一阶KKT条件小结

- 线性可行方向L_S 受优化问题的可行域的表示方式影响
- 切锥 $T(S)$由可行域决定
- 切锥不容易计算，但反映了约束集合的本质
- 线性可行方向容易计算，但是不能反应可行域的本质
- 引入MFCQ或LICQ，可以确保切锥与线性可行方向相等，从而推出KKT条件。
- Fritz-John条件无需任何约束规范性条件。
- KKT条件需要MFCQ或者LICQ保证。当然还有其他形式的规范性条件，我们这里不做介绍。

二阶最优条件

再次考虑约束问题(52)，即

$$\begin{array}{ll} \min & f(\mathbf{x}) \\ \text{s.t.} & g_i(\mathbf{x}) \geq 0, i = 1, \cdots, m \\ & h_j(\mathbf{x}) = 0, j = 1, \cdots, \ell \end{array}$$

假设x^*是满足KKT条件的点，并且切锥与线性可行方向相等，即$T(x^* \mid S) = L_g \cap L_h$。则$\forall d \in T(x^* \mid S)$

$$d^T \nabla f(x^*) = \sum_{i \in \mathcal{I}(x^*)} \lambda_i^* \nabla g_i(x^*)^T d + \sum_{j=1}^{\ell} \mu_j^* \nabla h(x^*)^T d \geq 0.$$

此时一阶条件无法判断x^*是否是最优值点.

- 若$d^T \nabla f(x^*) = 0$, 则需要利用二阶信息来进一步判断在其可行邻域内的目标函数值.
- 拉格朗日函数在这些方向上的曲率即可用来判断x^*的最优性.
- 这里引入临界锥来精确刻画这些方向.

二阶最优条件：临界锥

定义 (临界锥)
设 (x^*, λ^*, μ^*) 是满足KKT条件的KKT对, 定义临界锥为

$$C(x^*, \lambda^*, \mu^*) = \{d \in T(x^* \mid S) | \nabla g_i(x^*)^T d = 0, \forall i \in \mathcal{I}(x^*) 且 \lambda_i^* > 0\}$$

其中 $F(x^*)$ 为点 x^* 处的线性化可行方向锥.

1. 临界锥是线性化可行方向锥 $L_g \cap L_h$ 的子集.
2. 沿着临界锥中的方向进行优化, 所有等式约束和 $\lambda_i^* > 0$ 对应的不等式约束(此时这些不等式均取等)都会尽量保持不变.
3. 当 $d \in C(x^*, \lambda^*, \mu^*)$ 时, $\forall i = 1, ..., m, j = 1, ..., \ell$ 有 $\lambda_i^* \nabla g_i(x^*)^T d = 0$, $\mu_j^* \nabla h_j(x^*)^T d = 0$, 故

$$d^T \nabla f(x^*) = \sum_{i=1}^{m} \lambda_i^* \nabla g_i(x^*)^T d + \sum_{j=1}^{\ell} \mu_j^* \nabla h(x^*)^T d = 0$$

4. 临界锥定义了依据一阶导数不能判断是否为下降或上升方向的线性化可行方向, 必须使用高阶导数信息加以判断.

二阶最优条件

定理 (二阶最优性条件)
必要性: 假设 x^* 是问题的一个局部最优解, 并且 $T(x^* \mid S) = L_g \cap L_h$ 成立. 令 (x^*, λ^*, μ^*) 满足KKT条件, 那么

$$d^T \nabla^2_{xx} L(x^*, \lambda^*, \mu^*) d \geq 0$$
$$\forall d \in C(x^*, \lambda^*, \mu^*)$$

充分性: 假设在可行点 x^* 处, 存在一个拉格朗日乘子 λ^*, 使得 (x^*, λ^*) 满足KKT条件. 如果

$$d^T \nabla^2_{xx} L(x^*, \lambda^*, \mu^*) d > 0$$
$$\forall d \in C(x^*, \lambda^*, \mu^*), d \neq 0$$

那么 x^* 为问题的一个严格局部极小解.

二阶最优条件

回顾无约束优化问题的二阶最优性条件:

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

必要条件: 若 x^* 是 f 的一个局部极小点, 则 $\nabla f(x^*)=0$, $\nabla^2 f(x^*)\succeq 0$
充分条件: 若 $\nabla f(x^*)=0$, $\nabla^2 f(x^*)\succ 0$, 则 x^* 是 f 的一个局部极小点
约束优化问题的二阶最优性条件也要求某种“正定性”, 但只需要考虑临界锥 $C(x^*,\lambda^*,\mu^*)$ 中的向量而无需考虑全空间的向量.
有些教材中将其称为“投影半正定性”.

二阶最优条件:例子

$$\min\ x_1^2 + x_2^2$$
$$\text{s.t.}\ \frac{x_1^2}{4} + x_2^2 - 1 = 0$$

其拉格朗日函数为

$$L(x, \lambda) = x_1^2 + x_2^2 + \lambda(\frac{x_1^2}{4} + x_2^2 - 1)$$

该问题可行域在任意一点 $x = (x_1, x_2)^T$ 处的线性化可行方向锥为

$$L_S(x) = \{(d_1, d_2) | \frac{x_1}{4} d_1 + x_2 d_2 = 0\}$$

因为只有一个等式约束且其对应函数的梯度非零, 故有LICQ成立, 于是

$$L_S(x) = T(x \mid S)$$

若 (x, λ) 为KKT对, 由于无不等式约束, 故

$$C(x, \lambda) = F(x)$$

可以计算出其4 个KKT对

$$(x^T, \lambda) = (2, 0, -4), (-2, 0, -4), (0, 1, -1), (0, -1, -1)$$

二阶最优条件:例子

考虑第一个KKT对 $y = (2, 0, -4)^T$, 计算可得

$$\nabla^2_{xx} L(y) = \begin{bmatrix} 0 & 0 \\ 0 & -6 \end{bmatrix}$$

$$C(y) = \{(d_1, d_2) | d_1 = 0\}$$

取 $d = (0, 1)$, 则

$$d^T \nabla^2_{xx} L(y) d = -6 < 0$$

因此 y 不是局部最优点. 类似地, 对第三个KKT对 $z = (0, 1, -1)$,

$$\nabla^2_{xx} L(z) = \begin{bmatrix} \frac{3}{2} & 0 \\ 0 & 0 \end{bmatrix}$$

$$C(z) = \{(d_1, d_2) | d_2 = 0\}$$

对于任意的 $d = (d_1, 0)$ 且 $d_1 \neq 0$,

$$d^T \nabla^2_{xx} L(z) d = \frac{3}{2} d_1^2 > 0$$

因此, z 为一个严格局部最优点.

理论上可以用最优性条件求“非线性规划问题”的最优解，但在实践中并不切实可行。

在求解最优化问题时最常用的计算方法是“迭代下降算法”。

闭映射

算法映射：算法$\mathcal{A}$是定义在空间$\mathbf{X}$上的**点到集**的映射，即对每一点$\mathbf{x}^{(k)} \in \mathbf{X}$, 经算法$\mathcal{A}$作用后产生一个点集$\mathcal{A}(\mathbf{x}^{(k)}) \subset \mathbf{X}$, 任意选择一个点$\mathbf{x}^{(k+1)} \in \mathcal{A}(\mathbf{x}^{(k)})$作为$\mathbf{x}^{(k)}$ 的后续点。
引入所谓算法的闭性，其实质是点到点映射的连续性的推广。

设$\mathbf{X}$和$\mathbf{Y}$分别是空间$\mathbb{E}^p$和$\mathbb{E}^q$中的非空闭集， $\mathcal{A} : \mathbf{X} \longrightarrow \mathbf{Y}$为点到集的映射。如果$\mathbf{x}^{(k)} \in \mathbf{X}, \mathbf{x}^{(k)} \to \mathbf{x}, \mathbf{y}^{(k)} \in \mathcal{A}(\mathbf{x}^{(k)}), \mathbf{y}^{(k)} \to \mathbf{y}$ 蕴涵着$\mathbf{y} \in \mathcal{A}(\mathbf{x})$, 则称映射$\mathcal{A}$在$\mathbf{x} \in \mathbf{X}$处是闭的。

解集合

为研究算法的收敛性，首先要明确解集合的概念。

在许多情况下，要使算法产生的点列收敛于全局最优解是极为困难的。因此，一般把满足某些条件的点集定义为解集合。当迭代点属于这个集合时，就停止迭代。

例如，在无约束最优化问题中，可以定义解集合为

$$\Omega = \{\bar{\mathbf{x}} \mid \|\nabla f(\bar{\mathbf{x}})\| = 0\},$$

在约束最优化问题中，解集合取为

$$\Omega = \{\bar{\mathbf{x}} \mid \bar{\mathbf{x}}\text{是KKT点}\}.$$

下降函数

一般地，下降算法总是与某个函数在迭代过程中函数值的减小联系在一起，因此需要给出下降函数的概念。

设$\Omega \subset \mathbf{X}$为解集合，$\mathcal{A}$为$\mathbf{X}$上的一个算法映射，$\psi(\mathbf{x})$是定义在$\mathbf{X}$上的连续实函数，若满足

当$\mathbf{x} \notin \Omega$且$\mathbf{y} \in \mathcal{A}(\mathbf{x})$时，$\psi(\mathbf{y}) < \psi(\mathbf{x})$

当$\mathbf{x} \in \Omega$且$\mathbf{y} \in \mathcal{A}(\mathbf{x})$时，$\psi(\mathbf{y}) \leq \psi(\mathbf{x})$

则称ψ是关于解集合Ω和算法$\mathcal{A}$的下降函数。

算法收敛性

设Ω为解集合，$\mathcal{A}$为$\mathbf{X}$上的算法映射。若以任意初始点$\mathbf{x}^{(0)} \in \mathbf{Y} \subset \mathbf{X}$ 出发，算法产生的序列的任一收敛子列的极限属于解集合，则称算法映射 $\mathcal{A}$在$\mathbf{Y}$ 上收敛于解集合Ω.

算法收敛性

定理： 设$\mathcal{A}$为$\mathbf{X}$上的一个算法，Ω为解集合，给定初始点 $\mathbf{x}^{(0)} \in \mathbf{X}$, 进行如下迭代：如果$\mathbf{x}^{(k)} \in \Omega$, 则停止迭代；否则取 $\mathbf{x}^{(k+1)} \in \mathcal{A}(\mathbf{x}^{(k)})$, $k := k + 1$, 重复以上过程。这样产生迭代序列$\{\mathbf{x}^{(k)}\}$. 又设：

1. 序列$\{\mathbf{x}^{(k)}\}$含于$\mathbf{X}$的紧子集中；
2. 存在一个连续函数ψ, 它是关于Ω和$\mathcal{A}$的下降函数；
3. 映射$\mathcal{A}$在Ω的补集上是闭的。

则序列$\{\mathbf{x}^{(k)}\}$的任一收敛子列的极限属于Ω.

算法收敛性定理证明

证明： 先证序列$\{\mathbf{x}^{(k)}\}$对应的下降函数值数列$\{\psi(\mathbf{x}^{(k)})\}$有极限。

$\{\mathbf{x}^{(k)}\}$含于$\mathbf{X}$的紧子集，因此有收敛子列$\{\mathbf{x}^{(k)}\}_K$, 设其极限为$\mathbf{x} \in \mathbf{X}$. 由$\psi$的连续性得，
$\psi(\mathbf{x}^{(k)}) \longrightarrow \psi(\mathbf{x}), k \in K$. 即对$\forall \epsilon > 0, \exists N$使得当$k \geq N$时，
有$0 < \psi(\mathbf{x}^{(k)}) - \psi(\mathbf{x}) < \epsilon, k \in K$. 特别地，$0 < \psi(\mathbf{x}^{(N)}) - \psi(\mathbf{x}) < \epsilon$.

又由ψ的下降性知，$\psi(\mathbf{x}^{(k)}) - \psi(\mathbf{x}^{(N)}) < 0, \forall k > N$. 于是有

$$0 < \psi(\mathbf{x}^{(k)}) - \psi(\mathbf{x}) = \psi(\mathbf{x}^{(k)}) - \psi(\mathbf{x}^{(N)}) + \psi(\mathbf{x}^{(N)}) - \psi(\mathbf{x}) < \epsilon, \forall k > N.$$

即得$\lim\limits_{k \to \infty} \psi(\mathbf{x}^{(k)}) = \psi(\mathbf{x})$.

算法收敛性定理证明

证明（继）： 下证$\mathbf{x} \in \Omega$（反证法）。

假设$\mathbf{x} \notin \Omega$, 考虑序列$\{\mathbf{x}^{(k+1)}\}_K$. 由于它包含于紧集，所以也存在收敛子列$\{\mathbf{x}^{(k+1)}\}_{\bar{K}}, \bar{K} \subset K$, 且设其极限为$\bar{\mathbf{x}} \in \mathbf{X}$. 显然（同理前述证明）$\lim\limits_{k\to\infty} \psi(\mathbf{x}^{(k+1)}) = \psi(\bar{\mathbf{x}})$. 由$\psi(\mathbf{x}^{(k)})$极限的唯一性知，

$$\psi(\mathbf{x}) = \psi(\bar{\mathbf{x}}).$$

另外，对$k \in \bar{K} \subset K$有

$$\mathbf{x}^{(k)} \longrightarrow \mathbf{x}, \mathbf{x}^{(k+1)} \in \mathcal{A}(\mathbf{x}^{(k)}), \mathbf{x}^{(k+1)} \longrightarrow \bar{\mathbf{x}}.$$

算法$\mathcal{A}$在Ω的补集上是闭的，$\mathbf{x} \notin \Omega$, 因此$\mathcal{A}$在$\mathbf{x}$处是闭的，即有$\bar{\mathbf{x}} \in \mathcal{A}(\mathbf{x})$.

由于ψ是解集合Ω和算法$\mathcal{A}$的下降函数， $\mathbf{x} \notin \Omega$, 则有$\psi(\bar{\mathbf{x}}) < \psi(\mathbf{x})$. 这显然矛盾，所以$\mathbf{x} \in \Omega$.

实用收敛准则

在迭代下降算法里，当$\mathbf{x}^{(k)} \in \Omega$时才终止迭代。在实践中许多情况下，这是一个取极限的过程，需要无限次迭代。因此为解决实际问题，需要规定一些实用的终止迭代过程的准则，一般称为收敛准则或停机准则。

常用的收敛准则有以下几种：

1. $\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\| < \varepsilon$ or $\dfrac{\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\|}{\|\mathbf{x}^{(k)}\|} < \varepsilon$
2. $f(\mathbf{x}^{(k)}) - f(\mathbf{x}^{(k+1)}) < \varepsilon$ or $\dfrac{f(\mathbf{x}^{(k)}) - f(\mathbf{x}^{(k+1)})}{|f(\mathbf{x}^{(k)})|} < \varepsilon$
3. $\|\nabla f(\mathbf{x}^{(k)})\| < \varepsilon$

在这里，ε为事先给定的充分小的正数。除此之外，还可以根据收敛定理，制定出其它的收敛准则。

收敛速率

评价算法优劣的标准之一是收敛的快慢，通常称为收敛速率。

一般定义如下：设序列$\{\mathbf{x}^{(k)}\}$收敛与$\mathbf{x}^*$, 满足

$$0 \leq \varlimsup_{k\to\infty} \frac{\|\mathbf{x}^{(k+1)} - \mathbf{x}^*\|}{\|\mathbf{x}^{(k)} - \mathbf{x}^*\|^p} = \beta < \infty \tag{72}$$

的非负数p的上确界称为序列$\{\mathbf{x}^{(k)}\}$的收敛阶。

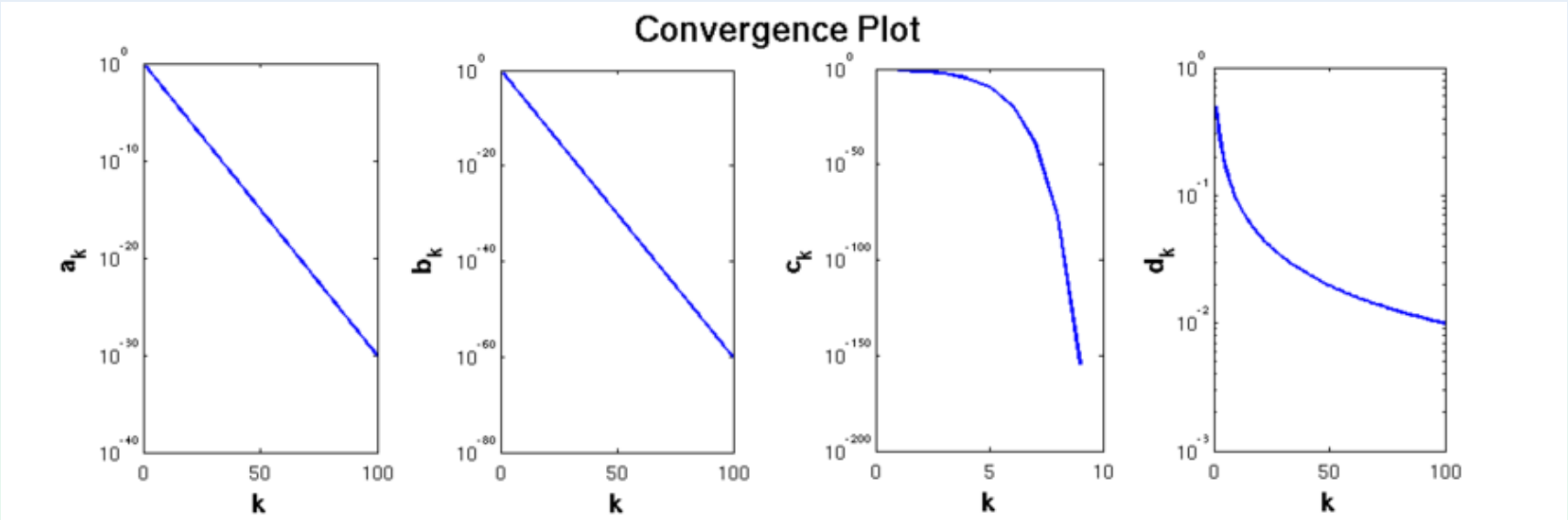
若在定义式(72)中，$p = 1$且$\beta < 1$, 则称序列是（收敛比β）线性收敛的。例如 $\mathbf{x}^{(k)} = \frac{1}{2^k}$

若在定义式(72)中，$p > 1$, 或者$\{p = 1, \beta = 0\}$, 则称序列是超线性收敛的。

若在定义式(72)中，$p = 2$, 或则称序列是二次收敛的。例如 $\mathbf{x}^{(k)} = \frac{1}{2^{2^k}}$

收敛速率

次线性收敛（sublinear）：一般形如 $\mathbf{x}^{(k)} = \mathcal{O}(1/k), \mathcal{O}(1/k^2)$的序列。



上述图像，分别表示线性收敛、超线性收敛、二次收敛和次线性收敛。

计算复杂性

最优化可以追溯到十分古老的极值问题，然而它成为一门独立的学科是在二十世纪四十年代末，当时Dantzig提出了求解一般线性规划问题的单纯形法。此后各种最优化问题的理论及应用研究得到迅速发展，特别是线性规划由于其模型的普遍性和实用性，相关算法的进展引起广泛的重视。

随着实际问题的规模越来越大以及在计算机技术的推动下，人们开始从复杂性角度研究线性规划和非线性规划的算法。

迭代方法

最优化方法通常采用迭代方法求问题的最优解，其基本思想是：

给定一个初始点$\mathbf{x}^{(0)} \in \mathbb{R}^n$，按照某一迭代规则产生一个点列$\{\mathbf{x}^{(k)}\}$，使得当$\{\mathbf{x}^{(k)}\}$是有穷点列时，其最后一个点是最优化模型问题的最优解，当$\{\mathbf{x}^{(k)}\}$是无穷点列时，它有极限点且其极限点是最优化模型问题的最优解。

迭代方法

一个好的迭代算法应具备的典型特征是：

迭代点$\mathbf{x}^{(k)}$能稳定地接近局部极小点$\mathbf{x}^*$的小邻域，然后迅速收敛于$\mathbf{x}^*$. 一般地，对于某种算法我们需要证明其迭代点列$\mathbf{x}^{(k)}$的聚点（即子列的极限点）为一局部极小点。在实际计算中，当指定的收敛准则满足时，迭代即终止。

搜索方向与步长因子

设$\mathbf{x}^{(k)}$为第k次迭代点，$\mathbf{d}^{(k)}$为第k次搜索方向，α_k为第k次步长因子，则第$k+1$次迭代为：

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \alpha_k \mathbf{d}^{(k)}. \tag{73}$$

从上述迭代格式可以看出，不同的搜索方向和不同的步长策略构成不同的方法。

搜索方向与步长因子

在最优化方法中，搜索方向$\mathbf{d}^{(k)}$一般选取的是某价值函数 (merit function) ψ 在$\mathbf{x}^{(k)}$处的下降方向，即$\mathbf{d}^{(k)}$满足

$$\nabla\psi(\mathbf{x}^{(k)})^T\mathbf{d}^{(k)} < 0. \tag{74}$$

步长因子的确定一般归结为解一维最优化问题

$$\min_{\alpha\geq 0}\psi(\mathbf{x}^{(k)} + \alpha\mathbf{d}^{(k)}) \tag{75}$$

最优化迭代算法的基本结构之一

(a) 给定初始点$\mathbf{x}^{(0)}$

(b) 计算搜索方向$\mathbf{d}^{(k)}$, 即构造某价值函数ψ在$\mathbf{x}^{(k)}$点处的下降方向作为搜索方向;

(c) 确定步长因子α_k, 使该价值函数值有某种程度的下降;

(d) 迭代更新，令$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \alpha_k \mathbf{d}^{(k)}$.

(e) 判断停机准则，若$\mathbf{x}^{(k+1)}$满足某种终止条件，则停止迭代，得到近似最优解$\bar{\mathbf{x}} = \mathbf{x}^{(k+1)}$. 否则，返回(b)重复以上步骤。

最优化迭代算法的基本结构之二

(a) 给定初始点$\mathbf{x}^{(0)}$

(b) 构造某价值函数ψ在$\mathbf{x}^{(k)}$附近（如一定半径内）的二次近似模型；

(c) 求解该近似模型得到$\mathbf{s}^{(k)}$作为更新位移向量；

(d) 迭代更新，令$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \mathbf{s}^{(k)}$.

(e) 判断停机准则，若$\mathbf{x}^{(k+1)}$满足某种终止条件，则停止迭代，得到近似最优解$\bar{\mathbf{x}} = \mathbf{x}^{(k+1)}$. 否则，返回(b)重复以上步骤。

本章作业

习题5.1： 对于问题(52), 若$\bar{x}$是局部最优解，并且$\bar{x}$满足MFCQ条件。证明如下结论：

1. 对于$\forall d \in T(\bar{x} \mid S)$, 我们有$\nabla f(\bar{x})^T d \geq 0$。
2. 考虑下面的线性规划问题：

$$\max(-\nabla f(\bar{x}))^T d \tag{76}$$

$$\text{s.t.} \quad \nabla g_j(\bar{x})^T d \geq 0, j \in \mathcal{I}(\bar{x}) \tag{77}$$

$$\nabla h_i(\bar{x})^T d = 0, i = 1, 2, \ldots, \ell. \tag{78}$$

写出其对偶问题，并利用强对偶定理证明，在MFCQ条件下的KKT条件成立。

本章作业

习题5.2： 给定 $c \in \mathbb{R}^n$, $\rho > 0$ 计算如下问题最优解：

$$\min_{w} c^T w + \rho \sum_{i=1}^{n} w_i \log(w_i) \tag{79}$$

$$\text{s.t.} \quad \sum_{i=1}^{n} w_i = 1, w_i \geq 0, i = 1, \ldots, n. \tag{80}$$

注： $\log w_i$ **表示以** e**为底，令**$0 \log 0 = 0$.

无约束最优化问题

$$\min_{\mathbf{x}\in\mathbb{R}^n} \quad f(\mathbf{x}) \tag{81}$$

其目标函数f是定义在$\mathbb{R}^n$上的实值函数，决策变量$\mathbf{x}$的可取值之集合是全空间 $\mathbb{R}^n$.

无约束问题的最优性条件

一阶必要条件：设目标函数$f(\mathbf{x})$在点$\bar{\mathbf{x}}$处可微，若$\bar{\mathbf{x}}$ 是局部极小点，则$\nabla f(\bar{\mathbf{x}}) = 0$.

二阶必要条件：设目标函数$f(\mathbf{x})$在点$\bar{\mathbf{x}}$处二次可微，若$\bar{\mathbf{x}}$是局部极小点，则$\nabla f(\bar{\mathbf{x}}) = 0$, 并且Hesse矩阵$\nabla^2 f(\bar{\mathbf{x}}) \geq 0$.

二阶充分条件：设目标函数$f(\mathbf{x})$在点$\bar{\mathbf{x}}$处二次可微，若$\nabla f(\bar{\mathbf{x}}) = 0$且$\nabla^2 f(\bar{\mathbf{x}}) > 0$, 则$\bar{\mathbf{x}}$是局部极小点。

充要条件：设$f(\mathbf{x})$是定义在$\mathbb{R}^n$上的可微凸函数，则$\bar{\mathbf{x}}$是整体极小点（全局最优解）的充要条件是$\nabla f(\bar{\mathbf{x}}) = 0$.

定义

对于无约束问题，满足$\nabla f(\bar{x}) = 0$的点$\bar{x}$称为稳定点。稳定点可分为两种情况：（1）局部最小值点；（2）鞍点。

例如 $f(x) = x^3$, $\bar{x} = 0$为鞍点。
对于凸问题来说，我们求解得到某=一个稳定点$\bar{x}$即得到全局最优点。对于非凸问题，我们同样认为，得到稳定点解并非全局最优或局部最优，因此还需进一步判断。

梯度类求解算法

梯度向量$\nabla f(\mathbf{x})$是函数f在点$\mathbf{x}$处增加最快的方向，故它成为最优化时的重要工具。实际上针对无约束最优化问题，大家所知的求解算法中大多属于下面的梯度方法类。

GRADIENT （梯度法类）

(0) 初始化：选取适当的初始点$\mathbf{x}^{(0)} \in \mathbb{R}^n$, 令$k := 0$.

(1) 计算搜索方向：利用适当的正定对称阵H_k计算搜索方向向量$\mathbf{d}^{(k)} := -H_k \nabla f(\mathbf{x}^{(k)})$. （如果$\nabla f(\mathbf{x}^{(k)}) = \mathbf{0}$, 则结束计算）

(2) 确定步长因子：解一维最优化问题$\min\limits_{\alpha \geq 0} f(\mathbf{x}^{(k)} + \alpha \mathbf{d}^{(k)})$, 求出**步长**$\alpha = \alpha_k$, 令$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \alpha_k \mathbf{d}^{(k)}$, $k := k + 1$, 回到第(1)步。

注：在机器学习领域，步长通常被称为学习率（learning rate）。

构造搜索方向：负梯度方向

无约束最优化问题：$\min\limits_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x})$

$$f(\mathbf{x}) = f(\mathbf{x}^{(k)}) + \nabla f(\mathbf{x}^{(k)})^T(\mathbf{x} - \mathbf{x}^{(k)}) + O(\|\mathbf{x} - \mathbf{x}^{(k)}\|^2) \tag{82}$$

取负梯度方向

$$\mathbf{d}^{(k)} = -\nabla f(\mathbf{x}^{(k)}),$$

则当α_k足够小时，总能使

$$f(\mathbf{x}^{(k)} + \alpha_k \mathbf{d}^{(k)}) < f(\mathbf{x}^{(k)}).$$

构造搜索方向：牛顿方向

$$
\begin{aligned}
f(\mathbf{x}) = \quad & f(\mathbf{x}^{(k)}) + \nabla f(\mathbf{x}^{(k)})^T(\mathbf{x} - \mathbf{x}^{(k)}) \\
& + \frac{1}{2}(\mathbf{x} - \mathbf{x}^{(k)})^T \nabla^2 f(\mathbf{x}^{(k)})(\mathbf{x} - \mathbf{x}^{(k)}) + O(\|\mathbf{x} - \mathbf{x}^{(k)}\|^3)
\end{aligned} \tag{83}
$$

取搜索方向

$$
\mathbf{d}^{(k)} = -G_k^{-1} \nabla f(\mathbf{x}^{(k)}),
$$

其中$G_k = \nabla^2 f(\mathbf{x}^{(k)})$为函数$f$在$\mathbf{x}^{(k)}$点处的Hesse矩阵。

确定步长因子：一维搜索

在迭代格式中，沿着下降方向$\mathbf{d}^{(k)}$, 通过解一维最优化问题

$$\min_{\alpha \geq 0} \varphi(\alpha) = f(\mathbf{x}^{(k)} + \alpha \mathbf{d}^{(k)}) \tag{84}$$

确定步长因子的方法称为**一维搜索**(Line Search).

确定步长因子：一维搜索

若以问题(84)的最优解为步长，此时称为**精确一维搜索**(Exact Line Search).

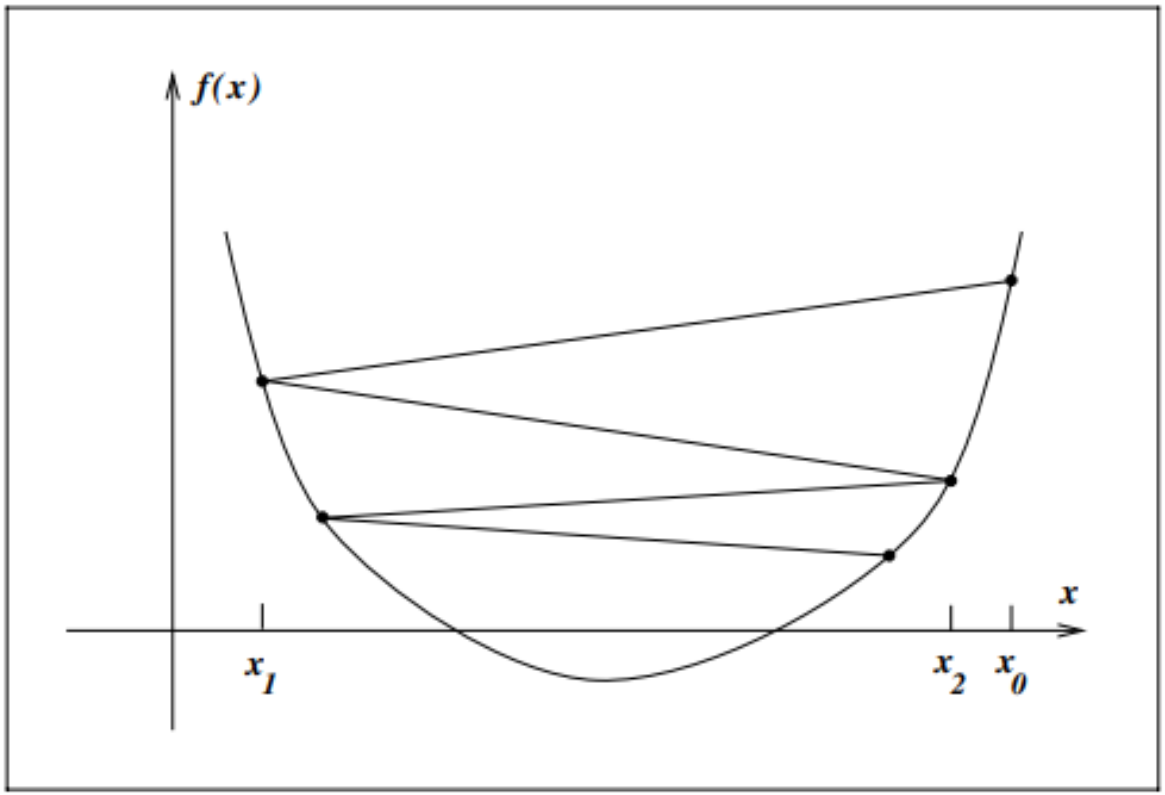
经常用到的精确一维搜索有黄金分割法和插值迭代法。即使说是精确一维搜索，通过有限次计算求出问题(84)的严密解一般也是不可能的，实际上在得到有足够精度的近似解时，就采用它作为步长。

确定步长因子：一维搜索

在实际计算中，往往不是求解一维最优化问题(84)，而是找出满足某些适当条件的粗略近似解作为步长，此时称为**非精确一维搜索**(Inexact Line Search).

与精确一维搜索相比，在很多情况下采用非精确一维搜索可以提高整体计算效率。

确定步长因子：一维搜索



上图中，由于步长 α选择较大，迭代产生了左右震荡。反之，若是步长太小，那么算法收敛速度非常缓慢。线搜索的目标就是，使得沿着下降方向**d**，每次的函数值满足充分下降（sufficient decrease）条件。

Backtracking linesearch

最简单、常见的线搜索条件为Backtracking linesearch方法:

1. 选取 $\gamma \in (0,1)$ 和 $c \in (0,1)$
2. 选择最小的整数$t \geq 0$, 使得

$$f(\mathbf{x}^{(k)} + \gamma^t \mathbf{d}^{(k)}) \leq f(\mathbf{x}^{(k)}) + c\gamma^t \nabla f(\mathbf{x}^{(k)})^T \mathbf{d}^{(k)} \tag{85}$$

3. 令 $\alpha_k = \gamma^t$, 更新 $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \alpha_k \mathbf{d}^{(k)}$

(85)被称为 Armijo-Goldstein不等式。故backtracking也被称为 Armijo-backtracking linesearch. γ通常选取0.9或者0.5.

t的存在性:

$$\lim_{\alpha \downarrow 0} \frac{f(\mathbf{x}^{(k)} + \alpha \mathbf{d}^{(k)}) - f(\mathbf{x}^{(k)})}{\alpha} = f'(\mathbf{x}^{(k)}; \mathbf{d}^{(k)}) < cf'(\mathbf{x}^{(k)}; \mathbf{d}^{(k)}) < 0.$$

故，存在$\bar{\alpha} > 0$，使得

$$\frac{f(\mathbf{x}^{(k)} + \alpha \mathbf{d}^{(k)}) - f(\mathbf{x}^{(k)})}{\alpha} \leq cf'(\mathbf{x}^{(k)}; \mathbf{d}^{(k)}), \quad \forall \alpha \in (0, \bar{\alpha}) \tag{86}$$

确定步长因子：一维搜索

令

$$\varphi(\alpha) = f(x + \alpha d).$$

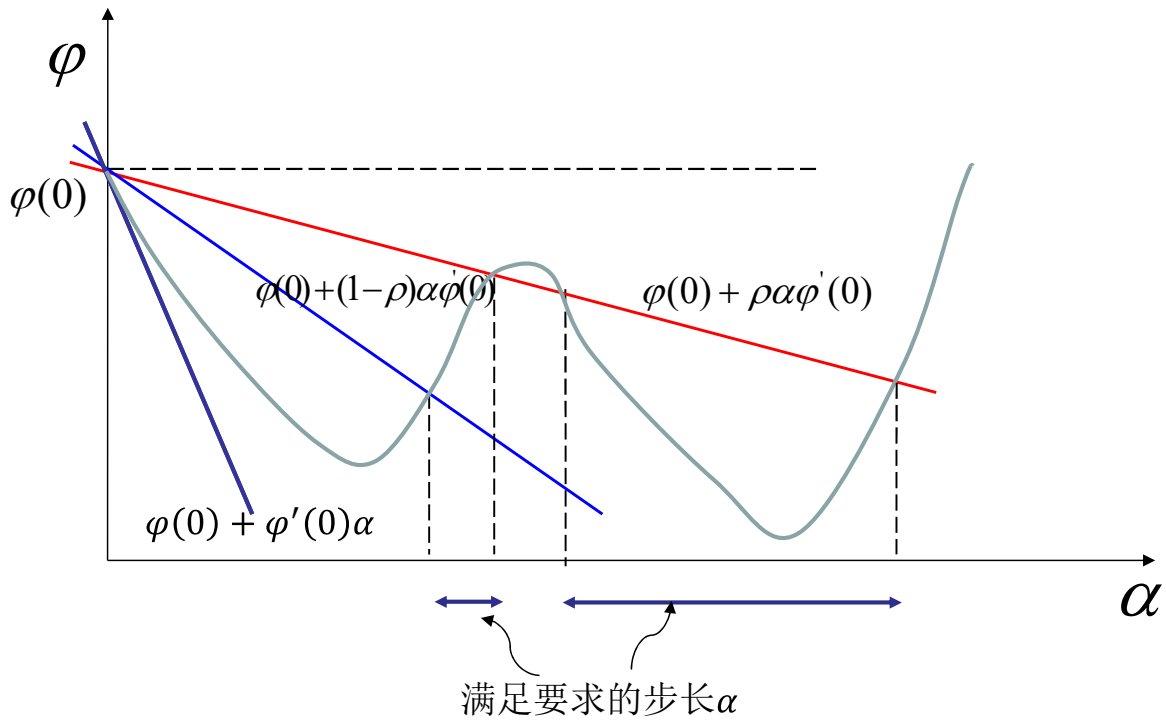
我们有 $\varphi'(\alpha) = \nabla f(x + \alpha d)^T d$.
Armijo-Goldstein conditions:

$$\varphi(\alpha) \leq \varphi(0) + \rho\alpha\varphi^{'}(0) \tag{87}$$

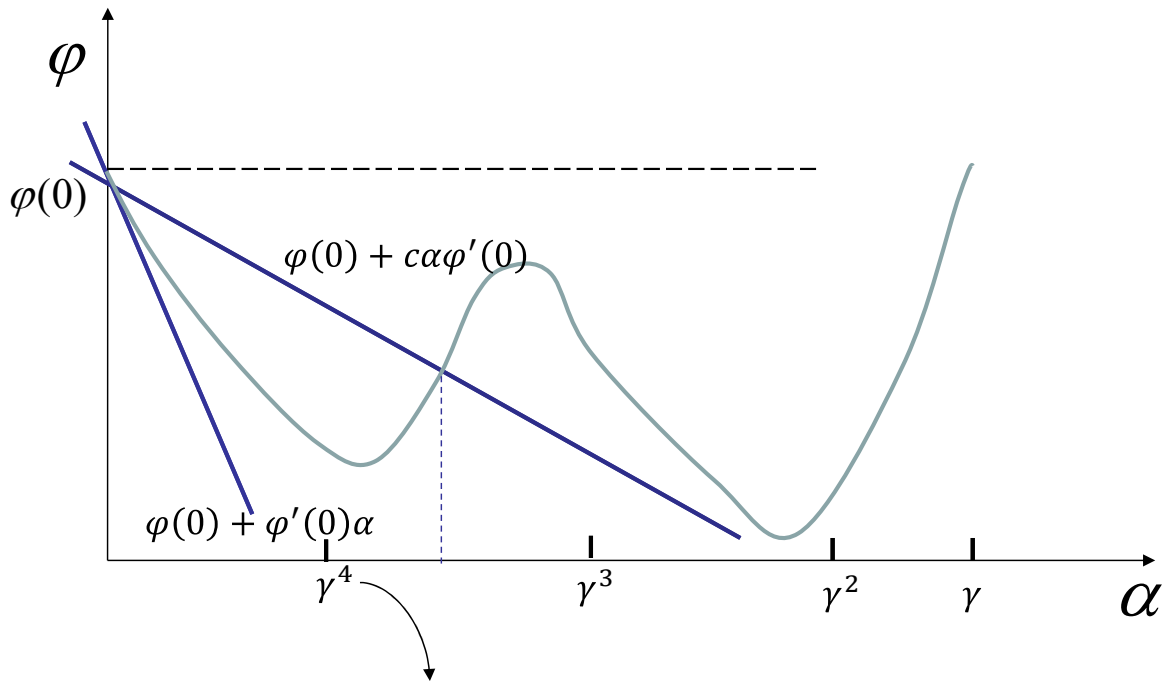
$$\varphi(\alpha) \geq \varphi(0) + (1-\rho)\alpha\varphi^{'}(0) \tag{88}$$

其中$\rho \in (0, 1/2)$是一个固定参数。

确定步长因子：Armijo-Goldstein 条件图例



确定步长因子：一维搜索



满足要求的步长$t=4,\alpha=\gamma^4$

确定步长因子：backtracking linesearch 图示

Wolfe(1968)-Powell(1976) conditions:

$$\varphi(\alpha) \leq \varphi(0) + c_1 \alpha \varphi^{'}(0) \tag{89}$$

$$\varphi^{'}(\alpha) \geq c_2 \varphi^{'}(0) \tag{90}$$

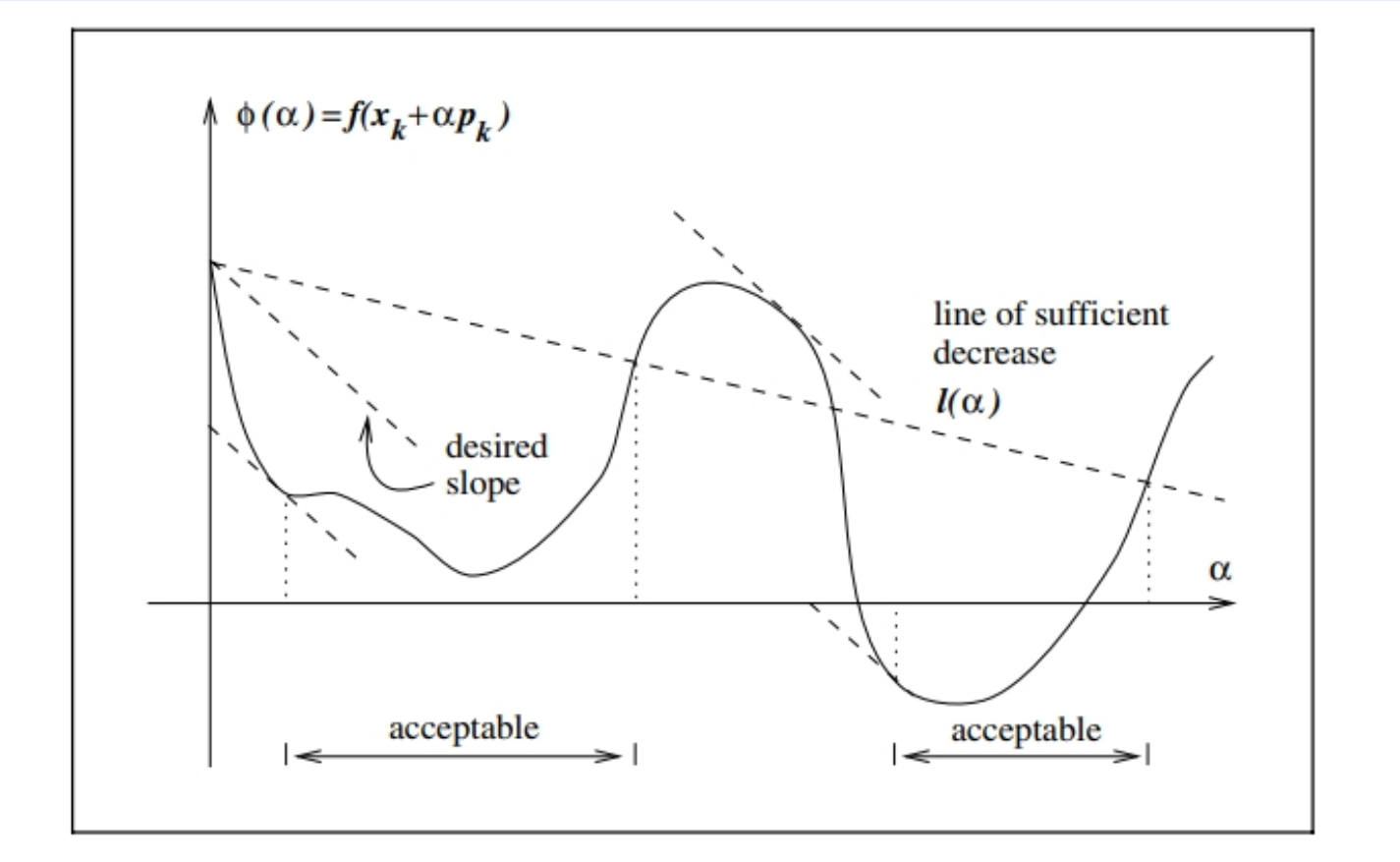
其中$0 < c_1 < c_2 < 1$是固定参数。
考虑问题

$$\min_{\alpha \in \mathbb{R}} \varphi(\alpha),$$

(90) 使得 $\varphi'(\alpha)$更接近0，也被称为弱Wolfe-Powell条件。

确定步长因子：Wolfe-Powell 条件

Wolfe(1968)-Powell(1976) conditions:



图片来源： Numerical optimization. By Jorge Nocedal and Stephen J. Wright.

确定步长因子：Wolfe-Powell 条件

在很多实际算法中，式(90)常被强化的双边条件所取代（也被称为强Wolfe-Powell条件。）

$$|\varphi^{'}(\alpha)| \leq c_2|\varphi^{'}(0)|, 0 < c_1 < c_2 < 1. \tag{91}$$

此条件排除了$\varphi^{'}(\alpha)$为非常大的正数情况。

Wolfe-Powell条件存在性： 若问题(81)中 f连续可微。令$\mathbf{d}^{(k)}$是下降方向，并且假设f沿着射线方向 $\{\mathbf{x}^{(k)} + \alpha\mathbf{d}^{(k)} \mid \alpha > 0\}$ 有下界。那么一定存在 $0 < c_1 < c_2 < 1$使得 (89)、(90) 或(91)成立。

证明： 因为f沿着射线方向 $\{\mathbf{x}^{(k)} + \alpha\mathbf{d}^{(k)} \mid \alpha > 0\}$ 有下界，$l(\alpha) := f(\mathbf{x}^{(k)}) + \alpha c_1 \nabla f(\mathbf{x}^{(k)})^T\mathbf{d}^{(k)}$单调减小至 $-\infty$, 可知$\varphi(\alpha)$与$l(\alpha)$至少有一个交点。设 $\bar{\alpha}$是最小的交点。有下述不等式成立

$$f(\mathbf{x}^{(k)} + \alpha_1\mathbf{d}^{(k)}) \leq f(\mathbf{x}^{(k)}) + c_1\alpha_1\nabla f(\mathbf{x}^{(k)})^T\mathbf{d}^{(k)}, \quad \forall\alpha_1 \in (0, \bar{\alpha}).$$

由中值定理可知，存在 $\alpha_2 \in (0, \bar{\alpha})$使得 $f(\mathbf{x}^{(k)} + \bar{\alpha}\mathbf{d}^{(k)}) - f(\mathbf{x}^{(k)}) = \bar{\alpha}\nabla f(\mathbf{x}^{(k)} + \alpha_2\mathbf{d}^{(k)})^T\mathbf{d}^{(k)}$ 因为 $l(\bar{\alpha}) = \varphi(\bar{\alpha})$, 即 $f(\mathbf{x}^{(k)} + \bar{\alpha}\mathbf{d}^{(k)}) = f(\mathbf{x}^{(k)}) + c_1\bar{\alpha}\nabla f(\mathbf{x}^{(k)})^T\mathbf{d}^{(k)}$, 我们有

$$\nabla f(\mathbf{x}^{(k)} + \alpha_2\mathbf{d}^{(k)})^T\mathbf{d}^{(k)} = c_1\nabla f(\mathbf{x}^{(k)})^T\mathbf{d}^{(k)} > c_2\nabla f(\mathbf{x}^{(k)})^T\mathbf{d}^{(k)}, 1 > c_2 > c_1 > 0.$$

因此$0 < c_1 < c_2 < 1, \alpha_1 \in (0, \bar{\alpha}), \alpha_2 \in (0, \bar{\alpha})$满足(89)、(90).

注意到$\nabla f(\mathbf{x}^{(k)} + \alpha_2\mathbf{d}^{(k)})^T\mathbf{d}^{(k)} < 0$,因此(91)也成立。

确定步长因子：Wolfe-Powell 条件

基于Wolfe-Powell准则的非精确一维搜索算法：

(0) 给定初始一维搜索区间$[0, \alpha_0]$, 以及$c_1 \in (0, 1/2)$, $c_2 \in (c_1, 1)$.
计算$\varphi_0 = \varphi(0) = f(\mathbf{x}^{(k)})$, $\varphi_0^{'} = \varphi^{'}(0) = \nabla f(\mathbf{x}^{(k)})^T \mathbf{d}^{(k)}$.
并令$a_1 = 0, a_2 = \alpha_0, \varphi_1 = \varphi_0, \varphi_1^{'} = \varphi_0^{'}$.
选取适当的$\alpha \in (a_1, a_2)$.

通常来说c_1比较小，例如$c_1 = 10^{-3}$. $c_2 = 0.9$.

确定步长因子：Wolfe-Powell 条件

基于Wolfe-Powell准则的非精确一维搜索算法：

(1) 计算$\varphi = \varphi(\alpha) = f(\mathbf{x}^{(k)} + \alpha\mathbf{d}^{(k)})$. 若$\varphi(\alpha) \leq \varphi(0) + c_1\alpha\varphi'(0)$, 则转到第(2)步。否则，由$\varphi_1, \varphi_1', \varphi$构造两点二次插值多项式$p^{(1)}(t)$, 并得其极小点

$$\hat{\alpha} = a_1 + \frac{1}{2}\frac{(a_1 - \alpha)^2\varphi_1'}{(\varphi_1 - \varphi) - (a_1 - \alpha)\varphi_1'}.$$

于是置$a_2 = \alpha, \alpha = \hat{\alpha}$, 重复第(1)步。

确定步长因子：Wolfe-Powell

基于Wolfe-Powell准则的非精确一维搜索算法：

(2) 计算$\varphi^{'} = \varphi^{'}(\alpha) = \nabla f(\mathbf{x}^{(k)} + \alpha\mathbf{d}^{(k)})^T\mathbf{d}^{(k)}$. 若$\varphi^{'}(\alpha) \geq c_2\varphi^{'}(0)$, 则输出$\alpha_k = \alpha$, 并停止搜索。否则，由$\varphi, \varphi^{'}, \varphi_1^{'}$构造两点二次插值多项式$p^{(2)}(t)$, 并得其极小点

$$\hat{\alpha} = \alpha - \frac{(a_1 - \alpha)\varphi^{'}}{\varphi_1^{'} - \varphi^{'}}.$$

于是置$a_1 = \alpha, \alpha = \hat{\alpha}, \varphi_1 = \varphi, \varphi_1^{'} = \varphi^{'}$, 返回第(1)步。

[作业6.1：请写出上述基于Wolfe-Powell准则的非精确一维搜索算法中插值多项式$p^{(1)}(t), p^{(2)}(t)$的具体表达式。]

全局收敛

从任意初始点出发，如果某迭代算法产生的点列的极限（聚点），在适当假定下可保证恒为问题的最优解（或者稳定点），则称该迭代法具有全局收敛性(Global Convergence).

与此相对，如果仅在解的附近选取初始点时，才可以保证所生成的点列收敛于该解，则称这样的迭代法有局部收敛性(Local Convergence).

全局收敛

为了证明迭代法的下降性，我们应尽量避免搜索方向与负梯度方向几乎正交的情形，即要求$\mathbf{d}^{(k)}$偏离$\mathbf{g}^{(k)} = \nabla f(\mathbf{x}^{(k)})$的正交方向远一些。否则，${\mathbf{g}^{(k)}}^T \mathbf{d}^{(k)}$接近于零，$\mathbf{d}^{(k)}$几乎不是下降方向。

为此，我们假设$\mathbf{d}^{(k)}$与$-\mathbf{g}^{(k)}$的夹角θ_k满足

$$\theta_k \leq \frac{\pi}{2} - \mu, \ \forall k \tag{92}$$

其中$\mu > 0$（与k无关）。
显然$\theta_k \in [0, \pi/2)$, 其定义为

$$\cos\theta_k = \frac{-{\mathbf{g}^{(k)}}^T \mathbf{d}^{(k)}}{\|\mathbf{g}^{(k)}\|\|\mathbf{d}^{(k)}\|} = \frac{-{\mathbf{g}^{(k)}}^T \mathbf{s}^{(k)}}{\|\mathbf{g}^{(k)}\|\|\mathbf{s}^{(k)}\|} \tag{93}$$

这里$\mathbf{s}^{(k)} = \alpha_k \mathbf{d}^{(k)} = \mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}$.

全局收敛

下面给出各种步长准则下的下降算法的全局收敛性结论。

全局收敛性定理： 假设$f(x)$有下界，即 $f(x) > -\infty, \forall x \in \mathbb{R}^n$. 设$f(\mathbf{x})$在包含水平集 $L(\mathbf{x}^{(0)}) = \{\mathbf{x} \mid f(\mathbf{x}) \leq f(\mathbf{x}^{(0)})\} \subset \mathcal{N}$的开集 $\mathcal{N}$上连续可微。同时，梯度$\nabla f(\mathbf{x})$ 在 $\mathcal{N}$上是李氏（Lipschitz continuous)连续的，即存在 $L > 0$，使得

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

下降算法的搜索方向$\mathbf{d}^{(k)}$与$-\nabla f(\mathbf{x}^{(k)})$ 之间的夹角θ_k满足式(92), 其中步长α_k由Wolfe-Powell (89),(90)确定。那么，$\nabla f(\mathbf{x}^{(k)}) \to \mathbf{0}$.

全局收敛

全局收敛性证明：为了记号简洁，我们记所有的k, $g_k = \nabla f(\mathbf{x}^{(k)})$, $f_k = f(\mathbf{x}^{(k)})$. 由梯度李氏连续性可知：

$$\|g_{k+1} - g_k\| \leq L\|x_{k+1} - x_k\| = L\alpha_k\|\mathbf{d}^{(k)}\|$$

由(90)可知，

$$(g_{k+1} - g_k)^T\mathbf{d}^{(k)} \geq (c_2 - 1)g_k^T\mathbf{d}^{(k)}.$$

结合上述两个不等式，我们有

$$\alpha_k \geq \frac{c_2 - 1}{L}\frac{g_k^T\mathbf{d}^{(k)}}{\|\mathbf{d}^{(k)}\|^2}. \qquad \text{(步长有下界)}$$

带入(89),我们有

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

根据(93), 我们有

$$f_{k+1} \leq f_k - c_1\frac{1 - c_2}{L}\cos^2\theta_k\|g_k\|^2.$$

全局收敛性证明（续）：

将上述不等式，对$k = 0, \ldots, T$相加，可得

$$f_{T+1} \leq f_0 - c_1 \frac{1 - c_2}{L} \sum_{k=0}^{T} \cos^2 \theta_k \|g_k\|^2.$$

因为f有下界，故

$$\sum_{k=0}^{T} \cos^2 \theta_k \|g_k\|^2 < +\infty.$$

上式对任意T成立，故

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

又因为(92), 存在 $\delta > 0$, 使得

$$\cos \theta_k \geq \delta.$$

所以

$$\lim_{k \to \infty} \|\nabla f(\mathbf{x}^{(k)})\| = 0.$$

全局收敛

[习题6.2：证明：假设$f(x)$有下界，即 $f(x) > -\infty, \forall x \in \mathbb{R}^n$. 设$f(\mathbf{x})$在包含水平集 $L(\mathbf{x}^{(0)}) = \{\mathbf{x} \mid f(\mathbf{x}) \leq f(\mathbf{x}^{(0)})\} \subset \mathcal{N}$的开集 $\mathcal{N}$上连续可微。同时，梯度$\nabla f(\mathbf{x})$ 在 $\mathcal{N}$上是李氏（Lipschitz continuous)连续的。基于backtracking linesearch (85)的非精确一维搜索算法,在$\mathbf{d}^{(k)} = -\nabla f(\mathbf{x}^{(k)})$情况下的全局收敛性。]