最优化算法

主讲教师：董庆兴

华中师范大学 信息管理学院
qxdong@mail.ccnu.edu.cn

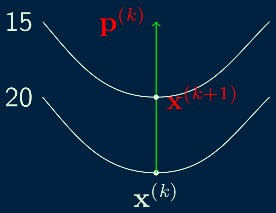
2017 年 12 月 5 日

大纲

1. 梯度法

2. 牛顿法

下降算法



- 假定已经迭代到 $\mathbf{x}^{(k)}$，如果此时没有下降方向（沿任何方向移动都无法使目标函数值减小），则 $\mathbf{x}^{(k)}$ 是一个局部极小点，迭代停止
- 如果从 $\mathbf{x}^{(k)}$ 出发至少有一个方向是下降方向 $\mathbf{p}^{(k)}$，则沿该方向迈进适当一步，得到下一个迭代点 $\mathbf{x}^{(k+1)}$ 并使得 $f(\mathbf{x}^{(k+1)}) < f(\mathbf{x}^{(k)})$
- 相当于在射线 $\mathbf{x} = \mathbf{x}^{(k)} + \lambda\mathbf{p}^{(k)}$ 上选定新点 $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \lambda_k\mathbf{p}^{(k)}$，其中 λ_k 叫做步长因子,$\mathbf{p}^{(k)}$ 为搜索方向

下降方向

下降方向

假设 $f:\mathbb{R}^n \to \mathbb{R}$，对于 $\mathbf{x} \in \mathbf{dom}\ f$，使得对于任意 $\bar{\alpha} > 0, \mathbf{d} \in \mathbb{R}^n$, 有 $f(\mathbf{x}+\alpha\mathbf{d}) < f(\mathbf{x}), \alpha \in (0, \bar{\alpha})$ 则 $\mathbf{d}$ 为 f 的一个下降方向

由泰勒展开可知 $f(\mathbf{x}+\alpha\mathbf{d}) = f(\mathbf{x}) + \alpha\nabla f(\mathbf{x})^{\mathrm{T}}\mathbf{d} + o(\alpha)$，因此满足 $\nabla f(\mathbf{x})^{\mathrm{T}}\mathbf{d} < 0$ 的 $\mathbf{d}$ 为 f 的一个下降方向

可行方向

假设 $f:\mathbb{R}^n \to \mathbb{R}$，对于 $\mathbf{x} \in \mathbf{dom}\ f$，若存在 $\alpha > 0, \mathbf{d} \in \mathbb{R}^n$, 使得 $f(\mathbf{x}+\alpha\mathbf{d}) \in \mathbf{dom}\ f$ 则 $\mathbf{d}$ 为 f 的一个可行方向

无约束最优化问题

$$
\begin{aligned}
\text{minimize} \quad & f(\mathbf{x}) \\
\text{subject to} \quad & \mathbf{x} \in \mathbb{R}^n
\end{aligned}
\tag{1}
$$

- 假定上述无约束最优化问题中，目标函数 $f(\mathbf{x})$ 有一阶连续偏导数，具有极小点 $\mathbf{x}^*$, 以 $\mathbf{x}^{(k)}$ 表示下降算法中极小点的第 k 次近似
- 则为了求其第 $k+1$ 次近似点 $\mathbf{x}^{(k+1)}$，需要沿某下降方向 $\mathbf{d}^{(k)}$ 取

$$
\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \alpha \mathbf{d}^{(k)}
$$

- 假设 $\mathbf{d}^{(k)}$ 的模一定（且不为 0），并设 $\nabla f(\mathbf{x}^{(k)}) \neq 0$（否则 $\mathbf{x}^{(k)}$ 是平稳点），下降方向 $\mathbf{d}^{(k)}$ 有无穷多个，那么该如何选择?

确定搜索方向

- 可知

$$f(\mathbf{x}^{(k)}+\alpha\mathbf{d}^{(k)})=f(\mathbf{x}^{(k)})+\alpha\nabla f(\mathbf{x}^{(k)})^{\mathrm{T}}\mathbf{d}^{(k)}+o(\alpha) \tag{2}$$

- 因此一个自然的选择是选择使得目标值得到尽可能大改善的方向，也就是使得 $\nabla f(\mathbf{x}^{(k)})^{\mathrm{T}}\mathbf{d}^{(k)}$ 最小的方向
- 展开可得

$$\nabla f(\mathbf{x}^{(k)})^{\mathrm{T}}\mathbf{d}^{(k)}=\|\nabla f(\mathbf{x}^{(k)})^{\mathrm{T}}\|\cdot\|\mathbf{d}^{(k)}\|\cos\left(\nabla f(\mathbf{x}^{(k)})^{\mathrm{T}},\mathbf{d}^{(k)}\right) \tag{3}$$

- 从而可知当 $\cos\left(\nabla f(\mathbf{x}^{(k)})^{\mathrm{T}},\mathbf{d}^{(k)}\right)=-1$ 时，式（3）取最小值。此时 $\angle\left(\nabla f(\mathbf{x}^{(k)})^{\mathrm{T}},\mathbf{d}^{(k)}\right)=180^\circ$
- 我们称这一方向 $\mathbf{d}^{(k)}=-\nabla f(\mathbf{x}^{(k)})^{\mathrm{T}}$ 为负梯度方向

确定搜索步长：近似最佳步长

- 由负梯度方向确定搜索方向之后，接下来就要确定搜索步长
- 若 $f(\mathbf{x})$ 有二阶连续偏导数，在 $\mathbf{x}^{(k)}$ 作 $f(\mathbf{x}^{(k)}-\lambda\nabla f(\mathbf{x}^{(k)})^{\mathrm{T}})$ 泰勒展开

$$f(\mathbf{x}^{(k)}-\lambda\nabla f(\mathbf{x}^{(k)})^{\mathrm{T}}) \approx f(\mathbf{x}^{(k)})-\nabla f(\mathbf{x}^{(k)})^{\mathrm{T}}\lambda\nabla f(\mathbf{x}^{(k)})+\frac{1}{2}\lambda\nabla f(\mathbf{x}^{(k)})^{\mathrm{T}}\mathbf{H}(\mathbf{x}^{(k)})\lambda\nabla f(\mathbf{x}^{(k)})$$

- 对 λ 求导并令其等于 0，可得近似最佳步长

$$\lambda_k=\frac{\nabla f(\mathbf{x}^{(k)})^{\mathrm{T}}\nabla f(\mathbf{x}^{(k)})}{\nabla f(\mathbf{x}^{(k)})^{\mathrm{T}}\mathbf{H}(\mathbf{x}^{(k)})\nabla f(\mathbf{x}^{(k)})} \tag{4}$$

- 有时将搜索方向规格化为 $\mathbf{d}^{(k)}=\frac{-\nabla f(\mathbf{x}^{(k)})^{\mathrm{T}}}{\|\nabla f(\mathbf{x}^{(k)})^{\mathrm{T}}\|}$，则规格化的步长为

$$\lambda_k=\frac{\nabla f(\mathbf{x}^{(k)})^{\mathrm{T}}\nabla f(\mathbf{x}^{(k)})\|\nabla f(\mathbf{x}^{(k)})^{\mathrm{T}}\|}{\nabla f(\mathbf{x}^{(k)})^{\mathrm{T}}\mathbf{H}(\mathbf{x}^{(k)})\nabla f(\mathbf{x}^{(k)})} \tag{5}$$

近似最佳步长法例题

近似最佳步长法例题法例题

取初始点 $\mathbf{x}^{(0)}=(0,0)^{\mathrm{T}},\epsilon=0.1$。采用近似最佳步长法求解下面的最优化问题

$\min f(\mathbf{x})=(x_1-1)^2+(x_2-1)^2$

- 极值点为 $(1,1)^{\mathrm{T}}$，有

$$\mathbf{d}^{(k)}=-\nabla f(\mathbf{x}^{(k)})=-\left(2(x_1^{(k)}-1),2(x_2^{(k)}-1)\right)^{\mathrm{T}},\mathbf{H}(\mathbf{x}^{(k)})=\begin{pmatrix}2&0\\0&2\end{pmatrix}$$

- 有 $\|\nabla f(\mathbf{x}^{(0)})\|=\sqrt{(-2)^2+(-2)^2}=2\sqrt{2}>\epsilon$, 从而可得
- 在 $\mathbf{x}^{(k)}$ 作 $f(\mathbf{x}^{(k)}-\lambda\nabla f(\mathbf{x}^{(k)})^{\mathrm{T}})$ 泰勒展开并对 λ 求导并令其等于 0 可得

$$\lambda_0=\frac{\nabla f(\mathbf{x}^{(0)})^{\mathrm{T}}\nabla f(\mathbf{x}^{(0)})}{\nabla f(\mathbf{x}^{(0)})^{\mathrm{T}}\mathbf{H}(\mathbf{x}^{(0)})\nabla f(\mathbf{x}^{(0)})}=\frac{(-2,-2)\begin{pmatrix}-2\\-2\end{pmatrix}}{(-2,-2)\begin{pmatrix}2&0\\0&2\end{pmatrix}\begin{pmatrix}-2\\-2\end{pmatrix}}=\frac{8}{16}=\frac{1}{2}$$

近似最佳步长法例题

- $\mathbf{x}^{(1)} = \mathbf{x}^{(0)} + \lambda_0 \mathbf{d}^{(0)} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} - \frac{1}{2} \begin{pmatrix} -2 \\ -2 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$
- 有 $\|\nabla f(\mathbf{x}^{(1)})\| = \sqrt{0^2 + 0^2} = 0 < \epsilon$，所以 $\mathbf{x}^{(1)}$ 即为极小点
- 由本例可知，对于等值线是圆的问题来说，不管初始点在哪里，负梯度方向总是指向圆心，因此一次迭代即可得到最小值点

确定搜索步长：最速下降法

- 由负梯度方向$\mathbf{d}^{(k)} = -\nabla f(\mathbf{x}^{(k)})^{\mathrm{T}}$ 确定搜索方向之后，接下来就要确定搜索步长
- 最速下降法的搜索步长采用作线性搜索或者一维搜索，这样确定的步长即为最优步长
- 就是使得目标函数值下降最多的 λ_k，也就是
  $\lambda_k = \arg\min f(\mathbf{x}^{(k)} - \lambda \nabla f(\mathbf{x})^{\mathrm{T}})$
- 求解以 λ 为变量的一元函数 $\phi(\lambda) = f(\mathbf{x}^{(k)} - \lambda \nabla f(\mathbf{x})^{\mathrm{T}})$ 的极小点 λ_k

最速下降法步骤

1. 给定初始点 $\mathbf{x}^{(0)} \in \mathbb{R}^n$，精度 $\epsilon > 0$，$k \leftarrow 0$
2. 若 $\|\nabla f(\mathbf{x}^{(k)})\| \leq \epsilon$，则算法终止，得到解 $\mathbf{x}^{(k)}$。否计算 $\mathbf{d}^{(k)} = -\nabla f(\mathbf{x}^{(k)})$
3. 由线性搜索确定步长 $\lambda_k = \arg\min f(\mathbf{x}^{(k)} - \lambda \nabla f(\mathbf{x}^{(k)})^{\mathrm{T}})$
4. 令 $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \lambda_k \mathbf{d}^{(k)}, k = k + 1$。转步骤 2

最速下降法例题

最速下降法例题

取初始点 $\mathbf{x}^{(0)}=(2,1)^{\mathrm{T}}, \epsilon=0.01$。采用最速下降法求解下面的最优化问题

$\min f(\mathbf{x})=\frac{1}{2}x_1^2+x_2^2$

- 求导计算可得，极值点为 $(0,0)^{\mathrm{T}}$，直接计算可得 $\mathbf{d}^{(k)}=-\nabla f(\mathbf{x}^{(k)})=-(x_1^{(k)},2x_2^{(k)})^{\mathrm{T}}$
- 有 $\|\nabla f(\mathbf{x}^{(0)})\|=\sqrt{2^2+2^2}=2\sqrt{2}>\epsilon$
- 令 $\phi(\lambda)=f(\mathbf{x}^{(k)}+\lambda\nabla f(\mathbf{x})^{\mathrm{T}})=\frac{1}{2}(x_1^{(k)}+\lambda d_1^{(k)})^2+(x_2^{(k)}+\lambda d_2^{(k)})^2$。从而令 $\phi'(\lambda)=0$，可得

$$\lambda_k=-\frac{x_1^{(k)}d_1^{(k)}+2x_2^{(k)}d_2^{(k)}}{\left(d_1^{(k)}\right)^2+2\left(d_2^{(k)}\right)^2}=\frac{\left(x_1^{(k)}\right)^2+4\left(x_2^{(k)}\right)^2}{\left(x_1^{(k)}\right)^2+8\left(x_2^{(k)}\right)^2}$$

- $\lambda_0=\frac{2}{3}$

最速下降法例题

- 从而可以计算 $\mathbf{x}^{(k+1)}$ 有

$$
\begin{aligned}
\mathbf{x}^{(k+1)} &= \mathbf{x}^{(k)} + \lambda_k \mathbf{d}^{(k)} \\
&= \mathbf{x}^{(k)} - \frac{\left(x_1^{(k)}\right)^2 + 4\left(x_1^{(k)}\right)^2}{\left(x_2^{(k)}\right)^2 + 8\left(x_2^{(k)}\right)^2} \nabla f(\mathbf{x}^{(k)}) \\
&= (x_1^{(k)}, x_2^{(k)})^{\mathrm{T}} - \frac{\left(x_1^{(k)}\right)^2 + 4\left(x_2^{(k)}\right)^2}{\left(x_1^{(k)}\right)^2 + 8\left(x_2^{(k)}\right)^2} \left(x_1^{(k)}, 2x_2^{(k)}\right)^{\mathrm{T}} \\
&= \left(\frac{4}{t_k^2+8} x_1^{(k)}, -\frac{t_k^2}{t_k^2+8} x_2^{(k)}\right)^{\mathrm{T}}
\end{aligned}
$$

其中 $t_k = \left|\frac{x_1^{(k)}}{x_2^{(k)}}\right|$

- 从而有 $\mathbf{x}^{(1)} = (\frac{2}{3}, -\frac{1}{3})$,$\mathbf{x}^{(2)} = (\frac{2}{9}, \frac{1}{9})$,$\mathbf{x}^{(3)} = (\frac{2}{27}, -\frac{1}{27}) \cdots$
- 观察可得通项公式

$$
\mathbf{x}^{(k)} = \left(\frac{1}{3}\right)^k \begin{pmatrix} 2 \\ (-1)^k \end{pmatrix}, k = 0, 1, \cdots
$$

最速下降法例题

- 由通项公式

$$\mathbf{x}^{(k)} = \left(\frac{1}{3}\right)^k \begin{pmatrix} 2 \\ (-1)^k \end{pmatrix}, k = 0, 1, \cdots$$

- 可知 $\{\mathbf{x}^{(k)}\} \to \mathbf{x}^* = (0, 0)^{\mathrm{T}}$，也就是算法产生的点列收敛于问题的解

梯度下降法的性质

- 如果目标函数等值线为同心圆或同心球面，则负梯度方向指向圆心或球心，因此从任意初始点出发，沿最速下降方向一步可达极小值点
- 由于负梯度方向的最速下降性，很容易使人们认为负梯度方向是理想的搜索方向，要指出的是 $\mathbf{x}$ 处的负梯度方向 $-\nabla f(\mathbf{x})$ 仅仅是 $\mathbf{x}$ 点附近才有最速下降性质，而对于整个极小化过程则未必
- 例如对于一般的二元二次函数而言，其等值线为椭圆，最速下降法趋近极小点时，其搜索路径呈直角锯齿状
- 因此在实际问题中，常将梯度法和别的方法联合起来用，在前期用梯度法，接近极小点时换别的收敛快的方法

牛顿法简介

- 牛顿法（Newton's method）又称为牛顿-拉弗森方法（Newton-Raphson method），它是一种在实数域和复数域上近似求解方程的方法。方法使用函数 $f(x)$ 的泰勒级数的前面几项来寻找方程 $f(x) = 0$ 的根
- 牛顿法最初由艾萨克 · 牛顿 (1643-1727) 在《流数法》（Method of Fluxions，1671 年完成，在牛顿死后的 1736 年公开发表）。约瑟夫 · 拉弗森也曾于 1690 年在 Analysis Aequationum 中提出此方法
- 在数值分析领域，牛顿法是一种通过迭代求解可微函数的根的方法。在优化领域，牛顿法用来给二次可微函数寻找其一阶微分等于零的驻点

牛顿法理论基础

- 若 $f(\mathbf{x})$ 有二阶连续偏导数，$\mathbf{x}^{(k)}$ 为其极小点的某一近似，作 $f(\mathbf{x}^{(k)})$ 的二阶泰勒展开

$$f(\mathbf{x}) \approx f(\mathbf{x}^{(k)}) + \nabla f(\mathbf{x}^{(k)})^{\mathrm{T}}\Delta(\mathbf{x}) + \frac{1}{2}\Delta(\mathbf{x})^{\mathrm{T}}\mathbf{H}(\mathbf{x}^{(k)})\Delta(\mathbf{x}) \tag{6}$$

  其中，$\Delta(\mathbf{x}) = \mathbf{x} - \mathbf{x}^{(k)}$

- 极值点梯度满足

$$\nabla f(\mathbf{x}) \approx \nabla f(\mathbf{x}^{(k)}) + \mathbf{H}(\mathbf{x}^{(k)})\Delta(\mathbf{x}) = 0 \tag{7}$$

- 从而

$$\mathbf{x} = \mathbf{x}^{(k)} - \mathbf{H}(\mathbf{x}^{(k)})^{-1}\nabla f(\mathbf{x}^{(k)}) \tag{8}$$

- 当 $f(\mathbf{x})$ 为二次函数，则式（6）的逼近是准确的，从任一点 $\mathbf{x}^{(k)}$ 出发，由式（8）只需一步即可求出 $f(\mathbf{x})$ 极小点

牛顿方向

- 当 $f(\mathbf{x})$ 不是二次函数时，则式（6）的逼近仅仅是近似表达式，由式（8）出发求出得到的极小点只是 $f(\mathbf{x})$ 极小点的近似
- 此时，人们常取 $\mathbf{d}^{(k)} = -\mathbf{H}(\mathbf{x}^{(k)})^{-1}\nabla f(\mathbf{x}^{(k)})$ 为搜索方向，这一方向就是牛顿方向
- 从而此时有 $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \lambda\mathbf{d}^{(k)}$
- 搜索步长可以采用最佳步长：$\lambda_k = \arg\min f(\mathbf{x}^{(k)} + \lambda\mathbf{d}^{(k)})$

牛顿法步骤

1. 给定初始点 $\mathbf{x}^{(0)} \in \mathbb{R}^n$，精度 $\epsilon > 0$，$k \leftarrow 0$
2. 若 $\|\nabla f(\mathbf{x}^{(k)})\| \leq \epsilon$，则算法终止，得到解 $\mathbf{x}^{(k)}$。否则计算 $\nabla f(\mathbf{x}^{(k)}) + \mathbf{H}(\mathbf{x}^{(k)})\mathbf{d}^{(k)} = 0$，求得 $\mathbf{d}^{(k)} = -\mathbf{H}(\mathbf{x}^{(k)})^{-1}\nabla f(\mathbf{x}^{(k)})$
3. 由线性搜索确定步长 $\lambda_k = \arg\min f(\mathbf{x}^{(k)} - \lambda \mathbf{H}(\mathbf{x}^{(k)})^{-1}\nabla f(\mathbf{x}^{(k)}))$
4. 令 $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \lambda_k \mathbf{d}^{(k)}, k = k+1$。转步骤 2

牛顿法例题

牛顿法例题

取初始点 $\mathbf{x}^{(0)} = (0,0)^{\mathrm{T}}$ 和 $(1,1)^{\mathrm{T}}, \epsilon = 0.01$。采用牛顿法求解下面的最优化问题

$\min f(\mathbf{x}) = \frac{1}{2}x_1^2 + x_2^2 - x_1x_2 - x_1$

- 求导计算可得，极值点为 $(2,1)^{\mathrm{T}}$，直接计算可得

$$\nabla f(\mathbf{x}^{(k)}) = \begin{pmatrix} x_1^{(k)} - x_2^{(k)} - 1 \\ -x_1^{(k)} + 2x_2^{(k)} \end{pmatrix}, \mathbf{H}(\mathbf{x}^{(k)}) = \begin{pmatrix} 1 & -1 \\ -1 & 2 \end{pmatrix}$$

-

$$\mathbf{d}^{(k)} = -\mathbf{H}(\mathbf{x}^{(k)})^{-1}\nabla f(\mathbf{x}^{(k)}) = -\begin{pmatrix} 1 & -1 \\ -1 & 2 \end{pmatrix}^{-1} \begin{pmatrix} x_1^{(k)} - x_2^{(k)} - 1 \\ -x_1^{(k)} + 2x_2^{(k)} \end{pmatrix}$$
$$= -\begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} x_1^{(k)} - x_2^{(k)} - 1 \\ -x_1^{(k)} + 2x_2^{(k)} \end{pmatrix} = -\begin{pmatrix} x_1^{(k)} - 2 \\ -x_2^{(k)} - 2 \end{pmatrix}$$

牛顿法例题

- 计算最佳步长，有

$$
\begin{aligned}
\phi(\lambda) &= f(\mathbf{x}^{(k)} + \lambda_k \mathbf{d}^{(k)}) \\
&= \frac{1}{2}(x_1^{(k)} + \lambda d_1^{(k)})^2 + (x_2^{(k)} + \lambda d_2^{(k)})^2 \\
&\quad - (x_1^{(k)} + \lambda d_1^{(k)})(x_2^{(k)} + \lambda d_2^{(k)}) - (x_1^{(k)} + \lambda d_1^{(k)})
\end{aligned}
$$

- 令 $\phi'(\lambda) = 0$，可得

$$
\lambda_k = \frac{(x_1^{(k)} - x_2^{(k)} - 1)\, d_1^{(k)} + (-x_1^{(k)} + 2x_2^{(k)})\, d_2^{(k)}}{\left(d_1^{(k)}\right)^2 + 2\left(d_2^{(k)}\right) - 2d_1^{(k)}\, d_2^{(k)}}
$$

牛顿法例题

$\mathbf{x}^{(0)}$	k	$\mathbf{x}^{(k)}$	$f(\mathbf{x}^{(k)})$	$\nabla f(\mathbf{x}^{(k)})$	$\mathbf{d}^{(k)}$	λ_k
$(0,0)^{\mathrm{T}}$	0	$(0,0)^{\mathrm{T}}$	0	$(-1,0)^{\mathrm{T}}$	$(2,1)^{\mathrm{T}}$	1
	1	$(2,1)^{\mathrm{T}}$	-1	$(0,0)^{\mathrm{T}}$		
$(1,1)^{\mathrm{T}}$	0	$(1,1)^{\mathrm{T}}$	$-\frac{1}{2}$	$(-1,1)^{\mathrm{T}}$	$(1,0)^{\mathrm{T}}$	1
	1	$(2,1)^{\mathrm{T}}$	-1	$(0,0)^{\mathrm{T}}$		

牛顿法性质

- 对于
$$f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^{\mathrm{T}}\mathbf{Q}\mathbf{x} + \mathbf{q}^{\mathrm{T}}\mathbf{x}$$
其中 $\mathbf{Q}$ 对称正定，则从任意初始点 $\mathbf{x}^{(0)}$ 出发，均最多经过一次迭代即可达到极小值点。
- 若一个算法用于求解严格凸二次函数极小值问题时，从任意初始点出发，算法经有限次迭代后可达最小值点，则称算法具有二次终止性。
- 牛顿法具有二次终止性，由例题可知，最速下降法不满足
- 牛顿法的优点是二次收敛性，比最速下降法更快，具有更好的全局判断力