最优化理论与方法

第9讲

下降算法的收敛性

下降算法的一般步骤

算法 2.1 (求解无约束问题的下降算法)

步 1 给定初始点 $x^{(0)} \in R^n$, 精度 $\epsilon > 0$. 令 $k := 0$.

步 2 若 $\|\nabla f(x^{(k)})\| \le \epsilon$, 则终止算法, 得解 $x^{(k)}$. 否则, 转步 3.

步 3 确定下降方向 $d^{(k)}$, 使得

$$\nabla f(x^{(k)})^T d^{(k)} < 0.$$

步 4 确定步长 $\alpha_k > 0$ 使得

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

步 5 令 $x^{(k+1)} = x^{(k)} + \alpha_k d^{(k)}$, $k := k + 1$. 转步 2.

由下降算法的结构知,为构造一个使用的下降算法,我们的工作主要有两部分：

已知近似最优解x_k,

1. 计算下降方向d_k满足: $\nabla f(x_k)^T d_k < 0$

基本的构造方法有两个:

(1) $d_k = -H_k \nabla f(x_k)$,其中H_k是某一对称正定矩阵

(2) $d_k = -a_k \nabla f(x_k) + b_k d_{k-1}$, 即$-\nabla f(x_k)$和$d_{k-1}$的线性组合

d_k的不同构造方式对应不同的最优化算法,具体将在以后各章介绍

2. 计算步长$\alpha_k > 0$满足: $f(x_k + \alpha_k d_k) < f(x_k)$

这一步主要通过线性搜索来完成——一元函数

$$\phi(\alpha) = f(x_k + \alpha\, d_k)$$

求极值,具体在下一节介绍

$$\boxed{x^{k+1} = x^k + \alpha^k d^k, \qquad k = 0, 1, \ldots}$$

步长 α^k 的选取：线性搜索

计算步长α_k的两种线性搜索：精确搜索和非精确搜索：

精确搜索：α_k是一维优化问题的解

$$\min_{\alpha>0} f(x_k + \alpha d_k)$$

即α_k满足

$$\nabla f(x_k + \alpha_k d_k)^{\mathrm{T}} d_k = 0$$

非精确搜索：α_k按照某种规则计算使之满足

$$f(x_k + \alpha_k d_k) < f(x_k)$$

即函数值有一定程度下降.

精确线搜索（最优步长）小结

$$\alpha_k = \arg\min_{\alpha \geq 0} f(x_k + \alpha d_k)$$

- ✔ 目标函数沿搜索方向下降量达到最大。

理想化策略，计算量大，很难取到。

- ✔ 便于算法理论分析（如二次终止性、收敛速率等）
- ✘ 对全局最优解不起关键作用

二、 非精确线性搜索— Armijo 型线性搜索和 Wolfe-Powell 型线性搜索

Armijo型线性搜索：给定 $\sigma_1 \in (0,1)$，计算 α_k 满足

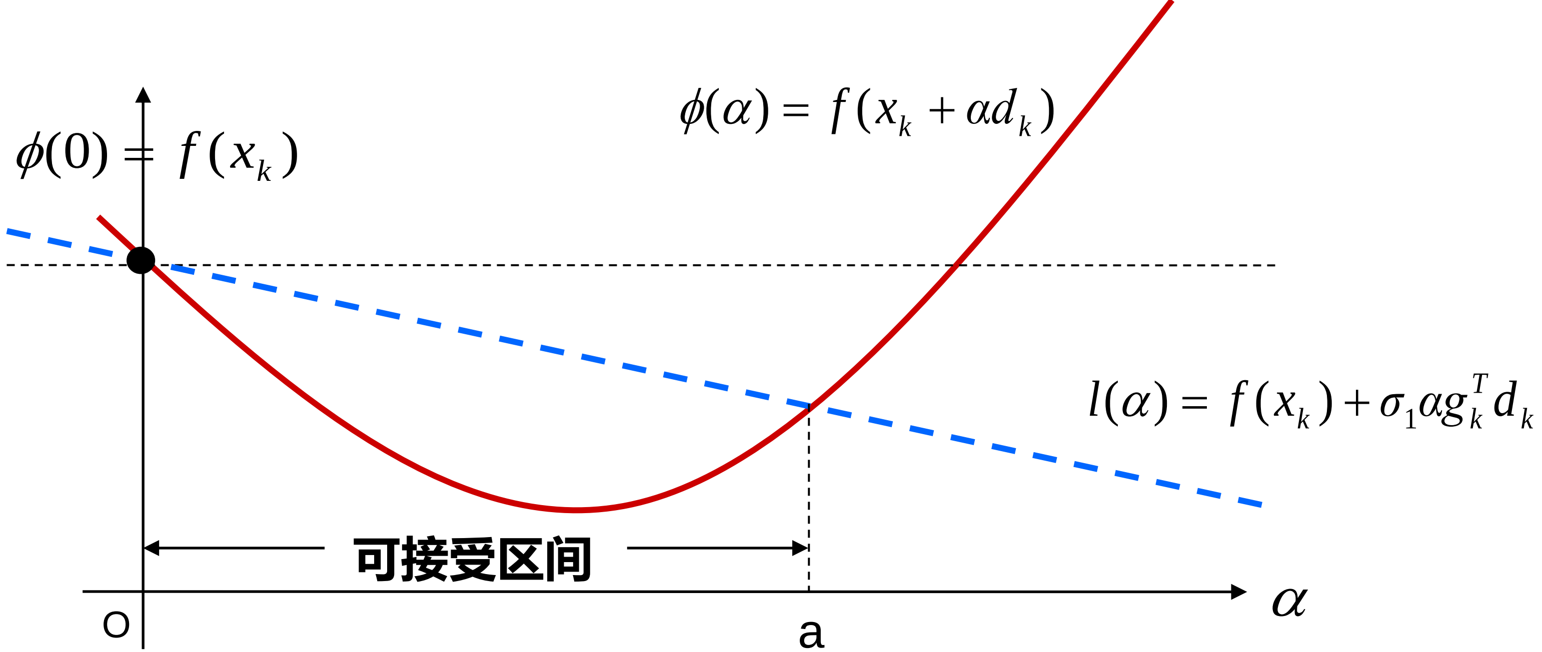
$$f(x_k+\alpha_k d_k) \le f(x_k)+\sigma_1\alpha_k \nabla f(x_k)^{\mathrm{T}} d_k \qquad (2.6)$$

令 $\phi(\alpha) = f(x_k+\alpha d_k)$，(2.6)等价于

$$\phi(\alpha_k) \le \phi(0)+\sigma_1\alpha_k\phi'(0)$$

Wolfe - Powell线性搜索：给定常数 $0<\sigma_1<\sigma_2<1$

$$\begin{cases} f(x_k+\alpha_k d_k) \le f(x_k)+\sigma_1\alpha_k\nabla f(x_k)^T d_k \\ \nabla f(x_k+\alpha_k d_k)^{\mathrm{T}} d_k \ge \sigma_2 \nabla f(x_k)^T d_k \end{cases} \qquad (2.7)$$



算法2.3 (Armijo 型线性搜索)

步0 若 $\alpha_k = 1$ 满足(2.6), 则取 $\alpha_k = 1$. 否则转下一步;

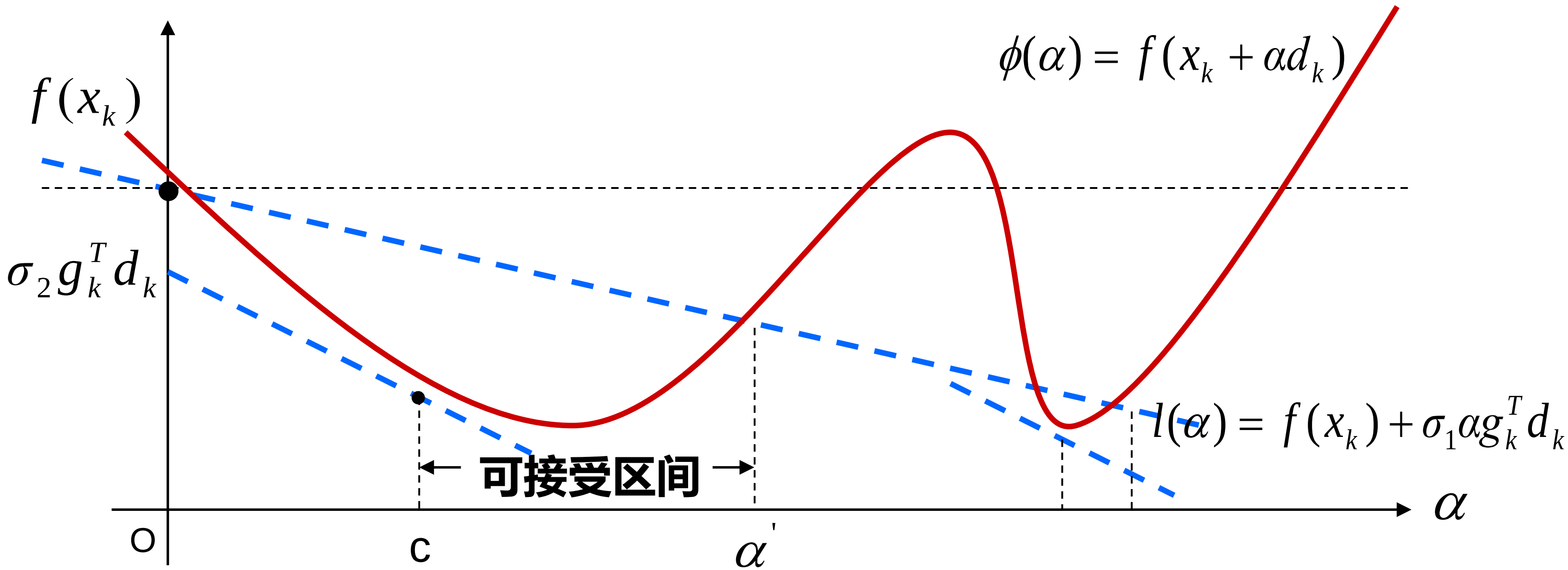
步1 给定常数 $\beta > 0, \rho \in (0,1)$. 令 $\alpha_k = \beta$;

步2 若 α_k 满足(2.6), 则得到步长 α_k, 终止计算. 否则转步3;

步3 令 $\alpha_k := \rho\alpha_k$, 转步2.

Wolfe - Powell 线性搜索：给定常数 $0<\sigma_1<\sigma_2<1$

$$\begin{cases} f(x_k+\alpha_k d_k) \le f(x_k)+\sigma_1\alpha_k\nabla f(x_k)^T d_k \\ \nabla f(x_k+\alpha_k d_k)^T d_k \ge \sigma_2\nabla f(x_k)^T d_k \end{cases} \quad (2.7)$$



从而得到了另一个重要的非精确搜索－Wolfe - Powell 搜索。

具体实现如下：

Wolfe - Powell 线性搜索：给定常数 $0<\sigma_1<\sigma_2<1$

$$\begin{cases} f(x_k+\alpha_k d_k)\le f(x_k)+\sigma_1\alpha_k\nabla f(x_k)^T d_k \\ \nabla f(x_k+\alpha_k d_k)^{\mathrm{T}} d_k \ge \sigma_2\nabla f(x_k)^T d_k \end{cases} \qquad (2.7)$$

步0: 若 $\alpha_k=1$ 满足(2.7), 取若 $\alpha_k=1$; 否则转下一步；

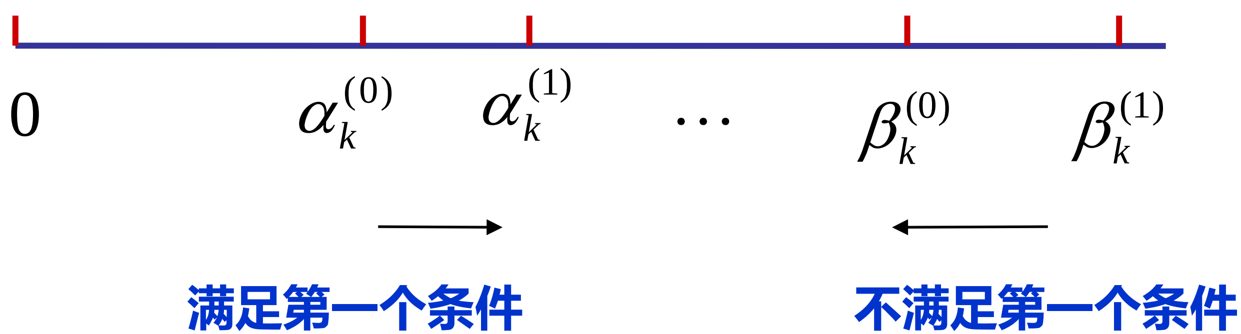
步2: 给定常数 $\beta>0, \rho, \rho_1\in(0,1)$. 令 $\alpha_k^{(0)}$ 是集合 $\{\beta\rho^j \mid (j$ 可正可负$)\}$ 中满足第一个不等式的最大者. 令 $i=0$;

步2: 若 $\alpha_k^{(i)}$ 满足第二个条件，则取 $\alpha_k=\alpha_k^{(i)}$; 否则, 令 $\beta_k^{(\mathrm{i})}=\rho^{-1}\alpha_k^{(i)}$.

步3: 令 $\alpha_k^{(i+1)}$ 是集合

$$\{\alpha_k^{(i)}+\rho_1^{\mathrm{j}}(\beta_k^{(\mathrm{i})}-\alpha_k^{(i)}),\ j=0,1,2,\cdots\}$$

中使得(2.7)第一个不等式成立的最大者，令 $i=i+1$, 转步2.



全局收敛性定理含义

1) 算法产生的点列收敛于问题的稳定点（或解）：

$$\lim_{k\to\infty} x^{(k)} = x^*, \quad \nabla f(x^*) = 0.$$

2) 算法产生的点列的每一个极限点都是问题的稳定点（或解）：

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

3) 算法产生的点列中有一个极限点是问题的稳定点（或解）：

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

$$1) \quad \Rightarrow \quad 2) \quad \Rightarrow \quad 3)$$

三种全局收敛性的等价条件

定理（三种全局收敛性的等价条件）

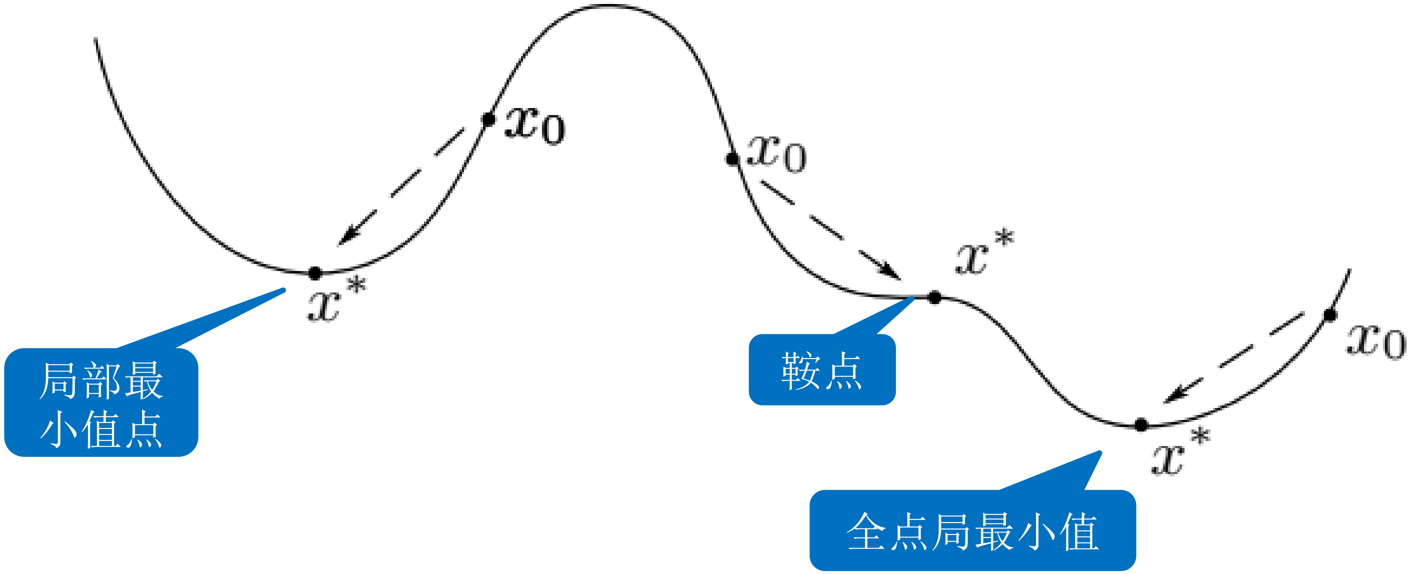
设$\{x^{(k)}\}$是下降算法产生的点列. 若f是一致凸函数，则三种收敛性等价. 即由

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

可推得：$\{x^{(k)}\}$收敛于函数f的唯一全局最小值点.

证明： 由条件知：存在子列$\{x^{(k)}\}_K \to x^*$, $\nabla f(x^*) = 0$. 由于f是一致凸函数，因此x^*是问题的全局最小值点.
另一方面, 函数值序列$\{f(x^{(k)})\}$单调递减，因此$\{f(x^{(k)})\} \to f(x^*)$. 从而，$\{x^{(k)}\}$的每一个极限点处的函数值都是$f(x^*)$，即$f$的全局最小值. 即$\{x^{(k)}\}$的每一个极限点都是$f$的全局最小值点. 但$f$的全局最小值点唯一，因此，$\{x^{(k)}\} \to x^*$. □

下降算法产生的聚点满足 $\lim_{k\to\infty}\|\nabla f(x_k)\|=0$ ，　该聚点可能不是最小值点，也可能不是最大值点，但最差是鞍点。



为求全局最优值点，可取多个初始点进行计算，最后取结果中目标函数值最小的当做全局最小值点。

四、下降算法的全局收敛性和超线性收敛性

现在我们来研究下降算法2.1的全局收敛性和超线性收敛速度.

记向量d_k和$-\nabla f(x_k)$的夹角为θ_k, 则有

$$\cos\theta_k = \frac{-\nabla f(x_k)^{\mathrm{T}} d_k}{\|\nabla f(x_k)\| \, \|d_k\|} \tag{2.12}$$

从下面的定理可以看到，下降算法的收敛性与夹角θ_k有重要的关系:

首先我们给出下面的基本假设:

假设2.4.1 函数$f(x)$连续可微有下界且$\nabla f(x)$Lipschitz 连续, 即存在常数$L>0$, 使得

$$\|\nabla f(x)-\nabla f(y)\| \le L\|x-y\|, \quad \forall x, y \in R^n$$

定理2.4.1-2.4.2 设假设2.4.1成立，序列$\{x_k\}$由算法2.1产生，其中步长α_k由精确搜索或Wolfe-Powell搜索产生，则

$$\sum_{k=0}^{\infty} \|\nabla f(x_k)\|^2 \cos^2\theta_k < +\infty \qquad (2.13)$$

特别地，若存在常数$\delta > 0$使得$\cos\theta_k \geq \delta$，则

$$\lim_{k\to\infty} \|\nabla f(x_k)\| = 0 \qquad (2.14)$$

如果(2.14)成立，我们认为算法是全局收敛的.

该定理表明：算法的收敛性与夹角有关

由于Wolfe-Powell包含精确搜索，我们只需证明定理对Wolfe-Powell搜索成立

证明： 由Wolfe - Powell搜索条件及假设，我们有

$$-(1-\sigma_2)\nabla f(x_k)^T d_k \le [\nabla f(x_k+\alpha_k d_k)-\nabla f(x_k)]^T d_k$$
$$\le \alpha_k L \| d_k \|^2$$

即可得

$$\alpha_k \ge -\frac{1-\sigma_2}{L}\frac{\nabla f(x_k)^T d_k}{\| d_k \|^2} = -c_1 \frac{\nabla f(x_k)^T d_k}{\| d_k \|^2} \qquad (2.15)$$

这里$c_1 = \dfrac{1-\sigma_2}{L}$. 进一步由Wolfe - Powell搜索第一个条件，

得

$$f(x_k+\alpha_k d_k) - f(x_k) \le -\sigma_1 c_1 \frac{[\nabla f(x_k)^T d_k]^2}{\| d_k \|^2}$$
$$= -\sigma_1 c_1 \| \nabla f(x_k) \|^2 \cos^2 \theta_k \qquad (2.16)$$

将上面的不等式左右两边从$k=0$到∞相加，并注意到f有下界得(2.13)成立. 由无穷级数收敛的必要条件即可得(2.14)成立

定理2.4.3 设假设2.4.1成立，序列$\{x_k\}$由算法2.1产生，其中步长α_{k}由Armijo 搜索产生，且存在常数$C>0$，使得

$$\|\nabla f(x_k)\| \le C\|d_k\| \tag{2.17}$$

则定理2.4.1的结论成立

定理2.4.4 设假设2.4.1成立,序列$\{x_k\}$由算法2.1产生,其中步长α_{k}由精确搜索或Wolfe-Powell搜索或Armijo搜索确定且(2.17)成立. 若进一步

$$\liminf_{\text{k}\to\infty} \cos\theta_k > 0 \tag{2.18}$$

则

$$\liminf_{\text{k}\to\infty} \|\nabla f(x_k)\| = 0 \tag{2.19}$$

特别地,（2.19）成立，若存在常数$\eta>0$,使得

$$\prod_{i=0}^{k-1} \cos\theta_i \ge \eta^k \tag{2.20}$$

练习

- 习题2:

作　业: 5，9, 17， 22,

思考题: 16, 18,19

- 14(用最速下降法求解无约束问题的最优解,其中步长分别用Armijo型和Wolfe-Powell型线性搜索法确定.**请编写程序实现算法，分析参数对算法收敛性的影响.)**

课后：预习下降算法的收敛速度定理的证明。