**拟牛顿算法**

## 一、模型

## 二、算法描述

迭代公式：

其中，为拟牛顿方向。

拟牛顿法是求解非线性优化问题的有效方法之一，于1959年由美国Argonne国家实验室的物理学家W.C. Davidon 提出，它是一类每步迭代计算量少而又保持超线性收敛的牛顿型迭代法。其后，1963年R. Fletcher 和M.J.D. Powebl 给出了称为DFP-秩2-拟牛顿法，1964年C.G. Broyden 给出了秩1拟牛顿法。方法的收敛性是在20世纪60年代末至70年代末逐渐被证明。拟牛顿法不需要二阶导数的信息，有时比牛顿法更有效。如今，优化软件中包含了大量的拟牛顿算法用来解决无约束、约束和大规模的优化问题。

具体步骤，首先构造目标函数在当前迭代的二次模型

这里，是一个对称正定矩阵，于是取此二次模型的最优解作为搜索方向，且得到新的迭代点：

其中步长要求满足Wolfe条件，即满足

这样的迭代与牛顿法相似，区别在于用代替Hessian阵，所以拟牛顿法最关键的地方就是每一步迭代中矩阵的更新。现在假设得到一个新的迭代，并得到一个新的二次模型：

我们尽可能地利用上一步的信息来选取。具体来讲，我们要求

即

这个公式被称为割线方程，也称为拟牛顿方程或拟牛顿条件。

下面介绍两种常用的拟牛顿算法：BFGS和DFP算法。

## 三、DFP拟牛顿法

DFP拟牛顿法亦称DFP校正方法，是第一个拟牛顿法，有Davidon1959年提出，后经Fletcher和Powell解释和改进，命名时以三人名字的首字母命名。

对于拟牛顿方程

令，可得

在DFP校正方法中，假设

DFP校正公式：

其中，。

**3.1 DFP校正公式的详细推导如下**

由上面可以简化拟牛顿方程为

将带入上式，并令，得到

即

上式中参数和解的可能性有很多，取特殊情形，，，得

进一步，设

从而有，，，。可得

从而，

因此的更新公式（DFP校正公式）

**3.2 DFP拟牛顿法的算法框架**

**定理：**设对称正定，对称正定的充要条件是。

**证明：**必要性是显然的。因为，若正定，则显然有。

下证充分性。因为且正定。由DFP校正公式，对任意的，我们有

因对称正定，故存在对称正定矩阵矩阵，使得，从而利用Cauchy-Schwarz不等式得

不难发现，上式等式成立的充要条件是存在实数，使得，即。故而，若不等式(3.2.2)式严格成立，则由(3.2.1)式可得

否则，若(3.2.2) 等式成立，即存在，使得，则由(3.2.1)和(3.2.2)时可得

故对任意的，，总有，因此，对称正定。证毕。

在确定步长时，搜索准则可取Armijo准则，Wolfe准则等。在利用Armijo搜索准则时并不一定满足，此时可以对DFP校正公式做调整如下：

* DFP拟牛顿法（最优步长）算法框架如下

步0 给定初始点，终止误差。初始n阶对称正定矩阵，令

步1 计算，若，停止，输出作为近似极小点。

步2 计算搜索方向

步3 计算步长

令。

步4 确定，由DFP公式给出，即

令，转步1。

* DFP拟牛顿法（Armijo搜索）算法框架如下

步0 给定参数，，初始点，终止误差。

初始对称正定矩阵，令

DFP算法：

步1 计算，若，停止，输出作为近似极小点。

步2 计算搜索方向

步3 设是满足下列不等式的最小非负整数m，

令，。

步4 确定，若，则由DFP公式给出，即

否则，，转步1。

**3.3 收敛性分析**

在本节主要分析证明在最优步长规则下的DFP的全局收敛性和超线性收敛性。下面是拟牛顿算法的全局收敛性分析所需要的假设条件。

**假设3.3.1**

1. 在上二阶连续可微。
2. 存在M>m>0，使对任意的，

第二个假设要求目标函数在水平集内是一致凸函数，所以目标函数在该集合内有唯一极小值点。

**引理3.3.1** 设满足假设3.3.1线搜索步长规则下的下降算法产生迭代点列。则下述数列有界

**证明：**由Cauchy-Schwarz不等式，。故只须证明和均有界即可。

事实上，由于

所以

由假设3.3.1中的（2），

因而，，即

下面对进行估计。由及假设3.3.1中的（2）得

故

证毕。

**引理3.3.2** 设满足假设3.3.1，则最优步长规则下的下降算法产生的点列对应的级数和收敛。

证明：在区间[0,1]上定义函数

由假设3.3.1，对任意的，，而精确线搜索意味着.进而由假设3.3.1得

其中，。取得

上边两式关于k求和得

其中，是的极小值。于是，收敛。

利用引理3.3.1，可得级数收敛。

**引理3.3.3** 在假设1之下，设是的极小值，则对任意的，

证明：由于函数在水平集内是凸函数，故

令并利用得

证毕。

借助以上引理，可建立最优步长规则下的DFP方法的全局收敛性

**定理3.3.1**设满足假设3.3.1，则最优步长规则下的DFP方法产生的点列收敛到其最优值点。

**证明：**为方便说明，先给出和的DFP校正公式

对的表达式两边求迹得

由于，及，上式右端的中间可写成

考察最后一项的分母，利用校正公式和得

其中，第三、四个等式利用了

对上式求倒数得，

利用（3.3.2）和（3.3.3），可将（1）写成关于的迹的一个递推关系式：

从而

根据引理3.3.1，存在正常数使

下面对上式右端的中间项进行估计。

由DFP校正公式知，

由于正定，故上式右端为正。从而由引理3.3.1知存在与k无关的数，使

不妨设由

及

并结合（3.3.5），（3.3.6）得

再利用Cauchy-Schwarz不等式得

进而由（3.3.8）得

基于以上分析，下用反证法证明命题的结论。反设算法产生的点列不收敛于的唯一最小值点，则存在，使对所有的k，

由于目标函数是一致凸的，由，有

利用目标函数在水平集上有下界知数列单调有界。从而在时，，进而。故由（3.3.9）和（3.3.10）知，对充分大的k，

再结合（3.3.4），对于充分大的k成立

设 为矩阵的n个特征根，则，，···，为矩阵的特征根。由矩阵论的知识，

利用（3.3.11）式得，这是不可能的。该矛盾说明收敛于。证毕。

接下来是拟牛顿算法的超线性收敛性分析所需要的假设条件以及DFP算法在最优步长规则下的超线性收敛性。

**假设3.3.2**

1. 为二阶连续可微函数；
2. 在局部极小值点的Hessian矩阵正定；
3. 存在点的邻域，使在该邻域内Lipschitz连续。

**引理3.3.4** 设二阶连续可微，Hess阵 Lipschitz连续（常数为L）。则对任意的，

令

若非奇异，则对任意的，存在与x有关的常数，使当时，有

**证明：**对任意的，由积分的性质及Cauchy-Schwarz不等式，

得证。

对任意的及满足的，利用式，

取即得的右端不等式。

另一方面，利用范数的性质及得

取得的左端不等式。证毕。

基于此引理，可建立取单位步长的拟牛顿算法的局部超线性收敛性。

**定理 3.3.2** 设目标函数满足假设3.3.2，为一非奇异矩阵序列。若迭代过程

产生的无穷点列收敛于的局部最小值点，则点列超线性收敛到的充分必要条件是

**证明：**由知

结合式知，对任意的，

充分性 若成立，由，利用上述两式可得

由于非奇异和，由引理3.3.4知，存在及，使对任意的有

从而，

其中，。这样式即为

所以

这说明点列超线性收敛到。

必要性 设超线性收敛到。由引理3.3.4知存在及，使对任意的有

由于超线性收敛，故由上式得

再由（因为点列超线性收敛到），故有

再结合式和式可知式成立。证毕。

**定理3.3.3** 设满足定理3.3.2中的假设，是非奇异矩阵序列。假定由迭代过程产生的点列收敛到的局部最小值点。如果式成立，则超线性收敛到的充要条件是收敛到1。

**证明：**必要性 设超线性收敛到。由定理3.3.2，

再由得

由于，故上式可写成

而由非奇异和引理3.3.4知，存在及，使对任意的，

又因超线性收敛，由超线性收敛的定义和，得。

充分性 设收敛到1.由和成立。从而由定理3.3.2知超线性收敛到。证毕。

下面讨论在假设下最优步长规则下的拟牛顿DFP算法的超线性收敛性。

**定理3.3.4** 设满足假设3.3.2，是非奇异矩阵序列。假定最优步长规则下的迭代过程产生的点列收敛到的局部最小值点。则当步长有界且式成立时，，从而超线性收敛到。

**证明：**根据定理3.3.3，只须证明即可。由，

利用和的连续性，

于是

利用最优步长规则的性质和及步长有界，上式可以写成

而由式知，

两式相减得

利用假设3.3.2的(2)推知收敛到1。证毕。

**引理3.3.5** 设是非奇异对称阵，若对任意的，

则对任意n阶方阵E，下述结论成立。

（1）；

（2）；

（3），其中，

**证明：**将两边平方得

所以

再利用得

这样，

结合(3.3.23)，(3.3.24)得

得（1）的左端不等式。

其次，由于

结合得

由此得（1）的右端不等式。

对（2），利用矩阵的Frobenius范数的性质，对任意的和矩阵，

从而，利用（1）得

（2）得证。

对（3），由于

由（2）只需证

事实上，

利用（1）便得要证明的结论。容易知道。证毕。

**引理3.3.6** 设非负数列和满足

则收敛。

**证明：**首先证明数列有界。定义

显然，。利用对数函数的性质得

这表明存在某个常数，使对任意的k，。再由得

因此，

利用和的有界性，知存在使对任意的，。

下面用反证法证明收敛。主要思路是假设若它不收敛，则存在的两个子列和，它们分别收敛于两个不同的极限和。先证明，然后利用对称性得，从而由。此矛盾说明收敛。

事实上，对，由，

令得

再令，得。从而。证毕。

**引理3.3.7** 设满足假设3.3.2，并记。若DFP方法产生的点列满足

则存在常数，使得

**证明：**记，。由的DFP校正公式

得

再利用和Frobenius范数的性质得

其中，

下面利用上述三式估计。

首先，由及得

其次，由引理3.3.4和题设得

再由引理3.3.5得

于是，依次有

因此，

将上述式子代入得，其中，。证毕。

令，。在引理3.3.7中的假设下，若收敛，由引理3.3.6可知数列收敛，即存在使得

若DFP方法产生的点列收敛到最优解，则当k充分大时式成立。由于，所以下述条件可保证收敛，

虽然引理3.3.7提供的估计是可以保证数列收敛，但我们并不关注其是否收敛于0，而是借助其收敛性验证成立。为此，需对做进一步估计。

**引理3.3.8** 设满足假设3.3.2，存在正常数，，使得DFP方法产生的矩阵序列满足

其中，

**证明：**由引理3.3.7中得

记

则

由引理3.3.7的证明过程知存在使得

记，则可写成

下对进行估计。由于

利用引理3.3.5中的（3）得

再利用引理3.3.5中的（3），

于是，利用和得

令，，将上式代入即得命题结论。证毕。

利用上述引理可建立带最优步长的DFP方法的超线性收敛性。

**定理3.3.5** 设满足假设3.3.2，若最优步长规则下的DFP方法产生的点列收敛到最优解，步长有界且满足时，则超线性收敛到。

**证明：**由知，引理3.3.8中的不等式可写成

两边对k求和得

由知，。从而由知，有界。因此，

再由，若的某个子列收敛到零，则整个数列收敛到零。从而，

由定理3.3.4，结论得证。否则，若存在使对任意的k，，则由知在时，。而此时，

令得

这样，无论哪种情况，根据定理3.3.4都可以推出超线性收敛到。证毕。

**3.4 程序和算法实践**

## 四、BFGS拟牛顿法

BFGS是Broyden Fletder，Goldfarb Shanno在1970年发明，至今仍被认为是最好的拟牛顿算法，具有全局收敛性和超线性收敛速度。

**4.1 BFGS算法的更新公式**

为了推导BFGS算法，需用到对偶或互补的概念，前边已讨论过拟牛顿法Hessian阵之逆矩阵的近似矩阵需要满足方程

或

基于上述条件，可以构造的更新公式（Hessian矩阵逆矩阵的近似），如秩1算法和DFP算法。当然，也可以直接构造Hessian矩阵的近似矩阵。上述两方程十分相似，区别在于和互换。因此，给定关于的更新公式，交换和的位置，并将替换为，就可得到的更新公式。

在BFGS算法中，矩阵对应着DFP算法的，满足这种结果的两类公式称为对偶或互补的。

已知DFP算法中关于即Hseeian阵逆矩阵的近似矩阵的更新公式

利用互补概念，可得即Hessian阵的近似矩阵之更新公式为：

为获得Hessian阵的逆矩阵的近似矩阵，只需对求逆。

**4.2 谢尔曼-莫星森-渥德需公式（SMW-Sherman-Morrison-Woodbury）**

**引理（SMW）** 如果矩阵非奇异，是列向量，满足，则非奇异，其逆矩阵可用表示，即

（SM）

（SMW）

对应应用2次SMW引理，可得BFGS算法中的更新公式

下面给出的推导过程。

令，由引理（SMW），易求

其中。那么，

BFGS算法保持了拟牛顿法的一切性质，包括共轭方向的性质，也能使得近似矩阵一直保持正定。当迭代过程中一维搜索的精度不够高时，BFGS算法仍比较稳健。这一过程有助于将计算资源从追求高精度的一维搜索中释放出来，就效率而言，BFGS算法要远超DFP算法。

对于严格凸二次函数，DFP算法和BFGS算法有相同的数值效果，但对于一般的非线性函数，这两种算大表现出较大差异。理论分析和数值结果表明，BFGS算法有较强的自我校正能力，即一旦出现对目标函数的Hessian矩阵的你近似效果很差的情况，那么BFGS校正公式会在几步之内把它矫正好，而DFP校正公式要差一些，不过BFGS校正公式的这种能力也仅限于特定的步长，如Wolfe步长。

**4.3 BFGS算法流程**

BFGS算法（I） （Armijo步长规则）

步0 给定参数，，初始点，终止误差。

初始对称正定矩阵，令。

步1 计算，若，停止，输出作为近似极小点。

步2 计算搜索方向

步3 设是满足下列不等式的最小非负整数m，

令，。

步4 计算。

步5确定，计算

令，转步1。

上述算法中计算搜索方向通常要求解线性方程组进行。然而，更一般的做法是用Sherman-Morrison公式，查找给出与间的关系式：

或更进一步写成

重复利用上面步骤，我们可将BFGS算法更改如下。为了避免出现矩阵求逆运算，就将改为。整个算法中不再需要求解线性代数方程组，由矩阵向量运算即可。

BFGS算法（II）：（Armijo步长规则）

步0 给定参数，，初始点，终止误差。

初始对称正定矩阵，令。

步1 计算，若，停止，输出作为近似极小点。

步2 计算搜索方向

步3 设是满足下列不等式的最小非负整数m，

令，。

步4计算。

步5 确定，计算

令，转步1。

BFGS算法：（Wolfe步长规则）

步0 给定参数，，，初始点，终止误差。

初始对称正定矩阵，令。

步1 计算，若，停止，输出作为近似极小点。

步2 计算搜索方向

步3 设是同时满足下列不等式的最小非负整数m，

令，。

步4计算。

步5 确定，计算

令，转步1。

注：非精确搜索为拟牛顿法研究，由1976年Powell开始，他证明了带Wolfe搜索的BFGS算法的全局收敛性和超线性收敛性。

**4.4 L-BFGS算法**

在BFGS算法中，需用到一个的矩阵，当N很大时，存储这个矩阵将非常耗内存。例如，考虑N为10万时，且用double型（8字节）存储，需要多大的内存呢？我们来看

，考虑对称性，可将为一半，这对于一般的服务器是很难承受的。在机器学习问题中，像10万这样的规模还只能算是中小规模。那么，是否可以通过对BFGS算法进行改进，从而减少其迭代过程所需的内存开销呢？L-BFGS算法就是为此而设计的。

L-BFGS（Limited-memory BFGS或Limited-storage BFGS）算法是对BFGS算法进行改进（近似），基本思想是：不存储完整的矩阵，而是存储计算过程中的向量序列，需要矩阵时，利用向量序列的计算来代替。而且，向量序列也不是全存，而是固定存最新的m个。每次计算时，只利用最新的m个和。这样，我们将存储由原来的降到了。

下面讨论L-BFGS算法的具体实现过程。首先，考虑BFGS算法（II）中的迭代公式

记，，则上式可写成

如果给定初始矩阵（通常取），则依次可得

一般地，有

由上式可见，计算需用到，因此，若从，开始连续地存储m组的话，只能存到，，即只能一次计算，，···，，那么如何计算呢？

自然，要丢掉一些最早生成的向量，具体讲，计算时，可以丢掉，保存。但是舍弃一些向量后，就只能近似计算了，当时，可以构造近似计算公式。

注意，千万不要被上面（\*\*）冗长负责的形式所吓倒。事实上，有BFGS算法流程易知，的作用仅为计算的搜索方向。若能利用（\*\*）设计出一种计算得快速算法，则有效提升算法销量。Nocedal于1980年给出了一种算法，具体如下。

L-BFGS算法（计算的快速算法）

步1 初始化

；；

步2 （后向循环）

For Do

{

;

; *//*需存下来

;

}

步3 （前向循环）

For Do

{

;

; *//*需存下来

;

}

•Nocedal J. “Updating quasi-Newton matrices with limited storage”. Mathematics computation, 1980, 35(151):773-782

**4.5 收敛性分析**

**定理4.5.1** （Wolfe准则下的线搜索法收敛定理）

设是由Wolfe步长规则下的下降算法产生的序列，有下界且对任意的，在水平集

上存在且一致连续，若下降方向满足条件，则采用Wolfe准则求搜索步长时，有。

**定理 4.5.2** 设是由BFGS校正公式产生的非奇异矩阵序列，为满足Wolfe 准则的步长。若满足假设3.3.1，那么由BFGS算法产生的序列全局收敛到的极小点。

**证明：**根据定理4.5.1，我们只需验证搜索方向与负梯度方向的夹角满足条件。注意到

以下只需证明存在，使得上式定义的满足即可。

由的定义，可得

故有

利用假设3.3.1(2)可得

，即 (4.5.2)

由（4.5.1）式得

不难发现，由假设3.3.1(2)有

因此由（4.5.3）式可推得。结合（4.5.2）式，我们有

对BFGS公式

两边求迹得

为求的行列式，先给出一个秩-2校正行列式的计算式：

其中。该式的推导过程如下：容易有

从而由引理（SMW）得

为了便于应用上述公式，我们将BFGS公式写成如下形式：

利用上述公式（4.5.5）得

记

则

于是由（4.5.6）有

关于对称正定矩阵B，定义函数

则有事实上，设为B的特征值，则

由（4.5.4），（4.5.7）及（4.5.8）得

上式的最后一个不等式利用了函数在区间上的非正性及和，且正常数.于是有

下面证明数列。用反证法。若结论不成立，则对上述的常数，存在使得对所有的，有

由（4.5.9）式得

矛盾。这样便存在的无穷子列和数，使得对所有的，有。于是根据定理4.5.1 可以推出。因为在水平集上是严格凸函数，其稳定点与全局极小点是一致的也是唯一的，从而点列收敛到目标函数的唯一极小值点。证毕。

**定理4.5.3** 设满足假设3.3.2，是非奇异矩阵序列。设对某个，以1做试探步长（即）的Wolfe步长准则下的迭代过程产生的点列收敛到的局部最小值点。如果式成立，则当k充分大时，，从而超线性收敛到。

**证明：**根据定理3.3.3，只需证明对充分大的k，步长即可。

由于，故由知，

所以

即

从而由正定知，存在使对充分大的k，

下面验证当k充分大时，，函数满足Wolfe步长规则。

利用Taylor展开式和得

其中，不等式由和得到。这说明，时，Wolfe步长规则的第一式成立。

利用梯度函数的Taylor展开式和目标函数的二阶连续可微性得

结合，得

这样，时，Wolfe步长准则的第二式成立。结论得证。证毕。

根据上述结论，要建立以1做试探步长的Wolfe步长规则下的BFGS方法的超线性收敛性，只须证明成立即可，也就是证明BFGS方法产生的拟牛顿方向趋于牛顿方向。

下面将证明以1做试探步长的Wolfe步长规则下的BFGS方法产生的点列满足，然后利用定理4.5.3建立该算法的超线性收敛性。对于假设条件，除了需要满足假设3.3.2之外，同样要求迭代点列满足式。

**引理 4.5.1**设满足假设3.3.2，BFGS方法产生的点列满足

其中，为常数。

**证明：**为简单起见，记

由于，其中，

所以

而由假设3.3.2中的（3），

取，即得

证毕。

**定理4.5.4** 设满足假设3.3.2，设以1作为试探步长的Wolfe步长规则下的BFGS方法产生的点列收敛到最优解且满足式，则超线性收敛到最优解。

**证明：**借用上面引理证明过程中的记号，并记

由引理4.5.1，

所以，

将两边平方并利用得

于是

由的定义，

再由和，

由知。而由知存在正数，对于充分大的k，

利用函数关于非正知，

由，可设对充分大的k，.从而

结合知，对充分大的k，

另一方面，在的BFGS校正公式

两边分别左乘和右乘得

对正定对称阵，定义函数

由定理4.5.2中的讨论知。

利用容易计算

与类似的推导有

另外，由，及的表达式依次得到

从而由知

结合得

对上式两边关于k求和并利用得

由于上式方括号里面的值非正，而且对于所有的k，，所以

而该两式意味着

由于

从而由得

再由定理4.5.3得命题结论。证毕。

**4.6 程序和算法实践**