模型备注——规划、最优化部分

2017年11月23日

目录

1 导言

本文档旨在从梳理表格中所枚举模型间的关系，主要对象是规划论中的部分模型与最优化主题中的迭代算法。

本文档的主要参考资料是Stephen Boyd编著的教材Convex Optimization。

本文档没有重点介绍凸集、凸函数等过于基础的部分，只旨在说明模型列表中一些知识点的关系。

2 凸优化理论

一般的优化问题（目标函数，不等式约束，等式约束）形式为：

minimize $f_0(x)$

subject to $f_i(x) \leq b_i, i = 1, .., m$

subject to $f_i(x) = b_i, i = m, .., m + p$

在凸优化中，我们限定$f_i, i = 0, ...m$为凸函数，而等式约束是仿射函数$Ax = b$。

2.1 线性规划

线性规划是凸优化的特例，因为仿射函数是凸函数：

minimize $c^T x + d$

subject to $Gx \leq h$

subject to $Ax = b$

其中不等式约束条件为广义的（定义在非负锥上），即对于所有分量成立。

2.2 二次规划(QP)

当二次规划中的二次系数是半正定矩阵时，二次规划是凸优化的特例，因为此时海森矩阵是半正定的：

minimize $\frac{1}{2}x^T Px + q^T x + r$

subject to $Gx \leq h$

subject to $Ax = b$

2.3 整数规划

因为整数约束导致约束条件函数往往不可微，一般可将整数规划放宽为其他的凸优化问题并得到原问题解的下界。

例子：考虑一个Boolean线性规划，优化变量x的分量被限制为$\{0,1\}$之一：

$$\begin{aligned}&\text{minimize } c^T x\\&\text{subject to } Ax \le b\\&\text{subject to } x_i \in \{0,1\}\end{aligned}$$

将其松弛为：

$$\begin{aligned}&\text{minimize } c^T x\\&\text{subject to } Ax \le b\\&\text{subject to } x_i \le 1\\&\text{subject to } -x_i \le 0\end{aligned}$$

则转化为一个一般的线性规划问题（Convex Optimization，习题4.15）。

2.4 对偶

从一个优化问题出发：

$$\begin{aligned}&\text{minimize } f_0(x)\\&\text{subject to } f_i(x) \le b_i, i = 1,..,m\\&\text{subject to } f_i(x) = b_i, i = m,..,m+p\end{aligned}$$

假设该问题的最优解为p^*，当x^*时取到。

构造拉格朗日函数：

$$L(x,\lambda,v) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} v_i h_i(x), \lambda \ge 0$$

拉格朗日对偶函数定义为：

$$g(\lambda, v) = \inf_x L(x,\lambda,v)$$

易证：

$$g(\lambda, v) \le p^*$$

即一个问题的拉格朗日对偶函数的值给出了原问题解答的一个下界。

拉格朗日函数提出的动机是将变量违背约束的程度负相关地添加到目标函数中，如果将不等式约束的拉格朗日乘子λ替换成一个示性变量，当不等式约束被违反时返回无穷大，否则返回零，则此时对偶问题等价于原问题。

因为拉格朗日对偶函数的值给出了原问题解的下界，所以可以通过取它的极大值来得到一个尽可能紧的下界：

$$\text{maxmize } g(\lambda, v)$$
$$\text{subject to } \lambda \geq 0$$

对于比较复杂的问题，我们有信心认为对偶问题比原问题更“容易”求解，因为如上所示的问题是一个凸优化问题。

例子：考虑原问题为一个朴素二次规划时：

$$\text{minimize } x^T x$$
$$\text{subject to } Ax = b$$

它的对偶问题为：

$$\text{maxmize } -\tfrac{1}{4} v^T A A^T v - b^T v$$

拉格朗日对偶问题的解是原问题的下界，当这个下界是确切的时候，即对偶问题的解给出原问题解时，称此时存在强对偶性，强对偶性成立的条件（Slater条件）此处不加论证地给出：一个凸优化问题中，如果可行点使得非仿射的不等式约束严格成立，则强对偶性成立。

2.5 KKT条件

设对偶问题的最优解在λ^*, v^*时取到，为了使得：

$$f_0(x^*) = g(\lambda^*, v^*) \leq f_0(x^*) + \sum_{i=1}^{m} \lambda_i^* f_i(x^*) + \sum_{i=1}^{p} v_i^* h_i(x^*)$$

显然必须有第一个和式为零（第二个和式显然为零），因为该和式的每一项非负，所以每一项均为零，由此得出称为KKT的优化条件，即存在x^*, λ^*, v^*使得：

$$f_i(x^*) \leq 0, i = 1, ..., m$$
$$h_i(x^*) = 0, i = 1, ..., p$$
$$\lambda_i^* \geq 0, i = 1, ...m$$
$$\lambda_i^* f_i(x^*) = 0, i = 1, ...m$$
$$\nabla f_0(x^*) + \textstyle\sum_{i=1}^{m} \lambda_i^* \nabla f_i(x^*) + \sum_{i=1}^{p} v_i^* \nabla h_i(x^*) = 0$$

当原问题为凸优化时强对偶性成立。

例子：考虑无等式约束的一个凸优化问题：

minimize $f_0(x)$

subject to $f_i(x) \leq b_i, i = 1, .., m$

我们证明此时对于任何可行点x，存在$\nabla f_0(x^*)^T(x - x^*) \geq 0$：

证明：由KKT条件给出：

$$\nabla f_0(x^*) + \sum_{i=1}^{m} \lambda_i^* \nabla f_i(x^*) = 0$$

所以目标不等式等价于：

$$\sum_{i=1}^{m} \lambda_i^* \nabla f_i(x^*)^T(x - x^*) \leq 0$$

对于λ_i^*分情况讨论：

$\lambda_i^* = 0$时，$\lambda_i^* \nabla f_i(x^*)^T(x - x^*) \leq 0$成立；

$\lambda_i^* > 0$时，KKT条件给出$f_i(x^*) = 0$，如果f_i在x^*的导数大于零，则由x的可行性给出$\lambda_i^* \nabla f_i(x^*)^T(x - x^*) \leq 0$成立，导数的其他情况类似易证。

对于一个凸函数而言，$\nabla f_0(x^*)^T(x - x^*) \geq 0$意味着$x^*$为全局最优解。

因为可微凸函数本身有：

$$f_0(y) \geq f_0(x) + \nabla f_0(x)^T(y - x)$$

所以KKT条件给出了无等式约束的凸优化问题的最优解。（Convex Optimization，习题5.31）

3 凸优化应用

以最小二乘法为首的一系列问题都属于凸优化的范畴，事实上，我们所熟知的最大似然估计也仅仅成立于概率密度函数的对数——凹性质上（正态分布、指数分布、均匀分布、Wishart分布等等，显然对数——凹性对于线性指数族全部成立，所以可以肆无忌惮地使用似然概率的对数的相反数作为优化目标，它是一个凸函数）。

相关的内容将由其他文档介绍。

4 凸优化求解算法

本部分介绍最基础的求解算法，受制于篇幅，不对于算法收敛性进行严格证明。

4.1 梯度下降法

首先考虑一个无约束的优化问题，我们假设目标函数为可微凸函数：

$$\text{minimize } f_0(x)$$

此时在最优值p^*的x^*需要满足$\nabla f_0(x^*) = 0$，当x^*的值不能直接解析地得到时，我们希望诉诸于迭代的方法。

形式上说，我们希望构造一个点列$x^(k)$，其中：

$$x^{(k+1)} = x^{(k)} + t^{(k)}\Delta x^{(k)}$$

其中$t^{(k)}$和$\Delta x^{(k)}$分别代表一次迭代的步长和方向，我们希望：

$$f(x^{(k+1)}) < f(x^{(k)})$$

再次由凸性条件：

$$f(y) \geq f(x) + \nabla f(x)^T(y - x)$$

我们必须要有：

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

换言之，构造点列时的迭代方向必须目标函数的负梯度成锐角。步长t^k的选择算法本身也有多种。

显然地，负梯度本身是符合迭代方向条件的，此时的迭代方法称作梯度下降法。

4.2 下降步长选择

步长t的选择至少有两种方法：

第一种是精确搜索，即：

$$t^{(k)} = argmin\left\{f(x^{(k)} + t\Delta x^{(k)})\right\}$$

该方法一般在等式右侧可以解析的解时才能使用。

第二种是回溯搜索，事先选定$0 < \alpha < 0.5, 0 < \beta < 1$，初始化$t = 1$：

if $f(x^{(k)} + t\Delta x^{(k)}) > f(x) + \alpha t \nabla f(x)^T \Delta x$:

then $t = t * \beta$

α的取值范围是为了收敛性证明时的便利，此时不加进一步说明。

4.3 最速下降法

将$f(x+v)$在x处进行泰勒展开：

$$f(x+v) \approx f(x) + \nabla f(x)^T v$$

现在我们试图选择一个v使得$\nabla f(x)^T v$尽量地小，虽然我们可以取v为任意大使得这一项任意小，但是这会使得泰勒展开失效，所以必须限制v的大小本身，一般采取范数限制，此时我们有（nsd for Normalized Steepest Descendent）：

$$\Delta x_{nsd} = argmin \left\{ \nabla f(x)^T v \middle| \|v\| \leq 1 \right\}$$

也即一个范数球向负梯度方向的最大投影的值。

为了说明最速下降法的意义，我们考虑一个这样的情景：对于原空间做坐标变换使得$\bar{x} = P^{\frac{1}{2}} x$，此时由$P$定义的原始空间的范数等价于新空间中的Euclid范数，即：$\|x\|_P = \|\bar{x}\|_2$，同时$\bar{f}(\bar{x}) = f(P^{-\frac{1}{2}}\bar{x}) = f(x)$

在新的变换后空间执行梯度下降法：

$$\Delta \bar{x} = -\nabla \bar{f}(\bar{x}) = -P^{-\frac{1}{2}} f(x)$$

将这一方向变换到原始空间：

$$\Delta x = P^{-\frac{1}{2}} \Delta \bar{x} = -P^{-1} \nabla f(x)$$

换言之，使用P定义的范数进行最速下降法等同于将原始的梯度下降法进行了一种线性变换，也等同于在一个新的空间中进行原始的梯度下降法。

最速下降法之所以能提高收敛效率是因为倘若我们取P为最优解时的海森矩阵$\hat{H}$，则新空间中的汉森矩阵为：

$$\hat{H}^{-\frac{1}{2}} \nabla^2 f \hat{H}^{-\frac{1}{2}} \approx I$$

这意味着新空间中的条件数近似于最小值1，使得收敛速度变快。

（*）一般在界定了初始点后，我们有$mI \leq \nabla^2 f \leq MI$（该不等式由半正定矩阵锥定义），此时定义条件数为$\frac{M}{m}$，可以认为梯度下降法的收敛速度满足：

$$f(x^{(k+1)}) - p^* \leq (1 - \frac{m}{M})(f(x^{(k)}) - p^*)$$

4.4 Newton下降法

Newton步径定义为：

$$\Delta x_{nt} = -\nabla^2 f(x)^{-1} \nabla f(x)$$

可以认为最速下降法是取一个固定矩阵P为最优处海森矩阵的新空间内的梯度下降法，而Newton方法是取此时的海森矩阵为最优处海森矩阵近似的新空间内的梯度下降法。

也可以将$f(x+v)$在x处进行泰勒展开二阶：

$$f(x+v) \approx f(x) + \nabla f(x)^T v + \frac{1}{2} v^T \nabla^2 f(x) v$$

最小化右侧关于v的二次型，同样可得Newton步径。

在梯度下降方法中，我们取$\|\nabla f(x^{(k)})\| < \epsilon$为终止迭代的条件，类似地，在Newton方法中定义Newton减量：

$$\lambda(x) = (\nabla f(x)^T \nabla^2 f(x)^{-1} \nabla f(x))^{\frac{1}{2}}$$

当某次迭代后，Newton减量的数值小于阈值时终止迭代。

4.5 等式约束下的Newton方法

本节我们分析Newton方法在等数约束凸优化中的形式，首先考虑一个二次规划问题：

$$\begin{aligned} &\text{minimize } f_0(x) = \tfrac{1}{2} x^T P x + q^T x + r \\ &\text{subject to } Ax = b \end{aligned}$$

KKT条件给出：

$$Ax^* = b$$

$$Px^* + q + A^T v^* = 0$$

写成矩阵形式：

$$\begin{pmatrix} P & A^T \\ A & 0 \end{pmatrix} \begin{pmatrix} x^* \\ v^* \end{pmatrix} = \begin{pmatrix} -q \\ b \end{pmatrix}$$

当最左侧的矩阵（又称为KKT矩阵）非奇异时，可直接解得x^*。

对于一个一般的凸优化问题，我们将目标函数做二阶泰勒展开：

$$\text{minimize } f(x) + \nabla f(x)^T v + \tfrac{1}{2} v^T \nabla^2 f(x) v$$
$$\text{subject to } Av = 0$$

此时的KKT矩阵方程为（此时对偶变量为w）：

$$\begin{pmatrix} \nabla^2 f(x) & A^T \\ A & 0 \end{pmatrix} \begin{pmatrix} \Delta x_{nd} \\ w \end{pmatrix} = \begin{pmatrix} -\nabla f(x) \\ 0 \end{pmatrix}$$

由上式出发计算点列迭代方向，类似于无约束情形，我们记Newton减量为：

$$\lambda(x) = (\nabla f(x)^T \nabla^2 f(x)^{-1} \nabla f(x))^{\frac{1}{2}}$$

并仍旧以其为收敛的判定条件，可以证明：

$$f(x) - \inf \left\{ f(x) + \nabla f(x)^T v + \frac{1}{2} v^T \nabla^2 f(x) v | A(x+v) = b \right\} = \frac{\lambda(x)^2}{2}$$

证明：将$v = \Delta x_{nt}$代入取得最小值，并在KKT矩阵方程中得到：

$$\nabla^2 f(x) \Delta x_{nt} = -\nabla f(x) - A^T w$$

所以要证等式的左侧等于：

$$-\nabla f(x)^T \Delta x_{nt} + \frac{1}{2} \Delta x_{nt}^T \nabla f(x) + \frac{1}{2} (A \Delta x_{nt})^T w$$

因为$A\Delta x_{nt} = 0$消除最后一项，故左式等于：

$$-\frac{1}{2} \nabla f(x)^T \Delta x_{nt}$$

再由KKT矩阵方程：

$$\nabla^2 f(x) \Delta x_{nt} + A^T w = -\nabla f(x)$$

$$\Delta x_{nt}^T \nabla^2 f(x) \Delta x_{nt} + (A \Delta x_{nt})^T w = -\nabla f(x)^T \Delta x_{nt}$$

即：

$$\lambda(x)^2 = -\nabla f(x)^T \Delta x_{nt}$$

（Convex Optimization，习题10.6）

至此得到了在等式约束凸优化问题中寻找方向和判定收敛与否的Newton方法。

4.6 不等式约束下的Newton方法

和适用拉格朗日函数将一个约束优化问题转化为一个无约束优化问题一样，我们试图用某种方法近似地将不等式约束转化为等式约束并利用上一节的方法求解。

回顾一般凸优化问题的KKT条件：

$f_i(x^*) \leq 0, i = 1, ..., m$

$\lambda_i^* \geq 0, i = 1, ...m$

$\lambda_i^* f_i(x^*) = 0, i = 1, ...m$

$Ax^* = b$

$\nabla f_0(x^*) + \sum_{i=1}^m \lambda_i^* \nabla f_i(x^*) + A^T v^* = 0$

这里我们放宽限制，将上述第三个条件替换为：

$\lambda_i^* f_i(x^*) = \frac{1}{t}, i = 1, ...m, t > 0$

并记此时的最优取值为t的函数$x^*(t)$，利用上式消去λ_i，则KKT条件的最后一个条件化为：

$$t\nabla f_0(x^*(t)) + \sum_{i=1}^m \frac{1}{-f_i(x^*(t))} \nabla f_i(x^*(t)) + A^T v^*(t) = 0$$

记：

$$\phi(x) = -\sum_{i=1}^m \log(-f_i(x))$$

那么我们已经把原始问题近似成：

minimize $tf_0(x) + \phi(x)$

subject to $Ax = b$

易证此时目标函数的凸性，同时，当$t \to 0$时，近似问题逼近原问题，因为二者的KKT条件等价（$f_i(x) \leq 0$隐式包含在了$\phi(x)$的定义中）。

当t给定时，可以使用上一节所述的等式约束下的Newton方法寻找$x^*(t)$。

对于原问题的求解，不仅需要构造固定t时的点列，更需要一个$x^*(t^{(k)})$点列。

一个称作障碍方法的算法每一次在外层迭代中增大t的值，而在内层迭代中用等式约束Newton法寻找$x^*(t)$，并将$x^*(t^{(k)})$作为$t^{(k+1)}$时Newton方法的初始值。

5 其他材料

对于凸优化基础概念的一个简要重述可见：
http://cs229.stanford.edu/section/cs229-cvxopt.pdf
对于对偶性的一个简要重述可见：
http://cs229.stanford.edu/section/cs229-cvxopt2.pdf
希望了解更多详情可参考原教材：
https://web.stanford.edu/~boyd/cvxbook/bv_cvxbook.pdf