tau} \ln c_i(\mathbf{x}) + \frac{1}{2\mu} \sum_{i\in\epsilon} c_i(\mathbf{x})^2$$

然后做无约束优化。

容易发现，$\mu \sum_{i\in\tau} \ln c_i(\mathbf{x})$ 就是障碍函数的惩罚项，$\frac{1}{2\mu} \sum_{i\in\epsilon} c_i(\mathbf{x})^2$ 是外点罚函数的惩罚项，只不过用 $\sigma = \frac{1}{2\mu} \to \infty$，$\mu \to 0$ 代替了 $\sigma \to \infty$。算法步骤如下所示：

算法 7.2 (障碍函数方法)

步 1 给定初始内点 x_0, $\mu_1, \varepsilon_1 > 0$, $\varepsilon > 0$, $k := 1$.

步 2 以 x_{k-1} 为初始点，求 $x(\mu_k) = \arg\min \mathrm{B}(x, \mu_k)$，其迭代当 $\|\nabla_x \mathrm{B}(x(\mu_k), \mu_k)\| \leqslant \varepsilon_1$ 时停止.

步 3 当 $\mu_k \sum_{i\in I} c_i(x(\mu_k))^{-1} \leqslant \varepsilon$ 时，迭代停止；否则 $x_k = x(\mu_k)$，选 $\mu_{k+1} < \mu_k$, $k := k+1$, 转步 2.

障碍函数法的数值困难问题

当 $\mu \to 0$ 时，$\nabla^2_{\mathbf{xx}} B(\mathbf{x}, \mu)$ 的病态性越来越严重，极小化 $B(\mathbf{x}, \mu)$ 的数值困难性也越来越大。

证明. 以对数障碍函数为例，

$$\nabla^2_{\mathbf{xx}} B_L(\mathbf{x}, \mu_k) = \nabla^2_{\mathbf{xx}} f(\mathbf{x}) - \sum_{i \in \tau} \frac{\mu_k}{c_i(\mathbf{x})} \nabla^2 c_i(\mathbf{x}) + \sum_{i \in \tau} \frac{\mu_k}{c_i^2(\mathbf{x})} \nabla c_i(\mathbf{x}) \nabla c_i(\mathbf{x})^\top$$

令 $\mu_i^{(k)} = \dfrac{\mu_k}{c_i(\mathbf{x}_k)}$，可以注意到当 $\mu_k \to 0$ 时，$\lambda_i^{(k)} \to \lambda_i^*$，$\nabla^2_{\mathbf{xx}} B_L(\mathbf{x}, \mu_k)$ 展开的前两项近似于 Lagrangian 函数的 Hessian 矩阵，即

$$\nabla^2_{\mathbf{xx}} B_L(\mathbf{x}, \mu_k) \approx \nabla^2_{\mathbf{xx}} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}^*) + \sum_{i \in \tau} \frac{(\lambda_i^*)^2}{\mu_k} \nabla c_i(\mathbf{x}) \nabla c_i(\mathbf{x})^\top$$

当 $\mu_k \to 0$ 时，$\nabla^2_{\mathbf{xx}} B_L(\mathbf{x}, \mu_k)$ 的条件数会趋于无穷，这就是病态性的来源。 □

5.2.3 增广 Lagrangian 函数法

外点罚函数和障碍函数法的主要问题在于，两种方法分别都需要令 $\sigma \to \infty$ 和 $\mu \to 0$ 才能让最优解逼近约束优化问题的最优解。但 $\sigma \to \infty$ 或 $\mu \to 0$ 会导致无约束优化问题 Hessian 矩阵的病态性。我们希望构造某个函数，不需要无穷大的 σ_k 或 μ_k，就能使得无约束优化最优解逼近约束优化问题的最优解。假设这个函数是 $\phi(\mathbf{x}, \sigma)$，我们希望满足 $\mathbf{x}^* = \arg\min_{\mathbf{x}} \phi(\mathbf{x}, \sigma)$ 就是约束优化问题的最优解。根据一阶/二阶条件，其必须满足

$$\begin{aligned} \nabla_{\mathbf{x}} \phi(\mathbf{x}^*, \sigma) &= 0 && \text{(一阶必要条件)} \\ \forall \mathbf{d},\ \mathbf{d}^\top \nabla_{\mathbf{xx}} \phi(\mathbf{x}^*, \sigma) \mathbf{d} &\geq 0 && \text{(二阶充分条件)} \end{aligned}$$

如果取 $\phi(\mathbf{x}, \sigma)$ 为

1. 外点罚函数：**对于等式约束优化问题**：$\nabla_{\mathbf{x}} P(\mathbf{x}^*, \sigma) = g(\mathbf{x}^*) + \sigma \sum_{i \in \epsilon \cup \tau} c_i(\mathbf{x}) a_i(\mathbf{x}) = g(\mathbf{x}^*) \neq 0$，不满足一阶必要条件。
2. Lagrangian 函数：一阶必要条件显然满足，二阶必要条件只在**可行方向**上满足！

由此有一个修改 Lagrangian 函数的思路：对于**等式约束优化问题**，不改变 Lagrangian 函数在 $\mathbf{x}^*$ 可行方向的值，但**提高**其沿着**不可行方向**的值，直到 Lagrangian 函数沿着任意方向的二阶方向导数都是正的。相当于当变量离开**等式约束**时就增加一个惩罚项，加大 Lagrangian 函数的函数值。

等式约束优化问题的增广 Lagrangian 函数

对于等式约束优化问题

$$\min_{\mathbf{x} \in \mathbb{R}^n} f(\mathbf{x}) \quad \text{s.t.} \quad c_i(\mathbf{x}) = 0,\ i \in \epsilon$$

定义增广 Lagrangian 函数

$$\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \sigma) = f(\mathbf{x}) - \sum_{i \in \epsilon} \lambda_i c_i(\mathbf{x}) + \frac{\sigma}{2} \sum_{i \in \epsilon} c_i(\mathbf{x})^2$$

其中取合适的 $\sigma > 0$ 使得 $\nabla_{\mathbf{xx}} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \sigma)$ 正定。

[注]：这里 $\sigma>0$ 就不用取到 $\sigma\to\infty$ 了。

在增广 Lagrangian 方法中，$\boldsymbol{\lambda}^*$ 也是通过迭代得到的。假设 σ_k,λ_k 给定，$\mathbf{x}_k=\arg\min_{\mathbf{x}}\mathcal{L}(\mathbf{x},\boldsymbol{\lambda}_k,\sigma_k)$，那么

$$0=\nabla_{\mathbf{x}}\mathcal{L}(\mathbf{x}_k,\boldsymbol{\lambda}_k,\sigma_k)=g(\mathbf{x}_k)-\sum_{i\in\epsilon}\lambda_i^{(k)}a_i(\mathbf{x}_k)+\sigma_k\sum_{i\in\epsilon}c_i(\mathbf{x}_k)a_i(\mathbf{x}_k)$$
$$\Rightarrow g(\mathbf{x}_k)=\sum_{i\in\epsilon}(\lambda_i^{(k)}-\sigma_k c_i(\mathbf{x}_k))a_i(\mathbf{x}_k)$$

当迭代步数足够大时，$g(\mathbf{x}_k)\approx g(\mathbf{x}^*),a_i(\mathbf{x}_k)\approx a_i(\mathbf{x}^*)$，又根据 KKT 条件知道

$$g(\mathbf{x}^*)=\sum_{i\in\epsilon}\lambda_i^*a_i(\mathbf{x}^*)$$

所以 $\lambda_i^*\approx\lambda_i^{(k)}-\sigma_k c_i(\mathbf{x}_k)$。于是 Lagrangian 乘子的迭代公式为

$$\lambda_i^{(k+1)}=\lambda_i^{(k)}-\sigma_k c_i(\mathbf{x}_k)$$

如果约束中存在不等式约束，**可以将其中的不等式约束转化为等式约束，再用求解等式约束的增广 Lagrangian 函数方法求解**。

一般约束优化问题的增广 Lagrangian 法

对于一般约束优化问题

$$\min_{\mathbf{x}\in\mathbb{R}^n}f(\mathbf{x})\quad\text{s.t.}\quad c_i(\mathbf{x})=0,\ i\in\epsilon,c_i(\mathbf{x})\geq 0,\ i\in\tau$$

定义如下增广 Lagrangian 函数

$$\min_{\mathbf{x}}\mathcal{L}(\mathbf{x},\boldsymbol{\lambda},\boldsymbol{\mu},\sigma)=f(\mathbf{x})-\sum_{i\in\epsilon}\lambda_i c_i(\mathbf{x})+\frac{\sigma}{2}\sum_{i\in\epsilon}c_i(\mathbf{x})^2+\sum_{i\in\tau}\phi_i$$

其中

$$\phi_i=\begin{cases}-\dfrac{1}{2\sigma}\lambda_i^2 & c_i(\mathbf{x})\geq\dfrac{\lambda_i}{\sigma}\\ -\dfrac{1}{2\sigma}\lambda_i^2+\dfrac{1}{2\sigma}(\sigma c_i(\mathbf{x})-\lambda_i)^2 & c_i(\mathbf{x})<\dfrac{\lambda_i}{\sigma}\end{cases}=\frac{1}{2\sigma}\left(\max\{0,\lambda_i-\sigma c_i(\mathbf{x})\}^2-\lambda_i^2\right)$$

乘子的迭代服从乘子的迭代为

- **等式约束的乘子**：$\lambda_i^{(k+1)}=\lambda_i^{(k)}-\sigma_k c_i(\mathbf{x}_k)$
- **不等式约束的乘子**：$\lambda_i^{(k+1)}=\lambda_i^{(k)}-\sigma_k[c_i(\mathbf{x}_k)-s_i^{(k)}]=\max\{\lambda_i^{(k)}-\sigma_k c_i(\mathbf{x}_k),0\}$

[注]：增广 Lagrangian 法的求解步骤如下：(1) 先列出增广 Lagrangian 函数，求解 $\dfrac{\partial\mathcal{L}}{\partial\mathbf{x}}$，将 $\mathbf{x}^*$ 用 $\sigma,\boldsymbol{\lambda}^{(k)}$ 表示出来 (如果有不等式约束，求梯度时还要分类讨论；一般而言 σ 可以先取定值不用迭代)；(2) 迭代乘子，找到 $\boldsymbol{\lambda}^{(k+1)}$ 和 $\boldsymbol{\lambda}^{(k)}$ 的递推关系，算出不动点 $\boldsymbol{\lambda}^*$；(3) 将 $\boldsymbol{\lambda}^*$ 带回 $\mathbf{x}^*$，算出最终值。

证明. 先引入**松弛变量 s_i 将原始优化问题转化为如下松弛问题**

$$\min_{\mathbf{x},\mathbf{s}}f(\mathbf{x})\quad\text{s.t.}\begin{cases}c_i(\mathbf{x})=0,\ i\in\epsilon,\\ c_i(\mathbf{x})-s_i=0,\ i\in\tau,\\ s_i\geq 0,\ i\in\tau\end{cases}$$

定义增广 Lagrangian 函数

$$\hat{\mathcal{L}}(\mathbf{x},\mathbf{s},\boldsymbol{\lambda},\sigma)=f(\mathbf{x})-\sum_{i\in\epsilon}\lambda_i' c_i(\mathbf{x})+\frac{\sigma}{2}\sum_{i\in\epsilon}c_i(\mathbf{x})^2+\sum_{i\in\tau}\hat{\phi}_i$$

其中 $\hat{\phi}_i=-\lambda_i(c_i(\mathbf{x})-s_i)+\frac{\sigma}{2}(c_i(\mathbf{x})-s_i)^2$，然后做以下子问题

$$\min_{\mathbf{x},\mathbf{s}}\hat{\mathcal{L}}(\mathbf{x},\boldsymbol{\lambda},\boldsymbol{\mu},\sigma)\quad \text{s.t. } s_i\geq 0,\ i\in\tau$$

再对其化简。对 $\mathbf{s}$ 取极小：$\frac{\partial\hat{\mathcal{L}}}{\partial s_i}=0\Rightarrow s_i^*=c_i(\mathbf{x})-\frac{\lambda_i}{\sigma}$。将 $s_i=\max\{c_i(\mathbf{x})-\frac{\lambda_i}{\sigma},0\}$ 带入 $\hat{\phi}_i$ 得到

$$\phi_i=\begin{cases}-\dfrac{\lambda_i^2}{2\sigma} & c_i(\mathbf{x})\geq\dfrac{\lambda_i}{\sigma}\\ -\lambda_i c_i(\mathbf{x})+\dfrac{\sigma}{2}c_i(\mathbf{x})^2 & c_i(\mathbf{x})<\dfrac{\lambda_i}{\sigma}\end{cases}$$

从而最终的增广 Lagrangian 无约束优化问题为

$$\min_{\mathbf{x}}\mathcal{L}(\mathbf{x},\boldsymbol{\lambda},\boldsymbol{\mu},\sigma)=f(\mathbf{x})-\sum_{i\in\epsilon}\lambda_i c_i(\mathbf{x})+\frac{\sigma}{2}\sum_{i\in\epsilon}c_i(\mathbf{x})^2+\sum_{i\in\tau}\phi_i$$

□

算法步骤如下所示：

算法流程

算法7.3(增广Lagrange函数方法)

1. 给定初始点x_0，λ_1，$\sigma_1>0$，$\rho>1$，$\epsilon>0$，$\epsilon_1>0$，$k\leftarrow 1$；

2. 以x_{k-1}为初始点，求解$x_k=\arg\min\Phi(x,\lambda_k,\sigma_k)$，当$\|\nabla_x\Phi(x_k,\lambda_k,\sigma_k)\|\leq\epsilon_1$时停止迭代；

$$\left[\sum_{i\in\varepsilon}c_i^2(x_k)+\sum_{i\in\tau}(c_i(x_k)-s_i^k)^2\right]^{\frac{1}{2}}\leq\epsilon$$

3. 若$[\sum_{i\in\varepsilon}c_i^2(x_k)+\sum_{i\in\tau}\min\{c_i(x_k),\frac{\lambda_i^k}{\sigma_k}\}^2]^{1/2}\leq\epsilon$，则迭代停止，输出$x_k$；

4. 对于等式约束乘子$\lambda_i^{k+1}=\lambda_i^k-\sigma_k c_i(x_k)$，对于不等式约束乘法子$\lambda_i^{k+1}=-\sigma_k\min\{c_i(x_k)-\frac{\lambda_i^k}{\sigma_k},0\}$；

5. $\sigma_{k+1}=\rho\sigma_k$，$k\leftarrow k+1$，转到第2步。

5.2.4 增广 Lagrangian 函数法和罚函数法的区别

出于简便，我们只考虑等式优化的情况。回顾外点罚函数方法的收敛性定理，我们知道当 $\sigma_k\to\infty$ 时

$$-\sigma_k c_i(\mathbf{x}_k)\to\lambda_i^*\Rightarrow c_i(\mathbf{x}_k)\to-\frac{\lambda_i^*}{\sigma_k},\ i\in\epsilon$$

而根据增广 Lagrangian 函数法中乘子的迭代公式，我们知道

$$\lambda_i^{(k)}-\sigma_k c_i(\mathbf{x}_k)\to\lambda_i^*\Rightarrow c_i(\mathbf{x}_k)\to-\frac{(\lambda_i^*-\lambda_i^{(k)})}{\sigma_k},\ i\in\epsilon$$

在外点罚函数法中，为了使约束条件 $c_i(\mathbf{x})$ 逼近 0，只能令 $\sigma_k\to\infty$，而在增广 Lagrangian 函数法中，σ_k 不需要取到 ∞，只要其充分大，使得 λ_i^* 和 $\lambda_i^{(k)}$ 充分接近，就能使得 $c_i(\mathbf{x}_k)$ 逼近 0。这就是增广 Lagrangian 函数法的优势所在。举个例子：对于优化问题

$$\min f(\mathbf{x})=2x_1^2-x_2^2+x_1-x_2,\ \text{s.t. } x_1-x_2=0$$

在 $\lambda^* = 1$ 处的 Lagrangian 函数为

$$\mathcal{L}(\mathbf{x}, \lambda = 1, \sigma) = (2 + \frac{\sigma}{2})x_1^2 - \sigma x_1 x_2 + (\frac{\sigma}{2} - 1)x_2^2$$

可以计算发现，只要 $\sigma > 4$，$\nabla_{\mathbf{xx}}\Phi(\mathbf{x}, \lambda = 1, \sigma)$ 就是正定的。

6 总结表格

优化方法	线搜索准则	收敛性	收敛速度	迭代方向	步长	注
最速下降法	任何	全局收敛	线性收敛	$-\mathbf{g}_k$	$\alpha_k = \frac{\mathbf{g}_k^\top \mathbf{g}_k}{\mathbf{g}_k^\top \mathbf{G}\mathbf{g}_k}$(正定二次函数)	收敛速度与条件数大小相关
坐标下降法	精确线搜索	不一定收敛	线性收敛	坐标轴	$\alpha_i = \min_\alpha f(x_k^{(i)} - \alpha\frac{\partial f(\mathbf{x}_k)}{\partial x^{(i)}})$	收敛速度通常慢于最速下降
BB 法	/	/	/	$-\mathbf{g}_k$	$\alpha_k^{\text{BB1}} = \frac{\mathbf{g}_{k-1}^\top \mathbf{G}\mathbf{g}_{k-1}}{\mathbf{g}_{k-1}^\top \mathbf{G}^2\mathbf{g}_{k-1}}$ $\alpha_k^{\text{BB2}} = \frac{\mathbf{g}_{k-1}^\top \mathbf{g}_{k-1}}{\mathbf{g}_{k-1}^\top \mathbf{G}\mathbf{g}_{k-1}}$	比最速下降最小梯度快
基本 Newton 法	无	局部收敛	二次收敛	$-\mathbf{G}_k^{-1}\mathbf{g}_k$	/	
阻尼 Newton 法	精确线搜索/ Goldstein/ Wolfe	全局收敛	二次收敛	$-\mathbf{G}_k^{-1}\mathbf{g}_k$	/	全局收敛性要求 $\mathbf{u}^\top G(\mathbf{x})\mathbf{u} \geq \beta\|\mathbf{u}\|^2, \forall \mathbf{u} \in \mathbb{R}^n$ 二次收敛速度要求 $\mathbf{x}^*$ 处 $G(\mathbf{x})$Lipschitz 连续
混合法	/	/	/	$\mathbf{G}_k$ 正定非奇异: $-\mathbf{G}_k^{-1}\mathbf{g}_k$ $\mathbf{G}_k$ 奇异或几乎正交: $-\mathbf{g}_k$ $\mathbf{G}_k$ 负定非奇异: $\mathbf{G}_k^{-1}\mathbf{g}_k$	/	如果多次使用负梯度方向收敛速度会趋于负梯度方法的收敛速度
SR1	/	二次终止	/	$-\mathbf{B}_k^{-1}\mathbf{g}_k$ $-\mathbf{H}_k\mathbf{g}_k$	/	Broyden 族
BFGS/DFP	精确线搜索/ Wolfe	全局收敛 二次终止	超线性收敛	$-\mathbf{B}_k^{-1}\mathbf{g}_k$ $-\mathbf{H}_k\mathbf{g}_k$	/	全局收敛性要求 $m\|\mathbf{z}\|^2 \leq \mathbf{z}^\top \mathbf{G}(\mathbf{x})\mathbf{z} \leq M\|\mathbf{z}\|^2$ Broyden 族
共轭梯度法	精确线搜索	二次终止	/	$\mathbf{d}_i^\top \mathbf{G}\mathbf{d}_j = 0, i \neq j$ $\mathbf{g}_k^\top \mathbf{d}_i = 0, i = 0, \cdots, k-1$	/	
线性共轭梯度法	精确线搜索	二次终止	/	$-\mathbf{g}_k + \frac{\mathbf{g}_k^\top \mathbf{g}_k}{\mathbf{g}_{k-1}^\top \mathbf{g}_{k-1}}\mathbf{d}_{k-1}$	$\alpha_k = \frac{\mathbf{g}_k^\top \mathbf{g}_k}{\mathbf{d}_k^\top \mathbf{G}\mathbf{d}_k}$	
FR 方法	精确线搜索/ 强 Wolfe	全局收敛	/	$-\mathbf{g}_k + \frac{\mathbf{g}_k^\top \mathbf{g}_k}{\mathbf{g}_{k-1}^\top \mathbf{g}_{k-1}}\mathbf{d}_{k-1}$	/	
PRP 方法	/	全局收敛	/	$-\mathbf{g}_k + \frac{\mathbf{g}_k^\top (\mathbf{g}_k - \mathbf{g}_{k-1})}{\mathbf{g}_{k-1}^\top \mathbf{g}_{k-1}}\mathbf{d}_{k-1}$	/	
Gauss-Newton 法	/	局部收敛	线性收敛	$-(\mathbf{J}_k^\top \mathbf{J}_k)^{-1}\mathbf{J}_k^\top \mathbf{r}_k$	/	要求 $\|[(\mathbf{J}^*)^\top (\mathbf{J}^*)]^{-1}\|\|\mathbf{S}^*\| < 1$ $\|\mathbf{S}^*\| \uparrow$ 收敛速度减慢
	/	局部收敛	二次收敛	$-(\mathbf{J}_k^\top \mathbf{J}_k)^{-1}\mathbf{J}_k^\top \mathbf{r}_k$	/	线性收敛要求 $\|\mathbf{S}^*\| = 0$
阻尼 Gauss-Newton 法	Wolfe	全局收敛	线性收敛	$-(\mathbf{J}_k^\top \mathbf{J}_k)^{-1}\mathbf{J}_k^\top \mathbf{r}_k$	/	
外点罚函数法	/	全局收敛	/	/	/	$\lambda_i^* = \lim_{k\to\infty} -\sigma_k c_i(\mathbf{x}_k)$ $\sigma \to \infty$ 时病态性严重
障碍函数法	/	全局收敛	/	/	/	/ $\mu \to 0$ 时病态性严重

约束优化重要公式

线性可行方向集与线性可行不变方向集

$$\begin{aligned}\mathcal{F}^* = \{\mathbf{d} : a_i(\mathbf{x})^\top \mathbf{d} = 0, i \in \epsilon,\\ a_i(\mathbf{x})^\top \mathbf{d} \geq 0, i \in \tau(\mathbf{x}^*)\}\\ \mathcal{F}_1^* = \{\mathbf{d} : (\mathbf{g}^*)^\top \mathbf{d} = 0, i \in \mathcal{F}^*\}\end{aligned}$$

所有的罚函数构造和增广 Lagrangian 函数构造

$$P(\mathbf{x}, \sigma) = f(\mathbf{x}) + \frac{\sigma}{2}\left(\sum_{i\in\epsilon} c_i(\mathbf{x})^2 + \sum_{i\in\tau} \min\{0, c_i(\mathbf{x})\}^2\right)$$

$$B_I(\mathbf{x}, \mu) = f(\mathbf{x}) + \mu \sum_{i\in\tau} \frac{1}{c_i(\mathbf{x})}, \quad B_L(\mathbf{x}, \mu) = f(\mathbf{x}) - \mu \sum_{i\in\tau} \ln c_i(\mathbf{x})$$

$$B(\mathbf{x}, \mu) = f(\mathbf{x}) - \mu \sum_{i\in\tau} \ln c_i(\mathbf{x}) + \frac{1}{2\mu} \sum_{i\in\epsilon} c_i(\mathbf{x})^2$$

$$\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \sigma) = f(\mathbf{x}) - \sum_{i\in\epsilon} \lambda_i c_i(\mathbf{x}) + \frac{\sigma}{2} \sum_{i\in\epsilon} c_i(\mathbf{x})^2$$

$$\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}, \sigma) = f(\mathbf{x}) - \sum_{i\in\epsilon} \lambda_i c_i(\mathbf{x}) + \frac{\sigma}{2} \sum_{i\in\epsilon} c_i(\mathbf{x})^2 + \sum_{i\in\tau} \frac{1}{2\sigma} \left(\max\{0, \lambda_i - \sigma c_i(\mathbf{x})\}^2 - \lambda_i^2\right)$$

乘子迭代

$$\begin{aligned}\lambda_i^{(k+1)} &= \lambda_i^{(k)} - \sigma_k c_i(\mathbf{x}_k), i \in \epsilon\\ \lambda_i^{(k+1)} &= \max\{\lambda_i^{(k)} - \sigma_k c_i(\mathbf{x}_k), 0\}, i \in \tau\end{aligned}$$