第三讲：最速下降法和牛顿法

最速下降法

算法1

给定$x_0 \in R^n$, $0 \leq \varepsilon \ll 1$;

for $k = 0,\ 1,\ \cdots$

计算搜索方向$p_k = -\nabla f(x_k)$;

计算步长α_k, 使得 $f(x_k + \alpha_k p_k) = \min\limits_{\alpha \geq 0} f(x_k + \alpha p_k)$;

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

if $\|\nabla f(x_k)\| \leq \varepsilon$

stop;

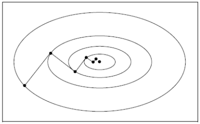
end (if)

end (for)

最速下降法

性质

- 负梯度方向是下降最快的方向，但仅是局部性质 。
- 如果采用精确线搜索方法，由于$\varphi'(\alpha_k) = 0$, 则会出现$p_{k+1}^T p_k = 0$. (当目标函数的等值线是一个扁长椭球时，会出现"锯齿现象"，下降十分缓慢)



- 总体是线性收敛的.

总体收敛性定理

Theorem (Zoutendijk)

- 考虑任何一个线搜索算法, 如果p_k是一个下降方向，步长因子α_k满足Wolfe 条件.
- 假设$f(x)$ 在 $\mathcal{N} \equiv \{x| : f(x) \leq f(x_0)\}$上连续可微并且有下界，其中$x_0$ 是迭代起始点.
- 假设 ∇f 在 $\mathcal{N}$ 上 Lipschitz 连续，即存在常数 $L > 0$ 使得

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|, \quad \forall x, y \in \mathcal{N}. \tag{2.1}$$

- **Then**

$$\sum_{k\geq 0} \cos^2(\theta_k)\|\nabla f(x_k)\|^2 < \infty \tag{2.2}$$

which is called Zoutendijk condition.

Convergence of Line Search Methods

Remark

- 如果使用 Goldstein 条件 strong Wolfe 条件，类似结果也成立
- The Zoutendijk condition (2.2) implies that

$$\cos^2(\theta_k)\|\nabla f(x_k)\|^2 \to 0. \tag{2.3}$$

- 可以推导 global convergence 结果.

总体收敛性定理

负梯度方向

Remark

- 如果 p_k 满足与`$\nabla f(x_k)$的夹角 $\theta_k < 90^\circ$, 即存在正常数$\delta > 0$,

$-\nabla f(x_k)$

$$\cos\theta_k \geq \delta > 0, \forall k \tag{2.4}$$

It follows immediately from (2.3) that

$$\lim_{k\to\infty} \|\nabla f(x_k)\| = 0. \tag{2.5}$$

- 即$\|\nabla f(x_k)\| \to 0$, 只要p_k不靠近 $-\nabla f(x_k)$的正交方向。

最速下降法收敛速度

首先考虑二次函数$f(x) = x^T Gx$, 其中G对称正定。

二次函数最速下降法的收敛速度定理

假设λ_1和λ_n是G的最大和最小特征值. 设x°是问题的解, 则最速下降法的收敛,速度至少是线性的, 并且满足

$$\frac{f(x_{k+1}) - f(x^*)}{f(x_k) - f(x^*)} \leq \frac{(\kappa-1)^2}{(\kappa+1)^2} = \frac{(\lambda_1-\lambda_n)^2}{(\lambda_1+\lambda_n)^2}, \tag{2.6}$$

$$\frac{\|x_{k+1} - x^*\|}{\|x_k - x^*\|} \leq \sqrt{\kappa}\frac{(\kappa-1)}{(\kappa+1)}, \tag{2.7}$$

其中$\kappa = \lambda_1/\lambda_n \geq 1$ (证明)

梯度法推广

两点步长梯度法

- 基本思想： $x_{k+1} = x_k - D_k \nabla f(x_k)$ 其中 $D_k = \alpha_k I$.
- 可以使 D_k 具有拟牛顿性质，即计算 α_k 使得

$$\min \ \|s_{k-1} - D_k y_{k-1}\| \text{ 或者 } \min \ \|D_k^{-1} s_{k-1} - y_{k-1}\| \tag{2.10}$$

其中 $s_{k-1} = x_k - x_{k-1}$, $y_{k-1} = \nabla f(x_k) - \nabla f(x_{k-1})$

- 可以求出 α_k，

$$\alpha_k = \frac{s_{k-1}^T y_{k-1}}{\|y_{k-1}\|^2}, \text{ 或者 } \alpha_k = \frac{\|s_{k-1}\|^2}{s_{k-1}^T y_{k-1}}$$

→ 二次函数求解

两点步长梯度法

The Barzilai-Borwein gradient method

给定$x_0 \in R^n$, $0 \le \varepsilon \ll 1$;
for $k = 0,\ 1,\ \cdots$
  计算搜索方向$p_k = -\nabla f(x_k)$;
  计算步长α_k，如果$k = 0$, 进行线搜索，
  如果$k \neq 0$, $\alpha_k = \frac{s_{k-1}^T y_{k-1}}{\|y_{k-1}\|^2}$，或者 $\alpha_k = \frac{\|s_{k-1}\|^2}{s_{k-1}^T y_{k-1}}$;
  $x_{k+1} = x_k + \alpha_k p_k$;
  if $\|\nabla f(x_k)\| \le \varepsilon$
    stop;
  end (if)
end (for)

两点步长梯度法

性质

- 不需要做线搜索(k=0除外)，
- 没有矩阵乘以向量运算，因此计算量小，
- 本质上是梯度法, 但是收敛速度要更快.
- Barzilai and Borwein 证明了对于二次目标函数问题，该算法是R-superlinearly收敛.
- 对于非二次目标函数, α_k有可能很大或很小，因此我们需要给α_k设置上下界.

牛顿法

$f(x)$二次可微, $x_k \in R^n$, Hessian$\nabla^2 f(x_k)$正定，有二次Taylor展开

$$f(x_k + s) \approx f(x_k) + \nabla f(x_k)^T s + \frac{1}{2} s^T \nabla^2 f(x_k) s. \tag{2.11}$$

极小化二次近似函数, 可得到$s = -[\nabla^2 f(x_k)]^{-1} \nabla f(x_k)$, 于是

$$x_{k+1} = x_k - [\nabla^2 f(x_k)]^{-1} \nabla f(x_k) \tag{2.12}$$

记$p_k = -[\nabla^2 f(x_k)]^{-1} \nabla f(x_k)$, $\alpha_k = 1$, 则迭代格式可以写为$x_{k+1} = x_k + p_k$.

nx1 . 方向 步长

牛顿法

Remark

记$\nabla^2 f(x_k) = G_k$，$g_k = \nabla f(x_k)$.
迭代格式记为$x_{k+1} = x_k - G_k^{-1} g_k$

- 牛顿法可以看作在椭球范数$\|\cdot\|_{G_k}$下的最速下降法.
- 对于$f(x_k + s) \approx f(x_k) + g_k^T s$，$s$是如下极小化问题的解

$$\min_{s \in R^n} \frac{g_k^T s}{\|s\|}$$

  - 如果范数采用l_2，则$s_k = -g_k$.
  - 如果范数采用$\|\cdot\|_{G_k}$，
    则原问题等价于 $\min_{s \in R^n} \; g_k^T s$ where $\|s\|_{G_k} \le 1$.
    由于$(g_k^T s)^2 \le (g_k^T G_k^{-1} g_k)(s^T G_k s)$，则 $s_k = -G_k^{-1} g_k$.

牛顿法收敛性

如果f是二次函数，则牛顿迭代法一步就达到最优解.

牛顿法收敛定理

x_k (证明)

设f二次连续可微，$\cancel{x^*}$充分靠近x^*，$\nabla f(x^*)=0$，如果$\nabla^2 f(x^*)$正定,且Hessian矩阵$G(x)$满足Lipschitz条件，即存在$L>0$，对所有的(i,j)，

$$|G_{ij}(x)-G_{ij}(y)| \le L\|x-y\|. \tag{2.13}$$

则对一切k，牛顿迭代法得到的序列x_k收敛到x^*，并且具有二次收敛速度.

$$|G_{ij}(x)-G_{ij}(y)| \le L\|x-y\|$$

牛顿法

Remark G_k不正定

在牛顿法中可以增加步长因子的搜索过程，使得当G_k不正定的时候也收敛.

$$p_k = -G_k^{-1} g_k,\ x_{k+1} = x_k + \alpha_k p_k \tag{2.14}$$

算法2(带步长因子的牛顿法)

给定$x_0 \in R^n$, $0 \le \varepsilon \ll 1$;
for $k = 0,\ 1,\ \cdots$
  计算搜索方向$p_k = -[\nabla^2 f(x_k)]^{-1} f(x_k)$;
  计算步长α_k, 使得 $f(x_k + \alpha_k p_k) = \min\limits_{\alpha \ge 0} f(x_k + \alpha p_k)$;
  $x_{k+1} = x_k + \alpha_k p_k$;
  if $\|\nabla f(x_k)\| \le \varepsilon$
    stop;
  end (if)
end (for)

带步长因子牛顿法收敛性

定理

设f在开凸集D中二阶连续可微，又设对任意的$x_0 \in D$, 存在常数$m > 0$, 使得$f(x)$在水平集$L(x_0) = \{x | f(x) \leq f(x_0)\}$ 上满足

$$u^T \nabla^2 f(x) u \geq m \|u\|^2, \forall u \in R^n, x \in L(x_0). \tag{2.15}$$

则在精确一维搜索条件下，带步长因子的牛顿法产生的迭代点列$\{x_k\}$满足

1. 当x_k为有限点列，则对某个k，$\nabla f(x_k) = 0$.
2. 当x_k为无穷点列，$\{x_k\}$收敛到f的唯一极小值点.

Goldstein-Price修正牛顿法

- 如果G_k不正定，可以用$p_k = -\nabla f(x_k)$代替$p_k^N = -[\nabla^2 f(x_k)]^{-1}\nabla f(x_k)$，因此也会适当引入步长因子.
- 给定阈值$\eta > 0$

$$p_k = \begin{cases} -G_k^{-1} g_k, & if \cos\theta \geq \eta \\ -g_k, & otherwise. \end{cases}$$

负曲率方向 Negative curvature direction

负曲率方向属于修正牛顿法，是处理在$\nabla^2 f(x)$不定时，如何寻找搜索方向

定义 Eigenvalue

设$f: R^n \to R$在开集D上二次连续可微 Indefinite point

1. 如果$\nabla^2 f(x)$至少有一个负特征值，则称x为不定点
2. 如果x是一个不定点，若方向p满足 $p^T\nabla^2 f(x)p < 0$，则 称p为$f(x)$在x处的负曲率方向.
3. 如果x为不定点，

$$s^T\nabla f(x) \le 0, p^T\nabla f(x) \le 0, p^T\nabla^2 f(x)p < 0$$

则称(s,p)为不定点x处的下降对.

4. 如果x不是一个不定点，(s,p)满足

$$s^T\nabla f(x) < 0, p^T\nabla f(x) \le 0, p^T\nabla^2 f(x)p = 0$$

则称(s,p)为x处的下降对.

负曲率方向

Remark

- 一个下降对例子

$$s = -\nabla f(x), p = \begin{cases} 0, & 如果\nabla^2 f(x) \geq 0, \\ -\mathsf{sign}(u^T \nabla f(x))u, & otherwise, \end{cases}$$

其中u是$\nabla^2 f(x)$的负特征值对应的特征向量.

- 显然，当且仅当$\nabla f(x) = 0$且$\nabla^2 f(x)$半正定时，下降对不存在.
- 如果$\nabla f(x) = 0$, 负曲率方向是下降方向.
- 在一般点处$(\nabla f(x) \neq 0)$,
  - 如果负曲率方向p满足$p^T \nabla f = 0$， 则p和$-p$都是下降方向.
  - 如果负曲率方向p满足$p^T \nabla f \leq 0$， 则p是下降方向.
  - 如果负曲率方向p满足$p^T \nabla f \geq 0$， 则$-p$是下降方向.

Gill-Murray稳定牛顿法收敛性

定理：Gill-Murray稳定牛顿法收敛性

- 设f二次连续可微
- 存在$\bar{x} \in R^n$, 使得$L(\bar{x}) = \{x | f(x) \leq f(\bar{x})\}$ 为有界闭凸集
- 以上算法中$\varepsilon = 0$
- 则序列$\{x_k\}$满足
    1. 当$\{x_k\}$为有限序列， 最后一个点是稳定点
    2. 当$\{x_k\}$为无限序列，则必有聚点, 且所有聚点都是稳定点.

不精确的牛顿法

考虑非线性方程组 $F(x)=0$ ，其中$F: R^n \to R^n$，满足

- (1)存在解x^*，$F(x^*)=0$，
- (2)F在x^*的领域中连续可微，
- (3)$F'(x^*)$非奇异.

- 经典牛顿法：$F(x_k+s) \approx F(x_k)+F'(x_k)s=0$，$s=-(F'(x_k))^{-1}F(x_k)$
- 不精确牛顿法：$F'(x_k)s=-F(x_k)+r_k$，其中 $\frac{r_k}{\|F(x_k)\|} \leq \eta_k$.
  $r_k=F'(x_k)s+F(x_k)$表示残量，η_k是一个控制不精确程度的序列.

引理：F'在x^*附近非奇异

设$F: R^n \to R^n$在$x^* \in D$连续可微，$F'(x^*)$非奇异，则存在$\delta>0, \gamma>0, \epsilon>0$，使得当$\|y-x^*\|<\delta$，$y \in D$时，$F'$非奇异，且

$$\|F'(y)^{-1}\| \leq \gamma, \quad \|F'(y)^{-1}-F'(x^*)^{-1}\| \leq \epsilon.$$

证明

不精确的牛顿法收敛性

定理：线性收敛

设$F: R^n \to R^n$满足假设条件(1)(2)(3). 设η_k满足$0 \le \eta_k \le \eta_{\max} < t < 1$. 那么，对于某个$\epsilon > 0$, 如果$\|x_0 - x^*\| < \epsilon$, 则有不精确牛顿法产生的迭代序列$\{x_k\}$线性收敛到$x^*$,

$$\|x_{k+1} - x^*\|_* \le t\|x_k - x^*\|_*,$$

其中$\|y\|_* = \|F'(x^*)^{-1}y\|$. $=\|F'(x^*)y\|$

引理

设$\alpha = \max\{\|F'(x^*) + \frac{1}{2\beta}\|, 2\beta\}$, $\beta = \|F'(x^*)^{-1}\|$, 则当$\|y - x^*\|$充分小时，

$$\frac{1}{\alpha}\|y - x^*\| \le \|F(y)\| \le \alpha\|y - x^*\|.$$

定理: 超线性收敛

假设上述定理条件都成立，不精确牛顿法产生的序列$\{x_k\}$是收敛到x^*, 那么当且仅当$\|r_k\| = o(\|F(x_k)\|)$, $k \to \infty$时，$\{x_k\}$超线性收敛到x^*.