深度之眼
deepshare.net

数学基础—最优化基础

导师：Johnson

主要内容

主要内容

- **概率论**
  - 概率基础
    - 1.随机试验样本空间和随机事件
    - 2.概率的定义
    - 3.条件概率与乘法公式
    - 4.全概率公式与贝叶斯公式以及应用
    - 5.独立性
  - 随机变量与多维随机变量
    - 1.离散随机变量
    - 2.连续随机变量
    - 3.多维随机变量
    - 4.边缘分布和条件分布
    - 5.独立性
  - 期望与方差
    - 1.期望与方差的定义
    - 2.常用随机变量期望与方差的计算
    - 3.期望与方差的性质
    - 4.协方差与协方差矩阵
  - 参数的估计
    - 1.矩估计
    - 2.极大似然估计
    - 3.最大后验估计
- **最优化理论**
  - 无约束最优化
    - 1.无约束最优化的定义
    - 2.梯度下降，全批量梯度下降，随机梯度下降，小批量梯度下降
    - 3.梯度下降在深度学习中的应用
    - 4.牛顿下降法，两种角度的解释，牛顿法的二次收敛证明
  - 约束最优化
    - 1.等式约束的最优化
    - 2.不等式约束的最优化

主要内容

- 概率论
  - 概率基础
    - 1.随机试验样本空间和随机事件
    - 2.概率的定义
    - 3.条件概率与乘法公式
    - 4.全概率公式与贝叶斯公式以及应用
    - 5.独立性
  - 随机变量与多维随机变量
    - 1.离散随机变量
    - 2.连续随机变量
    - 3.多维随机变量
    - 4.边缘分布和条件分布
    - 5.独立性
  - 期望与方差
    - 1.期望与方差的定义
    - 2.常用随机变量期望与方差的计算
    - 3.期望与方差的性质
    - 4.协方差与协方差矩阵
  - 参数的估计
    - 1.矩估计
    - 2.极大似然估计
    - 3.最大后验估计
- 最优化理论
  - 无约束最优化
    - 1.无约束最优化的定义
    - 2.梯度下降，全批量梯度下降，随机梯度下降，小批量梯度下降
    - 3.梯度下降在深度学习中的应用
    - 4.牛顿下降法，两种角度的解释，牛顿法的二次收敛证明
  - 约束最优化
    - 1.等式约束的最优化
    - 2.不等式约束的最优化

主要内容

- **概率论**
  - 概率基础
    - 1.随机试验样本空间和随机事件
    - 2.概率的定义
    - 3.条件概率与乘法公式
    - 4.全概率公式与贝叶斯公式以及应用
    - 5.独立性
  - 随机变量与多维随机变量
    - 1.离散随机变量
    - 2.连续随机变量
    - 3.多维随机变量
    - 4.边缘分布和条件分布
    - 5.独立性
  - 期望与方差
    - 1.期望与方差的定义
    - 2.常用随机变量期望与方差的计算
    - 3.期望与方差的性质
    - 4.协方差与协方差矩阵
  - 参数的估计
    - 1.矩估计
    - 2.极大似然估计
    - 3.最大后验估计
- **最优化理论**
  - 无约束最优化
    - 1.无约束最优化的定义
    - 2.梯度下降，全批量梯度下降，随机梯度下降，小批量梯度下降
    - 3.梯度下降在深度学习中的应用
    - 4.牛顿下降法，两种角度的解释，牛顿法的二次收敛证明
  - 约束最优化
    - 1.等式约束的最优化
    - 2.不等式约束的最优化

无约束最优化

无约束最优化的定义

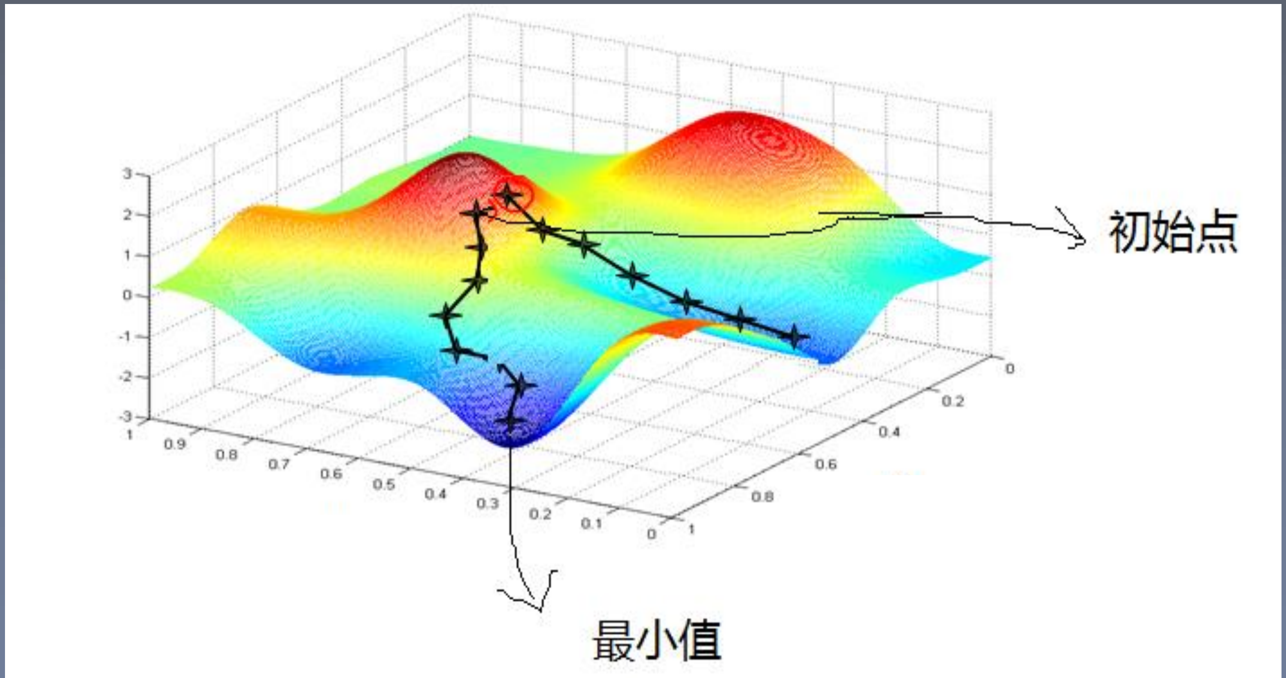
深度之眼
deepshare.net

无约束优化问题是机器学习中最普遍、最简单的优化问题。

$$x^* = \min_x f(x), x \in R^n$$

无约束最优化

梯度下降法



$$J(x,y)$$

$$t=0,(x_0,y_0)$$

$$t=1,(x_1,y_1)\quad \begin{bmatrix}x_1\\y_1\end{bmatrix}=\begin{bmatrix}x_0\\y_0\end{bmatrix}-\lambda\begin{bmatrix}\frac{\partial J}{\partial x}\\\frac{\partial J}{\partial y}\end{bmatrix}\Bigg|_{\substack{x=x_0\\y=y_0}}$$

$$\cdots\cdots\cdots$$

$$t=n,(x_n,y_n)\quad \begin{bmatrix}x_n\\y_n\end{bmatrix}=\begin{bmatrix}x_{n-1}\\y_{n-1}\end{bmatrix}-\lambda\begin{bmatrix}\frac{\partial J}{\partial x}\\\frac{\partial J}{\partial y}\end{bmatrix}\Bigg|_{\substack{x=x_{n-1}\\y=y_{n-1}}}$$

λ为步长，也叫学习率，是一个超参数。

终止条件：①$|J(x_n,y_n)-J(x_{n-1},y_{n-1})|<\varepsilon$

②$n>N$(最大迭代次数)

③$\left\|\begin{bmatrix}\frac{\partial J}{\partial x}\\\frac{\partial J}{\partial y}\end{bmatrix}\Bigg|_{\substack{x=x_n\\y=y_n}}\right\|<\varepsilon$

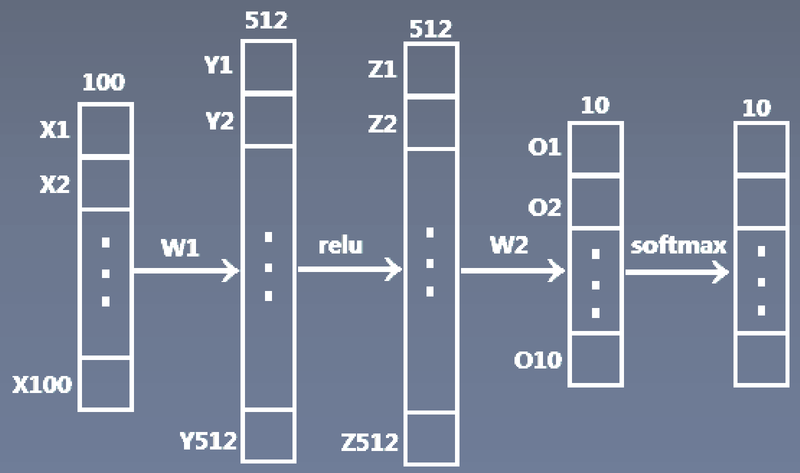
优点： 简单，计算量小

缺点： 陷入局部最优，易震荡，一阶收敛，收敛速度慢

无约束最优化

梯度下降法

梯度下降在深度学习的应用，随机梯度下降，全批量梯度下降，小批量梯度下降



设共有N个样本. $x^1, x^2, \cdots, x^N; x^i \in R^{100}$

$y^1, y^2, \cdots, y^N; y^i \in \{0,1,\cdots,9\}$

$$J(w_1, w_2) = J_1(w_1, w_2, x^1, y^1) + J_2(w_1, w_2, x^2, y^2) + \cdots + J_N(w_1, w_2, x^N, y^N)$$

$$= \sum_{i=1}^{N} J_i\left(w_1, w_2, x^i, y^i\right)$$

随机梯度下降：初始(w_1^0, w_2^0).

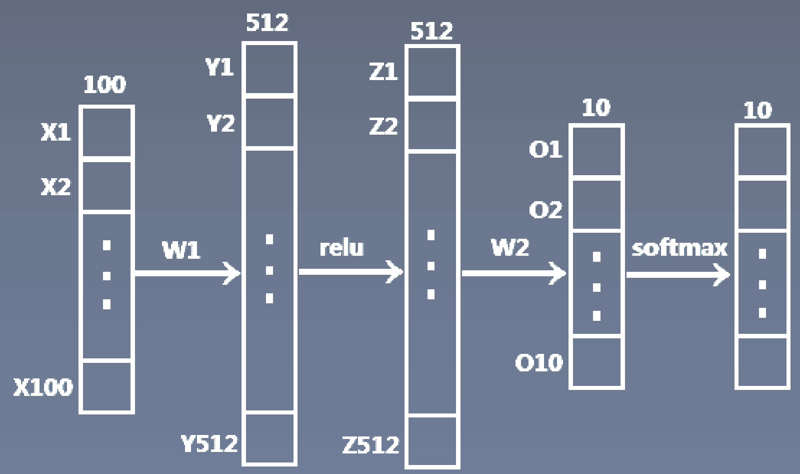
$$\begin{aligned} w_1^1 &= w_1^0 - \lambda \frac{\partial J_1}{\partial w_1}\Big|_{w_1 = w_1^0} \\ w_2^1 &= w_2^0 - \lambda \frac{\partial J_1}{\partial w_2}\Big|_{w_2 = w_2^0} \\ t &= 1 \end{aligned} \quad \Rightarrow \quad \begin{aligned} w_1^2 &= w_1^1 - \lambda \frac{\partial J_2}{\partial w_1}\Big|_{w_1 = w_1^1} \\ w_2^2 &= w_2^1 - \lambda \frac{\partial J_2}{\partial w_2}\Big|_{w_2 = w_2^1} \\ t &= 2 \end{aligned}$$

每次更新都用一个新的样本.

无约束最优化

梯度下降法

梯度下降在深度学习的应用，随机梯度下降，全批量梯度下降，小批量梯度下降



设共有N个样本. $x^1, x^2, \cdots, x^N; x^i \in R^{100}$

$y^1, y^2, \cdots, y^N; y^i \in \{0,1,\cdots,9\}$

$$J(w_1, w_2) = J_1(w_1, w_2, x^1, y^1) + J_2(w_1, w_2, x^2, y^2) + \cdots + J_N(w_1, w_2, x^N, y^N)$$

$$= \sum_{i=1}^{N} J_i\left(w_1, w_2, x^i, y^i\right)$$

全批量梯度下降：初始(w_1^0, w_2^0).

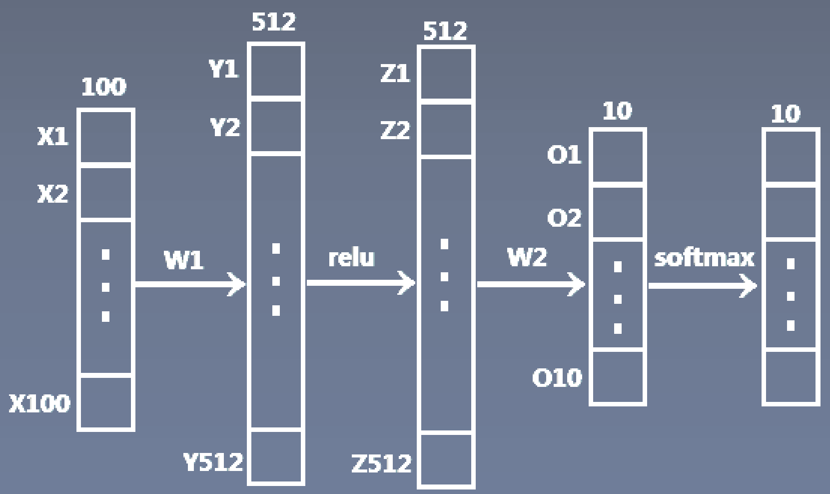
$$\begin{aligned} w_1^1 &= w_1^0 - \lambda \frac{\partial J}{\partial w_1}\Big|_{w_1 = w_1^0} \\ w_2^1 &= w_2^0 - \lambda \frac{\partial J}{\partial w_2}\Big|_{w_2 = w_2^0} \\ t &= 1 \end{aligned} \quad \Rightarrow \quad \begin{aligned} w_1^2 &= w_1^1 - \lambda \frac{\partial J}{\partial w_1}\Big|_{w_1 = w_1^1} \\ w_2^2 &= w_2^1 - \lambda \frac{\partial J}{\partial w_2}\Big|_{w_2 = w_2^1} \\ t &= 2 \end{aligned}$$

无约束最优化

深度之眼 deepshare.net

梯度下降法

梯度下降在深度学习的应用，随机梯度下降，全批量梯度下降，小批量梯度下降



设共有N个样本. $x^1, x^2, \cdots, x^N; x^i \in R^{100}$

$y^1, y^2, \cdots, y^N; y^i \in \{0,1,\cdots,9\}$

$$J(w_1, w_2) = J_1(w_1, w_2, x^1, y^1) + J_2(w_1, w_2, x^2, y^2) + \cdots + J_N(w_1, w_2, x^N, y^N)$$

$$= \sum_{i=1}^{N} J_i\left(w_1, w_2, x^i, y^i\right)$$

小批量梯度下降：初始(w_1^0, w_2^0).

$$w_1^1 = w_1^0 - \lambda \frac{\partial(J_1 + \cdots + J_m)}{\partial w_1}\bigg|_{w_1 = w_1^0} \qquad w_1^2 = w_1^1 - \lambda \frac{\partial(J_{m+1} + \cdots + J_{2m})}{\partial w_1}\bigg|_{w_1 = w_1^1}$$

$$\Rightarrow$$

$$w_2^1 = w_2^0 - \lambda \frac{\partial(J_1 + \cdots + J_m)}{\partial w_2}\bigg|_{w_2 = w_2^0} \qquad w_2^2 = w_2^1 - \lambda \frac{\partial(J_{m+1} + \cdots + J_{2m})}{\partial w_2}\bigg|_{w_2 = w_2^1}$$

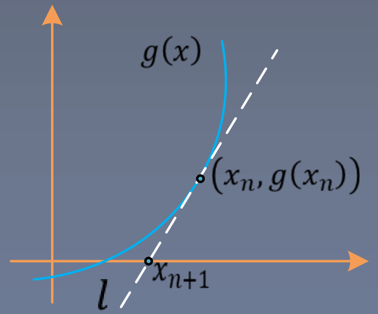
$$t = 1 \qquad\qquad t = 2$$

无约束最优化

牛顿法：两种解释

$$\min f(x)$$
$$f'(x) = 0$$

令 $g(x) = f'(x) \Rightarrow$ 求 $g(x) = 0$ 的解.



切线 $l: y - g(x_n) = g'(x_n)(x - x_n)$

令 $y = 0$

$$\Rightarrow x = x_n - \frac{g(x_n)}{g'(x_n)}$$

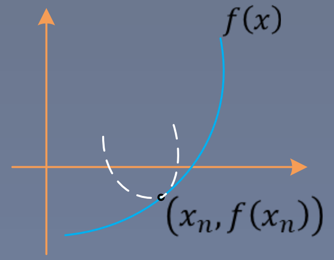
$$\Rightarrow x_{n+1} = x_n - \frac{f'(x_n)}{f''(x_n)}$$

$$f(x) = f(x_n) + f'(x_n)(x - x_n) + \frac{f''(x_n)}{2!}(x - x_n)^2 + \cdots$$

$$g(x)$$

$$g(x_n) = f(x_n)$$
$$g'(x_n) = f'(x_n)$$
$$g''(x_n) = f''(x_n)$$

$$g'(x) = f'(x_n) + f''(x_n)(x - x_n)$$
$$g''(x) = f''(x_n)$$



令 $g'(x) = 0 \Rightarrow \min g(x)$

$$f'(x_n) + f''(x_n)(x - x_n) = 0$$

$$x = x_n - \frac{f'(x_n)}{f''(x_n)} = x_{n+1}$$

推广到多元：$x_{n+1} = x_n - H^{-1}_{|x=x_n} \nabla f_{|x=x_n}$

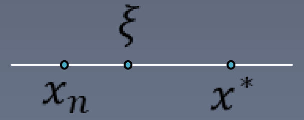
无约束最优化

牛顿法：收敛速度

$$x_{n+1} = x_n - \frac{f'(x_n)}{f''(x_n)}$$

令x^*为极值点，$f'(x^*) = 0$

$$|x_{n+1} - x^*| = \left|x_n - \frac{f'(x_n)}{f''(x_n)} - x^*\right| = \left|x_n - x^* - \frac{f'(x_n) - f'(x^*)}{f''(x_n)}\right|$$



$$= \left|x_n - x^* - \frac{(x_n - x^*)f''(\xi)}{f''(x_n)}\right| = |x_n - x^*|\left|\frac{f''(x_n) - f''(\xi)}{f''(x_n)}\right|$$

$$= |x_n - x^*|\left|\frac{(x_n - \xi)f'''(\eta)}{f''(x_n)}\right| < |x_n - x^*|^2\left|\frac{f'''(\eta)}{f''(x_n)}\right|$$

假设$f'''(x)$连续则有界

则 $\left|\frac{f'''(\eta)}{f''(x_n)}\right| < M$

所以$|x_{n+1} - x^*| < M|x_n - x^*|^2$二次收敛.

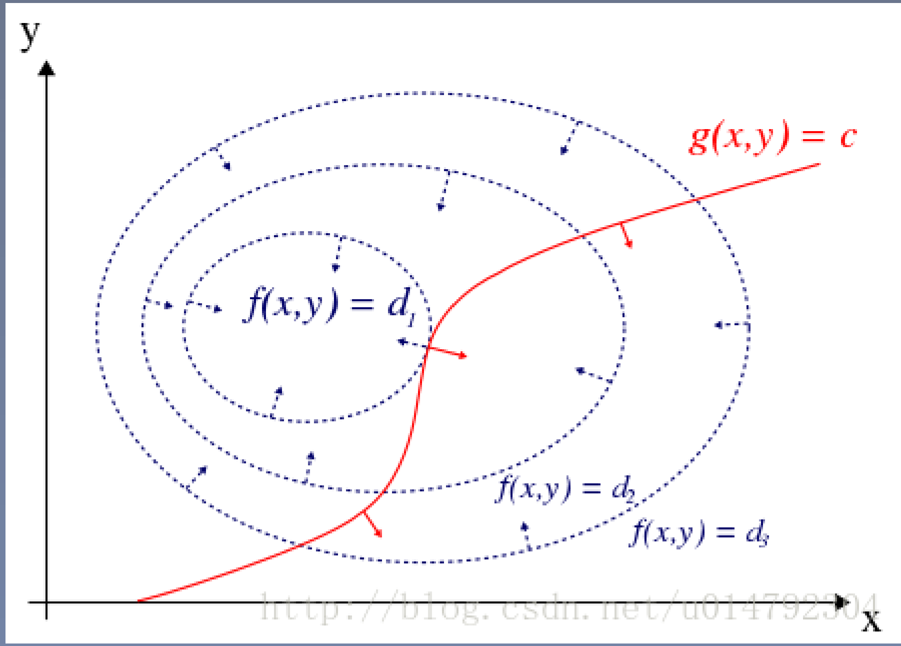
带约束最优化

等式约束

经典拉格朗日乘子法是下面的优化问题（注：x是一个向量）：

$$\min_x f(x)$$
$$s.t.\ g(x) = 0$$

直观上理解，最优解$x_{optimal}$一定有这样的性质，以x是二维变量为例：



$$\begin{cases} \nabla f(x) = \lambda \nabla g(x) \\ g(x) = 0 \end{cases}$$

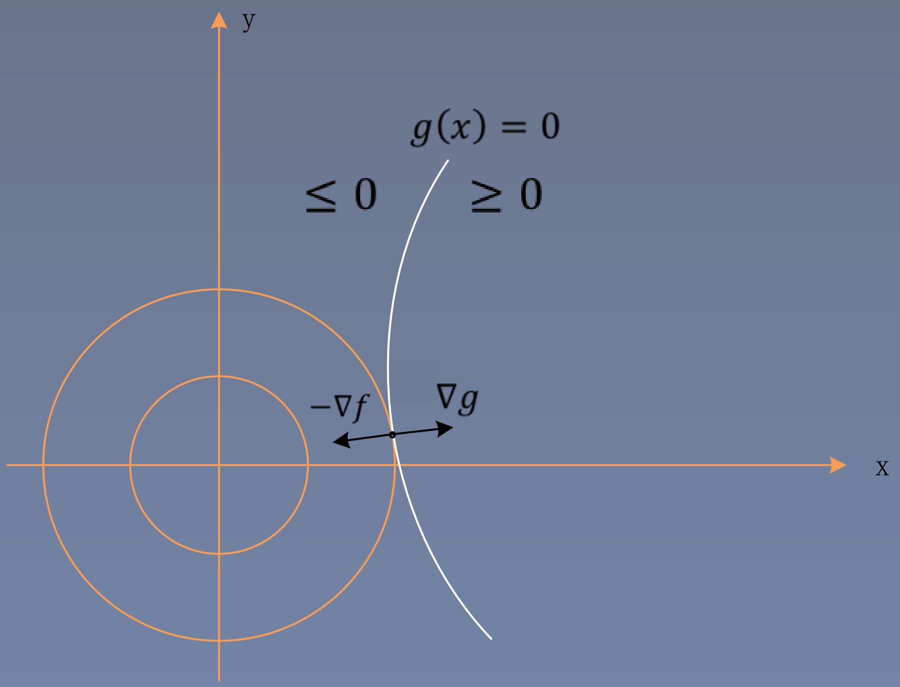
这时引入拉格朗日函数：

$$L(x, \lambda) = f(x) - \lambda g(x)$$

带约束最优化

不等式约束：形式1

$$\begin{cases} \min f(x) \\ g(x) \geq 0 \end{cases}$$



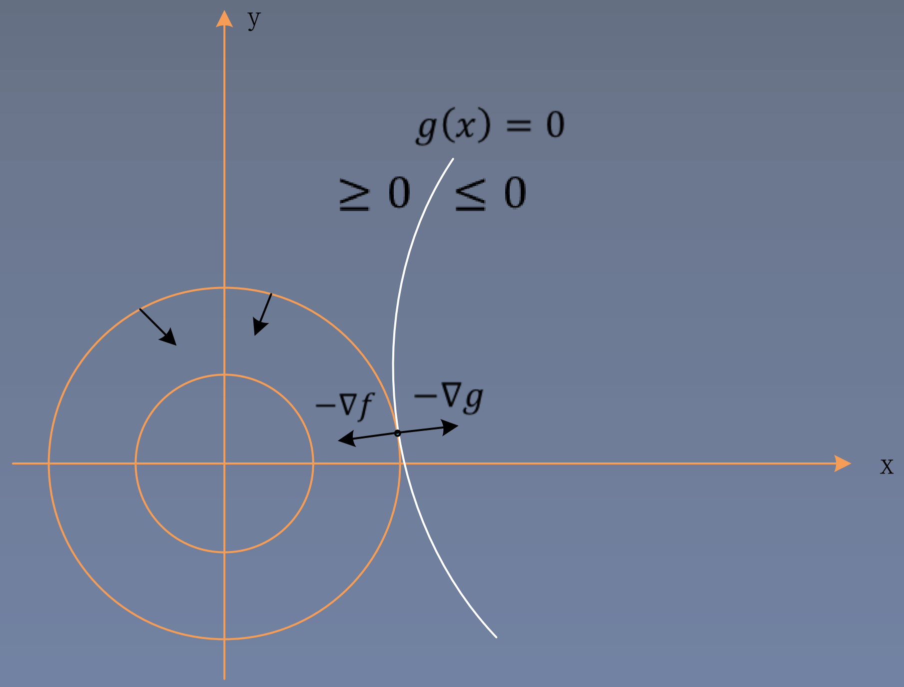
$$\nabla f = \lambda \nabla g \qquad \lambda > 0.$$

$$\begin{cases} \nabla f\big|_{x=x^*} = \lambda^* \nabla g\big|_{x=x^*} \\ \lambda^* \geq 0 \\ \lambda^* g(x^*) = 0 \end{cases}$$

带约束最优化

不等式约束：形式2

$$\begin{cases} \min f(x) \\ g(x) \le 0 \end{cases}$$



$$-\nabla f = \lambda(-\nabla g) \qquad \lambda < 0.$$

$$\begin{cases} \nabla f\big|_{x=x^*} = \lambda^* \nabla g\big|_{x=x^*} \\ \lambda^* \le 0 \\ \lambda^* g(x^*) = 0 \end{cases}$$

带约束最优化

混合问题

$$\begin{cases} \min f(x) \\ h_i(x) = 0 \qquad i = 1,2,\cdots,m; \\ g_i(x) \geq 0 \qquad i = 1,2,\cdots,n; \end{cases}$$

$$\begin{cases} \nabla f\Big|_{x^*} = \sum_{i=1}^{m} \lambda_i^* \nabla h_i\Big|_{x^*} + \sum_{i=1}^{n} \mu_i^* \nabla g_i\Big|_{x^*} \\ \mu_i^* \geq 0 \qquad i = 1,2,\cdots,n \\ h_i(x^*) = 0 \\ \mu_i^* g_i(x^*) = 0 \end{cases}$$

带约束最优化

不等式约束例子

例 求下列非线性规划问题的K-T点：

$$\min f(x) = 2x_1^2 + 2x_1x_2 + x_2^2 - 10x_1 - 10x_2;$$

$$s.t\begin{cases} x_1^2 + x_2^2 \le 5, \\ 3x_1 + x_2 \le 6. \end{cases}$$

解 将上述问题的约束条件改写为$g_i(x) \ge 0$的形式：

$$s.t\begin{cases} g_1(x) = -x_1^2 - x_2^2 + 5 \ge 0, \\ g_2(x) = -3x_1 - x_2 + 6 \ge 0. \end{cases}$$

设K-T点为$x^* = (x_1, x_2)^T$,有

$$\nabla f(x^*) = \begin{bmatrix} 4x_1 + 2x_2 - 10 \\ 2x_1 + 2x_2 - 10 \end{bmatrix},$$

$$\nabla g_1(x^*) = \begin{bmatrix} -2x_1 \\ -2x_2 \end{bmatrix},$$

$$\nabla g_2(x^*) = \begin{bmatrix} -3 \\ -1 \end{bmatrix}.$$

$$\begin{cases} 4x_1 + 2x_2 - 10 + 2\gamma_1 x_1 + 3\gamma_2 = 0, \\ 2x_1 + 2x_2 - 10 + 2\gamma_1 x_2 + \gamma_2 = 0, \\ \gamma_1(5 - x_1^2 - x_2^2) = 0, \\ \gamma_2(6 - 3x_1 - x_2) = 0, \\ \gamma_1 \ge 0, \\ \gamma_2 \ge 0. \end{cases}$$

$$\begin{cases} x_1 = 1 \\ x_2 = 2, \\ \gamma_1 = 1, \\ \gamma_2 = 0. \end{cases}$$

深度之眼
deepshare.net

联系我们:

电话：18001992849

邮箱：service@deepshare.net

Q Q： 2677693114

公众号

客服微信