**对部分二阶优化算法的研究报告**

丁昭旭，学号21xxxxxxxxx

刘 翔，学号21200211005

1. 引言

深度学习（DL）的兴起和快速发展推动了一阶优化方法的研究和应用，一些经典且高效的一阶优化算法如随机梯度下降法（SGD）等在机器学习（ML）领域中得到了大规模应用。但一阶优化算法并不是完美无缺的，该类优化算法存在着包括收敛性缓慢、对学习率等超参数设置敏感、高训练误差下的停滞、以及难以摆脱平坦区域和鞍点等难以避免的缺陷。这些缺陷导致的问题在高度不凸的情况下尤为突出。于是在本文，我们以牛顿算法和部分拟牛顿算法为例，对部分二阶优化算法展开研究。

一般而言的二阶优化方法包括牛顿方法（Newton）,高斯-牛顿方法（Gauss-Newton），拟牛顿方法（Quasi-Newton）。拟牛顿方法中又以BFGS（Broyden–Fletcher–Goldfarb–Shanno algorithm）和L-BFGS（Limited-memory BFGS）方法等应用最为广泛。

牛顿方法最早于1685年由John Wallis在《A Treatise of Algebra both Historical and Practical 》一书中提出。1690年，Joseph Raphson在《Analysis aequationum universalis》提出了一种更简明的描述。Raphson再次把牛顿的方法看作纯粹的代数方法，并把它的使用限制在多项式上，但他用逐次逼近来描述方法，而不是用牛顿使用的更复杂的多项式序列。 最后，在1740年，Thomas Simpson将牛顿方法描述为使用微积分求解一般非线性方程组的迭代方法，本质上也给出了上述描述。 在同一篇文章中，Simpson还对两个方程组进行了推广，并指出牛顿方法可以通过将梯度设为零来求解优化问题。

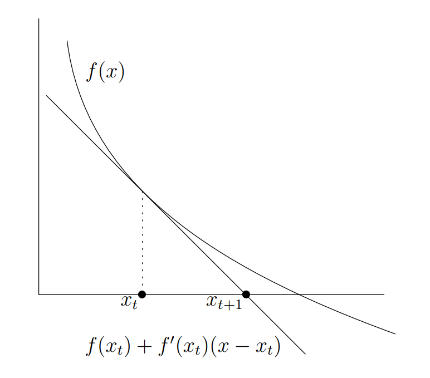
1879年，Arthur Cayley在Newton-Fourier假想问题中首次注意到将Newton方法推广到次数大于2且初值复杂的多项式复根的困难。 这为有理函数迭代理论的研究开辟了道路。

第一个拟牛顿算法是由Argonne国家实验室的物理学家William C.Davidon提出的。 他在1959年提出了第一个拟牛顿算法：DFP（Davidon–Fletcher–Powell formula）更新公式，后来由Fletcher和Powell在1963年推广，但现在很少使用。目前最常用的拟牛顿算法是SR1公式、BHHH方法、BFGS方法（由Broyden、Fletcher、Goldfarb和Shanno在1970年提出的方法）及其低内存扩展算法L-BFGS。Broyden类是DFP和BFGS方法的线性组合。

需要提出的是，虽然二阶优化方法的收敛速度相比于一阶方法有更大优势，由于其通常需要计算Hessian矩阵，对算力有着近乎苛刻的要求，二阶优化算法的发展因此受到较大限制。

1. 牛顿方法及其下降性分析

牛顿方法是数值计算中求解可微方程零点近似值常用的方法。一个简化的例子如图1所示，目标是求解可微函数在中的数值解，初值为。将在迭代初值处进行Tylor展开可得：



**图 1：**求解可微函数的零点数值解。

对上式舍去高阶项可得：

将目标值带入其中可得：

于是得到：

式(4)是舍弃了高阶项后的近似，实际上我们无法根据该式直接得到最终解，而需要将每一次计算的结果作为下一步的初值进行多次迭代。由此得到牛顿方法求可微函数零点数值解的迭代方程：

* 1. 牛顿方法在最优化上的应用

对于一阶可微函数，根据费马定理，其极值点满足，而最优化的目的即使求出这个极值点。考虑牛顿方法求可微函数零点数值解的过程，令有：

上式的迭代终点即是可微函数的极值点数值解且该式对高维空间的可微函数依然成立。

若令为牛顿方法求解最优化问题的搜索方向。对高维空间中的进行Tylor展开有：

舍去式中高阶项，并考虑以搜索方向为变量的单值函数：

为了使单次迭代下降的尽可能多，需要最小化：

于是有牛顿方法求解最优化问题的迭代方程如：

其中二阶导数项即是Hessian矩阵。

* 1. 牛顿方法与梯度下降方法的比较

对比式(10)所示的牛顿方法和梯度下降方法，可以发现，这两个方法都是基于当前迭代点的梯度信息进行搜索方向的选择的，只不过梯队下降法是在梯度的反方向上进行线搜得到下一个迭代点，而牛顿方法则是通过Hessian矩阵在梯度上进行线性变换得到搜索方向（甚至步长都不需要确定）。所以牛顿方法对函数在迭代点处的信息利用更加充分，直观来看，相比于梯度下降法，函数足够正定的情况下牛顿方法迭代得更加准确，收敛速率也会更快。

以正定二次型为例，当时有全局最优点。记迭代起始点为，有：

于是得到第一次迭代结果。

可以发现，只迭代一步，就恰好迭代到了全局最优点。因此，无论怎么选取迭代初值，二次型的Hessian矩阵的条件数有多么恶心，上述推导都保证了牛顿方法可以帮我们一步迭代到最优点。

事实上，所有基于梯度的迭代方程都可以写作如下形式：

其中对与梯度下降方法而言有，对于牛顿方法而言有。可以看到，牛顿方法的 ​是随着当前迭代点的变化动态变化的。因此，相比于梯度下降法，牛顿方法具有更加灵活的迭代过程。有的书上直接说牛顿方法就是“自适应的梯度下降法”。

* 1. 牛顿方法的下降性分析

事实上，虽然牛顿方法可是实现快速收敛，但其搜索并不总是下降的。已知牛顿方法求解最优化问题的搜索方向为，只有当，使得时，才认为其搜索方向是下降的。

令，有，，继而有

于是易知，当且仅当，也即可微函数的Hessian矩阵正定时，有，从而在的领域，，使得，也即。

为解决Hessian矩阵正定性对牛顿方法解决最优化问题的限制，研究人员提出了多个牛顿方法的变种，比如阻尼牛顿法、混合方法、LM方法等等。

阻尼牛顿法通过为牛顿搜索方向添加一个线性约束项保证其搜索方向的下降性：

混合方法结合牛顿方法和梯度下降法的特点，若牛顿方向可行，则使用牛顿方法进行迭代，若不行，则使用梯度下降法确定搜索方向：

LM方法则比较暴力，通过为Hessian矩阵加上一个项强行保证其正定：

只要足够大，新的Hessian矩阵一定是正定的。

* 1. 牛顿方法应用的主要阻碍

事实上，可微函数的Hessian矩阵的正定性和可逆性无法得到保证；即使通过各种限定条件和修正项使Hessian矩阵正定且可逆，计算和储存Hessian矩阵及其逆矩阵仍然是大规模机器学习模型中应用牛顿方法的最大阻碍。

由于每次迭代都需要计算Hessian矩阵及其逆矩阵，计算机求解逆矩阵的方法一般是求解线性方程组。n维向量的Hessian矩阵由个偏导数组成，于是求解 Hessian矩阵及其逆矩阵的时间复杂度高达。这种计算量对于规模稍大的（如参数量破万）机器学习模型而言都是不可接受的。

1. 拟牛顿方法

已知牛顿方法实际应用的主要阻碍在于Hessian矩阵的计算，难以应对较高维度、较大规模数据。于是一种显而易见的解决思路是近似求取矩阵，也就是拟牛顿方法(Quasi-Newton methods)。

记Hessian矩阵在第t步迭代的近似值为，期望。于是原始牛顿方法中的迭代式(10)变为：

进一步我们发现，牛顿方法迭代方程中使用的只有Hessian矩阵的逆，而不用具体算出Hessian矩阵。于是又记，则迭代式变为：

满足上述迭代式并且能够收敛的优化算法都可以被叫做拟牛顿法，而拟牛顿法的关键步骤也变成了怎样更新Hessian矩阵的逆的近似。

* 1. 几种典型的拟牛顿算法

记，有：

又记的更新策略为映射方式，接下来对SR1，DFP和BFGS三种拟牛顿方法进行简要介绍。

* + 1. SR1

SR1的全称为Symmetric Rank-One，由William C. Davidon在1956年首次提出。其逻辑比较简单：根据处的信息得到一个修正量直接加到当前步上得到更新后的，如下式：

于是SR1解决的主要问题就变成了的求解。由于期望，于是有：

由于Hessian矩阵是对称矩阵，可得对称。从而可表示为，于是迭代式(21)变为：

根据式(19)，将式(22)两侧乘以，有：

由于式(23)中是实数，于是：

记有：

将式(25)回代到式(23)有：

显然也是实数，于是式(26)又可以写作：

于是更加显然，当且仅当，上式成立：

将式(28)、(25)回代到式(22)有：

也即SR1算法更新的迭代公式。

虽然SR1算法可以将待优化函数的Hessian矩阵近似表达为从而减少计算量，但其得到的却无法保证其正定性。从而，SR1拟牛顿方法的优化方向不一定是下降的。

* + 1. DFP

DFP以其三位发现者Davidon, Fletcher, Powell的名字首字母大写命名，是第一个公认的拟牛顿方法。DFP的实现思路和SR1类似，都是将计算复杂的Hessian矩阵近似表示为一个对称矩阵来减少计算消耗，再使用一个对称矩阵修正量来更新。相比于SR1的，DFP拥有更大的更新自由量：，从而DFP算法的更新迭代公式为：

其中都是待定系数。

由式(19)，将式(30)两侧同乘以有：

对上式，令。方便起见，又令，于是有：

当且仅当，式(31)恒成立，于是有：

同理有：

将以上诸式诸值回代入式(30)，得到DFP更新迭代式：

* + 1. BFGS

BFGS算法由其四个提出者Broyden, Fletcher, Goldfard, Shanno的名字首字母命名，其推导过程和DFP算法类似，不同点是DFP通过矩阵推导，而BFGS通过矩阵推导而来：

由SMW公式：

令*, ，*，对式(37)取逆有：

也即BFGS算法的更新迭代公式。

* 1. 拟牛顿方法的下降性分析

依旧记当前点的下降方向为，我们期望的是迭代到时，能最大程度地比小。对进行Tylor展开：

舍去高阶项有：

其中，且默认以保证迭代式的下降性。于是相比小了。注意到当当前迭代点的位置被确定，其在可微函数*f(x)*上的梯度也被确定。于是我们希望通过改变搜索方向使尽可能变大。于是有了新的优化问题：

然而显而易见的是，当与的夹角固定，只要足够大，将无限大。因此需要给一个约束。为了与其原本的定义组成——Hessian矩阵，同时也是其在拟牛顿法中的近似产生联系，使用椭球范数对的值进行约束，且令约束值，于是式(21)所定义的优化问题成为：

由Cauchy-Schwartz不等式有：

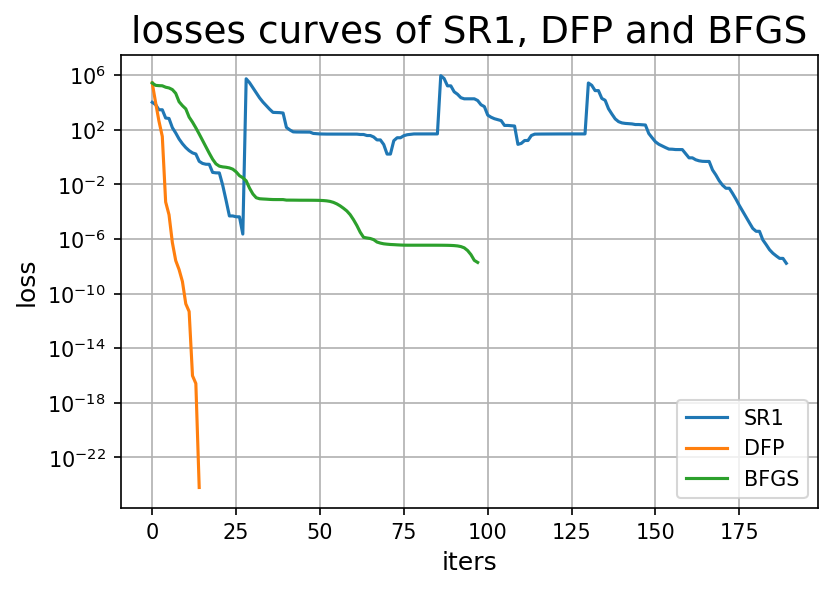
当且仅当等号成立。又有，于是由(18)，拟牛顿法是下最速下降法。

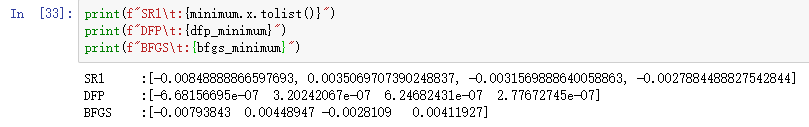
* 1. 三种拟牛顿算法的实验对比

分别使用SR1，DFP和BFGS三种方法优化以下函数：

其中，, 。已知该函数的最小值点为，于是可以设定当时停止迭代，且实验中令。

绘制三种拟牛顿算法loss曲线如下图：

并打印迭代终点如下图：

显然都接近，于是，都是收敛的。

****参考文献****