[凸优化笔记]从拉格朗日法说起

by：JackeyMiao

1 引言

断断续续历经数月，终于拜读完了凌青老师的《凸优化》课程，与其说是拜读，不如说是欣赏。课程整体知识体系庞大，各种定理对于初识优化的本人而言显得有些晦涩难懂，第一遍学习仅能留下些许印象，习得一层皮毛。

尽管整体课程难度较高，但其中各种巧妙精巧的证明实在是让人如沐春风、叹为观止。针对凸优化领域中一些精华的定理方法，凌老师善于从不同的角度入手，提供许多耳目一新的理解。

学习完整门课，希望能将一些印象深刻、关系密切的知识记录下来，梳理脉络后佐以自身理解，希望能“入门”优化领域，也希望分享自身拙见能抛砖引玉。由于本人初出茅庐、水平有限，内容若有不严谨或错误之处烦请多多包涵并指出。

2 拉格朗日乘子法

在优化领域中，拉格朗日乘子法是用于求解有约束条件的优化问题时的一种常见方式，接下来将从等式约束、不等式约束、混合约束三种情况分类讨论。

首先需要明确的是，拉格朗日乘子法的极值条件并非是充要条件，极值点必满足极值条件，但满足极值条件的并非一定是极值点，但是当问题为凸问题时，其为充要条件。

2.1 等式约束

考虑一个等式约束最优化问题：

$$\begin{cases} \min\ f(x) \\ s.t.\ h_i(x)=0 \quad i=1,2,\dots,m \end{cases}$$

可以构造一个拉格朗日函数：

$$L(x,\lambda_i)=f(x)+\sum_{i=1}^{m}\lambda_i h_i(x)$$

其中λ_i称为拉格朗日乘子，可得极值条件如下：

$$\begin{cases} \dfrac{\partial L}{\partial x}=\dfrac{\partial f}{\partial x}+\sum\limits_{i=1}^{m}\lambda_i\dfrac{\partial h_i}{\partial x}=0 \\ \dfrac{\partial L}{\partial \lambda_i}=h_i(x)=0 \qquad i=1,2,\dots,m \end{cases}$$

2.2 不等式约束

考虑一个不等式约束最优化问题：

$$\begin{cases} \min\ f(x) \\ s.t.\ g_j(x)\le 0 \quad j=1,2,\dots,p \end{cases}$$

其中$g_j(x)$为不等式约束，而我们现在已知等式约束下的拉格朗日乘子法，那么如果我们能将不等式约束化为等式约束就万事大吉了。

我们可以通过引入松弛变量的方式，将不等式约束化为等式约束，以约束$g_j(x)\le 0$为例，我们可以将其转化为：

$$g_j(x)+S_j=0$$

其中S_i为松弛变量，且满足$S_j \geq 0$，至此即可构造拉格朗日函数：

$$L(x, \nu_j) = f(x) + \sum_{j=1}^{p} \nu_j \cdot (g_j(x) + S_j)$$

其中ν_j为拉格朗日乘子，极值条件同2.1中。

2.3 混合约束

考虑一个混合约束最优化问题：

$$\begin{cases} min\ f(x) \\ s.t.\ h_i(x) = 0 \quad i = 1, 2, \dots, m \\ \quad\ g_j(x) \leq 0 \quad j = 1, 2, \dots, p \end{cases}$$

我们可以构造对应的拉格朗日函数：

$$L(x, \lambda_i) = f(x) + \sum_{i=1}^{m} \lambda_i h_i(x) + \sum_{j=1}^{p} \nu_j \cdot (g_j(x) + S_j)$$

2.4 证明推导

考虑一个二维等式约束问题：

$$\begin{cases} \min\ f(x_1, x_2) \\ s.t.\ h(x_1, x_2) = 0 \end{cases}$$

极值点需要满足的条件有：

- 目标函数$f(X) = f(x_1, x_2)$沿着约束曲线$h(x_1, x_2) = 0$的切线s方向的方向导数$\frac{\partial f}{\partial s} = 0$，即

$$\frac{\partial f}{\partial s} = \frac{\partial f}{\partial x_1} \times \frac{\partial x_1}{\partial s} + \frac{\partial f}{\partial x_2} \times \frac{\partial x_2}{\partial s} = 0 \tag{1}$$

  也即目标函数梯度$[\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}]^T$在$s$方向的投影为$\mathbf{0}$

- 约束函数$h(X) = h(x_1, x_2)$沿着约束曲线$h(x_1, x_2) = 0$的切线s方向的方向导数$\frac{\partial h}{\partial s} = 0$，即

$$\frac{\partial h}{\partial s} = \frac{\partial h}{\partial x_1} \times \frac{\partial x_1}{\partial s} + \frac{\partial h}{\partial x_2} \times \frac{\partial x_2}{\partial s} = 0 \tag{2}$$

  也即约束函数梯度$[\frac{\partial h}{\partial x_1}, \frac{\partial h}{\partial x_2}]^T$在$s$方向的投影为$\mathbf{0}$

故由式(1)和式(2)可推出下列等式，并令其比值为$-\lambda$

$$\frac{\partial f}{\partial x_1} / \frac{\partial h}{\partial x_1} = \frac{\partial f}{\partial x_2} / \frac{\partial h}{\partial x_2} = -\lambda$$

其中λ即拉格朗日乘子，式子的几何意义是目标函数的梯度方向与约束函数的梯度方向平行。

进一步，我们可以得到下列方程组：

$$\begin{cases} \frac{\partial f}{\partial x_1} - \lambda \frac{\partial h}{\partial x_1} = 0 \\ \frac{\partial f}{\partial x_2} - \lambda \frac{\partial h}{\partial x_2} = 0 \\ h(x_1, x_2) = 0 \end{cases}$$

解方程组我们可以求出x_1^*、x_2^*和λ^*，其中x_1^*和x_2^*即为原问题的极值点，基于这个极值条件，我们可以构造拉格朗日函数：

$$L(x_1, x_2, \lambda) = f(x_1, x_2) - \lambda \cdot h(x_1, x_2)$$

令其偏微分为0，有：

$$\begin{cases} \dfrac{\partial L}{\partial x_1} = \dfrac{\partial f}{\partial x_1} - \lambda \dfrac{\partial h}{\partial x_1} = 0 \\ \dfrac{\partial L}{\partial x_2} = \dfrac{\partial f}{\partial x_2} - \lambda \dfrac{\partial h}{\partial x_2} = 0 \\ \dfrac{\partial L}{\partial \lambda} = h(x_1, x_2) = 0 \end{cases}$$

故上式与原问题的极值条件等价，因此我们通过求解引入的拉格朗日函数的无约束极值，从而可以等价求解等式约束下目标函数$f(X)$的极值。

（若约束为不等式约束，可将不等式约束化为等式约束）

3 对偶理论

所谓对偶理论，即研究原问题（Primal）与对偶问题（Dual）之间的对偶关系的理论。

对偶理论在优化领域是一个十分重要的理论，在经济学等领域也有着广泛的应用，人们经常尝试通过求解对偶问题来确定原问题的一些性质。尤其是在线性规划问题中，通过求解对偶问题的解，可以直接给出原问题的解。

3.1 对偶函数及对偶问题定义

对于一个一般的优化问题：

$$\begin{cases} \min\ f(x) \\ s.t.\ f_i(x) \le 0 \quad i = 1, 2, \dots, m \\ \quad h_j(x) = 0 \quad j = 1, 2, \dots, p \end{cases}$$

我们可以构造拉格朗日函数（**Lagrangian Function**）：

$$L(x, \lambda, \nu) = f(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{j=1}^{p} \nu_j h_j(x)$$

其中λ_i和ν_j分别为不等式与等式约束的拉格朗日乘子。

进一步，我们可以定义对偶函数（**Lagrange Dual Function**）：

$$g(\lambda, \nu) = \inf_{x \in D} L(x, \lambda, \nu)$$

至此，可能会有疑问，这么一个函数有什么用呢？我们不妨利用约束条件对他进行一下放缩，观察一下他的性质。

$$\begin{aligned} g(\lambda, \nu) &= \inf_{x \in D} L(x, \lambda, \nu) \\ &\le f(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{j=1}^{p} \nu_j h_j(x) \\ &\le f(x) \end{aligned} \tag{3}$$

我们可以发现，对偶函数$g(\lambda, \nu)$为原问题的最优值p^*提供了一个下界，且在特定的条件下，可以取到等号。那么在这个时候我们就拥有了很好的性质，也就是之后会提到的强对偶性。

因此，我们可以将对偶问题描述为

$$\max g(\lambda, \nu)$$

求解该问题即可为原问题求解一个最“紧”的下界。

此外，对偶函数还有一个很好的性质：

不论原函数的凹凸性，对偶函数必为凹函数，对偶问题必为凸问题。

我们知道，凹问题都是比较“难”的问题，而凸问题由于具有许多良好的特点更利于求解，那么由上述性质，我们常常通过求解对偶问题，将凹问题转化为凸问题来分析、求解。

3.2 强/弱对偶性定义

上文曾提到，在特定的条件下式(3)可以取等号，该条件就被称为强对偶性，此时满足原问题最优值与对偶问题最优值等价，为了定义强/弱对偶性，我们不妨设原问题最优值为p^*，对偶问题最优值为d^*，那么我们可以定义如下：

弱对偶性：$d^* \leq p^*$

强对偶性：$d^* = p^*$

对偶间隙：$p^* - d^*$

有了定义，为了更好的利用这些性质，我们自然希望能知道该如何判定一个问题是否有强对偶性，因此**Slator**条件便应运而生。

3.3 Slator条件

在给出Slator条件之前，我们需要明确的是Slator是针对凸问题而言的，同样不是充要条件，而是充分不必要条件，即：

凸问题若满足**Slator**条件，则强对偶性成立；反之，凸问题若强对偶，其不一定满足**Slator**条件。

接下来我们给出Slator条件的数学描述，假定有凸问题：

$$\begin{cases} \min\ f(x) \\ s.t.\ f_i(x) \leq 0 \quad i = 1, 2, \dots, m \\ Ax = b \end{cases}$$

那么当$\exists x \in RelintD,\ s.t.\ f_i(x) < 0,\ i = 1, \dots, m$且$Ax = b$时，强对偶成立，有$p^* = d^*$。

故**Slator**条件可表述为：若可行域D中能寻出一点，使得不等式约束严格成立，则强对偶。

但这一条件并非那么好满足，因此在此基础上，我们有了弱**Slator**条件，其表述为：若不等式约束为仿射约束，只要可行域非空，必有强对偶成立。

顺理成章，我们可以得到如下推论：线性规划问题若可行，必有强对偶性成立。因此，在线性规划中，通过对偶问题来求解原问题的最优值十分常见。

4 鞍点定理

趁热打铁，在第3节中我们对对偶理论进行了简单的介绍和理解，现在，我们再来介绍一下鞍点定理，鞍点定理与拉格朗日乘子法、对偶理论均有着密切的联系，通过鞍点定理我们可以从不同的角度来理解上述理论方法。在后文第7节中，也会给出如何从鞍点定理的角度理解拉格朗日乘子法，在本节中，先介绍鞍点定理和对偶理论的一些联系。

给定一个优化问题：

$$\begin{cases} \min\ f(x) \\ s.t.\ f_i(x) \leq 0 \quad i = 1, 2, \dots, m \\ h_j(x) = 0 \quad j = 1, 2, \dots, p \end{cases}$$

易得其对偶函数：

$$g(\lambda, \nu) = \inf_x L(x, \lambda, \nu) = \inf\{f(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{j=1}^{p} \nu_j h_j(x)\}$$

对乘子λ_i求极大可得对偶问题（满足约束时可取到最大值，否则为正无穷，因为λ_i可以无限大）：

$$\sup_{\lambda\geq 0}\{f(x)+\sum_{i=1}^{m}\lambda_i f_i(x)\}=\begin{cases}f(x), & f_i(x)\leq 0\\ +\infty, & otherwise\end{cases}$$

故原问题最优值可表示为

$$p^*=\inf_x\sup_{\lambda\geq 0}L(x,\lambda)$$

对偶问题最优值可表示为

$$d^*=\sup_{\lambda\geq 0}\inf_x L(x,\lambda)$$

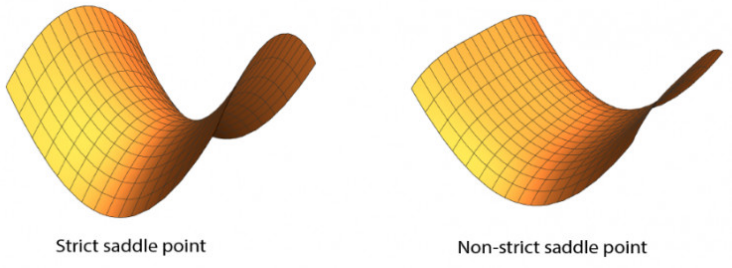
因此，强对偶性可表述为

$$d^*=\sup_{\lambda\geq 0}\inf_x L(x,\lambda)=\inf_x\sup_{\lambda\geq 0}L(x,\lambda)=p^*$$

弱对偶性可表述为（max-min不等式）

$$d^*=\sup_{\lambda\geq 0}\inf_x L(x,\lambda)\leq\inf_x\sup_{\lambda\geq 0}L(x,\lambda)=p^*$$

到这一步，可能很容易把人绕晕，那么我们该如何理解这个不等式呢？矮子里的高个不如高个里的矮子。



进一步，我们可以将$\inf_x L(x,\lambda)$的结果记为$L(\widetilde{x},\lambda)$，将$\sup_{\lambda\geq 0}L(x,\lambda)$的结果记为$L(x,\widetilde{\lambda})$，从而我们就可以给出鞍点的定义：

$$\exists(\widetilde{x},\widetilde{\lambda})\qquad s.t.\quad\sup_{\lambda}L(\widetilde{x},\lambda)=L(\widetilde{x},\widetilde{\lambda})=\inf_x L(x,\widetilde{\lambda})$$

可见，结合马鞍的形状，鞍点的定义十分生动形象，故不难得出得到以下推论：

拉格朗日函数存在鞍点的等价条件为强对偶性成立，且$(\widetilde{x},\widetilde{\lambda})$同时为原问题与对偶问题的最优解。

这也就是著名的鞍点定理，当然这一定理也并非是凭空捏造的，其充要性是可以被严格证明的，限于篇幅，此处便不展开证明。

5 KKT条件

现在，我们已经对拉格朗日乘子法和对偶理论有了初步的了解，随后就该介绍凸优化领域非常重要的条件之一——**KKT**条件。

KKT条件将**Lagrange**乘子法所处理涉及等式的约束优化问题推广至不等式，同时是非线性规划最优解的必要条件。接下来，让我们从对偶理论的角度引出KKT条件，依旧给定一个凸优化问题如下：

$$\begin{cases}\min\ f(x)\\ s.t.\ f_i(x)\leq 0 \quad i=1,2,\dots,m\\ \quad h_j(x)=0 \quad j=1,2,\dots,p\end{cases}$$

可构造拉格朗日函数如下：

$$L(x,\lambda,\nu)=f(x)+\sum_{i=1}^{m}\lambda_i f_i(x)+\sum_{j=1}^{p}\nu_j h_j(x)$$

5.1 Primal/Dual Feasibility

设(x^*,λ^*,ν^*)为最优解，由约束条件可得：

$$\begin{cases} f_i(x^*)\le 0 & \textit{Primal Feasibility} \\ h_j(x^*)=0 & \textit{Primal Feasibility} \\ \lambda^*\ge 0 & \textit{Dual Feasibility} \end{cases}$$

5.2 Complementarity Slackness

假设强对偶性成立，且f_i，h_j均可微，则：

$$\begin{aligned} f(x*) &= g(\lambda^*,\nu^*) \\ &= \inf_x\{f(x)+\sum_{i=1}^{m}\lambda_i^* f_i(x)+\sum_{j=1}^{p}\nu^* h_j(x)\} \\ &\le f(x^*)+\sum_{i=1}^{m}\lambda_i^* f_i(x^*)+\sum_{j=1}^{p}\nu^* h_j(x^*) \\ &\le f(x^*) \end{aligned}$$

可以神奇的发现，我们得到了一个不等式$f(x^*)\le f(x^*)$，显然可以取到等号，那么也顺理成章地可以证得

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

结合之前推得的*primal/dual feasibility*条件$f_i(x^*)\le 0$，$h_j(x^*)=0$，$\lambda^*\ge 0$，易得如下关系：

$$\begin{cases} \lambda_i^*>0 & \rightarrow \quad f_i(x^*)=0 \\ f_i(x^*)<0 & \rightarrow \quad \lambda_i^*=0 \end{cases}$$

故可得KKT条件中的互补松弛条件：

$$\lambda_i^* f_i(x^*)=0 \qquad \textit{Complementarity Slackness}$$

5.3 Stationality Condition

由于原问题满足强对偶，可知

$$\left.\frac{\partial L(x,\lambda^*,\nu^*)}{\partial x}\right|_{x=x^*}=0$$

$$\nabla f(x^*)+\sum_{i=1}^{m}\lambda_i^*\nabla f_i(x^*)+\sum_{j=1}^{p}\nu^*\nabla h_j(x^*)=0 \qquad \textit{Stationality Condition}$$

5.4 总结

至此，我们推出了KKT条件的所有四个条件：

$$\begin{cases} \textit{Primal Feasibility} \\ \textit{Dual Feasibility} \\ \textit{Complementarity Slackness} \\ \textit{Stationality Condition} \end{cases}$$

同时，**KKT**条件仅仅是充分非必要条件，而非充要条件，最优解必定满足**KKT**条件，但满足**KKT**条件并非是最优解，其充要性当且仅当优化问题可微、凸且满足强对偶时成立。

6 罚函数

至此，对于优化问题的等式约束和不等式约束，我们分别讲解了拉格朗日乘子法和KKT条件。除此之外呢，也必然逃不过罚函数方法，罚函数方法同样也是将有约束问题化为无约束问题来求解，接下来将从等式约束和线性不等式约束两种情况来介绍。

6.1 等式约束可微凸优化问题

首先给出优化问题形式如下：

$$\begin{cases} \min\ f(x) \\ s.t.\ Ax-b=0 \end{cases}$$

那么通过罚函数法，我们可将约束罚入目标函数，其中α为罚因子，从而可以得到新的问题形式：

$$\min f(x)+\frac{\alpha}{2}||Ax-b||_2^2 \quad ,\alpha \geq 0 \tag{4}$$

设新问题的最优解为$\widetilde{x}$,，求其极值条件，令其一阶微分为0，可得下式：

$$\nabla f(\widetilde{x})+\alpha(A\widetilde{x}-b)=0 \tag{5}$$

我们可以构造一个问题：

$$\min f(x)+\alpha(A\widetilde{x}-b)^T(Ax-b) \tag{6}$$

对其求极值条件，令其对x的一阶微分为0，同样可得下式：

$$\nabla f(\widetilde{x})+\alpha(A\widetilde{x}-b)=0 \tag{7}$$

显然式(5)与式(7)完全一样，故可证得问题(4)与问题(6)有相同解的结论，我们先将得到的结论放在一边，给出原问题的对偶函数：

$$g(\nu)=\inf_x\{f(x)+\nu(Ax-b)\} \tag{8}$$

将其形式与问题(6)对比，可知当$\nu=\alpha(A\widetilde{x}-b)^T$时，其与问题(6)等价，又因其具有强对偶性，我们可以得到：

$$\begin{aligned} f(x^*)=p^*=d^* &= \max_\nu g(\nu) \\ &\geq g(\alpha(A\widetilde{x}-b)^T) \\ &= f(\widetilde{x})+\frac{\alpha}{2}||A\widetilde{x}-b||_2^2 \\ &\geq f(\widetilde{x}) \end{aligned}$$

其中，$f(x^*)$为原问题的最优值，$f(\widetilde{x})$为罚函数方法得到的最优值，那么根据上述推导，我们可以得到一个十分有意思的结论：罚函数得到的最优值总小于或等于原问题最优值，因此我们可以将罚函数理解成对原问题做了一个松弛。

我们也可以从另一个角度来思考：

- 当$\alpha=0$时，相当于无约束问题
- 当$\alpha\to+\infty$时，为了$f(x^*)=p^*=d^*\geq f(\widetilde{x})+\frac{\alpha}{2}||A\widetilde{x}-b||_2^2$成立，那么必有$A\widetilde{x}=b$即满足约束条件

故当α较小时，问题会在距离约束条件较远处求解；当α较大时，问题会在约束附近求解。

在应用罚函数法时，我们可以先取$\alpha=0$求得初始点x，再以其为起点，逐步增加罚因子α，使其向约束不断靠近。

6.2 带线性不等式约束的可微凸优化问题（log-barrier法）

思路与**6.1**基本一致，尽在问题的构造上略有不同。

首先给出优化问题形式如下：

$$\begin{cases} \min\ f(x) \\ s.t.\ Ax \geq b \end{cases}$$

同6.1中，我们也可以将不等式约束罚入目标函数，从而得到新的问题形式：

$$\min f(x) - \sum_{i=1}^{m} u \cdot log(a_i^T x - b_i) \quad , a_i^T x - b_i > 0 \tag{9}$$

其中a_i表示A的第i行，且该问题仍为凸问题。

设新问题的最优解为$\widetilde{x}$,，求其极值条件，令其一阶微分为0，可得下式：

$$\nabla f(\widetilde{x}) - \sum_{i=1}^{m} u \cdot \frac{a_i}{a_i^T \widetilde{x} - b_i} = 0 \tag{10}$$

我们可以构造一个问题：

$$\min f(x) - \sum_{i=1}^{m} u \cdot \frac{a_i^T x - b_i}{a_i^T \widetilde{x} - b_i} \tag{11}$$

同6.1中，可对问题(11)的目标函数求一阶微分，同样可证得问题(9)与问题(11)有相同解的结论，给出原问题的对偶函数：

$$g(\lambda) = \inf_x \{f(x) + \sum_{i=1}^{m} \lambda_i (b_i - a_i^T x)\} \tag{12}$$

将其形式与问题(11)对比，可知当$\lambda = \frac{u}{a_i^T \widetilde{x} - b_i}$时，其与问题(11)等价，又因其具有强对偶性，我们可以得到：

$$\begin{aligned} f(x^*) = p^* = d^* &= \max_{\lambda} g(\lambda) \\ &\geq g(\frac{u}{a_i^T \widetilde{x} - b_i}) \\ &= f(\widetilde{x}) - \sum_{i=1}^{m} u \cdot log(a_i^T \widetilde{x} - b_i) \\ &\geq f(\widetilde{x}) \end{aligned}$$

至此，同样完成了线性不等式约束下的罚函数证明，可以得到与等式约束下相同的结论。

6.3 拉格朗日乘子法vs罚函数

在学习的过程中，我曾一直疑惑拉格朗日乘子法与罚函数法有何不同，现分享一些自身的理解，希望能有所帮助。

首先，从形式上看，二者都是将约束转化后加入了目标函数中，从而实现了将有约束问题向无约束问题的转化。而从理解层面上，二者实际上是可以相互解释的，在后文第7节也会介绍如何从罚函数的角度去阐述拉格朗日乘子法，因此我们可以认为这两种方法有着密切的联系。但这并不意味着二者是完全等价的，二者在思路以及性质上仍有一定的差异。

罚函数在某种意义上可以理解成通过拉格朗日乘子法得到的对偶函数，与此同时存在一个特殊的拉格朗日乘子可以使对偶函数与罚函数相等。此前，已经证明了罚函数问题（等效对偶）的最优值不会超过原问题的最优值，因此可以认为罚函数求解是对原问题的一个松弛，其松弛程度由罚因子大小来控制，而利用拉格朗日乘子法求解则是等价的。若强对偶性成立，那么求解罚函数问题会等效于求解原问题。

进一步，我们也可以从因子更新的角度来理解，拉格朗日乘子法与罚函数法分别有拉格朗日乘子与罚因子，在求解迭代的过程中，二者是会不断更新的。对于拉格朗日乘子而言，其本质是原问题的对偶变量，在求解的过程中，乘子的更新是遵循某些特定规则的（如梯度下降等）；而罚函数法的罚因子则通常以固定步长或固定倍率进行更新，实现松弛的逐渐收紧，使得解离约束条件越来越近。

7 不同角度理解拉格朗日乘子法

此节呼应第4节鞍点定理及第6节罚函数中所提，对拉格朗日乘子法的多角度理解进一步补充，希望能加强对整个知识体系的理解，接下来会分别从鞍点角度与罚函数角度来对拉格朗日乘子法进行阐述。

首先给出有约束优化问题，问题为凸，且目标函数可微，标准问题如下：

$$\begin{cases} \min\ f(x) \\ s.t.\ Ax = b \end{cases}$$

易得其KKT条件如下：

$$\begin{cases} Ax^* = b \\ \nabla f(x^*) + A^T\nu^* = 0 \end{cases}$$

7.1 鞍点角度理解

先列出拉格朗日函数：

$$L(x,\nu) = f(x) + \nu(Ax - b)$$

根据鞍点定义，以及拉格朗日函数$L(x,\nu)$，可得鞍点：

$$\begin{aligned} (x^*,\nu^*) &= \arg\max_{\nu}\min_{x} L(x,\nu) \\ &= \arg\min_{x}\max_{\nu} L(x,\nu) \end{aligned}$$

若强对偶性成立，鞍点即最优点，则可得下列方程组：

$$\begin{cases} x^* = \arg\min_{x} L(x,\nu^*) & (13) \\ \nu^* = \arg\max_{\nu} L(x^*,\nu) & (14) \end{cases}$$

分别为已知ν^*求x^*和已知x^*求ν^*，这种已知（固定）一个变量，再求另一个变量的最优，本质上是求$\mathbf{L}$的单变量最优。

7.1.1 已知ν^*求x^*

优化问题为问题(13)，使用梯度下降法求解，梯度下降的迭代公式如下：

$$x^{k+1} = x^k + \alpha^k \cdot d^k$$

其中α^k指在第k步时的步长（可用exact/inexact line search进行求取，此处不做展开），d^k为第k步时的方向导数，此处取负梯度方向，由于是求L的单变量最优，故对L的x求一阶微分代入即可，可得下列迭代公式：

$$\begin{aligned} & x^{k+1} &=& \ x^k + \alpha^k \cdot (-\nabla f(x^k) - A^T\nu^*) \\ (\nu^* \approx \nu^k) \rightarrow\ & x^{k+1} &=& \ x^k - \alpha^k \cdot (\nabla f(x^k) + A^T\nu^k) \end{aligned}$$

7.1.2 已知x^*求ν^*

优化问题为问题(14)，同理可得（此处为求极大，故d^k为关于变量ν的梯度方向，而非关于变量x的负梯度方向）：

$$\begin{aligned} & \nu^{k+1} &=& \ \nu^k + \alpha^k \cdot (Ax^* - b) \\ (x^* \approx x^k) \rightarrow\ & \nu^{k+1} &=& \ \nu^k + \alpha^k \cdot (Ax^k - b) \end{aligned}$$

7.1.3 总结

经7.1.1与7.1.2的推导，利用拉格朗日乘子法对问题求解的过程可以概括为下列迭代方程：

$$\begin{cases} x^{k+1} = x^k - \alpha^k \cdot (\nabla f(x^k) + A^T \nu^k) \\ \nu^{k+1} = \nu^k + \alpha^k \cdot (Ax^k - b) \end{cases}$$

因此，拉格朗日乘子法的本质可以理解为利用梯度下降法求取鞍点，而由鞍点定理，强对偶条件下鞍点即最优点。

7.2 罚函数角度理解

上文中，我们已经给出了原问题的KKT条件：

$$\begin{cases} Ax^* = b \\ \nabla f(x^*) + A^T \nu^* = 0 \end{cases}$$

而求解原问题等价于求解其**KKT**条件，因此，我们可以将KKT条件改写为罚函数形式：

$$\min P(x, \nu) = \frac{1}{2} \left|\left| Ax - b \right|\right|_2^2 + \frac{1}{2} \left|\left| \nabla f(x) + A^T \nu \right|\right|_2^2 \tag{15}$$

在7.1中，我们的优化问题可以描述为：

$$\begin{cases} x^* = \arg\min\limits_x L(x, \nu^*) \\ \nu^* = \arg\max\limits_\nu L(x^*, \nu) \end{cases} \tag{16}$$

易知，问题(15)与问题(16)是等价的，其中问题(15)的梯度下降方向为：

$$d(x^k, \nu^k) = \begin{bmatrix} -\nabla f(x^k) - A^T \nu^k \\ Ax^k - b \end{bmatrix} \tag{17}$$

对问题(16)同取负梯度方向，有：

$$-\nabla P(x^k, \nu^k) = -\begin{bmatrix} A^T(Ax^k - b) + \nabla^2 f(x^k)(\nabla f(x^k) + A^T \nu) \\ A(\nabla f(x^k) + A^T \nu^k) \end{bmatrix} \tag{18}$$

我们对式(17)与式(18)，也即问题(15)与问题(16)的方向导数做内积，并观察其性质：

$$d(x^k, \nu^k) \cdot -\nabla P(x^k, \nu^k) = (\nabla f(x^k) + A^T \nu^k)^T \nabla^2 f(x^k)(\nabla f(x^k) + A^T \nu^k) \tag{19}$$

其满足二次型形式，故其正负取决于$\nabla^2 f(x^k)$的正负定性质，因此当满足下列条件时，满足式(19)大于等于0：

$$\begin{cases} \nabla^2 f(x^k) \succ 0 \\ \nabla f(x^k) + A^T \nu^k \neq 0 \end{cases} \tag{20}$$

故满足条件(20)即可使得式(19)大于等于0，也即下降方向与负梯度方向夹角小于等于**90°**，从而可证明罚函数的下降方向即原问题目标函数的下降方向。

7.3 总结

可见，在凸优化领域，“没有知识是一座孤岛”，所有的知识都是紧密相连、环环相扣的。希望如此从鞍点和罚函数的角度出发去解释拉格朗日乘子法，能够提供一个崭新的思路，这也应征了知识与知识之间是可以相互解释、推导的，这种互相耦合的关系也正是将数学之美体现的淋漓尽致~

至此，本文从拉格朗日乘子法着手，推导出了对偶理论，进一步引出了**KKT**条件。随后介绍了罚函数法，将其与拉格朗日乘子法作了对比，并尝试从鞍点定理和罚函数的角度来解释拉格朗日乘子法，提供了不同的视角。

文章至此本应结束了，但仍想再提提增广拉格朗日法，作为对拉格朗日乘子法和罚函数法的一个补充。

8 增广拉格朗日法

拉格朗日乘子法在实际应用中其实并不广泛，主要是由于在梯度下降的过程中，其α值取值困难，导致拉格朗日法鲁棒性差、收敛慢、解不稳定，因此增广拉格朗日法应运而生！

8.1 基本形式

现有问题形如：

$$\begin{cases} \min\ f(x) \\ s.t.\ Ax=b \end{cases}$$

可写出其增广拉格朗日函数为：

$$L_c(x,\nu)=f(x)+\nu^T(Ax-b)+\frac{c}{2}\left|\left|Ax-b\right|\right|_2^2$$

若利用梯度下降法迭代求解，推导过程与第7节中拉格朗日法推导过程类似，此处直接给出结论，增广拉格朗日法表达形式为：

$$\begin{cases} x^{k+1}=\arg\min\limits_x L_c(x,\nu^k) \\ \nu^{k+1}=\nu^k+c\cdot(Ax^{k+1}-b) \end{cases}$$

8.2 增广拉格朗日法的可行性分析

现有问题形如：

$$\begin{cases} \min\ f(x) \\ s.t.\ Ax=b \end{cases} \tag{P1}$$

写出增广拉格朗日函数法形式下的问题：

$$\begin{cases} \min\ f(x)+\frac{c}{2}\left|\left|Ax-b\right|\right|_2^2 \\ s.t.\ Ax=b \end{cases} \tag{P2}$$

欲证明其可行性，即证明问题(P1)与问题(P2)等价，即证明二者最优解一致，即：

$$\begin{cases} x_1^*=x_2^* \\ \nu_1^*=\nu_2^* \end{cases}$$

假设其满足强对偶，则KKT条件与原问题等价，故分别列写两个问题的KKT条件。

对于问题(P1)，KKT条件有：

$$\begin{cases} \nabla_{x_1}L_c(x_1^*,\nu_1^*)=0 \\ \nabla_{x_1}\{f(x_1^*)+\nu_1^T(Ax_1^*-b)\}=0 \end{cases} \tag{21}$$

对于问题(P2)，KKT条件有：

$$\begin{cases} \nabla_{x_2}L_c(x_2^*,\nu_2^*)=0 \\ \nabla_{x_2}\{f(x_2^*)+\nu_2^T(Ax_2^*-b)\}+cA^T(Ax_2^*-b)=0 \end{cases} \tag{22}$$

由于x_2^*作为问题(P2)的最优解，必满足约束条件，即$Ax_2^*-b=0$，故显然问题(P1)与问题(P2)完全一致，因此可证明增广拉格朗日法解即原文题解，方法可行，证毕。

8.3 拉格朗日法vs增广拉格朗日法

列写出梯度下降法下，二者的迭代方程。

拉格朗日法：

$$\begin{cases} x^{k+1} = x^k - \alpha^k \cdot (\nabla f(x^k) + A^T \nu^k) \\ \nu^{k+1} = \nu^k + \alpha^k \cdot (Ax^k - b) \end{cases}$$

其中x^k与ν^k分别为原问题（Primal）和对偶问题（Dual）的变量，因此拉格朗日法是**Primal**与**Dual**的同步优化。

增广拉格朗日法：

$$\begin{cases} x^{k+1} = \arg\min\limits_x L_c(x, \nu^k) \\ \nu^{k+1} = \nu^k + c \cdot (Ax^{k+1} - b) \end{cases}$$

而在增广拉格朗日法中，我们可以将x^{k+1}带入ν^{k+1}的迭代式中，可得：

$$\nu^{k+1} = \nu^k + c \cdot (A \cdot \arg\min_x L_c(x, \nu^k) - b)$$

因此，增广拉格朗日法并非是同步优化，而是只对ν^k做更新，x^k仅为更新过程中的中间变量。

8.4 相关性质

性质**1**:

若$\nu = \nu^*$，则$\forall c > 0$，$x^* = \arg\min_x L_c(x, \nu^*)$

其表明得到ν^*后，必可以算出x^*，只要满足条件$c > 0$

性质**2**:

若$c \to +\infty$，则$\forall \nu$，$x^* = \arg\min_x L_c(x, \nu^*)$

其表明当$c \to +\infty$时，必可以求出x^*，说明了其优良的鲁棒性。

9 结语

文章中的绝大多数内容均来自于凌青老师讲授的《凸优化》课程，十分感谢凌青老师！

“道阻且长，行则将至”，希望该文章的总结能成为本人打开“优化”领域大门的钥匙，以改善如今“管中窥豹”的现况。