优化（Optimization）

前提：$f(x)$，$g(x)$ 都必须是凸函数

(1) 无约束求解

$$\min_x f(x)$$

$\Rightarrow$ 求 $\nabla_x f(x)=0$ 即可

(2) 有约束求解

① 等式约束

$$\begin{cases} \min\limits_x f(x) & (1) \\ s.t.\ g(x)=0 & (2) \end{cases}$$

即：在满足 $g(x)=0$ 的前提下，求 $f(x)$ 的极小值

$\Downarrow$

构造拉格朗日函数（Lagrange Function），目的是将 (1) 式、(2) 式合并求解

拉格朗日乘子（Lagrange Multiplier）

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

$$\Rightarrow \begin{cases} \dfrac{\partial L(x,\lambda)}{\partial x}=\nabla_x f(x)+\lambda\nabla_x g(x)=0 \\ \dfrac{\partial L(x,\lambda)}{\partial \lambda}=\underbrace{g(x)=0}_{\text{约束条件}} \end{cases} \Rightarrow -\nabla_x f(x)=\lambda\nabla_x g(x)$$

因为我们求最小值，所以使用的是"负梯度"
（若求 $f(x)$ 的最大值，则需沿"梯度方向"变化）

$\lambda>0$

当两者的梯度平行时，才可能有共同的解

$\Downarrow$

总结为

在等式约束条件下，求解最小值问题可以转化为：

$$\begin{cases} -\nabla_x f(x)=\nabla_x g(x) \\ \lambda>0 \\ g(x)=0 \end{cases}$$

若有多个等式约束，则为

$$\begin{cases} \min\limits_x f(x) & & (1) \\ s.t.\ g_i(x)=0 & i=1,2,3,\dots,m & (2) \end{cases} \Rightarrow \begin{cases} -\nabla_x f(x)=\nabla_x g_i(x) \\ \lambda_i>0 \\ g_i(x)=0 \end{cases} \quad i=1,2,3,\cdots,m$$

$$\Rightarrow L(x,\vec{\lambda})=f(x)+\sum_{i=1}^{m}\lambda_i g_i(x)$$

② 不等式约束

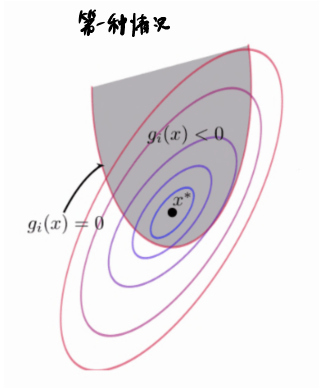
$$\begin{cases} \min f(x) \\ s.t.\ g(x) \le 0 \end{cases}$$

$\Downarrow$ 有两种情况

[第一种情况] $f(x)$ 的最小值正好落在 $g(x)$ 的约束范围内

$\Rightarrow$ 相当于没有这个约束条件

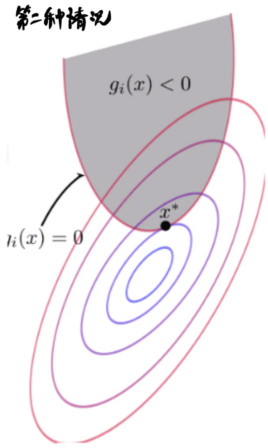
即 $$\begin{cases} -\nabla_x f(x) = 0 \\ g(x) \le 0 \end{cases}$$



[第二种情况] $f(x)$ 的最小值落在 $g(x)$ 的约束范围外面

$\Rightarrow$ 这时就变成了等式约束 $\longrightarrow$ 放弃了全局最优解，去寻找 $g(x)=0$ 条件下的解

即 $$\begin{cases} -\nabla_x f(x) = \lambda \nabla_x g(x) \\ \lambda > 0 \\ g(x) = 0 \end{cases}$$



$\Downarrow$ 将两种情况合并，便得到KKT条件
(Karush-Kuhn-Tucker conditions)

$$\begin{cases} L(x,\lambda) = f(x) + \lambda g(x) \\ s.t. \\ (1)\quad -\nabla_x f(x) = \lambda \nabla_x g(x) \\ (2)\quad \lambda \ge 0 \\ (3)\quad \lambda g(x) = 0 \\ (4)\quad g(x) \le 0 \end{cases}$$

注意：这四个条件并非要同时满足

当 $\lambda = 0 \Rightarrow \begin{cases} -\nabla_x f(x) = 0 \\ g(x) \le 0 \end{cases} \Rightarrow$ 即：第一种情况

当 $g(x) = 0 \Rightarrow \begin{cases} -\nabla_x f(x) = \nabla_x g(x) \\ \lambda > 0 \\ g(x) = 0 \end{cases} \Rightarrow$ 即：第二种情况

注 主要参考了北航秦曾昌《机器学习》，以及我自己的一些理解。可能有误。

B.O