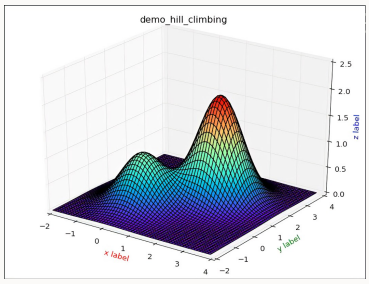
最优化理论与方法

Lin Wenjie

Nanjing Normal University

版本：1.0

更新：2025 年 10 月 24 日



目录

Part I

最优化理论基础

1 无约束最优化问题

无约束最优化问题考虑下面这样的问题形式：

$$\min_x f(x),$$

其中 $x \in \mathbb{R}^n$，$f : \mathbb{R}^n \to \mathbb{R}$ 是一个光滑函数。这个问题的解是什么呢？一般来说，最优化问题需要求解如下的三种极小点。

定义 1.1 (极小点)

- 全局极小点: 若对所有的 x，都有 $f(x^*) \leq f(x)$，则点 x^* 是全局极小点。
- 局部极小点: 若存在 x^* 的一个邻域 $\mathcal{N}$，使得对所有的 $x \in \mathcal{N}$，都有 $f(x^*) \leq f(x)$，则点 x^* 是局部极小点。
- 严格局部极小点: 若存在 x^* 的一个邻域 $\mathcal{N}$，使得对所有的 $x \in \mathcal{N}$ 且 $x \neq x^*$，都有 $f(x^*) < f(x)$，则点 x^* 是严格局部极小点（也称为强局部极小点）。

注 1.1 对于一个优化问题，最好当然可以找到全局极小点，实在不行也可以退而求其次找到局部极小点或严格局部极小点。

如何判断这个点是极小点呢？如果完全根据定义，那就需要比较其附近的所有点，对于数值计算来说不现实。不过对于一个光滑函数我们有更高效的方法验证某个点是否是极小点，主要用到的工具是光滑函数的泰勒展开。

定理 1.1 (泰勒定理) 设 $f : \mathbb{R}^n \to \mathbb{R}$ 是连续可微的，且 $p \in \mathbb{R}^n$。则存在 $t \in (0, 1)$，使得

$$f(x + p) = f(x) + \nabla f(x + tp)^T p.$$

此外，若 f 是二次连续可微的，则存在 $t \in (0, 1)$，使得

$$\nabla f(x + p) = \nabla f(x) + \int_0^1 \nabla^2 f(x + tp) p \, dt,$$

且

$$f(x + p) = f(x) + \nabla f(x)^T p + \frac{1}{2} p^T \nabla^2 f(x + tp) p.$$

定理 1.2 (一阶必要条件) 若 x^* 是局部极小点，且 f 在 x^* 的一个开邻域内连续可微，则 $\nabla f(x^*) = 0$。

证明. 假设 $\nabla f(x^*) \neq 0$。令 $p = -\nabla f(x^*)$，注意到 $p^T \nabla f(x^*) = -\|\nabla f(x^*)\|^2 < 0$。由于 ∇f 在 x^* 附近连续，存在 $T > 0$，使得对所有 $t \in [0, T]$，有

$$p^T \nabla f(x^* + tp) < 0.$$

对任意 $\bar{t} \in (0, T]$，由泰勒定理可知，存在 $t \in (0, \bar{t})$，使得

$$f(x^* + \bar{t}p) = f(x^*) + \bar{t}p^T \nabla f(x^* + tp).$$

因此，对所有 $\bar{t} \in (0, T]$，有 $f(x^* + \bar{t}p) < f(x^*)$。故 x^* 不是局部极小点，矛盾。

注 1.2

- 若 $\nabla f(x^*) = 0$，则称 x^* 为稳定点。根据定理 1.2，任何局部极小点都为稳定点。
- 我个人更习惯通过极限的保号性得到该结论：

  证明. 如果 $\nabla f(x^*) \neq 0$, 令 $g(t) = f(x^* - t\nabla f(x^*))$, 则 $g'(0) = -\|\nabla f(x^*)\|^2 < 0$，即

  $$\lim_{t \to 0^+} \frac{g(t) - g(0)}{t} = g'(0) < 0$$

  根据保号性，存在 $\delta > 0$, 使得 $\forall t \in (0, \delta)$, 有

  $$\frac{g(t) - g(0)}{t} < 0$$

  于是 $g(t) < g(0)$, 即 $f(x^* - t\nabla f(x^*)) < f(x^*), \forall t \in (0, \delta)$, 这与局部极小点的定义矛盾

定理 1.3 (二阶必要条件) 若 x^* 是 f 的局部极小点，且 $\nabla^2 f$ 在 x^* 的一个开邻域内存在且连续，则 $\nabla f(x^*) = 0$ 且 $\nabla^2 f(x^*)$ 是半正定的。

证明. 由定理 1.2 知 $\nabla f(x^*) = 0$。假设 $\nabla^2 f(x^*)$ 不是半正定的。此时可选取向量 p 使得 $p^T \nabla^2 f(x^*) p < 0$，又因 $\nabla^2 f$ 在 x^* 附近连续，故存在 $T > 0$，使得对所有 $t \in [0, T]$，有 $p^T \nabla^2 f(x^* + tp) p < 0$。

对 x^* 做泰勒展开，对所有 $\bar{t} \in (0, T]$ 及某个 $t \in (0, \bar{t})$，有

$$f(x^* + \bar{t}p) = f(x^*) + \bar{t}p^T \nabla f(x^*) + \frac{1}{2}\bar{t}^2 p^T \nabla^2 f(x^* + tp) p < f(x^*).$$

故 x^* 不是局部极小点，矛盾。

定理 1.4 (二阶充分条件) 假设 $\nabla^2 f$ 在 x^* 的一个开邻域内连续，且 $\nabla f(x^*) = 0$ 且 $\nabla^2 f(x^*)$ 是正定的。则 x^* 是 f 的严格局部极小点。

证明. 由于 Hessian 矩阵在 x^* 处连续且正定，存在 $r > 0$，使得对于开球 $\mathcal{D} = \{z \mid \|z - x^*\| < r\}$ 内的所有 x，$\nabla^2 f(x)$ 保持正定。取任意非零向量 p 满足 $\|p\| < r$，则 $x^* + p \in \mathcal{D}$，于是

$$\begin{aligned} f(x^* + p) &= f(x^*) + p^T \nabla f(x^*) + \frac{1}{2} p^T \nabla^2 f(z) p \\ &= f(x^*) + \frac{1}{2} p^T \nabla^2 f(z) p, \end{aligned}$$

其中 $z = x^* + tp$ 对某个 $t \in (0, 1)$ 成立。由于 $z \in \mathcal{D}$，故 $p^T \nabla^2 f(z) p > 0$，因此 $f(x^* + p) > f(x^*)$，得证。

注 1.3 注意，二阶充分条件不是必要的：点 x^* 可能是严格局部极小点，但却不满足充分条件。一个简单的例子是函数 $f(x) = x^4$，其点 $x^* = 0$ 是严格局部极小点，但在该点处 Hessian 矩阵为零（因此不是正定的）。

定理 1.5 当 f 是凸函数时，任何局部极小点 x^* 都是 f 的全局极小点。又若 f 可微，则任何稳定点 x^* 都是 f 的全局极小点。

证明. 假设 x^* 是局部极小点但不是全局极小点，则存在点 $z \in \mathbb{R}^n$ 使得 $f(z) < f(x^*)$。考虑连接 x^* 和 z 的线段，即

$$x = \lambda z + (1 - \lambda) x^*, \quad \text{对某个 } \lambda \in (0, 1].$$

由 f 的凸性，有

$$f(x) \leq \lambda f(z) + (1 - \lambda) f(x^*) < f(x^*).$$

x^* 的任意邻域 $\mathcal{N}$ 都包含线段的一部分，因此总存在 $x \in \mathcal{N}$ 满足上式，故 x^* 不是局部极小点，矛盾。

又若 f 可微，假设 x^* 不是全局极小点，点 $z \in \mathbb{R}^n$ 使得 $f(z) < f(x^*)$。由凸性，有

$$\begin{aligned} \nabla f(x^*)^T (z - x^*) &= \frac{d}{d\lambda} f(x^* + \lambda(z - x^*)) \Big|_{\lambda=0} \\ &= \lim_{\lambda \downarrow 0} \frac{f(x^* + \lambda(z - x^*)) - f(x^*)}{\lambda} \\ &\leq \lim_{\lambda \downarrow 0} \frac{\lambda f(z) + (1 - \lambda) f(x^*) - f(x^*)}{\lambda} \\ &= f(z) - f(x^*) < 0. \end{aligned}$$

因此，$\nabla f(x^*) \neq 0$，故 x^* 不是稳定点，矛盾。

注 1.4 本节的讨论概括起来就是：

- 对于一个一般的无约束最优化问题，一个最好的情况是可以得到全局极小点；
- 如果无法全局极小点，可以退而求其次去得到局部极小点或严格局部极小点；

- 对于一个光滑函数，可以用局部极小点的一阶和二阶必要条件以及严格局部极小点的充分条件来判断极小点；
- 对于一个凸函数，局部极小点与全局极小点是等价的；
- 对于一个光滑凸函数，只要求得稳定点就可以得到全局极小点。

注 1.5 上述的讨论一般都基于其目标函数可微的情况，对于非光滑的目标函数，通常用次梯度（梯度的推广）来判断极小点，这里暂不做讨论。

2 一般的最优性条件

上一节考虑了可微背景下的最优性条件，本节考虑不可微的最优性条件，设 $\mathbb{E}$ 是一个赋范空间。

2.1 凸优化的最优性条件

定理 2.1 (Fermat 最优性条件) 设 $f:\mathbb{E}\to(-\infty,\infty]$ 是适当凸函数，则

$$x^*\in\operatorname{argmin}\{f(x):x\in\mathbb{E}\}$$

当且仅当 $0\in\partial f(x^*)$。

证明.

$$\begin{aligned}
&x^*\in\operatorname{argmin}\{f(x):x\in\mathbb{E}\}\\
\iff& f(x)\geqslant f(x^*),\quad \forall x\in\operatorname{dom}(f),\\
\iff& f(x)\geqslant f(x^*)+\langle 0,x-x^*\rangle,\quad \forall x\in\operatorname{dom}(f),\\
\iff& 0\in\partial f(x^*)
\end{aligned}$$

考虑约束优化问题

$$\min\{f(x):x\in C\},$$

其中 f 是广义实值凸函数，且 $C\subseteq\mathbb{E}$ 是凸集。

定理 2.2 (凸约束优化的充要最优性条件) 设 $f:\mathbb{E}\to(-\infty,\infty]$ 是适当凸函数，$C\subseteq\mathbb{E}$ 是凸集且满足 $\operatorname{ri}(\operatorname{dom}(f))\cap\operatorname{ri}(C)\neq\varnothing$。则 $x^*\in C$ 是上述问题的最优解当且仅当

$$\text{存在 } g\in\partial f(x^*) \text{ 使得 } -g\in N_C(x^*).$$

证明.

$$\min\{f(x) : x \in C\},$$

可改写为

$$\min_{x\in\mathbb{E}} f(x) + \delta_C(x).$$

由于 $\mathrm{ri}(\mathrm{dom}(f)) \cap \mathrm{ri}(C) \neq \varnothing$，根据次微分的和法则，对任意 $x \in C$，有

$$\partial(f + \delta_C)(x) = \partial f(x) + \partial\delta_C(x) = \partial f(x) + N_C(x).$$

于是，利用定理 2.1, $x^* \in C$ 是最优解当且仅当 $0 \in \partial f(x^*) + N_C(x^*)$，即当且仅当

$$\text{存在 } g \in \partial f(x^*) \text{ 使得 } -g \in N_C(x^*).$$

注 2.1 进一步由法锥的定义得到:$x^* \in C$ 是最优解当且仅当

$$\text{存在 } g \in \partial f(x^*) \text{ 使得对任意 } x \in C,\ \langle g, x - x^*\rangle \geq 0.$$

这就是变分不等式。

2.2 非凸优化的最优性条件

定理 2.3 (复合问题的最优性条件) 设 $f : \mathbb{E} \to (-\infty, \infty]$ 是适当函数，$g : \mathbb{E} \to (-\infty, \infty]$ 是适当凸函数，且满足 $\mathrm{dom}(g) \subseteq \mathrm{int}(\mathrm{dom}(f))$。考虑问题

$$\min_{x\in\mathbb{E}} f(x) + g(x).$$

(a)（必要条件）若 $x^* \in \mathrm{dom}(g)$ 是问题 (P) 的局部最优解，且 f 在 x^* 处可微，则

$$-\nabla f(x^*) \in \partial g(x^*).$$

(b)（凸问题的充要条件）假设 f 是凸函数。若 f 在 $x^* \in \mathrm{dom}(g)$ 处可微，则 x^* 是问题 (P) 的全局最优解当且仅当上述条件成立。

证明.

(a) 设 $y \in \mathrm{dom}(g)$。由 $\mathrm{dom}(g)$ 的凸性，对任意 $\lambda \in (0, 1)$，点 $x_\lambda = (1-\lambda)x^* + \lambda y$ 属于 $\mathrm{dom}(g)$；再由 x^* 的局部最优性，当 λ 足够小时，有

$$f(x_\lambda) + g(x_\lambda) \geq f(x^*) + g(x^*).$$

即

$$f((1-\lambda)x^* + \lambda y) + g((1-\lambda)x^* + \lambda y) \geq f(x^*) + g(x^*).$$

利用 g 的凸性，可得

$$f((1-\lambda)x^*+\lambda y)+(1-\lambda)g(x^*)+\lambda g(y)\geq f(x^*)+g(x^*),$$

整理得

$$\frac{f((1-\lambda)x^*+\lambda y)-f(x^*)}{\lambda}\geq g(x^*)-g(y).$$

令 $\lambda\to 0^+$，由 f 在 x^* 处可微, 于是 $f'(x^*;y-x^*)=\langle\nabla f(x^*),y-x^*\rangle$，因此对任意 $y\in\mathrm{dom}(g)$，有

$$g(y)\geq g(x^*)+\langle-\nabla f(x^*),y-x^*\rangle,$$

于是 $-\nabla f(x^*)\in\partial g(x^*)$。

(b) 额外假设 f 是凸函数。若 x^* 是问题 (P) 的最优解，由 (a) 已证条件成立。反之，若条件成立，则对任意 $y\in\mathrm{dom}(g)$，有

$$g(y)\geq g(x^*)+\langle-\nabla f(x^*),y-x^*\rangle.$$

由 f 的凸性，对任意 $y\in\mathrm{dom}(g)$，有

$$f(y)\geq f(x^*)+\langle\nabla f(x^*),y-x^*\rangle.$$

将两式相加，得对任意 $y\in\mathrm{dom}(g)$，有

$$f(y)+g(y)\geq f(x^*)+g(x^*),$$

即 x^* 是问题 (P) 的最优解。

注 2.2

- 当 f 可微，g 适当凸，我们称满足下面关系的 x^* 为稳定点：

$$-\nabla f(x^*)\in\partial g(x^*).$$

- 当 $g=\delta_C$（其中 $C\subseteq\mathbb{E}$ 是非空凸集）时，问题 (P) 变为

$$\min\{f(\mathbf{x}):\mathbf{x}\in C\}.$$

若 f 在 $\mathbf{x}^*\in C$ 处可微，则 $\mathbf{x}^*$ 是问题 (P) 的平稳点当且仅当

$$-\nabla f(\mathbf{x}^*)\in\partial\delta_C(\mathbf{x}^*)=N_C(\mathbf{x}^*),$$

根据法锥的定义，条件可改写为

$$\langle\nabla f(\mathbf{x}^*),\mathbf{x}-\mathbf{x}^*\rangle\geq 0\quad\forall\mathbf{x}\in C.$$

也就是说，求这个问题的稳定点相当于求解一个变分不等式问题。

2.3 KKT 条件

引理 2.4 考虑如下约束的优化问题

$$\begin{aligned} &\min\ f(x) \\ &\text{s.t. } g_i(x) \le 0,\ i = 1, 2, \ldots, m, \end{aligned}$$

其中 $f, g_1, g_2, \ldots, g_m : \mathbb{E} \to \mathbb{R}$ 是实值函数。假设问题的最小值有限且等于 $\bar{f}$。定义函数

$$F(x) \equiv \max\{f(x) - \bar{f}, g_1(x), g_2(x), \ldots, g_m(x)\}.$$

则原问题的最优解集与 F 的极小值点集相同。

证明. 设 X^* 是问题的最优解集。F 满足以下两个性质：

(i) 对任意 $x \notin X^*$，有 $F(x) > 0$；
设 $x \notin X^*$，分两种情况：
- 若 x 不可行，即存在 i 使得 $g_i(x) > 0$，由 F 的定义知 $F(x) > 0$；
- 若 x 可行但非最优，则对所有 i 有 $g_i(x) \le 0$ 且 $f(x) > \bar{f}$，因此 $F(x) > 0$。

(ii) 对任意 $x \in X^*$，有 $F(x) = 0$。
设 $x \in X^*$，则对所有 i 有 $g_i(x) \le 0$ 且 $f(x) = \bar{f}$，故 $F(x) = 0$。

上面分析可以看出, F 的极小值点集就是 $\{x : F(x) = 0\} = X^*$, 即原问题的最优解集与 F 的极小值点集相同。

定理 2.5 (Fritz-John 必要最优性条件) 考虑如下约束的优化问题

$$\begin{aligned} &\min\ f(x) \\ &\text{s.t. } g_i(x) \le 0,\ i = 1, 2, \ldots, m, \end{aligned}$$

其中 $f, g_1, g_2, \ldots, g_m : \mathbb{E} \to \mathbb{R}$ 是实值凸函数。设 x^* 是问题的最优解，则存在不全为零的 $\lambda_0, \lambda_1, \ldots, \lambda_m \ge 0$，使得

$$\begin{aligned} &0 \in \lambda_0 \partial f(x^*) + \sum_{i=1}^{m} \lambda_i \partial g_i(x^*) \\ &\lambda_i g_i(x^*) = 0,\ i = 1, 2, \ldots, m. \end{aligned}$$

证明. 设 x^* 是问题的最优解，记问题的最优值为 $\bar{f} = f(x^*)$。由定理 2.4，x^* 是问题

$$\min_{x \in \mathbb{E}} \{F(x) \equiv \max\{g_0(x), g_1(x), \ldots, g_m(x)\}\}$$

(其中 $g_0(x) = f(x) - \bar{f}$) 的最优解，显然 $F(x^*) = 0$。由于 F 是凸函数的最大值，故 F 是凸函数。利用定理 2.1，有

$$0 \in \partial F(x^*) = \operatorname{conv}\left(\bigcup_{i \in I(x^*)} \partial g_i(x^*) \right),$$

其中 $I(x^*) = \{i \in \{0, 1, \ldots, m\} : g_i(x^*) = 0\}$。结合上式可知，存在 $\lambda_i \geq 0$ $(i \in I(x^*))$ 且 $\sum_{i \in I(x^*)} \lambda_i = 1$，使得

$$0 \in \sum_{i \in I(x^*)} \lambda_i \partial g_i(x^*).$$

由于 $g_0(x^*) = f(x^*) - \bar{f} = 0$，故 $0 \in I(x^*)$，因此上式可改写为

$$0 \in \lambda_0 \partial f(x^*) + \sum_{i \in I(x^*) \setminus \{0\}} \lambda_i \partial g_i(x^*).$$

对任意 $i \in \{1, 2, \ldots, m\} \setminus I(x^*)$，令 $\lambda_i = 0$，则两个条件都成立。最后，由于 $\sum_{i \in I(x^*)} \lambda_i = 1$，故 λ_i 不全为零。

如果要求 $\lambda_0 = 1$，则需要满足以下附加条件（称为 Slater 条件）：

$$\text{存在 } \bar{\mathbf{x}} \in \mathbb{E} \text{ 使得 } g_i(\bar{\mathbf{x}}) < 0,\ i = 1, 2, \ldots, m.$$

KKT 条件的充分性不需要额外假设（除凸性外），且无需借助弗里茨-约翰条件的结论即可轻松推导。

定理 2.6 (KKT 条件) *考虑如下的优化问题*

$$\min\ f(\mathbf{x})$$

$$\text{s.t. } g_i(\mathbf{x}) \leq 0,\ i = 1, 2, \ldots, m,$$

其中 $f, g_1, g_2, \ldots, g_m : \mathbb{E} \to \mathbb{R}$ 是实值凸函数。

(a) *设 x^* 是问题的最优解，且 Slater 条件成立，则存在 $\lambda_1, \ldots, \lambda_m \geq 0$，使得*

$$0 \in \partial f(x*) + \sum_{i=1}^{m} \lambda_i \partial g_i(x^*)$$

$$\lambda_i g_i(x^*) = 0,\ i = 1, 2, \ldots, m.$$

(b) *若可行点 $x^* \in \mathbb{E}$ 对某组 $\lambda_1, \lambda_2, \ldots, \lambda_m \geq 0$ 满足上述两个条件，则 x^* 是问题的最优解。*

证明.

(a) 由定理 2.5，存在不全为零的 $\tilde{\lambda}_0, \tilde{\lambda}_1, \ldots, \tilde{\lambda}_m \geq 0$，使得

$$0 \in \tilde{\lambda}_0 \partial f(x^*) + \sum_{i=1}^{m} \tilde{\lambda}_i \partial g_i(x^*)$$

且

$$\tilde{\lambda}_i g_i(x^*) = 0,\ i = 1, 2, \ldots, m.$$

只要证明 $\tilde{\lambda}_0 \neq 0$ 即可。假设 $\tilde{\lambda}_0 = 0$，则由上式得

$$0 \in \sum_{i=1}^{m} \tilde{\lambda}_i \partial g_i(\mathbf{x}^*),$$

即存在 $\xi_i \in \partial g_i(x^*)\ (i = 1, 2, \ldots, m)$，使得

$$\sum_{i=1}^{m} \tilde{\lambda}_i \xi_i = 0.$$

设 $\overline{x}$ 是满足 Slater 条件的点，对函数 $g_i\ (i = 1, 2, \ldots, m)$，由次梯度定义有

$$g_i(x^*) + \langle \xi_i, \overline{x} - x^* \rangle \leq g_i(\overline{x}),\ i = 1, 2, \ldots, m.$$

将第 i 个不等式乘以 $\tilde{\lambda}_i \geq 0$ 并对 $i = 1, 2, \ldots, m$ 求和，得

$$\sum_{i=1}^{m} \tilde{\lambda}_i g_i(x^*) + \left\langle \sum_{i=1}^{m} \tilde{\lambda}_i \xi_i, \overline{x} - x^* \right\rangle \leq \sum_{i=1}^{m} \tilde{\lambda}_i g_i(\overline{x}).$$

但是左边两项都为 0，故 $\sum_{i=1}^{m} \tilde{\lambda}_i g_i(\overline{\mathbf{x}}) \geq 0$。但 Slater 条件中 $g_i(\overline{x}) < 0$ 且 $\tilde{\lambda}_i \geq 0$（不全为零），于是矛盾。因此 $\tilde{\lambda}_0 > 0$。

令 $\lambda_i = \frac{\tilde{\lambda}_i}{\tilde{\lambda}_0}\ (i = 1, 2, \ldots, m)$，则

$$0 \in \partial f(x*) + \sum_{i=1}^{m} \lambda_i \partial g_i(x^*)$$

$$\lambda_i g_i(x^*) = 0,\ i = 1, 2, \ldots, m.$$

(b) 假设可行点 x^* 对某组 $\lambda_1, \lambda_2, \ldots, \lambda_m \geq 0$ 满足两个条件。设 $\overline{x}$ 是可行点，即 $g_i(\overline{x}) \leq 0\ (i = 1, 2, \ldots, m)$，我们证明 $f(\overline{x}) \geq f(x^*)$。

定义函数

$$h(x) = f(x) + \sum_{i=1}^{m} \lambda_i g_i(x),$$

则 h 是凸函数。于是 KKT 条件告诉我们

$$0 \in \partial h(x^*),$$

结合定理 2.1，x^* 是 h 在 $\mathbb{E}$ 上的极小值点。

于是

$$f(x^*) = f(x^*) + \sum_{i=1}^{m} \lambda_i g_i(x^*) = h(x^*) \leq h(\overline{x}) = f(\overline{x}) + \sum_{i=1}^{m} \lambda_i g_i(\overline{x}).$$

由于 $\lambda_i \geq 0$ 且 $g_i(\overline{x}) \leq 0$，故 $\sum_{i=1}^{m} \lambda_i g_i(\overline{x}) \leq 0$，因此

$$f(x^*) \leq f(\overline{x}) + \sum_{i=1}^{m} \lambda_i g_i(\overline{x}) \leq f(\overline{x}),$$

即 x^* 是最优解。

注 2.3 我们将满足以下条件的 x^* 称为 KKT 点：

$$\begin{aligned} &0 \in \partial f(x*) + \sum_{i=1}^{m} \lambda_i \partial g_i(x^*) \\ &\lambda_i g_i(x^*) = 0,\ i = 1, 2, \ldots, m, \\ &g_i(x^*) \leq 0,\ i = 1, 2, \ldots, m. \end{aligned}$$

3 最优化基本算法框架

最优化算法是一种数值算法，从初始点 x_0 开始，按一定的规则 (算法) 得到点列 $\{x^k\}$，从而得到解点 x^* 的近似点。

最优化算法根据 x^k 处 (甚至 $x^0, x^1, \ldots, x^{k-1}$ 处) 的信息 (如函数值、导数信息等) 得到下一个迭代点 x^{k+1}. 一般来说有两种类型的策略：**线搜索 (Line Search) 与信赖域 (Trust Region)**.

- **线搜索**:
  在线搜索策略中，算法选择一个方向 p_k，并从当前迭代点 x_k 沿该方向搜索，以找到函数值更小的新迭代点。沿 p_k 移动的距离 (步长 α_k) 可通过近似求解以下问题来确定：

  $$\alpha_k = \text{argmin}_{\alpha > 0} f(x_k + \alpha p_k).$$

  然后令 $x_{k+1} = x^k + \alpha_k p_k$. 若精确求解 α_k 固然很好，但精确极小化可能成本高昂且通常得不偿失。

- **信赖域**:
  在信赖域策略中，每一次迭代需要构造一个模型函数 m_k，该函数是在当前点 x_k 附近对目标函数 f 的相似。由于当 x 远离 x_k 时，模型 m_k 可能无法很好地近似 f，因

此我们将对 m_k 极小点的搜索限制在 x_k 周围的某个区域 (即信赖域) 内。信赖域策略通过近似求解以下子问题来找到候选步长 p_k：

$$p_k = \text{argmin}_p m_k(x_k + p), \quad \text{其中 } x_k + p \text{ 位于信赖域内.}$$

若 p_k 未使 f 产生足够的下降，则判定信赖域过大，会缩小并重新求解该子问题。通常，信赖域是一个由 $\|p\| \leq \Delta$ 定义的球，其中 $\Delta > 0$ 称为信赖域半径。

f 的近似模型 m_k 通常定义为如下形式的二次函数：

$$m_k(x_k + p) = f_k + p^T \nabla f_k + \frac{1}{2} p^T B_k p,$$

其中 f_k 和 ∇f_k 选取为在点 x_k 处的函数值和梯度值，使得 m_k 和 f 在当前迭代点 x_k 处一阶一致, 矩阵 B_k 是 $\nabla^2 f_k$ 或其近似矩阵。

- **区别：**
  线搜索和信赖域方法的区别在于**选择移动到下一个迭代点的方向和距离的顺序**。线搜索先固定方向 p_k，然后确定合适的距离，即步长 α_k。在信赖域中，首先选择最大距离——信赖域半径 Δ_k，然后在该距离约束下寻找能实现最佳改进的方向和步长, 若该效果不理想，我们就减小距离度量 Δ_k 并再次尝试。

对此有两个主要问题：线搜索方法中搜索方向 p_k 的选择，以及信赖域方法中 B_k 的选择。

3.1 线搜索的搜索方向

3.1.1 下降方向

最速下降方向为 $-\nabla f_k$，它是线搜索方法中搜索方向的直观选择，是从 x_k 出发使 f 下降最快的方向。事实上，令 $h(\alpha) = f(x_k + \alpha p)$, 为了让其在 x_k 处有最快的下降方向，则需要 $\frac{d}{d\alpha} h(0) = \nabla f_k^\top p$ 最小，即

$$\min_p \nabla f_k^\top p, \quad \text{s.t. } \|p\| = 1.$$

由于 $\nabla f_k^\top p = \|p\| \|\nabla f_k\| \cos\theta = \|\nabla f_k\| \cos\theta$（其中 θ 是 p 与 ∇f_k 的夹角），易知当 $\cos\theta = -1$ 时取得极小值，此时

$$p = -\nabla f_k / \|\nabla f_k\|.$$

除此之外，**任何下降方向**——与 $-\nabla f_k$ 的夹角严格小于 $\pi/2$ 的方向，也可以作为线搜索方向。这是因为只要步长足够小，都能保证使 f 减小。事实上，此时 $\frac{d}{d\alpha} h(0) = \nabla f_k^\top p < 0$，则根据可微函数的性质对足够小的 $\epsilon > 0$，有 $f(x_k + \epsilon p_k) < f(x_k)$。

3.1.2 牛顿方向

另一种线搜索方向是**牛顿方向**。该方向由 $f(x_k+p)$ 的二阶泰勒级数近似推导而来，即

$$f(x_k+p) \approx f_k + p^T\nabla f_k + \frac{1}{2}p^T\nabla^2 f_k p \overset{\text{def}}{=} m_k(p).$$

假设 $\nabla^2 f_k$ 是正定的，通过寻找使 $m_k(p)$ 极小的向量 p 可得到牛顿方向，即：

$$p_k^{\mathrm{N}} = -(\nabla^2 f_k)^{-1}\nabla f_k.$$

注 3.1 牛顿方向实际上是利用对目标函数的二阶近似得到的搜索方向，而下降方向是通过对目标函数的一阶近似得到的。从这个角度上来看，如果忽略求使用牛顿方向涉及到的矩阵运算带来的计算成本，牛顿方向比下降方向更有优势。

当 $\nabla^2 f_k$ 正定时，有

$$\nabla f_k^T p_k^{\mathrm{N}} = -p_k^{\mathrm{N}T}\nabla^2 f_k p_k^{\mathrm{N}} \le -\sigma_k\|p_k^{\mathrm{N}}\|^2$$

其中 $\sigma_k > 0$ 为 $\nabla^2 f_k$ 的最小特征值, 故牛顿方向也是下降方向。但是与一般的下降方向不同的是，选取牛顿方向是步长一般可以直接取 1, 不必再进行线搜索 (这是因为求解牛顿方向时已经蕴含了对步长的求解)。

3.1.3 拟牛顿方向

如果 $\nabla^2 f_k$ 或者 $\nabla^2 f_k^{-1}$ 不易求解怎么办？通常可以使用 $\nabla^2 f_k$ 的近似矩阵 B_k 代替 $\nabla^2 f_k$，且在每一步后更新 B_k 。更新利用了如下近似：

$$\nabla^2 f_k(x_{k+1}-x_k) \approx \nabla f_{k+1} - \nabla f_k,$$

即选择的 Hessian 近似 B_{k+1}，要求它满足以下割线方程：

$$B_{k+1}s_k = y_k,$$

其中

$$s_k = x_{k+1}-x_k, \quad y_k = \nabla f_{k+1} - \nabla f_k.$$

另外通常会对 B_{k+1} 施加额外条件，例如：

- 对称性（由精确 Hessian 的对称性启发）;
- 要求连续近似 B_k 和 B_{k+1} 之间的差具有低秩。

以下是两种更新公式：

- 秩一公式: $B_{k+1} = B_k + \frac{(y_k-B_ks_k)(y_k-B_ks_k)^T}{(y_k-B_ks_k)^Ts_k}$

- BFGS 公式: $B_{k+1} = B_k - \frac{B_k s_k s_k^T B_k}{s_k^T B_k s_k} + \frac{y_k y_k^T}{y_k^T s_k}$

注 3.2 注意，秩一公式中矩阵 B_k 和 B_{k+1} 之间的差是秩一矩阵, BFGS 公式中是秩二矩阵。这两种更新都满足割线方程，且都保持对称性。可以证明，当初始近似 B_0 正定且 $s_k^T y_k > 0$ 时，BFGS 更新会生成正定近似。

由于搜索方向定义为 $p_k = -B_k^{-1}\nabla f_k$，于是更节约的办法是直接对 $H_K = B_k^{-1}$ 进行更新：

$$H_{k+1} = (I - \rho_k s_k y_k^T) H_k (I - \rho_k y_k s_k^T) + \rho_k s_k s_k^T, \quad \rho_k = \frac{1}{y_k^T s_k}.$$

此时，搜索方向可通过 $p_k = -H_k \nabla f_k$ 计算。

3.1.4 共轭梯度方向

非线性共轭梯度方向形式为

$$p_k = -\nabla f(x_k) + \beta_k p_{k-1},$$

其中 $\beta_k \in \mathbb{R}^n$ 使得 p_k 和 p_{k-1} 是共轭的。

注 3.3 共轭梯度方法最初用于求解线性方程组 $Ax = b$（其中系数矩阵 A 对称正定）。求解该线性系统的问题等价于极小化凸二次函数

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

如果将其推广到扩展到更一般的无约束极小化问题就是非线性共轭梯度法。一般来说，非线性共轭梯度方向比最速下降方向有效得多，且计算几乎同样简单。这些方法虽未达到牛顿或拟牛顿方法的快速收敛速度，但优势在于避免存储矩阵。

3.2 信赖域子问题的近似模型

信赖域子问题一般取：

$$\min_p m_k(x_k + p) = f_k + p^T \nabla f_k + \frac{1}{2} p^T B_k p, \quad \text{s.t.} \|p\| \le \Delta_k.$$

若令 $B_k = 0$，并使用欧几里得范数，则信赖域子问题变为

$$\min_p f_k + p^T \nabla f_k \quad \text{s.t. } \|p\|_2 \le \Delta_k,$$

上述问题的解可写成：

$$p_k = -\frac{\Delta_k \nabla f_k}{\|\nabla f_k\|}.$$

注 3.4 这是一个简单的最速下降步，步长由信赖域半径确定；此时信赖域和线搜索方法本质上是相同的。

如果选择 B_k 为精确 Hessian $\nabla^2 f_k$。由于信赖域限制 $\|p\|_2 \leq \Delta_k$，即使 $\nabla^2 f_k$ 不正定，子问题也保证有解。

若二次模型函数 m_k 中的矩阵 B_k 由拟牛顿近似定义，则得到信赖域拟牛顿法。

Part II

基础算法

4 线搜索方法：Line Search Methods

线搜索方法的每次迭代都会计算一个搜索方向 p_k，然后决定沿该方向移动多远 (即步长 α_k)。迭代公式为

$$x_{k+1} = x_k + \alpha_k p_k$$

其中 α_k 称为步长。p_k 和 α_k 的有效选择是线搜索方法的关键。

4.1 步长的选择

在选择步长时, 当然希望 α_k 可以显著降低 f 的值，但是同时又不能花费过多时间。比如可以选择 $\phi(\cdot)$ 的全局极小点：

$$\phi(\alpha) = f(x_k + \alpha p_k), \quad \alpha > 0$$

但一般来说，确定这个值的计算成本很高。更实用的策略是执行**非精确线搜索准则**，以较小的成本获得使 f 足够降低的步长。

4.1.1 Wolfe Conditions

一个显然的目的是选择 α_k 使得 f 有充分的减少：

$$f(x_k + \alpha p_k) \leq f(x_k) + c_1 \alpha \nabla f_k^T p_k,$$

其中 $c_1 \in (0, 1)$。这个不等式就是**充分下降条件**，有时被称为 Armijo 条件。不等式右侧是一个线性函数，可记为 $l(\alpha)$，斜率为 $c_1 \nabla f_k^T p_k$ 为负。Wolfe Conditions 表明，只有当 $\phi(\alpha) \leq l(\alpha)$ 时，α 才被接受。在实际应用中，c_1 通常选得很小，例如 $c_1 = 10^{-4}$。

通常充分下降条件本身不一定确保算法取得合理进展，因为对于所有足够小的 α，该条件都能满足。为了排除过小的步长，通常会要求步长 α_k 满足**曲率条件**：

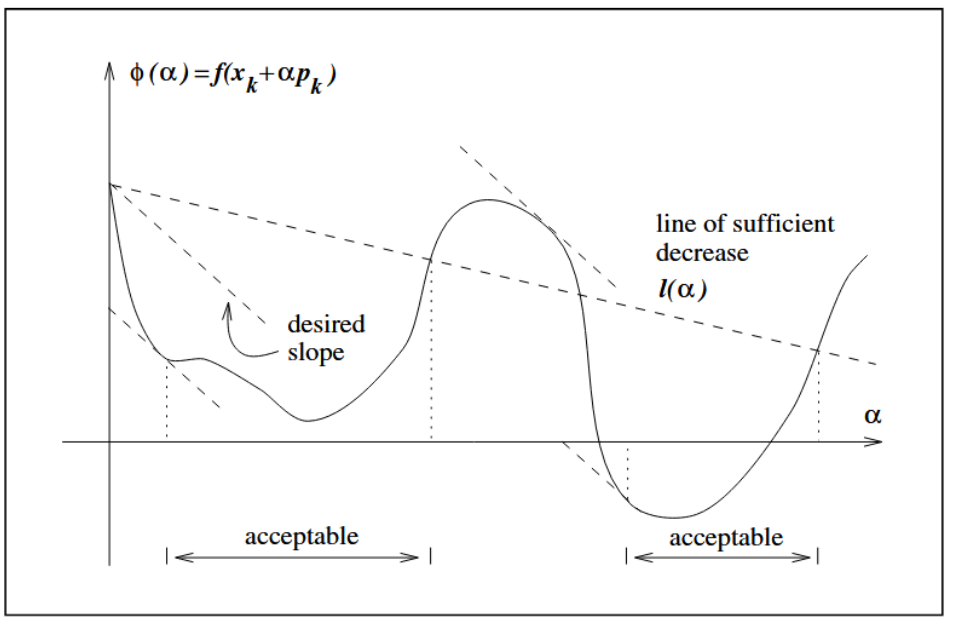
$$\nabla f(x_k + \alpha_k p_k)^T p_k \geq c_2 \nabla f_k^T p_k,$$

其中 $c_2 \in (c_1, 1)$。注意，左侧就是导数 $\phi'(\alpha_k)$，这个条件确保了 ϕ 在 α_k 处的斜率大于初始斜率 $\phi'(0)$ 的 c_2 倍。当搜索方向 p_k 由牛顿法或拟牛顿法选取时，c_2 一般为 0.9；当 p_k 由非线性共轭梯度法得到时，c_2 一般为 0.1。

充分下降条件和曲率条件都被统称为 Wolfe 条件:

$$\begin{aligned} f(x_k + \alpha p_k) &\leq f(x_k) + c_1 \alpha \nabla f_k^T p_k, \\ \nabla f(x_k + \alpha_k p_k)^T p_k &\geq c_2 \nabla f_k^T p_k, \end{aligned}$$

其中，$0 < c_1 < c_2 < 1$.



如果按照 Wolfe 条件，步长可能较大，从而远离附近的局部极小点，如上图。可以修改曲率条件，迫使 α_k 至少位于 ϕ 的局部极小值点或驻点的一个邻域内, 则得到强 Wolfe 条件:

$$\begin{aligned} f(x_k + \alpha_k p_k) &\leq f(x_k) + c_1 \alpha_k \nabla f_k^T p_k, \\ |\nabla f(x_k + \alpha_k p_k)^T p_k| &\leq c_2 |\nabla f_k^T p_k|, \end{aligned}$$

其中 $0 < c_1 < c_2 < 1$。区别在于强 Wolfe 条件不允许 $\phi'(\alpha_k)$ 超过 0 太多。

下面一个定理说明了对于一个连续可微函数一定存在步长区间满足 Wolfe 条件。

定理 4.1 *假设 $f : \mathbb{R}^n \to \mathbb{R}$ 是连续可微的。设 p_k 是在 x_k 处的下降方向，且假设 $f(x_k + \alpha p_k)$ 在 $\alpha > 0$ 时下方有界。那么若 $0 < c_1 < c_2 < 1$，存在满足 Wolfe 条件和强 Wolfe 条件的步长区间。*

证明. 注意到 $\phi(\alpha) = f(x_k + \alpha p_k)$ 对所有 $\alpha > 0$ 下方有界。由于 $c_1 \nabla f_k^T p_k < 0$，则直线 $l(\alpha) = f(x_k) + \alpha c_1 \nabla f_k^T p_k$ 下方无界，因此至少与 ϕ 的图像相交一次。设 $\alpha' > 0$ 是 α 的最小相交值，即

$$f(x_k + \alpha' p_k) = f(x_k) + \alpha' c_1 \nabla f_k^T p_k$$

考虑函数 $g(\alpha) = f(x_k + \alpha p_k) - f(x_k) - \alpha c_1 \nabla f_k^T p_k$, 因为 $g(0) = g(\alpha') = 0$ 且 $g'(0) = (1 - c_1)\nabla f_k^T p_k < 0$, 则根据连续可微函数的性质，充分下降条件对所有小于 α' 的步长成立。

由中值定理，存在 $\alpha'' \in (0, \alpha')$ 使得

$$f(x_k + \alpha' p_k) - f(x_k) = \alpha' \nabla f(x_k + \alpha'' p_k)^T p_k$$

结合上述两式，可得

$$\nabla f(x_k + \alpha'' p_k)^T p_k = c_1 \nabla f_k^T p_k > c_2 \nabla f_k^T p_k$$

因为 $c_1 < c_2$ 且 $\nabla f_k^T p_k < 0$。因此，α'' 满足 Wolfe 条件，。又由 f 的连续可微，存在以 α'' 的领域，在该领域内 Wolfe 条件成立。

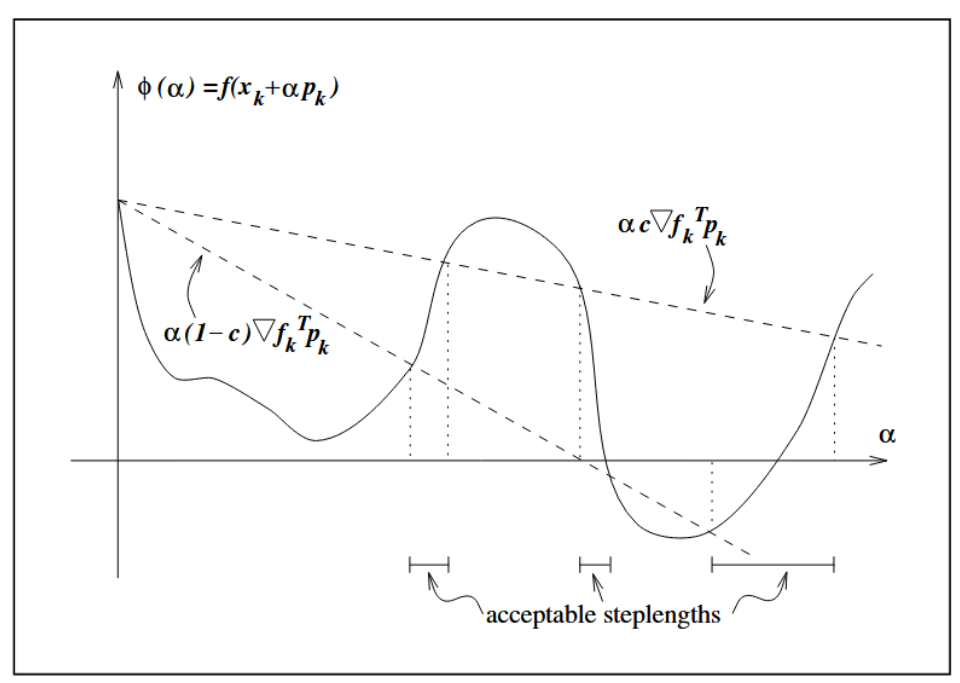
此外，由于 $\nabla f(x_k + \alpha'' p_k)^T p_k = c_1 \nabla f_k^T p_k$ 为负，显然强 Wolfe 条件在同一领域内也成立。

4.1.2 Goldstein 条件

另一种 Goldstein 条件也可以确保步长 α 实现充分下降且不会过短。Goldstein 条件为：

$$f(x_k) + (1 - c)\alpha_k \nabla f_k^T p_k \leq f(x_k + \alpha_k p_k) \leq f(x_k) + c\alpha_k \nabla f_k^T p_k$$

其中 $0 < c < 1/2$。其中第二个不等式是充分下降条件，而第一个不等式用于从下方控制步。



注 4.1 Goldstein 条件相较于 Wolfe 条件有一个缺点是，其第一个不等式可能会排除 ϕ 的所有极小值点。不过，Goldstein 条件和 Wolfe 条件的收敛理论十分相似。Goldstein 条件常用于牛顿型方法，但不太适用于保持正定 Hessian 近似的拟牛顿方法。

4.2 线搜索方法的收敛性

假设 p_k 为下降方向，即 $\nabla f_k^\top p_k < 0$, 因此 p_k 和最速下降方向 $-f_k$ 的夹角的余弦值为

$$\cos\theta_k = \frac{-\nabla f_k^\top p_k}{\|\nabla f_k\|\|p_k\|}.$$

定理 4.2 考虑迭代 $x_{k+1} = x_k + \alpha_k p_k$，其中 p_k 是下降方向且 α_k 满足 Wolfe 条件。假设 f 在 $\mathbb{R}^n$ 上下有界，且 f 在包含水平集 $\mathcal{L} \triangleq \{x : f(x) \le f(x_0)\}$ 的开集 $\mathcal{N}$ 上连续可微，其中 x_0 是迭代的起始点, 假设梯度 ∇f 在 $\mathcal{N}$ 上是 Lipschitz 连续的，即存在常数 $L > 0$，使得

$$\|\nabla f(x) - \nabla f(\tilde{x})\| \le L\|x - \tilde{x}\|, \quad \forall\, x, \tilde{x} \in \mathcal{N}.$$

则

$$\sum_{k\ge 0} \cos^2\theta_k \|\nabla f_k\|^2 < \infty.$$

证明. 由 Wolfe 条件第二个不等式与 Lipschitz 条件可得

$$(\nabla f_{k+1} - \nabla f_k)^T p_k \ge (c_2 - 1)\nabla f_k^T p_k,$$
$$(\nabla f_{k+1} - \nabla f_k)^T p_k \le \alpha_k L\|p_k\|^2.$$

于是可得

$$\alpha_k \geq \frac{c_2 - 1}{L} \frac{\nabla f_k^T p_k}{\|p_k\|^2}.$$

将此不等式代入 Wolfe 条件第二个不等式可得

$$f_{k+1} \leq f_k - c_1 \frac{1 - c_2}{L} \frac{(\nabla f_k^T p_k)^2}{\|p_k\|^2}.$$

即

$$c \cos^2 \theta_k \|\nabla f_k\|^2 \leq f_k - f_{k+1},$$

其中 $c = c_1(1 - c_2)/L$。求和可得 $\forall N \in \mathbb{N}^+$

$$\sum_{k=0}^{N} \cos^2 \theta_j \|\nabla f_j\|^2 \leq \frac{1}{c}(f_0 - f_{k+1}).$$

由于 f 下有界，所以有 $f_0 - f_{k+1}$ 小于某个正常数。取极限可得

$$\sum_{k=0}^{\infty} \cos^2 \theta_k \|\nabla f_k\|^2 < \infty.$$

注 4.2

- 我们称 $\sum_{k=0}^{\infty} \cos^2 \theta_k \|\nabla f_k\|^2 < \infty$ 为 Zoutendijk 条件.
- 由 Zoutendijk 条件可以得到：$\cos^2 \theta_k \|\nabla f_k\|^2 \to 0$. 因此如果 $\exists\, \delta > 0,\ \forall\, k,\ \cos\theta_k \geq \delta > 0$, 则 $\|\nabla f_k\| \to 0$. 也就是说，只要让搜索方向与负梯度的方向夹角始终小于 $\pi/2$, 就可以得到迭代点梯度的收敛性.
- 对于牛顿法或者拟牛顿法，$p_k = -B_k^{-1} f_k$，其中 B_k 为对称正定矩阵，则

  $$\cos\theta_k = \frac{\nabla f_k^\top B_k \nabla f_k}{\|\nabla f_k\| \|B_k^{-1} f_k\|} \geq \frac{\|B_k\|^{-1} \|\nabla f_k\|^2}{\|B_k^{-1}\| \|\nabla f_k\|^2} = \frac{\|B_k\|^{-1}}{\|B_k^{-1}\|},$$

  因此只要 $\|B_k\| \|B_k^{-1}\| \leq \delta^{-1}$ 即可.

4.3 线搜索方法的收敛率

4.3.1 最速下降法的收敛率

我们先考虑最基础的算法最速下降法的收敛率，先考虑最简单问题

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

其中 Q 是对称正定矩阵。梯度为 $\nabla f(x) = Qx - b$，问题的全局极小点 x^* 就是线性方程组 $Qx = b$ 的唯一解。通过对 $f(x_k - \alpha \nabla f_k)$ 求极小可得最速下降法的步长为:

$$\alpha_k = \frac{\nabla f_k^T \nabla f_k}{\nabla f_k^T Q \nabla f_k},$$

于是最速下降迭代式为

$$x_{k+1} = x_k - \left(\frac{\nabla f_k^T \nabla f_k}{\nabla f_k^T Q \nabla f_k} \right) \nabla f_k,$$

其中 $\nabla f_k = Qx_k - b$.

设 $||x||_Q^2 = x^T Qx$。利用关系 $Qx^* = b$，可以得到

$$\begin{aligned}
\frac{1}{2}||x - x^*||_Q^2 &= \frac{1}{2}x^\top Qx - x^{*\top} Qx + \frac{1}{2}x^{*\top} Qx^* \\
&= \frac{1}{2}x^\top Qx - b^\top x - \frac{1}{2}x^{*\top} Qx^* + x^{*\top} Qx^* \\
&= \frac{1}{2}x^\top Qx - b^\top x - \left(\frac{1}{2}x^{*\top} Qx^* - b^\top x^* \right) \\
&= f(x) - f(x^*),
\end{aligned}$$

因此该范数衡量了当前目标函数值与最优值之间的距离。由于 $\nabla f_k = Q(x_k - x^*)$，可推导出等式

$$\begin{aligned}
\frac{1}{2}\|x_{k+1} - x^*\|_Q^2 &= f(x^{k+1}) - f(x^*) \\
&= \frac{1}{2}\left(x^k - \alpha_k \nabla f(x^k)\right)^\top Q \left(x^k - \alpha_k \nabla f(x^k)\right) - b^\top (x^k - \alpha_k \nabla f(x^k)) - f(x^*) \\
&= \frac{1}{2}x^{k\top} Qx^k - b^\top x^k - f(x^*) + \frac{\alpha_k^2}{2}\nabla f(x^k)^\top Q \nabla f(x^k) - \alpha_k \nabla f(x^k)^\top (Qx^k - b) \\
&= \frac{1}{2}\|x_k - x^*\|_Q^2 + \frac{\alpha_k^2}{2}\nabla f(x^k)^\top Q \nabla f(x^k) - \alpha_k \nabla f(x^k)^\top \nabla f(x^k)
\end{aligned}$$

其中由于

$$\begin{aligned}
&\alpha_k^2 \nabla f(x^k)^\top Q \nabla f(x^k) - 2\alpha_k \nabla f(x^k)^\top \nabla f(x^k) \\
=&\frac{(\nabla f_k^T \nabla f_k)^2}{(\nabla f_k^T Q \nabla f_k)^2} \nabla f_k{}^\top Q \nabla f_k - 2\frac{(\nabla f_k^T \nabla f_k)^2}{\nabla f_k^T Q \nabla f_k} \\
=&-\frac{(\nabla f_k^T \nabla f_k)^2}{\nabla f_k^T Q \nabla f_k} \\
=&-\frac{(\nabla f_k^T \nabla f_k)^2}{(\nabla f_k^T Q \nabla f_k)(\nabla f_k^T Q^{-1} \nabla f_k)} \nabla f_k^T Q^{-1} \nabla f_k \\
=&-\frac{(\nabla f_k^T \nabla f_k)^2}{(\nabla f_k^T Q \nabla f_k)(\nabla f_k^T Q^{-1} \nabla f_k)} (x_k - x^*)^\top Q^\top Q^{-1} Q (x_k - x^*) \\
=&-\frac{(\nabla f_k^T \nabla f_k)^2}{(\nabla f_k^T Q \nabla f_k)(\nabla f_k^T Q^{-1} \nabla f_k)} \|x_k - x^*\|_Q^2
\end{aligned}$$

于是：

$$||x_{k+1} - x^*||_Q^2 = \left\{ 1 - \frac{(\nabla f_k^T \nabla f_k)^2}{(\nabla f_k^T Q \nabla f_k)(\nabla f_k^T Q^{-1} \nabla f_k)} \right\} ||x_k - x^*||_Q^2$$

这个表达式描述了每次迭代中 f 的精确下降量，上述讨论可以进一步得到下面这个定理：

定理 4.3

$$\|x_{k+1}-x^*\|_Q^2 \le \left(\frac{\lambda_n-\lambda_1}{\lambda_n+\lambda_1}\right)^2 \|x_k-x^*\|_Q^2$$

其中 $0<\lambda_1\le\lambda_2\le\cdots\le\lambda_n$ 是 Q 的特征值。

4.3.2 牛顿法的收敛率

考虑步长为 1 的牛顿法 $x_{k+1}=x_k+p_k^N=x_k-\nabla^2 f_k^{-1}\nabla f_k$

定理 4.4 假设 f 二阶可微，且 Hessian 矩阵 $\nabla^2 f(x)$ 在解 x^* 的邻域内 Lipschitz 连续，其中 x^* 满足 $\nabla f(x^*)=0$, $\nabla^2 f(x^*)$ 正定。考虑迭代 $x_{k+1}=x_k+p_k^N=x_k-\nabla^2 f_k^{-1}\nabla f_k$。则

(i) 若初始点 x_0 足够接近 x^*，则迭代序列收敛到 x^*；
(ii) $\{x_k\}$ 的收敛速度是二次的；
(iii) 梯度范数序列 $\{\|\nabla f_k\|\}$ 二次收敛到零。

证明. 由 $\nabla f_*=0$

$$x_k+p_k^{\mathrm{N}}-x^*=x_k-x^*-\nabla^2 f_k^{-1}\nabla f_k=\nabla^2 f_k^{-1}\left[\nabla^2 f_k(x_k-x^*)-(\nabla f_k-\nabla f_*)\right]$$

由于

$$\nabla f_k-\nabla f_*=\int_0^1 \nabla^2 f(x_k+t(x^*-x_k))(x_k-x^*)dt$$

我们有

$$\begin{aligned}\|\nabla^2 f(x_k)(x_k-x^*)-(\nabla f_k-\nabla f(x^*))\| &= \left\|\int_0^1[\nabla^2 f(x_k)-\nabla^2 f(x_k+t(x^*-x_k))](x_k-x^*)dt\right\| \\ &\le \int_0^1 \|\nabla^2 f(x_k)-\nabla^2 f(x_k+t(x^*-x_k))\|\|x_k-x^*\|dt \\ &\le \|x_k-x^*\|^2\int_0^1 Ltdt=\frac{1}{2}L\|x_k-x^*\|^2,\end{aligned}$$

其中 L 是 $\nabla^2 f(x)$ 在 x 接近 x^* 时的 Lipschitz 常数（**上面这个不等式很像 L-smooth 里面的下降引理，证明都差不多**）。

由于 $\nabla^2 f(x^*)$ 非奇异，存在半径 $r>0$，使得对所有满足 $\|x_k-x^*\|\le r$ 的 x_k，有 $\|\nabla^2 f_k^{-1}\|\le 2\|\nabla^2 f(x^*)^{-1}\|$。将其代入上面的式子，我们得到

$$\|x_k+p_k^{\mathrm{N}}-x^*\|\le L\|\nabla^2 f(x^*)^{-1}\|\|x_k-x^*\|^2=\tilde{L}\|x_k-x^*\|^2,$$

其中 $\tilde{L}=L\|\nabla^2 f(x^*)^{-1}\|$。选择 x_0 使得 $\|x_0-x^*\|\le\min(r,1/(2\tilde{L}))$，我们可以用数学归纳法推出该序列收敛到 x^*，且收敛速度是二次的。

通过利用关系 $x_{k+1}-x_k=p_k^{\mathrm{N}}$ 和 $\nabla f_k+\nabla^2 f_k p_k^{\mathrm{N}}=0$，我们得到

$$\begin{aligned}\|\nabla f(x_{k+1})\| &= \|\nabla f(x_{k+1})-\nabla f_k-\nabla^2 f(x_k)p_k^{\mathrm{N}}\| \\ &= \left\|\int_0^1 \nabla^2 f(x_k+tp_k^{\mathrm{N}})(x_{k+1}-x_k)dt-\nabla^2 f(x_k)p_k^{\mathrm{N}}\right\| \\ &\le \int_0^1 \|\nabla^2 f(x_k+tp_k^{\mathrm{N}})-\nabla^2 f(x_k)\|\|p_k^{\mathrm{N}}\|dt \\ &\le \frac{1}{2}L\|p_k^{\mathrm{N}}\|^2 \\ &\le \frac{1}{2}L\|\nabla^2 f(x_k)^{-1}\|^2\|\nabla f_k\|^2 \\ &\le 2L\|\nabla^2 f(x^*)^{-1}\|^2\|\nabla f_k\|^2,\end{aligned}$$

这证明了梯度范数二次收敛到零。

注 4.3 上述定理的条件其实是比较强的

4.3.3 拟牛顿法的收敛率

4.4 调整海瑟矩阵的牛顿法

牛顿法的迭代方向由以下方程聚决定：

$$\nabla^2 f(x_k)p_k^{\mathrm{N}}=-\nabla f(x_k)$$

但是这带来几个问题

- Hessian 矩阵 $\nabla^2 f(x)$ 可能不是正定的, 因此牛顿方向 p_k^{N} 可能不是下降方向
- 求解上述方程有时不是一件容易的事情

为了克服上述困难，考虑修改 Hessian 矩阵，修改后的 Hessian 矩阵是通过将正定对角矩阵或满矩阵添加到真实的 Hessian 矩阵 $\nabla^2 f(x_k)$ 中得到的：

Algorithm 1 带修正的线搜索牛顿法

Require: 初始点 x_0

1: **while** $k=0,1,2,\ldots$ **do**

2: 对矩阵 $B_k=\nabla^2 f(x_k)+E_k$ 进行分解，其中若 $\nabla^2 f(x_k)$ 是充分正定的，则 $E_k=0$；否则，选择 E_k 以确保 B_k 是充分正定的；

3: 求解 $B_k p_k=-\nabla f(x_k)$；

4: 令 $x_{k+1}\leftarrow x_k+\alpha_k p_k$，其中 α_k 满足 Wolfe、Goldstein 或 Armijo 回溯条件；

5: **end while**

注 4.4 通常不需要得到显示的 E_k，而是动态的修改。为了得到全局收敛结果，需要选择 E_k 使得 B_k 满足有界修改分解性质，即 $\{\nabla^2 f(x_k)\}$ 有界时

$$\kappa(B_k) = \|B_k\|\|B_k^{-1}\| \leq C, \quad \text{对某个 } C > 0 \text{ 和所有 } k = 0, 1, 2, \ldots,$$

它确保了修正后的矩阵 B_k 不会过于“病态”

定理 4.5 设 f 在开集 D 上二阶连续可微，假设算法 1的起始点 x_0 满足水平集 $\mathcal{L} = \{x \in D : f(x) \leq f(x_0)\}$ 是紧集。且有界修改分解性质成立，则有

$$\lim_{k\to\infty} \nabla f(x_k) = 0.$$

5 小记

5.1 一种证明迭代点列线性收敛的框架

设不动点迭代：

$$x^{k+1} = \Phi(x^k),$$

其中 $\Phi : \operatorname{dom}\Phi \subseteq \mathbb{R}^n \to \mathbb{R}^n$ 是某个（单值）映射。

设 X^* 表示 Φ 的非空不动点集。**令 $\psi : \operatorname{dom}\Phi \to \mathbb{R}_+$ 是 X^* 的残差函数，即 ψ 连续且 $\psi(x) = 0$ 当且仅当 $x \in X^*$**。假设存在三个正常数 η_1, η_2 和 δ，满足

- $\psi(x) - \psi(\Phi(x)) \geq \eta_1\|x - \Phi(x)\|^2, \quad \forall x \in \operatorname{dom}\Phi,$
- $\psi(\Phi(x)) \leq \eta_2\|x - \Phi(x)\|^2, \quad \forall x \in \operatorname{dom}\Phi \text{ 且} \|x - \Phi(x)\| \leq \delta.$

结合 $x^{k+1} = \Phi(x^k)$,, 则上面两个条件说明

- $\psi(x^k) - \psi(x^{k+1}) \geq \eta_1\|x^k - x^{k+1}\|^2$, 从这个意义上说，$\psi$ 一个评价函数。
- $\psi(x^{k+1}) \leq \eta_2\|x^k - x^{k+1}\|^2.$

定理 5.1 设 $\Phi : \operatorname{dom}\Phi \subseteq \mathbb{R}^n \to \mathbb{R}^n$ 是连续函数，其不动点集 X^* 非空。令 $\{x^k\}$ 由迭代 (12.6.1) 定义。若存在 X^* 的连续残差函数 $\psi : \operatorname{dom}\Phi \to \mathbb{R}_+$ 满足 (12.6.2) 和 (12.6.3)，则 $\{\psi(x^k)\}$ 至少以 R-线性速率收敛到 0，且序列 $\{x^k\}$ 至少以 R-线性速率收敛到 X^* 中的某个元素。

证明：由两个条件共同推出

$$\psi(x^{k+1}) \leq \frac{\eta_2}{\eta_1 + \eta_2}\psi(x^k),$$

于是得到 $\{\psi(x^k)\}$ 的 Q-线性收敛速率，且收敛到 0.

由第一个条件有

$$\eta_1\|x^k - x^{k+1}\|^2 \leq \psi(x^k) \leq \left(\frac{\eta_2}{\eta_1+\eta_2}\right)^k \psi(x^0),$$

从而得到

$$\|x^k - x^{k+1}\| \leq \sqrt{\frac{\psi(x^0)}{\eta_1}}\left(\sqrt{\frac{\eta_2}{\eta_1+\eta_2}}\right)^k.$$

因此，

$$\|x^k - x^{k+m}\| \leq \sqrt{\frac{\psi(x^0)}{\eta_1}}\sum_{j=k}^{k+m-1}\left(\sqrt{\frac{\eta_2}{\eta_1+\eta_2}}\right)^j.$$

故 $\{x^k\}$ 是 Cauchy 序列，从而收敛，记极限为 x^∞。由 ψ 的连续性且 $\lim \psi(x^k) = \Psi(x^\infty) = 0$，$x^\infty$ 属于 X^*；进一步，令 $m \to \infty$

$$\|x^k - x^\infty\| \leq \sqrt{\frac{\psi(x^0)}{\eta_1}}\frac{1}{1-\sqrt{\frac{\eta_2}{\eta_1+\eta_2}}}\left(\sqrt{\frac{\eta_2}{\eta_1+\eta_2}}\right)^k,$$

这表明 $\{x^k\}$ 至少以 R-线性速率收敛到 x^∞。

注 5.1 我觉得上述想法来源可能是一个特殊情况，就是 $\psi(x^k) = \|x^k - x^*\|^2$ 的时候，此时第一个条件其实就是我们熟悉的

$$\|x^{k+1} - x^*\|^2 \leq \|x^k - x^*\|^2 - \eta_1\|x^{k+1} - x^k\|^2$$

此时只要再结合第二个 $\|x^{k+1} - x^*\|^2 \leq \eta_2\|x^k - x^{k+1}\|^2$ 就可以得到 $\{x^k\}$ 的 Q-线性收敛。

而如果把第二个条件换成 $\|x^k - x^*\|^2 \leq \eta_3\|x^k - x^{k+1}\|^2$，且 $\eta_3 > \eta_1$，其实也可以得到 $\{x^k\}$ 的 Q-线性收敛。于是上述第二个条件可以换成

$$\psi(x^k) \leq \eta_3\|x^k - x^{k+1}\|^2, \eta_3 > \eta_1.$$