最优化技术原理及应用

王冰川
bingcwang@csu.edu.cn
https://intleo.csu.edu.cn/index.html
中南大学自动化学院

线搜索方法

主要参考中国科学院大学Xiao Wang老师PPT内容

主要内容

主要内容

算法框架

- 选择方向向量：p_k
- 从当前迭代 x_k 开始沿此方向搜索具有较低函数值的新迭代

$$x_{k+1} \leftarrow x_k + \alpha_k\, p_k,\ \ f(x_k) < f(x_{k+1})$$

- 在新迭代点，计算新的方向（p_{k+1}）和步长（α_{k+1}），并重复以上过程达到终止条件

重点

- 如何选择方向向量p_k和步长α_k
- 几种代表性线搜索方法的收敛性

主要内容

- Taylor's Theorem：假设$f:\mathcal{R}^n\rightarrow\mathcal{R}$连续可微，并且$p\in\mathcal{R}^n$,则有

$$f(x+p)=f(x)+\nabla f(x+tp)^T p \quad ,for\ some\ t\in(0,1)$$

如果f连续二阶可导，则有

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

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

$$for\ some\ t\in(0,1)$$

- 根据泰勒定理，给定搜索方向（p）和步长（α），可以得到

$$f(x_k+\alpha p)=f_k+\alpha p^T\nabla f_k+\frac{\alpha^2}{2}p^T\nabla^2 f(x_k+\theta p)p,\theta\in(0,\alpha)$$

- 可知f在x_k处沿着方向p的变化是$p^T\nabla f_k$，可以通过求解以下问题得到使得f下降最快的方向

$$\min_p p^T\nabla f_k,\qquad \|p\|=1$$

$p^T\nabla f_k=\|p\|\|\nabla f_k\|\cos\beta=\|\nabla f_k\|\cos\beta$，当$\cos\beta=-1$时，取最小，此时$p=-\nabla f_k/\|\nabla f_k\|$

- 因此，$-\nabla f_k$是f的最速下降方向

最速下降方向

- 在线搜索方法里面，最速下降方向$p_k = -\nabla f_k$是最常见的选择
- 使用最速下降方向$p_k = -\nabla f_k$的方法叫作最速下降法
- 最速下降法的优点之一在于只需要计算一阶导数
- 然而，在求解复杂优化问题时，该方法求解速度可能极慢
- 事实上，任意使得$p^T \nabla f_k < 0$的方向，只要搜索步长足够小，就能使得f下降，根据泰勒定理

$$f(x_k + \alpha p_k) = f_k + \alpha p_k^T \nabla f_k + O(\alpha^2)$$

$p_k^T \nabla f_k < 0$，当步长α足够小时，$f(x_k + \alpha p_k) < f_k$

牛顿方向

- 将$f(x_k+p)$使用泰勒级数展开，其二阶近似为

$$f(x_k+p) \approx f_k + p^T \nabla f_k + \frac{1}{2} p^T \nabla^2 f_k p \equiv m_k(p)$$

- 通过最小化$m(k)$可以得到牛顿方向

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

- 当$\nabla f_k \neq 0$且$\nabla^2 f_k$正定的时候，p_k^N是下降方向，因为

$$\nabla f_k^T p_k^N = -(p_k^N)^T \nabla^2 f_k p_k^N \leq -\sigma_k \left\| p_k^N \right\|^2, \exists \sigma_k > 0$$

- 当$f(x_k+p)$和$m_k(p)$越接近时，牛顿方向越可靠

- 对比$f(x_k+\alpha p)=f_k+\alpha p^T\nabla f_k+\frac{\alpha^2}{2}p^T\nabla^2 f(x_k+\theta p)p$和$f(x_k+p)\approx f_k+p^T\nabla f_k+\frac{1}{2}p^T\nabla^2 f_k p$，可知$\nabla^2 f(x_k+\theta p)$被$\nabla^2 f_k$替代，如果$\nabla^2 f_k$充分平滑，扰动误差是$O(\|p\|^3)$，因此$\|p\|$越小，$f(x_k+p)$和$m_k(p)$越接近，牛顿方向越可靠

- 与最速下降方法不同，牛顿法的搜索步长一般为1，只有当f下降不充分时，才对搜索步长进行调整

牛顿方向

- 当$\nabla^2 f_k$非正定时，$-(\nabla^2 f_k)^{-1}$可能不存在，或者$\nabla f_k^T p_k^N = -(p_k^N)^T \nabla^2 f_k p_k^N > 0$。在这些情况下，为了保证下降方向，可以进行正则化操作$p_k^N = -(\nabla^2 f_k + \sigma I)^{-1} \nabla f_k$
- 牛顿法一般二次收敛，达到解的邻域后，通常只需几次迭代即可收敛到高精度
- Hessian矩阵的显式计算有时可能是一个繁琐、容易出错且昂贵的过程
- 可以通过有限差分方法或者自动微分方法计算Hessian矩阵，避免手动计算

- 拟牛顿搜索方向是牛顿搜索方向的替代，其不需要计算Hessian 矩阵，但具有超线性（superlinear）收敛速度
- 使用B_k近似Hessian矩阵$\nabla^2 f_k$，拟牛顿方向如下：$p_k = -B_k^{-1}\nabla f_k$
- 在每个步骤之后使用额外知识更新B_k，事实上∇f_k可以提供有关 f 沿搜索方向的二阶导数的信息
- 通过泰勒定理$\nabla f(x+p) = \nabla f(x) + \int_0^1 \nabla^2 f(x+tp)pd_t$，可得$\nabla f(x+p) = \nabla f(x) + \nabla^2 f(x+p)p + \int_0^1 \nabla^2 f(x+tp)p - \nabla^2 f(x+p)pd_t \Rightarrow \nabla f_{k+1} = \nabla f_k + \nabla^2 f_{k+1}(x_{k+1}-x_k) + o(\|x_{k+1}-x_k\|)$

- 假设$\nabla^2 f_{k+1}$正定，上式可以近似为以下割线方程

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

- $\nabla^2 f_{k+1}(x_{k+1} - x_k) \approx \nabla f_{k+1} - \nabla f_k$，根据此式更新Hessian近似矩阵$B_{k+1}$

$$B_{k+1} s_k = y_k, \text{其中} s_k = x_{k+1} - x_k, y_k = \nabla f_{k+1} - \nabla f_k$$

- 此外，真正的Hessian矩阵是对称的，因此B_{k+1}通常满足对称约束，此外B_{k+1}和B_k的差异不能变化很大，也就是说$(B_{k+1} - B_k)$满足低秩约束

- symmetric-rank-one (SR1)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

$$B_{k+1}=B_k-\frac{B_ks_ks_k^TB_k}{s_k^TB_ks_k}+\frac{y_ky_k^T}{y_k^Ts_k}$$

- 两种更新方式都满足割线方程，且B_{k+1}对称

- 只要B_0正定，并且$s_k^Ty_k>0$，则BFGS更新中的B_{k+1}是正定的

- 最后可以通过求解线性方程组$p_k=-B_k^{-1}\nabla f_k$得到p_k

拟牛顿方法更新公式

将$H_k = -B_k^{-1}$，可以得到相应的更新公式

- symmetric-rank-one (SR1)公式：

$$H_{k+1} = H_k + \frac{(s_k - H_k y_k)(s_k - H_k y_k)^T}{(s_k - H_k y_k)^T y_k}$$

- BFGS 公式

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

- 最后可以通过$p_k = H_k \nabla f_k$得到p_k

- 在大多数线搜索方法中，p_k代表f的下降方向，$\nabla f_k^T p_k < 0$
- 最速下降方向，牛顿方向，拟牛顿方向可以总结如下

$$p_k = -B_k^{-1} \nabla f_k,\text{其中}B_k\text{是对称非奇异矩阵}$$

- 当B_k正定时，$\nabla f_k^T p_k = -\nabla f_k^T B_k^{-1} \nabla f_k < 0$，因此$p_k$是下降方向
- 在最速下降方法中，$B_k = I$
- 在牛顿法中，$B_k = \nabla^2 f(x_k)$
- 在拟牛顿法中，$B_k \approx \nabla^2 f(x_k)$，通过公式迭代更新

- 非线性共轭梯度更新

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

其中β_k是保证p_k和p_{k-1}保持共轭的一个标量，共轭是二次函数的最小化中一个重要的概念

- 非线性共轭方向比最速下降方向有效得多，并且计算起来几乎很简单。 这些方法没有达到牛顿法的快速收敛速度，但它们具有不需要存储矩阵的优点

主要内容

1. 概述
2. 搜索方向
3. 搜索步长
4. 全局收敛性
5. 收敛速率
6. 小结

- 搜索步长α_k的选择涉及到两个冲突的目标：1）能够使f充分下降；2） α_k的计算时间不能太长
- 可以通过最小化以下公式选择α_k

$$\phi(\alpha) = f(x_k + \alpha p_k)$$

事实上，这将导致$p_k^T \nabla f(x_k + \alpha_k p_k) = 0$

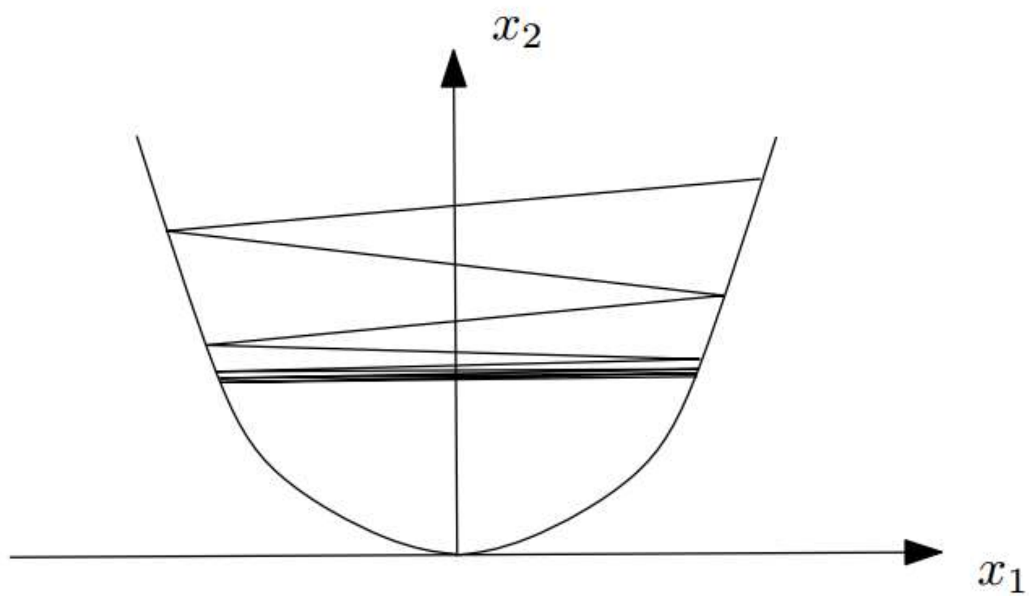
- 然而，最小化$\phi(\alpha)$并不容易，即使定位局部最优解也需要多次调用f和∇f

- 一些线搜索算法会尝试一系列 α 候选值，当满足某些条件时停止，并接受其中一个值

- 一般来说分为两步：在bracketing 阶段，寻找包含理想步长的区间；在bisection 或者 interpolation阶段，在区间内计算合适的步长

- 接下来，将讨论线搜索算法的各种终止条件，表明有效步长不需要位于单变量函数 $\phi(\alpha)$ 的最小值附近。

- 一般来说，满足$f(x_k + \alpha p_k) < f(x_k)$的$\alpha$都是可以考虑的



- 这一个单独的条件可能不能收敛到最优值x^*
- 因此为了保证收敛，需要强加一些充分下降条件（sufficient decrease condition）

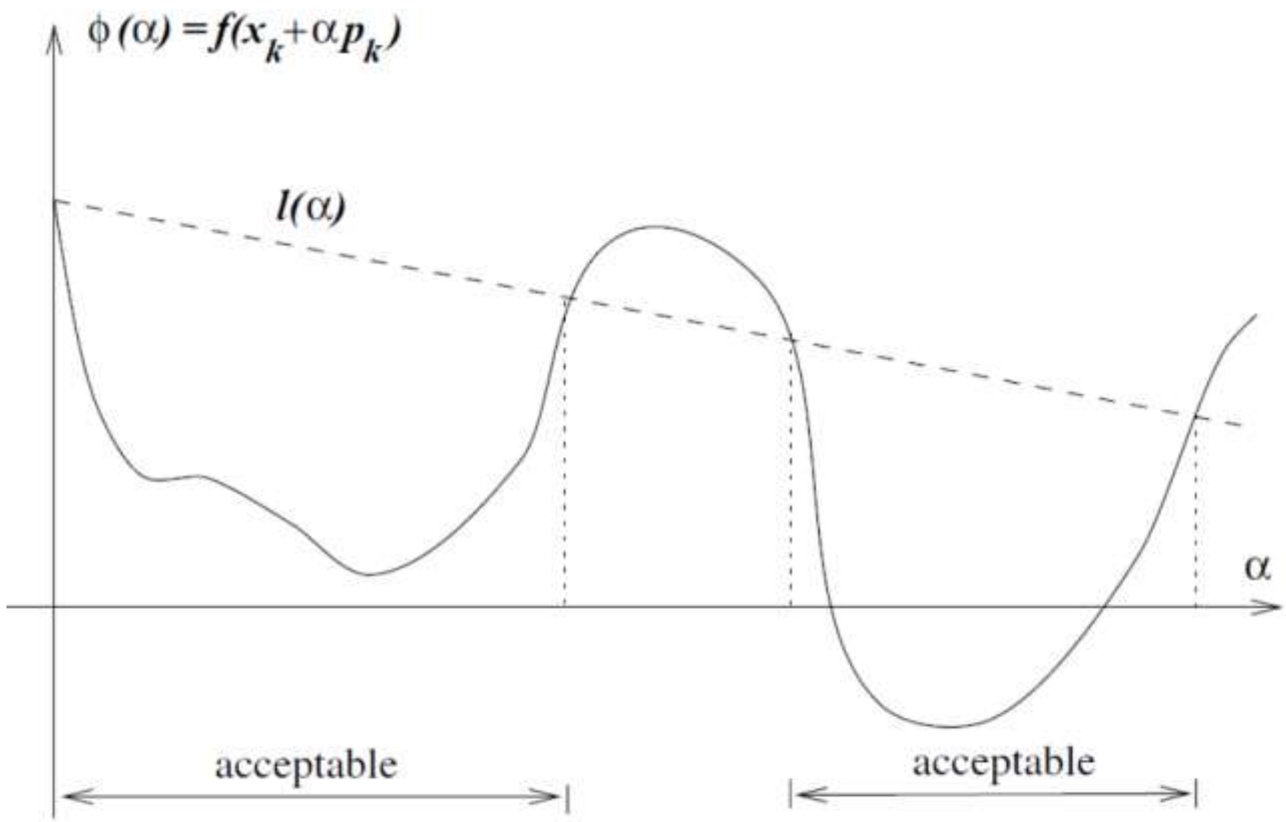
- Armijo条件

$$f(x_k + \alpha p_k) \leq f(x_k) + c_1 \alpha p_k^T \nabla f_k,\ c_1 \in (0,1)$$

- 一般来说p_k与∇f方向相反， $p_k^T \nabla f < 0$；此外$\alpha > 0$，因此该条件比原条件更强
- c_1不宜太大，否则难以找到满足该条件的α；通常来说取较小的值
- 整体来说， f的下降值与α和$p_k^T \nabla f$的大小成正比

Armijo condition

- 令$\phi(\alpha) = f(x_k + \alpha p_k)$，$l(\alpha) = f(x_k) + c_1 \alpha p_k^T \nabla f_k$



Curvature Condition

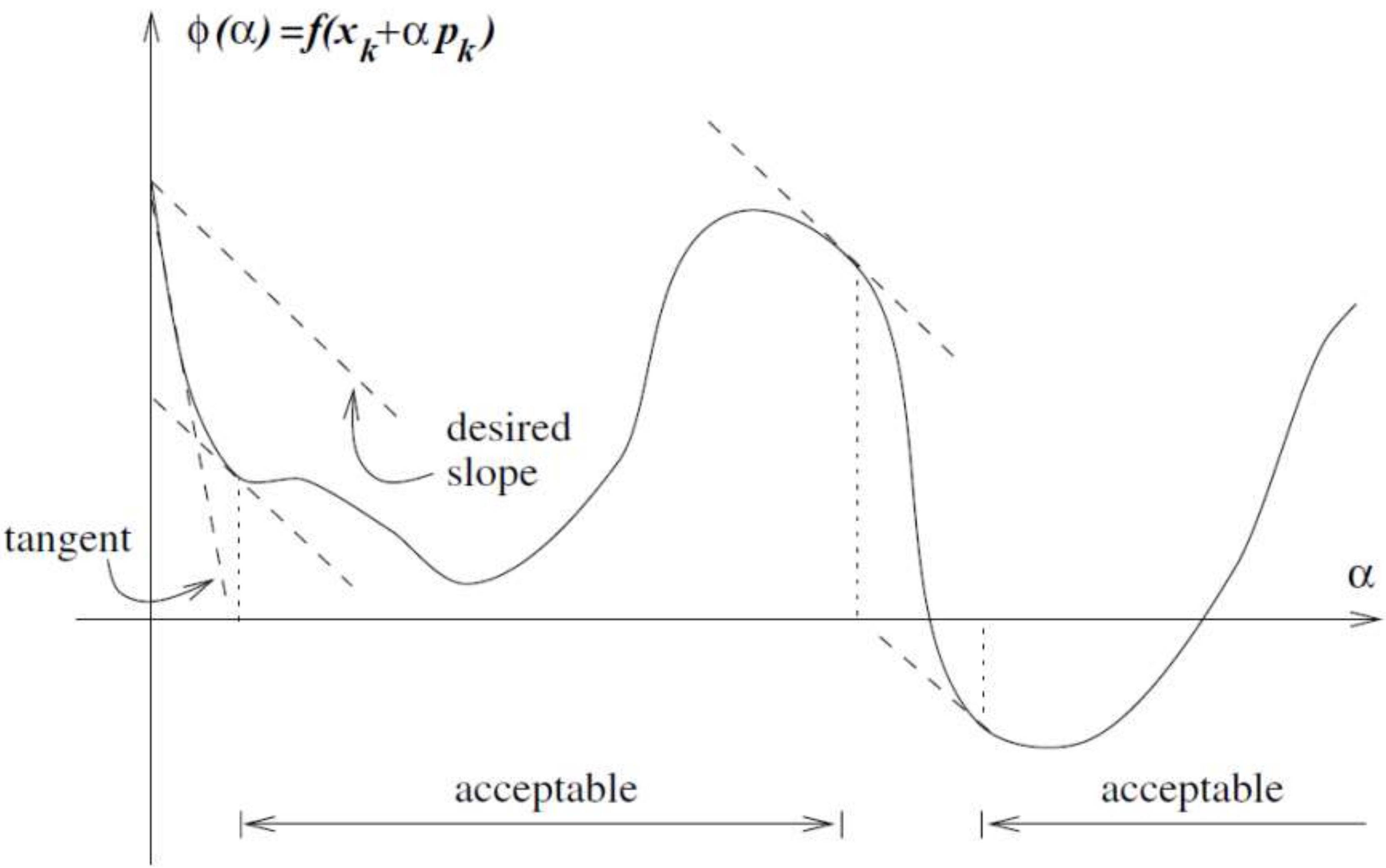
- 在Armijo condition中，较小的α都能满足该条件，太小的步长可能不利于算法收敛

- 附加一个曲率条件，避免太小的α

$$\nabla f(x_k + \alpha p_k)^T p_k \geq c_2 \nabla f_k^T p_k, c_2 \in (c_1, 1)$$

- $\phi'(\alpha) = \nabla f(x_k + \alpha p_k)^T p_k, \phi'(0) = \nabla f_k^T p_k$，因此有 $\phi'(\alpha) \geq c_2 \phi'(0)$

Curvature Condition



Curvature Condition

$$\nabla f(x_k+\alpha p_k)^T p_k \geq c_2 \nabla f_k^T p_k, c_2 \in (c_1, 1)$$

- 当该条件不满足时，说明$\nabla f(x_k+\alpha p_k)^T p_k$是比较大的负值，沿着$p_k$方向$f$能有较大的下降，因此可以继续寻找更合适的$\alpha$
- 当该条件满足时，说明$\nabla f(x_k+\alpha p_k)^T p_k$是轻微的负值或者正值，说明搜索已经到了极小值附近
- 当使用牛顿法和拟牛顿法时，c_2一般设置为0.9；当用共轭梯度法时，c_2一般设置为0.1

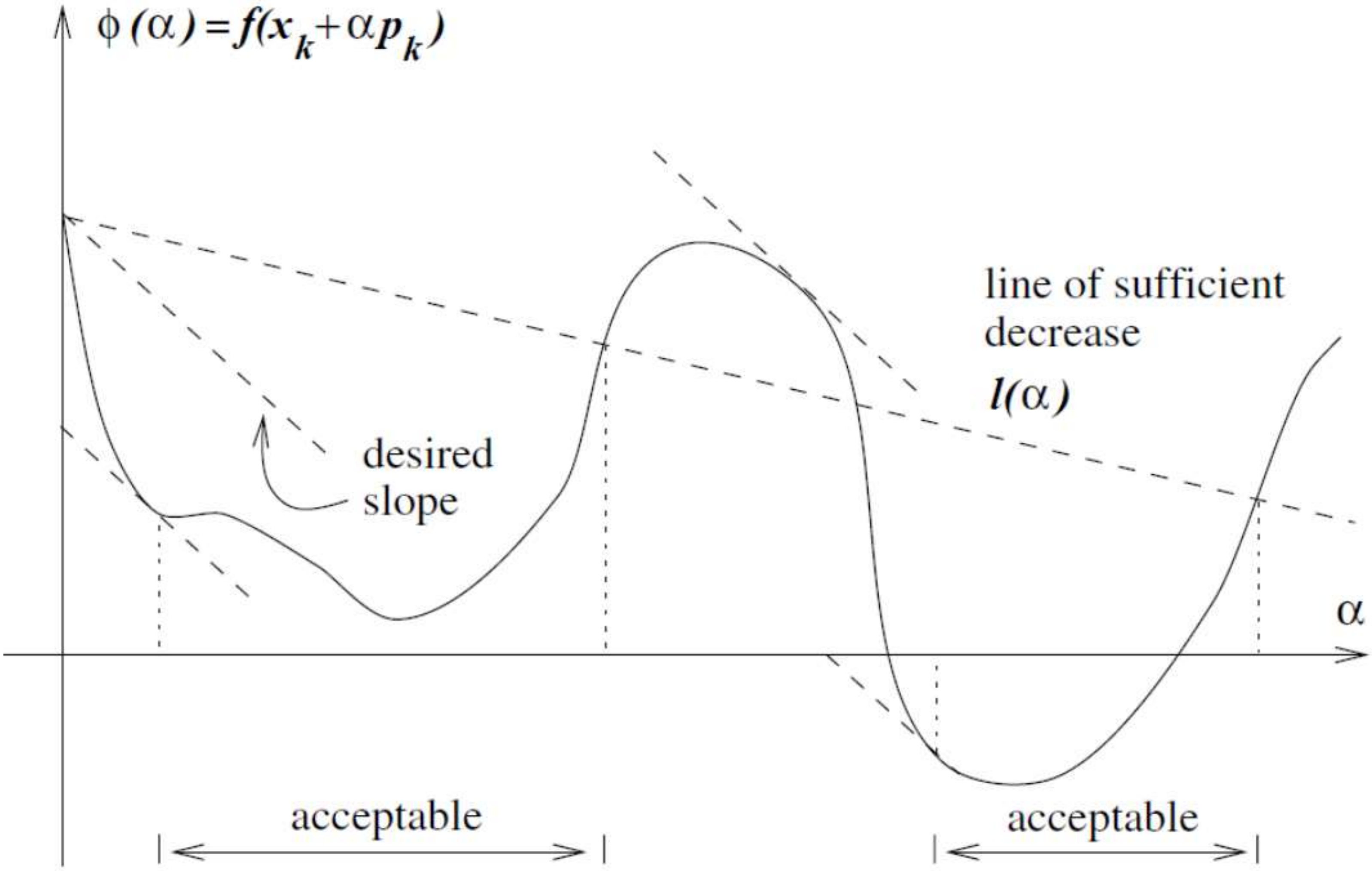
Wolfe Conditions

$$f(x_k + \alpha p_k) \le f(x_k) + c_1 \alpha p_k^T \nabla f_k, c_1 \in (0,1)$$

$$\nabla f(x_k + \alpha p_k)^T p_k \ge c_2 \nabla f_k^T p_k, c_2 \in (c_1, 1)$$

- Wolfe Conditions在广义上是尺度不变的：将目标函数乘以常数或对变量进行仿射变化不会改变它们。其可用于大多数线搜索方法，并且在拟牛顿方法的实现中尤其重要

Wolfe Conditions



- 原Wolfe Conditions中α可能离$\phi(\alpha)$极小值太远了，为此提出了强Wolfe Conditions

$$f(x_k+\alpha p_k)\le f(x_k)+c_1\alpha p_k^T\nabla f_k, c_1\in(0,1)$$

$$\left|\nabla f(x_k+\alpha p_k)^T p_k\right|\le c_2\left|\nabla f_k^T p_k\right|, c_2\in(c_1,1)$$

- 保证了$\nabla f(x_k+\alpha p_k)^T p_k$不会是一个很大的正值，也就是越过极小值太大距离

定理：假设$f:\mathcal{R}^n \to \mathcal{R}$连续可导， p_k是函数在x_k处的一个下降方向，并且f沿射线$\{x_k + \alpha p_k | \alpha > 0\}$方向有下界，$0 < c_1 < c_2 < 1$，则存在满足Wolfe Conditions和强Wolfe Conditions的步长区间

证明：$\phi(\alpha) = f(x_k + \alpha p_k)$，$l(\alpha) = f(x_k) + c_1 \alpha p_k^T \nabla f_k$，$p_k$是函数下降方向，则$l(\alpha)$是减函数；此外，$\phi(\alpha)$有下界，因此$\phi(\alpha)$和$l(\alpha)$必有交点，令$\alpha_{\min}$为最小的交点

$$f(x_k + \alpha_{\min} p_k) = f(x_k) + c_1 \alpha_{\min} p_k^T \nabla f_k$$

根据中值定理，存在$\alpha' \in (0, \alpha_{\min})$满足

$$f(x_k + \alpha_{\min} p_k) - f(x_k) = \alpha_{\min} \nabla f(x_k + \alpha' p_k)^T p_k$$

证明：$f(x_k+\alpha_{\min}p_k)=f(x_k)+c_1\alpha_{\min}p_k^T\nabla f_k$

$$f(x_k+\alpha_{\min}p_k)-f(x_k)=\alpha_{\min}\nabla f(x_k+\alpha' p_k)^T p_k$$

联立两式可以得到

$$\nabla f(x_k+\alpha' p_k)^T p_k=c_1p_k^T\nabla f_k$$

因为$p_k^T\nabla f_k\le 0$并且$0<c_1<c_2$，则有

$$\nabla f(x_k+\alpha' p_k)^T p_k=c_1p_k^T\nabla f_k\ge c_2p_k^T\nabla f_k$$

因为$\alpha_{\min}$是第一个交点和$\alpha'\in(0,\alpha_{\min})$，所以有

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

因此α'满足Wolfe conditions，此外$\nabla f(x_k+\alpha' p_k)^T p_k\le 0$，$\alpha'$满足强Wolfe conditions

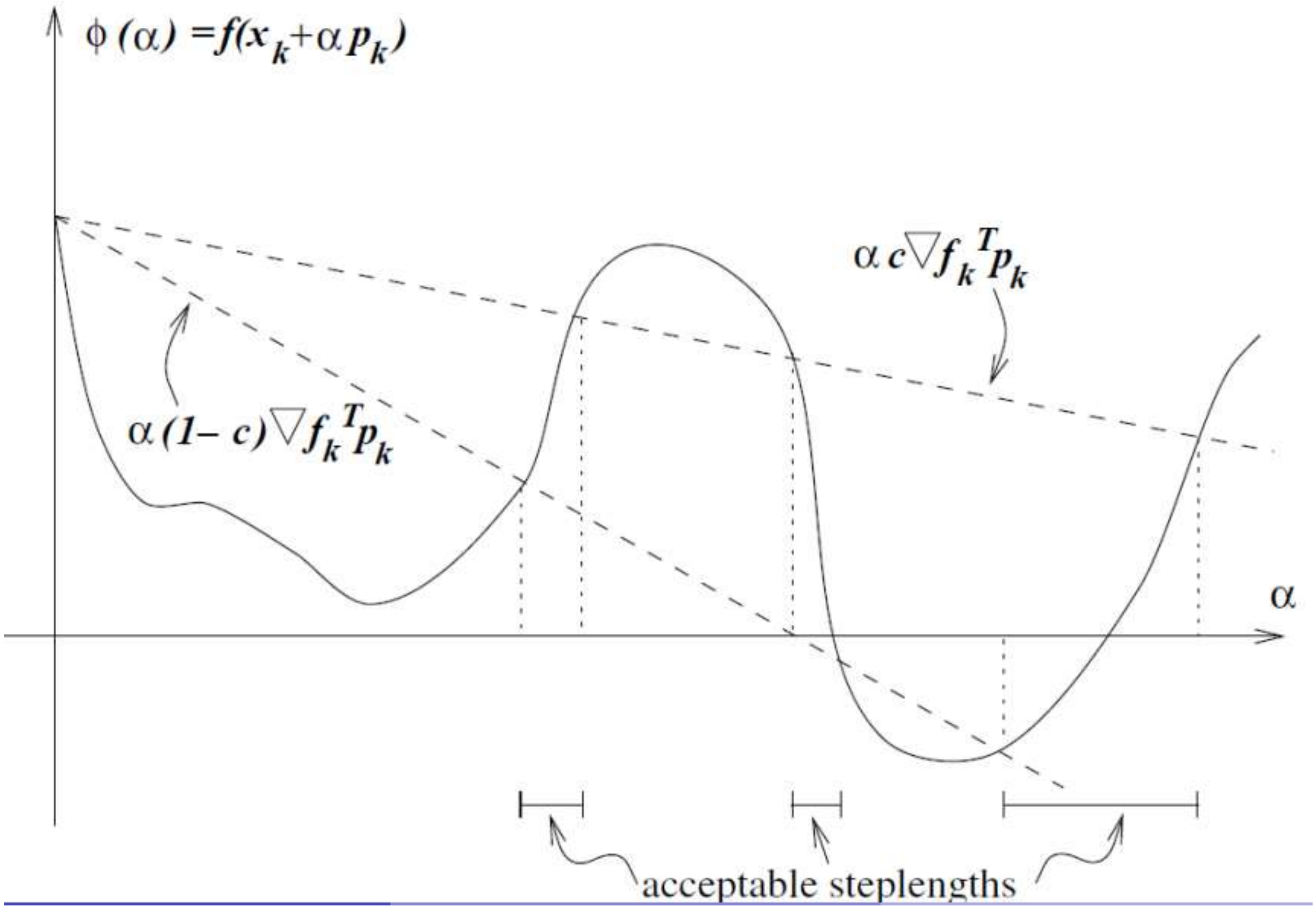
- Goldstein Conditions保证α使得函数充分下降的同时又不能太小

$$f(x_k + \alpha p_k) \le f(x_k) + c\alpha p_k^T \nabla f_k$$

$$f(x_k) + (1-c)\alpha p_k^T \nabla f_k \le f(x_k + \alpha p_k), c \in (0,1/2)$$

- 第一个公式是Armijo condition，第二个公式把一些比较小的α排除在外
- 可能把$\phi(\alpha)$的极小值排除在外， Goldstein Conditions和Wolfe Conditions条件有极大相似之处
- Goldstein Conditions在牛顿法中较常用，但是不太适合拟牛顿方法

Goldstein Conditions



- 选择合适的$\alpha > 0, c \in (0,1), \gamma \in (0,1)$
- 重复以下步骤直到$f(x_k + \alpha p_k) \leq f(x_k) + c\alpha p_k^T \nabla f_k$
- $\alpha \leftarrow \gamma\alpha$
- 最终设置$\alpha_k = \alpha$

- 在牛顿法和拟牛顿法中，α的初始值一般设置为1，当然在其他方法中略有不同
- 因为α逐渐变小最终都会满足终止条件
- 收敛因子γ在迭代过程中可以变化
- 回溯能够保证所选择的步长 α是某个固定值（初始值），或者在满足充分下降条件情况下不至于太小

- 基于以上知识，接下来介绍如何设置求得$\phi(\alpha)$的极小值，或者找到合适的α以满足前面介绍的一些条件
- 对于一些函数，可以用解析的方式求取$\phi(\alpha)$的极小值，例如$f(x)=\frac{1}{2}x^TQx-b^Tx$，通过直接求$\phi(\alpha)$极小可以得到$\alpha_k=(\nabla f_k^T p^k)/(p^kQp^k)$
- 对于无法解析求解的情况下，需要使用迭代算法。选择初始α_0，进行迭代直到达到终止条件（如Wolfe条件）或者α不存在
  - bracketing phase：找到包含合适步长的范围$[\bar{a},\bar{b}]$
  - selection phase : 定位最终步长

步长选择算法

- selection phase通常缩小步长区间，对之前步骤中收集的一些函数和导数信息进行插值，估计极小值的位置
- 以Armijo条件$\phi(\alpha_k) \le \phi(0) + c_1\alpha_k\phi'(0)$为例，如果$\alpha_0$满足此条件，则终止搜索，否则在$[0, \alpha_0]$之间搜索$\alpha_1$
- 利用已知信息$\phi(0), \phi'(0)$和$\phi(\alpha_0)$进行二次插值

$$\phi_q(\alpha) = \left(\frac{\phi(\alpha_0) - \phi(0) - \alpha_0\phi'(0)}{\alpha_0^2}\right)\alpha^2 + \phi'(0)\alpha + \phi(0)$$

- 取α_1使得$\phi_q(\alpha)$取极小值

$$\alpha_1 = -\frac{\phi'(0)\alpha_0^2}{2(\phi(\alpha_0) - \phi(0) - \alpha_0\phi'(0))}$$

步长选择算法

- 如果α_1满足终止条件，则终止搜索，否则在$[0, \alpha_1]$之间搜索α_2
- 利用已知信息$\phi(0), \phi'(0), \phi(\alpha_0)$和$\phi(\alpha_1)$进行三次插值

$$\phi_c(\alpha) = a\alpha^3 + b\alpha^2 + \phi'(0)\alpha + \phi(0)$$

$$\begin{pmatrix} a \\ b \end{pmatrix} = \frac{1}{\alpha_0^2\alpha_1^2(\alpha_1 - \alpha_0)} \begin{pmatrix} \alpha_0^2 & -\alpha_1^2 \\ -\alpha_0^3 & \alpha_1^3 \end{pmatrix} \begin{pmatrix} \phi(\alpha_1) - \phi(0) - \phi'(0)\alpha_1 \\ \phi(\alpha_0) - \phi(0) - \phi'(0)\alpha_0 \end{pmatrix}$$

- 取α_2使得$\phi_c(\alpha)$取极小值

$$\alpha_2 = -\frac{-b + \sqrt{b^2 - 3a\phi'(0)}}{3a}$$

- 重复使用$\phi(0), \phi'(0), \phi(\alpha_{k-2})$和$\phi(\alpha_{k-1})$进行三次插值直到$\alpha_k$满足终止条件

- 如果导数计算不用耗费过多资源，这些信息也可以用来进行插值
- 三次插值为具有显著曲率变化的函数提供了良好的模型，并且通常能够达到二次收敛速率
- 在使用牛顿法或者拟牛顿法时，初始步长一般设置为1，这将保证步长始终在[0,1]之间，进一步保证了收敛速率
- 对于一些不具有归一化搜索方向的方法，如最速下降法和共轭梯度法，使用有关问题和算法的当前信息估计初始步长则尤为重要

主要内容

1. 概述
2. 搜索方向
3. 搜索步长
4. 全局收敛性
5. 收敛速率
6. 小结

全局收敛性

- 所谓全局收敛性就是$\|\nabla f_k\|$的变化
- 定义：$\cos\theta_k = \dfrac{-\nabla f_k^T p_k}{\|\nabla f_k\| \|p_k\|}$
- Zoutendijk条件：$f:\mathcal{R}^n \to \mathcal{R}$有下界；$x_{k+1} = x_k + \alpha_k p_k$，$p_k$是下降方向，$\alpha_k$满足Wolfe条件；$f$在开邻域$\aleph$连续可微，并且$\mathcal{L} \equiv \{x | f(x) \le f(x_0)\}$，$x_0$是初始点；假设$\nabla f$在$\aleph$上满足Lipschitz连续条件，则有

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

简要证明：根据Lipschitz连续条件有

$$\|\nabla f_{k+1}-\nabla f_k\|\leq L\|x_{k+1}-x_k\|$$

$$\Longrightarrow (\nabla f_{k+1}-\nabla f_k)^T p_k\leq L\alpha_k\|p_k\|^2$$

根据曲率条件有

$$\nabla f_{k+1}^T p_k\geq c_2\nabla f_k^T p_k$$

结合两个不等式有

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

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

根据充分下降条件有

$$f_{k+1}\leq f_k+c_1\alpha_k\nabla f_k^T p_k$$

全局收敛性

简要证明：

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

$$\Rightarrow f_{k+1} \leq f_k - \frac{c_1(1-c_2)}{L}\cos^2\theta_k \|\nabla f_k\|^2$$

$$\Rightarrow f_{k+1} \leq f(x_0) - \frac{c_1(1-c_2)}{L}\sum_{i=0}^{k}\cos^2\theta_i \|\nabla f_i\|^2$$

因为f有下界，所以

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

全局收敛性

- 当使用 Goldstein 条件或强 Wolfe 条件代替 Wolfe 条件时，与该定理类似的结果成立
- Zoutendijk条件表明

$$\cos^2\theta_k\|\nabla f_k\|^2 \to 0$$

- 假如所选择的p_k保证θ_k远离90度，也就是搜索方向远离梯度正交方向，则存在一个正数δ使得满足

$$\cos\theta_k \geq \delta > 0$$

因此有$\lim_{k\to\infty}\|\nabla f_k\| = 0$

- 换句话说，只要搜索方向永远不会太接近梯度的正交方向，我们就可以确定梯度范数收敛于零。

全局收敛性

- 最速下降法中， $p_k=-\nabla f_k$，因此有$\cos\theta_k=1$，所以 $\lim_{k\to\infty}\|\nabla f_k\|=0$

- 对于牛顿法， $p_k=-B_k^{-1}\nabla f_k$，假设矩阵 B_k 是具有一致有界条件数的正定矩阵。 也就是说，存在一个常数 M，使得

$$\|B_k\|\left\|B_k^{-1}\right\|\leq M,\forall\, k$$

根据定义有

$$\cos\theta_k=\frac{-\nabla f_k^T p_k}{\|\nabla f_k\|\|p_k\|}=\frac{p_k^T B_k p_k}{\|p_k\|\|B_k p_k\|}$$

所以$\lim_{k\to\infty}\|\nabla f_k\|=0$

全局收敛性

$$\cos\theta_k = \frac{-\nabla f_k^T p_k}{\|\nabla f_k\|\|p_k\|} = \frac{p_k^T B_k p_k}{\|p_k\|\|B_k p_k\|} = \frac{p_k^T {B_k}^{1/2} {B_k}^{1/2} p_k}{\|p_k\|\|B_k p_k\|}$$

$$\Longrightarrow \ \cos\theta_k = \frac{\left\|{B_k}^{1/2} p_k\right\|^2}{\|p_k\|\|B_k p_k\|} \ge \frac{\|p_k\|^2}{\left\|{B_k}^{-\frac{1}{2}}\right\|^2 \|p_k\|\|B_k p_k\|}$$

$$\Longrightarrow \ \cos\theta_k \ge \frac{\|p_k\|^2}{\left\|{B_k}^{-1}\right\|\|p_k\|\|B_k p_k\|} \ge \frac{\|p_k\|^2}{\left\|{B_k}^{-1}\right\|\|B_k\|\|p_k\|^2}$$

$$\Longrightarrow \ \cos\theta_k \ge \frac{1}{\left\|{B_k}^{-1}\right\|\|B_k\|} = \frac{1}{M}$$

所以 $\lim\limits_{k\to\infty} \|\nabla f_k\| = 0$

全局收敛性

- 我们使用术语“全局收敛”来指代满足$\lim_{k\to\infty}\|\nabla f_k\|=0$的算法属性
- 对于以上提及的线搜索方法，$\lim_{k\to\infty}\|\nabla f_k\|=0$可以获得的最强的全局收敛结果。我们不能保证该方法收敛到最小值，而只能保证其被稳定点吸引
- 只有通过对搜索方向 p_k 强加另外的要求（例如，通过引入来自 Hessian 矩阵的负曲率信息），才能保证收敛到局部最小值
- 事实上，满足以下两个条件的算法都具有全局收敛性：1）每次迭代都产生一个下降方向；2）每m次迭代相当于一次最速下降，并且步长满足wolfe条件或者Goldstein条件
- 偶尔的最速下降步骤可能不会取得太大进展，但其至少保证了整体全局收敛

主要内容

收敛速率

- 在设计线搜索算法时，只要保证搜索方向p_k 不与梯度方向正交，或者定期采取最速下降步骤，就能保证算法全局收敛；但是但这样可能会导致算法收敛速度较低，例如最速下降法
- 此外，实现快速收敛的算法策略有时会与全局收敛的要求发生冲突。 例如，纯牛顿迭代在起始点接近极小值时会快速收敛，但当起始点远离极小值时，搜索步骤甚至可能不是下降方向
- 设计具有良好的全局收敛算法，并保证算法的收敛速度，是设计线搜索算法需要考虑的

定理：$f:\mathcal{R}^n \to \mathcal{R}$二阶连续可微，线搜索算法每次使用最速下降法产生搜索方向，使用精确线搜索产生搜索步长，算法收敛到x^*，$\nabla^2 f(x^*)$正定；定义γ满足

$$\gamma \in (\frac{\lambda_n - \lambda_1}{\lambda_n + \lambda_1}, 1)$$

其中，$\lambda_1 \leq \lambda_2, \dots, \leq \lambda_n$是$\nabla^2 f(x^*)$的特征值，则对于所有足够大的$k$有

$$f(x_{k+1}) - f(x^*) \leq \gamma^2 (f(x_k) - f(x^*))$$

证明过程见《numerical optimization》43-44页

- 一般来说，如果使用非精确的线搜索，收敛速度不会提高
- 即使 Hessian 条件数良好，最速下降法也可能给出令人无法接受的缓慢收敛速度
- 例如，假如条件数是$\frac{\lambda_n}{\lambda_1}=800$，$f(x_1)=1$，$f(x^*)=0$，定理表明，使用精确线搜索的最速下降法迭代一千次后，函数值仍约为 0.08左右

定理：$f:\mathcal{R}^n\rightarrow\mathcal{R}$二阶可微，在满足second-order sufficient conditions的解x^*的邻域，$\nabla^2 f(x)$满足Lipschitz条件；$x_{k+1}=x_k+p_k^N, p_k^N=-\nabla^2 f_k^{-1}\nabla f_k$；则有

1）如果起始点x_0充分接近x^*，则序列$\{x_k\}$收敛到x^*

2）$\{x_k\}$满足二次收敛速率

3）$\{\|\nabla f_k\|\}$二次收敛到0

牛顿法收敛速率

简要证明： $x_{k+1}-x^* = x_k + p_k^N - x^* = x_k - x^* + p_k^N$

$$\Rightarrow x_{k+1} - x^* = x_k - x^* - \nabla^2 f_k^{-1} \nabla f_k$$

$$\Rightarrow x_{k+1} - x^* = x_k - x^* - \nabla^2 f_k^{-1} (\nabla f_k - \nabla f_*)$$

$$\Rightarrow x_{k+1} - x^* = \nabla^2 f_k^{-1} [\nabla^2 f_k (x_k - x^*) - (\nabla f_k - \nabla f_*)]$$

根据泰勒定理有

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

结合两式可以得到

$$\|x_{k+1} - x^*\|$$
$$= \left\| \nabla^2 f_k^{-1} \int_0^1 [\nabla^2 f_k - \nabla^2 f\left(x_k + t(x_k - x^*)\right)](x_k - x^*) dt \right\|$$

牛顿法收敛速率

$$\le \left\|\nabla^2 f_k^{-1}\right\| \int_0^1 \left\|\nabla^2 f_k - \nabla^2 f\left(x_k + t(x_k - x^*)\right)\right\| \|(x_k - x^*)\| dt$$

$$\le \left\|\nabla^2 f_k^{-1}\right\| \int_0^1 Lt \|(x_k - x^*)\|^2 dt = \frac{L}{2}\left\|\nabla^2 f_k^{-1}\right\| \|(x_k - x^*)\|^2$$

（$\nabla^2 f(x)$满足Lipschitz条件）

因为$\nabla^2 f(x^*)$正定，所以在x^*存在领域使得$\left\|\nabla^2 f_k^{-1}\right\| \le 2\|\nabla^2 f_*^{-1}\|$，所以有

$$\|x_{k+1} - x^*\| \le L\|\nabla^2 f_*^{-1}\|\|(x_k - x^*)\|^2$$

因此$\{x_k\}$二次收敛到x^*

$\{\|\nabla f_k\|\}$二次收敛到0证明如下：

利用$x_{k+1}-x_k=p_k^N, \nabla f_k+\nabla^2 f_k p_k^N=0$可得

$$\|\nabla f_{k+1}\|=\left\|\nabla f_{k+1}-\nabla f_k-\nabla^2 f_k p_k^N\right\|$$

$$=\left\|\int_0^1 \nabla^2 f\left(x_k+tp_k^N\right)p_k^N dt-\nabla^2 f_k p_k^N\right\|$$

$$\leq \int_0^1 \left\|\nabla^2 f_k-\nabla^2 f\left(x_k+tp_k^N\right)\right\| \left\|p_k^N\right\| dt$$

$$\leq \frac{L}{2}\left\|p_k^N\right\|^2=\frac{L}{2}\left\|\nabla^2 f_k^{-1}\right\|^2 \|\nabla f_k\|^2$$

$$\leq 2L\|\nabla^2 f_*^{-1}\|^2 \|\nabla f_k\|^2$$

因此$\{\|\nabla f_k\|\}$二次收敛到0

拟牛顿法收敛速率

定理：$f:\mathcal{R}^n\to\mathcal{R}$二阶连续可微； $x_{k+1}=x_k+p_k$,$p_k=-B_k^{-1}\nabla f_k$； 对称正定矩阵B_k 使用拟牛顿公式更新，假设$\{x_k\}$收敛到x^*， $\nabla f(x^*)=0$， $\nabla^2 f(x^*)$正定，则$\{x_k\}$超线性收敛，当且仅当

$$\lim_{k\to\infty}\frac{\|(B_k-\nabla^2 f(x^*))p_k\|}{\|p_k\|}=0$$

证明过程见《numerical optimization》47页

主要内容

小结

- 在设计线搜索算法时，只要保证搜索方向p_k 不与梯度方向正交，或者定期采取最速下降步骤，就能保证算法全局收敛；但是但这样可能会导致算法收敛速度较低，例如最速下降法
- 此外，实现快速收敛的算法策略有时会与全局收敛的要求发生冲突。 例如，纯牛顿迭代在起始点接近极小值时会快速收敛，但当起始点远离极小值时，搜索步骤甚至可能不是下降方向
- 设计具有良好的全局收敛算法，并保证算法的收敛速度，是设计线搜索算法需要考虑的

Thanks for the attentions!

Q&A