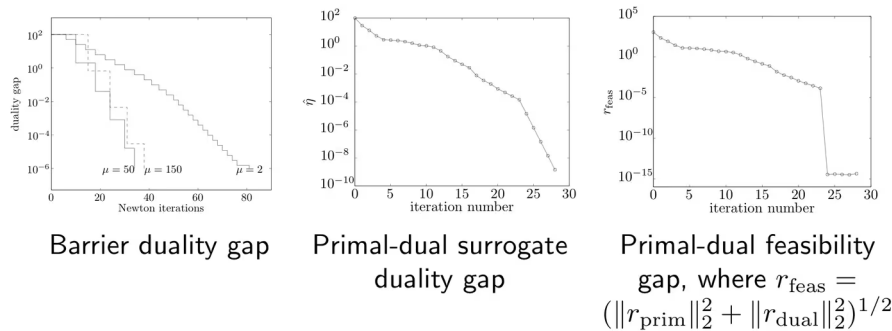


知乎

首发于  
一个大学生的日常笔记

Barrier method uses various values of  $\mu$ , primal-dual method uses  $\mu = 10$ . Both use  $\alpha = 0.01, \beta = 0.5$



## 凸优化|笔记整理（8）——内点法中的屏障法与原始-对偶方法，近端牛顿方法



学弱獐  
数学话题下的优秀答主

上一节笔记：

学弱獐：凸优化|笔记整理（7）——对偶性延伸：对偶范数，共轭函数，双对偶；再...  
162 赞同 · 15 评论 文章



大家好！

这一节我们主要谈一些二阶方法——**内点法**（Interior Method），如果还有空位的话，还会简单引入一下**近端牛顿方法**（Proximal Newton Method）。你可能要问明明只有一个方法，为什么要用“一些”？这是因为**内点法其实是一种方法的总称**，我们在《数值优化》的第A节

学弱獐：数值优化|笔记整理（A）——线性规划中的单纯形法与内点法  
255 赞同 · 13 评论 文章



和第C节

学弱獐：数值优化|笔记整理（C）——二次规划（下）：内点法；现代优化：罚项...  
194 赞同 · 17 评论 文章



分别提到过线性规划与二次规划问题的内点法。在这一节我们会提到两种内点法——**屏障法**（Barrier Method）和**原始-对偶方法**（Primal-Dual Method），它们与之前我们提到的方法的**思路非常相似，但是视角又略有不同**，因此值得我们再去谈一谈。

那么我们开始吧。

### 目录

- 屏障法
  - 线性约束下的屏障法
    - 算法细节与收敛性分析
- 原始-对偶方法
- 屏障法与原始-对偶方法的比较
- 近端牛顿方法引入

- CMU 10-725, *Convex Optimization*
- Boyd, Vandenberghe, *Convex Optimization*
- Nemirovski (2004), *Interior-point polynomial time methods in convex programming*, Chapter 4
- J. Nocedal, S. Wright (2006), *Numerical Optimization*
- S. Wright (1997), *Primal-dual interior-point methods*, Chapters 5 and 6
- J. Renegar (2001), *A mathematical view of interior-point methods*

## 屏障法

在说具体的方法之前, 我们先点出一个事实: **内点法就是迭代点在约束内部的方法**。之后我们介绍具体的方法的时候, 就会对这一句话理解的更为透彻。

**屏障法 (Barrier Method)** 希望解决的是下面这个问题

$$\begin{aligned} \min_x & f(x) \\ \text{s.t.} & h_i(x) \leq 0, i = 1, \dots, m \\ & Ax = b \end{aligned}$$

并且希望  $f, h_i (i = 1, \dots, m)$  都是凸函数且二阶可导, 且希望问题是一个具备强对偶性的问题。很明显这是一个凸问题。

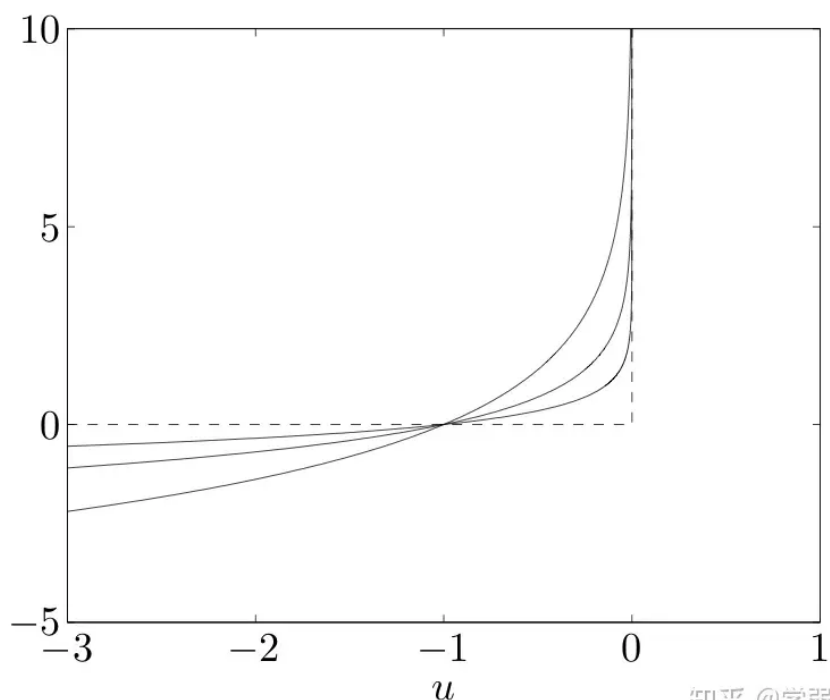
我们在之前提到过说, 这个问题可以转为一个无约束优化问题

$$\min_x f(x) + \sum_{i=1}^m I_{\{h_i(x) \leq 0\}}(x)$$

屏障法的关键就是找到一个函数, 能够把每一个约束所对应的示性函数  $I_{\{h_i(x) \leq 0\}}(x)$  都化为一个近似的, 但性质更好的函数。因为如果自变量  $x$  是一个标量, 那么把示性函数画出来, 它就像是一个屏障, 在  $h_i(x) = 0$  这一处, 突然函数值就从 0 变成了  $\infty$ 。而我们这里的比较好的逼近它的函数就是  $\frac{\log(-h_i(x))}{t}$ , 也就是说我们可以找到一个目标函数的近似

$$\min_x f(x) - \frac{1}{t} \sum_{i=1}^m \log(-h_i(x))$$

其中  $t$  越大, 逼近效果越好。下面这张图就说明了这两种函数的一个对应关系。



知乎 @学弱渣

方法要生效，函数要有定义，就必须要有**强对偶性**满足。当然了，新的问题依然是一个凸问题。

当然了，自然会有人问，为什么要做这样的逼近，**我直接解原问题不就完事了**？这当然是可以的，不过对于内点法这样是不适用的。这个原因在《数值优化》第A节中有解释过，即因为**解KKT条件的时候，矩阵不一定满秩**，当然了我们在后面提到相关的内容的时候，就直接引用《数值优化》的部分，不单独解释内点法的步骤了。

不管具体的技术细节，理论上来说一个 $t$ 可以对应一个解，假如说我们让 $t$ 连续变换，那么这个解也会连续变换，就会形成一条弧线。这一条弧线我们叫**中心路径**（central path）。

为了方便理解，我们用一个线性规划的例子来给出它的一个几何图形。

Example 1:  
考虑线性规划问题  $\min c^T x \quad s.t. \quad Dx \leq e$  的中心路径。

很显然线性规划问题对应的屏障法优化问题为

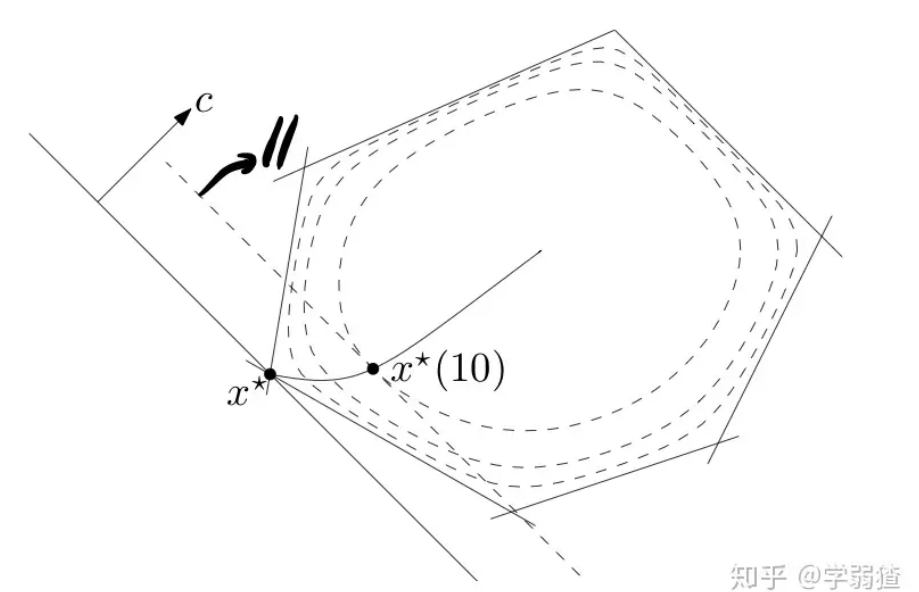
$$\min_x t c^T x - \sum_{i=1}^m \log(e_i - d_i^T x)$$

注意一般来说屏障法对应的问题是  $\min_x t f(x) + \phi(x)$ ，它其实和  $\min_x f(x) + \frac{\phi(x)}{t}$  所对应的解是相同的，只是去分母计算起来更方便一些。另外  $e_i, d_i^T$  就对应了  $e, D$  的行元素。

那么令梯度为0，我们可以得到

$$0 = tc - \sum_{i=1}^m \frac{1}{e_i - d_i^T x(t)} d_i$$

所以根据这个就可以解出  $x^*(t)$ ，当然了更重要的一个观察在于**梯度  $\nabla \phi(x^*(t))$  是与  $c$  平行的**。这可以见下面这张图。



这里的每一条虚线就对应了对数屏障函数的等高线，从内部出发到多边形顶点的那一条实线就是我们的中心路径。可以看出，如果在中心路径与等高线的交点处作等高线的法线（因为这是梯度的方向），那么这一条法线一定是与  $c$  这个向量平行的。而且你可以看出它确实是从**内部出发**逐步走到边界上的最优解的，所以也说明它确实是一种内点法。

### 线性约束下的屏障法

现在我们具体来看一类问题的屏障法求解方案，考虑问题

$$\min_x t f(x) + \phi(x)$$

知乎

首发于  
一个大学生的日常笔记

$$t \nabla f(x^*(t)) - \sum_{i=1}^m \frac{1}{h_i(x^*(t))} \nabla h_i(x^*(t)) + A^T w = 0$$
$$Ax^*(t) = b$$
$$h_i(x^*(t)) < 0, i = 1, \dots, m$$

注意因为问题只有等式约束（不等式约束都变成log-barrier了），所以没有互补松弛条件也没有对偶问题可行条件，**只有稳定性条件和原问题可行条件**。

注意到我们把问题给改掉了，所以这个问题的解如果不是原问题的解，那么其实很有可能就是在做无用功。所以我们自然希望说明的事情就是通过一些变换，使得**得到的屏障法目标函数的对偶问题的最优解也是原问题的对偶问题的最优解**（好长……）。所以我们将KKT条件的第一行同时除掉一个 $t$ ，然后就可以得到下面的结论。


Proposition 1:  
如果设 $u_i^*(t) = -\frac{1}{th_i(x^*(t))}, i = 1, \dots, m, v^*(t) = \frac{w}{t}$ ，那么 $u^*, v^*$ 是未经过屏障函数处理的原问题的对偶问题的解。

我们证明一下这个结论。首先对于原问题我们可以写出它的KKT条件为

$$L(x, u, v) = f(x) + \sum_{i=1}^m u_i h_i(x) + v^T (Ax - b)$$

我相信读者可以自己补出这个原问题的形式233，那么如果 $u^*, v^*$ 是这个原问题的对偶问题的解，则一定要求它在**对偶问题上**是可行解，且它处于**拉格朗日函数的鞍点处**（这个说法要看《凸优化》第6节

学弱渣：凸优化|笔记整理 (6) —— 对偶性：案例分析，强弱对偶性及理解，再看...  
191 赞同 · 14 评论 文章



对于对偶性的各种解释的部分）。第一个很好办，因为对偶问题的可行性条件只有 $u^* > 0$ 。根据 $h_i(x) < 0$ ，这个是显然的。第二个我们可以通过代入，利用屏障法问题的KKT条件，我们可以得到下面这个式子成立

$$\nabla f(x^*(t)) + \sum_{i=1}^m u_i^*(x^*(t)) \nabla h_i(x^*(t)) + A^T v^*(t) = 0$$

但是这个恰好就是 $L(x, u, v)$ 关于 $x$ 的一阶最优性条件，换句话说， $u^*, v^*$ 所在处可以使得 $x^*(t)$ 满足 $L(x^*, u^*, v^*) = \min_x L(x, u, v)$ 。这一条性质说明 $u^*, v^*$ 处于拉格朗日函数的鞍点处，也就是说结论得到了证明。

现在我们回头来看这句话，其实就说明一件事：**如果我们的 $u, v$ 的取法得当，那么它们不仅是屏障法对应优化问题的对偶问题的解，也完全可以使得它们属于原目标函数的对偶问题的解**，这个性质当然很重要，毕竟强对偶性把原问题与对偶问题的解都统一起来了。而且有了这个性质，我们就可以计算它的**对偶间隔**以判断迭代的收敛性。注意到我们有

$$g(u^*(t), v^*(t)) = f(x^*(t)) + \sum_{i=1}^m u_i^*(t) h_i(x^*(t)) + v^*(t)^T (Ax^*(t) - b)$$

那么根据 $x^*(t)$ 在原问题的约束条件， $u_i^*(t) h_i(x^*(t)) = -\frac{1}{t}$ ，就可以得到 $g(u^*(t), v^*(t)) = f(x^*(t)) - \frac{m}{t}$ ，这就说明了对偶间隔不会超过 $\frac{m}{t}$ 。这是一个比较好的判断准则，我们可以设置 $\frac{m}{t}$ 作为迭代的终止条件。而且也确实说明了，在 $t \rightarrow \infty$ 的时候， $x^*(t) \rightarrow x^*$ 为未经过屏障处理的原问题的最优解。

算法细节与收敛性分析

下面我们来说说屏障法的实操，事实上我们并不是取一个固定的 $t$ 然后来解这个优化问题，毕竟一开始就取一个很大的 $t$ ，那其实和直接解原问题也没啥区别。所以我们**取一系列的递增的 $t$ ，然后使用牛顿法来求逐步求解**，不过这里不是使用那个牛顿法的更新公式（毕竟那个是对**无约束优化问题**的求解方式），而是考虑对函数求一阶微分，因为如果说需要解一个方程组 $F(x) = 0$ ，那么根据

知乎

首发于  
一个大学生的日常笔记

$$Dh(x) = \begin{bmatrix} \dots \\ \nabla h_m(x)^T \end{bmatrix}.$$

具体来说, 在这里, 我们设

$$r(x, v) = \begin{bmatrix} \nabla f(x) + \sum_{i=1}^m (-\frac{1}{th_i(x)}) \nabla h_i(x) + A^T v \\ Ax - b \end{bmatrix} = 0$$

那么就是把两个等式条件组合成矩阵, 写成

$$\begin{bmatrix} H_{bar}(x) & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} \Delta x \\ \Delta v \end{bmatrix} = -r(x, v)$$

其中

$$H_{bar}(x) = \nabla^2 f(x) + \sum_{i=1}^m \frac{1}{th_i(x)^2} \nabla h_i(x) \nabla h_i(x)^T + \sum_{i=1}^m (-\frac{1}{th_i(x)}) \nabla^2 h_i(x)$$

比方说这里的 $A^T$ 是第一行第二列, 它就是第一个等式对 $v$ 求梯度的结果, **解这个方程组得到变化量, 然后做一次迭代, 就算运用了一次牛顿法**, 说它是牛顿法的原因, 主要在于**这个求解方式正是运用了牛顿法的设计思路**. 具体的可以看《数值优化》第5节

学弱渣：数值优化 (5) —— 信赖域子问题的求解, 牛顿法及其拓展

82 赞同 · 11 评论 文章



关于牛顿法的开头部分。

具体的步骤如下

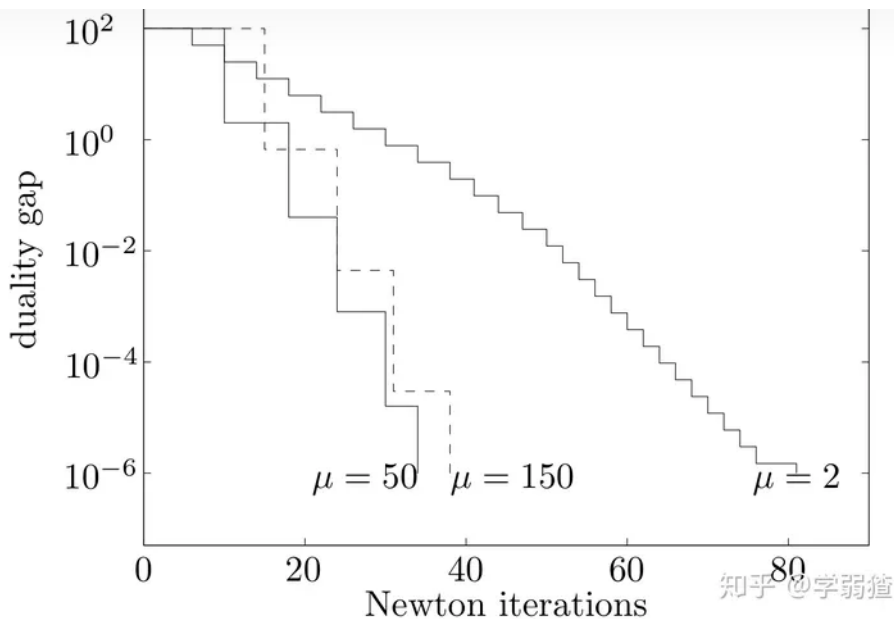
1. 设 $t^{(0)} > 0, \mu > 1, x^{(0)}$ 为初始值。
  2. 使用牛顿法迭代得到 $x^*$ 。
  3. 判断 $\frac{m}{t} \leq \epsilon$ 是否成立, 如果成立则停止迭代, 否则设 $t^{(k+1)} = \mu t$ 。
  4. 在下一步开始之前, 以上一步的最优值 $x^*$ 作为下一步的迭代初始点, 回到2。

我们可以看到, "逐步求解"的含义就是4中的迭代初始点的选择。因为问题是连续变化的, 所以每一次变化 $t$ , 都不会让最优解的位置变化的太多 (换句话说, 其实这就是一个不断的warm-up的过程), 那么使用之前一步的最优值作为当今的迭代值, 我们就有理由相信**它其实是接近最优点的**, 因此牛顿法直接进入二次收敛速度的概率就会大大增加。

说到这里一个容易想到的问题就是**如何选取 $\mu$ 和 $t^{(0)}$** , 因为 $\mu$ 取得大了或者小了都会有一些影响。取得大了, 每一步跨的步子太大了, 就会使得牛顿法的迭代步数变多。取得小了, 那可能需要很多步才能使得 $t$ 足够的大, 逼近真正的解 $x^*$ 。不过实际上这个问题不用太担心, 因为**实验上其实方法对于这两个参数的选取还是有足够的稳定性的**。

下面这张图展示了当 $n = 50$  ( $n$ 是数据的维数),  $m = 100$  ( $m$ 是约束的个数) 时的数值实验结果, 可以看出虽然 $\mu$ 的差距很大, 但是它们其实最终都能收敛到不错的结果。

知乎

首发于  
一个大学生的日常笔记

接下来我们来看看线性约束下的屏障法的收敛性分析，它的收敛性分析主要是如下的结果。

Theorem 1:

屏障法满足  $f(x^{(k)}) - f(x^*) \leq \frac{m}{\mu^k t^{(0)}}$ ，其中  $k$  是外层循环的迭代步数。

这里“外层循环”中的一步，指的是对于固定的  $t$  执行牛顿法到收敛，所以实际上涵盖了多步的牛顿法迭代。有的地方会把牛顿法迭代的每一步称为迭代的“内层循环”。

所以如果要求精度达到  $\epsilon$ ，我们就需要做  $\frac{\log(m/(t^{(0)}\epsilon))}{\log \mu}$  步的外层循环。虽然从这个步数来看， $\mu$  越大， $m$  越小，需要的步数越少，但这不代表收敛的越快。我们之前已经分析过， $\mu$  取得过大的话，内部的牛顿法所需要的步数就会变多，时间就会变长。这种分析的思想我们在《凸优化》第5节

学弱獐：凸优化|笔记整理 (5) —— 近端梯度法：性质，延伸与案例分析；对偶性：...

193 赞同 · 29 评论 文章



介绍矩阵补全案例的时候，已经提到过。

最后我们再提一个屏障法的处理细节。我们一开始说过对于屏障法解决的原优化问题，我们是需要强对偶性的，但是如何找到这一点呢？一般是考虑优化问题

$$\begin{aligned} \min_{x,s} \quad & s \\ \text{s.t.} \quad & h_i(x) \leq s, i = 1, \dots, m \\ & Ax = b \end{aligned}$$

我们可以用屏障法来解这个问题，可以看出其实只需要在优化过程中  $s < 0$  这个目标实现了，就可以提前终止，因为我们的目的就只是为了找到一个满足条件的点而已。

还有一个相关的问题是如果问题不是严格可行的，那么可能在什么约束下产生问题？我们可以通过下面的优化问题

$$\begin{aligned} \min_{x,s} \quad & 1^T s \\ \text{s.t.} \quad & h_i(x) \leq s_i, i = 1, \dots, m \\ & Ax = b, s \geq 0 \end{aligned}$$

来找到答案，那么事实上，一般认为当这个优化问题的最终得到的  $s$  哪一个分量  $s_i$  不是 0 而是一个正数，那么就认为这个分量对应的约束是影响因素。在这种情况下，如果依然希望使得问题满足强对偶性，就需要对这个约束做一些修改，删除等操作。



接下来我们来介绍一下**原始-对偶方法** (Primal-Dual Method), 它求解的问题依然是**屏障法对应的那个问题**, 对应了《数值优化》第A节, 第C节所提到的那个内点法(链接看最开始的部分), 也就是那两篇文章所提到的“**主对偶方法**”。不过在这里我们的分析视角略有不同, 并且我们也会把这个方法同上面的屏障法联系起来, **对于这个方法的一些动机也会说的更明确一些**。

要介绍这个方法, 我们要先介绍一下屏障法所引入的一个新的概念——**扰动KKT条件** (Perturbed KKT Conditions), 注意, 它**并不是**传统意义上的KKT条件, 是经过处理之后得到的一系列新的等式与不等式约束条件。具体来说就是

$$\begin{aligned} \nabla f(x) + \sum_{i=1}^m u_i \nabla h_i(x) + A^T v &= 0 \\ u_i h_i(x) &= -\frac{1}{t}, i = 1, \dots, m \\ h_i(x) &\leq 0, i = 1, \dots, m \\ u_i &\geq 0, i = 1, \dots, m \\ Ax &= b \end{aligned}$$

可以看出除了互补松弛条件被改掉了之后, 其它的式子都没有变。

有了这个之后, 我们可以把问题写成一个线性系统, 表示为

$$r(x, u, v) = \begin{bmatrix} \nabla f(x) + Dh(x)^T u + A^T v \\ -\text{diag}(u)h(x) - \frac{1}{t} \\ Ax - b \end{bmatrix} = 0$$

使用和上面相同的牛顿法, 可以得到

$$\begin{bmatrix} H_{pd}(x) & Dh(x)^T & A^T \\ -\text{diag}(u)Dh(x) & -\text{diag}(h(x)) & 0 \\ A & 0 & 0 \end{bmatrix} \begin{bmatrix} \Delta x \\ \Delta u \\ \Delta v \end{bmatrix} = - \begin{bmatrix} r_{dual} \\ r_{cent} \\ r_{prim} \end{bmatrix}$$

其中  $H_{pd}(x) = \nabla^2 f(x) + \sum_{i=1}^m u_i \nabla^2 h_i(x)$ 。  $r_{dual}, r_{cent}, r_{prim}$  分别对应了  $r(x, u, v)$  的第一, 二, 三行, 具体为什么起这些名字后面再说。

有了这个线性方程组之后, 我们解出变化量, 然后再更新迭代点就可以了。看似到此为止, 原始-对偶方法的核心已经介绍完了, **但是其实远没有结束**。它的隐患很多, 首先我们会发现, 当我们这么更新之后, **能够保证  $r(x, u, v) = 0$  吗?** 这显然是不可能的。如果这一点做不到, 相当于  **$x, u, v$  很可能不是可行解** (所以我们定义了  $f(x) + Dh(x)^T u + A^T v$  叫作  $r_{dual}$ , 是因为如果  $r_{dual} = 0$ , 说明  $u, v$  是对偶问题的可行解。  $r_{prim}$  类似, 而对于  $r_{cent}$ , 则是因为它是和中心路径 (central path) 有关的量), 既然都没办法确认点在可行解内, 就不可能再用屏障法的思路, 利用对偶间隔来判断迭代收敛了。其二就是这里的**步长**, 既然我们的迭代公式是诸如  $x^+ = x + s \Delta x$  这样的形式, 就必然需要知道这个  $s$  是什么, 这又和屏障法的处理方式不一样。

对于这两个问题, 我们自然需要提出解决方案。对于第一个, 想法其实很简单, 虽然我们的迭代点不一定是可行的解, 但是在这个迭代方法收敛的时候, 如果问题合适, **一定是在可行域内的** (否则就相当于说方程解不出来了)。因此事实上, 我们依然利用值

$$\eta = -h(x)^T u = -\sum_{i=1}^m u_i h_i(x)$$

来“近似”对偶间隔, 一般我们称它为**代理对偶间隔** (Surrogate Duality Gap), 那么可以看出, 在我们的  $x$  变化的时候,  $\eta$  也会变, 所以在**迭代点可行的时候, 它依然可以作为一个好的对偶间隔**。虽然在刚开始它不是, 但是我们“强行”认为它是对偶间隔, 所以通过这个式子, 我们可以认为  $t = \frac{m}{\eta}$  为我们每一步的  $t$ 。换句话说, 我们这里不再直接初始化  $t$ , 而是通过迭代点算出的  $\eta$  来近似的得到  $t$ 。那么对于每一步我们需要更新  $t$  的时候, 就根据上一步得到的  $t$  来更新即可。不同的是每一步根据  $\eta$  的不同, 会对  $t$  有额外的修改, 这个在算法流程中会看的清清楚楚。

第二个问题就是步长的选取, 很显然我们希望步长选取不能太离谱。注意我们的目的是为了解方程  $r(x, u, v) = 0$ , 那么对于  $r_{dual}$  和  $r_{prim}$  我们都有讨论过, 那么  $r_{cent}$  自然也需要有一定限制。这个限制就是扰动KKT条件中的

知乎

首发于  
一个大学生的日常笔记

$$s_{\max} = \min\{1, \min\{-u_i/\Delta u_i : \Delta u_i < 0\}\}$$

可以看出这个目标就是  $u + s\Delta u \geq 0$ , 那么还有两个目标是  $h_i(x^+) < 0, i = 1, \dots, m$  以及  $\|r(x^+, u^+, v^+)\|_2$  充分下降。所以我们考虑设  $s = \beta s$ , 直到下面两个条件成立

$$\begin{aligned} h_i(x^+) &< 0, i = 1, \dots, m \\ \|r(x^+, u^+, v^+)\|_2 &\leq (1 - \alpha s)\|r(x, u, v)\|_2 \end{aligned}$$

并且初始值设置为  $s = 0.999s_{\max}$ , 这里没有直接设置为  $s_{\max}$ , 自然是因为数值上怕有影响。

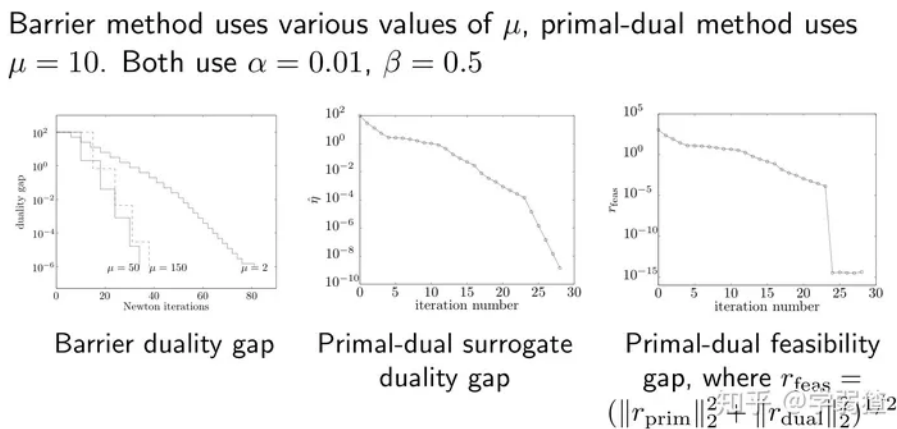
这两个问题解决了之后, 我们再把完整的算法流程放出来, 大家就很好理解了。

1. 设  $t = \frac{\mu m}{\eta^{(k-1)}}$ 。
2. 解线性方程组, 计算  $\Delta y$ 。
3. 计算步长  $s$ 。
4. 计算  $\eta^k = -h(x^{(k)})^T u^{(k)}$
5. 如果  $\eta^{(k)} \leq \epsilon$ ,  $(\|r_{\text{prim}}\|_2^2 + \|r_{\text{dual}}\|_2^2)^{1/2} \leq \epsilon$ , 那么停止迭代, 否则回到1。

可以看出这里的  $t$  不是事先给定的, 而是一开始根据上一步的  $\eta$  计算出来的。 $\eta$  在每一步也都会更新。5保证了我们的最终收敛的点一定距离可行解非常近, 所以那个时候使用代理对偶间隔也是合理的。

## 屏障法与原始-对偶方法的比较

这两种方法解决的是同一类型的问题, 那么它们究竟有什么不同呢? 这里我们可以首先看一下它们的迭代曲线。



总体来说, 原始-对偶方法的速度和准度都会更好一些。

当然了我们不可能只比较一个速度, 更重要的是看它们方法上的差异。事实上这两者都是使用牛顿法求解的, 只不过屏障法是人工设置了  $u, v$ , 并保证了在迭代中解一直是可行的。原始-对偶方法虽然也人工设置了  $u, v$ , 但却放松了“可行”的限制, 这就导致了两种不同的方案。但我们可以看出, 原始-对偶方法虽然限制更少, 但其实能得到的结果却往往更好, 这不得不说是个神奇的事情。

说到这里, 我们再提有一个很有趣的现象, 就是如果我们在原始-对偶问题中设置了步长  $s = 1$ , 会发生什么? 事实上, 注意到线性方程组其实告诉我们

$$\begin{aligned} A^T \Delta v + \Delta u &= -(A^T v + u - c) = -r_{\text{dual}} \\ A \Delta x &= -(Ax - b) = -r_{\text{prim}} \end{aligned}$$

那么假如说下一步, 我们有  $v^+ = v + \Delta v, u^+ = u + \Delta u, x^+ = x + \Delta x$ , 很容易计算出来新的结果为  $r_{\text{dual}}^+ = r_{\text{prim}}^+ = 0$ 。也就是说, 对于原始-对偶问题, 走一步真正的牛顿法 (也即步



事实上，拟牛顿法也可以认为是一种二阶方法，不过《数值优化》第6，7节已经详细介绍过这两种方法，所以我们这里就不多说了。

学弱渣：数值优化（6）——拟牛顿法：  
SR1，BFGS，DFP，DM条件  
239 赞同 · 12 评论 文章



学弱渣：数值优化（7）——限制空间的优  
化算法：LBFGS，LSR1  
120 赞同 · 15 评论 文章

$$x^k = 1, m = 2$$
$$\begin{pmatrix} 1 - \frac{2g_k^T}{g_k^T g_k} \\ \frac{1}{g_k^T g_k} \end{pmatrix} H_k^m \begin{pmatrix} 1 - \frac{2g_k^T}{g_k^T g_k} \\ \frac{1}{g_k^T g_k} \end{pmatrix}$$

### 近端牛顿方法引入

你们以为到此就结束了？那必不可能，我字数的kpi还没到呢.....哎，这段话我是不是之前说过一次？

还剩下一点篇幅，我们简单聊一下**近端牛顿方法**（Proximal Newton Method）的相关背景。通过这名字也容易看出，它是与近端梯度方法类似的，**通过近端算子求解特定非光滑问题的二阶方法**。

回顾一下近端梯度方法，我们一般会利用近端方法解决诸如

$$f(x) = g(x) + h(x)$$

这样的问题，其中 $g(x)$ 希望它可微且凸， $h(x)$ 希望它凸，但不一定需要连续。

对于近端梯度法，它是在解

$$x^+ = \arg \min_z \frac{1}{2t} \|x - t \nabla g(x) - z\|_2^2 + h(z)$$

根据这个目标我们定义了近端算子。但是其实我们也可以把它写成下面这样

$$x^+ = \arg \min_z \nabla g(x)^T (z - x) + \frac{1}{2t} \|z - x\|_2^2 + h(z)$$

设 $v = z - x$ 为当前步希望行走的方向向量，那么这样的话，可以把这个式子再转换一下，得到

$$x^+ = \arg \min_v \nabla g(x)^T v + \frac{1}{2t} \|v\|_2^2 + h(x + v)$$

为什么我们要转这么几步？这是因为**我们要把牛顿方法的思想运用过来**。牛顿法的含义就是在函数的二次逼近中，使用海塞矩阵而不是 $\frac{1}{t}I$ 这样的东西。所以我们改一下，就可以得到

$$\begin{aligned} v^+ &= \arg \min_v \nabla g(x^{(k-1)})^T v + \frac{1}{2} v^T H^{(k-1)} v + h(x^{(k-1)} + v) \\ x^{(k)} &= x^{(k-1)} + t_k v^{(k)} \end{aligned}$$

其中 $H^{(k-1)} = \nabla^2 g(x^{(k-1)})$ 是海塞矩阵，那么容易验证，它和求解下面这一个优化问题是等价的。

$$z^+ = \arg \min_z \frac{1}{2} \|x - H^{-1} \nabla g(x) - z\|_H^2 + h(z)$$

当然提醒一下这里的 $z = x + v$ ，且 $\|x\|_H^2 = x^T H x$ 。那么同样的，我们可以定义一个新的算子

$$\text{prox}_H(x) = \arg \min_z \frac{1}{2} \|x - z\|_H^2 + h(z)$$

和正常的近端算子相比，这个算子就是把普通的2-范数改成了这种正定矩阵范数。

所以总结一下，我们就可以得到近端牛顿方法的迭代公式

长体现在公式里就是 $t_k$ 。第二就是在新公式中，其实如果说 $h(z) = 0$ ，那么就与牛顿法没有区别了。第三则是在近端牛顿方法中， $g, h$ 的性质都会影响到问题的可解性，并且一般来说，这个近端算子不再具备解析解。这个时候究竟如何利用近端牛顿方法呢？这个我们放到之后再说。

小结

本节我们主要介绍了内点法中的两个经典方法——屏障法和原始-对偶方法，我们从KKT条件出发，以它们对对偶变量的不同的处理方法来介绍它们俩的区别与联系。除此之外我们简单介绍了一下近端牛顿方法，它也是以牛顿法作为基础的用来求解特定非光滑问题的工具。所以与其说是二阶方法的荟萃（如果包括一个字没提的拟牛顿法的话），倒不如说这一节是牛顿法的狂欢。

下一节我们会继续介绍近端牛顿方法，在介绍结束后，我们会继续介绍一些在机器学习，深度学习中更具有热度的几种方法。不过它们也更加高深，更加依赖矩阵论的一些工具，所以在这之前，可能我们需要一些基础知识来做一些warm-up~

一个大学生的日常笔记

本专栏为我的个人专栏，也是我学习笔记的主要生产地。任何笔记都具有著作权，不可随意转载和剽窃。  
个人微信公众号：**cha-diary**，你可以通过它来获得最新文章更新的通知。  
《一个大学生的日常笔记》专栏目录：[笔记专栏|目录](#)  
《GetDataWet》专栏目录：[GetDataWet|目录](#)  
想要更多方面的知识分享吗？可以关注专栏：一个大学生的日常笔记。你既可以在那里找到通俗易懂的数学，也可以找到一些杂谈和闲聊。也可以关注专栏：[GetDataWet](#)，看看在大数据的世界中，一个人的心路历程。我鼓励和我相似的同志们投稿于此，增加专栏的多元性，让更多相似的求知者受益~

发布于 2020-10-11 11:20

数学 凸优化 机器学习

写下你的评论...

21 条评论

默认 最新

知乎

首发于  
一个大学生的日常笔记

08-25

回复 喜欢



wpl

赞一个，从这里看懂了屏障法和原始对偶法的区别了。

08-09

回复 喜欢



tianyueh8erobot

“注意到如果这个方法要生效，函数要有定义，就必须要有强对偶性满足”，函数是否有定义与强对偶性的关系是什么？应该是严格可行？

01-19

回复 喜欢



知乎用户fMrKPH

请问数值优化的这些内容有具体的书籍吗，想系统学习一下！还望知道的小伙伴留个言，非常感谢！

2022-04-19

回复 喜欢



知乎用户fMrKPH · 学弱渣

好的谢谢

2022-05-03

回复 喜欢



学弱渣

看这个系列的第一篇文章哈

2022-04-19

回复 喜欢



恍恍惚惚

原始对偶内点法的初始解不一定可行，假设不可行时，那怎么保证选出当前的步长使不等式成立呢？此时可能使不等式的值更接近于0但不会小于0。

2022-03-06

回复 喜欢



随机森林里的剑龙

它在对偶问题上可行解，且它处于拉格朗日函数的鞍点处。不是鞍点，是驻点

2022-02-15

回复 喜欢



随机森林里的剑龙

Proposition 1少了t趋向于无穷，不然不满足互补松弛

2022-02-15

回复 喜欢



moon

楼主，想问下在例子1中梯度v和c平行怎么比较好理解？为啥c表示在图中是那样的？

2021-12-09

回复 喜欢



connor

请问博主，求出dx后可能下一步就变成不可行点了。如果用步长搜索解决这个问题的话，准则是什么？

2021-10-31

回复 喜欢



EIPSYCONGROO

写得很好，受教了

2021-06-13

回复 喜欢

点击查看全部评论 >

写下你的评论...

文章被以下专栏收录



一个大学生的日常笔记

你能找到的最通俗易懂的大学生数学与编程知识分享



Optimizers' Garden

优化&机器学习理论拾遗

知乎  
首发于  
一个大学生的日常笔记



凸优化笔记26：不动点迭代

周游                      发表于凸优化



凸优化笔记19：近似点梯度下降

周游                      发表于凸优化

前置知识：偏微分方程、泛函分析。参考书：Jean-Baptiste Hiriart-Urruty, Claude Lemarechal, Fundamentals of Convex Analysis, Springer, 2001 Cannarsa, Piermarco; Sinestrari...

Fiddi...                      发表于Fiddi...



凸优化  
PPA

周游