如何让神经网络收敛得更快

YWJ

2017年7月30日

1 Introduction

这篇文章在一个简单的模型上论证了使用特征去相关的样本可以让梯度下降训练神经网络的过程收敛得更快。主要内容有：

1. 第二小节论证了‘黑赛矩阵条件数’与‘使用梯度下降的优化收敛速度’的关系。
2. 第三小节介绍了文章 [4]如何论证‘一个简单的极大似然目标’在‘零点附近的黑塞阵条件数’与‘样本特征相关性’的关系。
3. 最后一小节讨论了这套理论仍然存疑的地方。

这些内容是所谓Batch Normalization的理论基石，而BN加速神经网络训练的效果是非常夸张的。单看提出BN的文章 [2]，只觉得作者脑洞清奇天赋异禀，但当你追溯BN上游的理论，你会发现BN的诞生有其必然的因素，就像N根默默无名的稻草被扔进框里后，第$N+1$根稻草有幸成为了压死骆驼的幸运儿。

你可以从https://github.com/Y-WJ/NLP-PLAYER/blob/master/%E4%B8%80%E4%BA%9B%E8%BE%85%E5%8A%A9%E6%96%87%E6%A1%A3/%E5%A6%82%E4%BD%95%E8%AE%A9%E7%A5%9E%E7%BB%8F%E7%BD%91%E7%BB%9C%E6%94%B6%E6%95%9B%E5%BE%97%E6%9B%B4%E5%BF%AB.pdf下载这份文档

2 从黑塞矩阵特征值的角度考察网络收敛速度

凸优化理论中的一个结论是：梯度下降算法的收敛速度与目标函数在极值点附近黑塞矩阵的条件数有关 [2]。为了便于讨论，这一小节将这个结论分解成下面三个问题：

1. 多元函数黑塞矩阵的几何意义。
2. 黑塞矩阵特征值的几何意义。
3. 黑塞矩阵条件数与网络收敛速度的关系。

2.1 多元函数黑塞矩阵的几何意义

矩阵可以看作一种线性变换的描述，一个确定的矩阵唯一对应着一种线性变换，如果把黑塞矩阵也看作一种变换，那么它描述的应该是是函数在“某一点到它邻域”的一阶导数的变换规则。对多元函数f来说，已知其在A点的黑塞阵$H_f(A)$ 和梯度向量$\nabla f(A)$，要计算其在B点的梯度向量$\nabla f(B)$,只需对黑塞阵沿路径积分：

$$\nabla f(B)-\nabla f(A)=\oint_{A\to B} H_f(x)dx$$

为了便于理解，这里举一个特殊情况下的例子，考虑$f(x,y)=x^2+y^2$,这个二元函数的特殊之处在于它的黑塞矩阵在全局上是个常数矩阵：

$$H_f(x)=\begin{pmatrix}2&0\\0&2\end{pmatrix}$$

因此对它的黑塞阵做路径积分相当于直接用路径向量乘以这个黑塞阵，考虑$A=(0,0)$,$b=(1,1)$的情况：

$$\nabla f(A)=(0,0)^T$$
$$\nabla f(B)=(2,2)^T$$

$$\begin{aligned}&\nabla f(A)+\oint_{A\to B}H_f(x)dx\\&=\nabla f(A)+H_f(x)\cdot\overrightarrow{AB}\\&=(0,0)^T+\begin{pmatrix}2&0\\0&2\end{pmatrix}\cdot(1,1)^T\\&=(2,2)^T=\nabla f(B)\end{aligned}$$

或者再虑这个黑塞阵不为常数阵的例子：$f(x,y)=x^3+y^3$：

$$H_f(x)=\begin{pmatrix}6x&0\\0&6y\end{pmatrix}$$

考虑从原点到点(a,b)的路径积分：

$$\begin{aligned}&\nabla f(a,b)=\nabla f(0,0)+\oint_{(0,0)\to(a,b)}H_f(x)dx\\&=(0,0)^T+\oint_{(0,0)\to(a,b)}\begin{pmatrix}6x&0\\0&6y\end{pmatrix}\begin{pmatrix}dx\\dy\end{pmatrix}\\&=\begin{pmatrix}\int_0^a 6xdx\\\int_0^b 6ydy\end{pmatrix}\\&=(3a^2,3b^2)^T\end{aligned}$$

这个结果与直接计算$f(x,y)$在(a,b)点的梯度是相等的。

2.2 黑塞矩阵特征值的几何意义

矩阵的特征值的几何意义是：被矩阵所变换的量在对应特征向量方向上的变化倍数。

一个简单的例子：向量a被矩阵A变换，

$$a=(1,2)^T$$
$$A=\begin{pmatrix}2&2\\2&2\end{pmatrix}$$

A的特征值与特征向量为：

$$\lambda_1=0,\alpha_1=(1,-1)^T$$
$$\lambda_2=4,\alpha_2=(1,1)^T$$

以两特征向量为基重新刻画向量a:

$$a=-0.5\alpha_1+1.5\alpha_2$$

在基的方向上以对应的特征值做放缩：

$$\widehat{a} = 0\alpha_1 + 6\alpha_2 = (6,6)^T = A \cdot a$$

那么黑塞矩阵的特征值的几何意义是很明显的：多元函数从‘某一点’到‘它沿对应的特征向量方向上的邻域’的一阶导数的变换系数。或者说函数在这一点沿特征向量的二阶方向导数。

2.3 黑塞矩阵条件数与网络收敛速度的关系

首先很明确的一点是：函数在某一点的凸性强弱由它在这一点各个方向上的二阶导数的绝对值刻画。或者说由多元函数在这一点的黑塞矩阵的特征值的绝对值刻画。

但麻烦的地方在于多元函数在某处各个方向上的二阶导各不相同，如果两个多元函数在同一点各个方向上的二阶导比较互有胜负，凸性的强弱似乎难以断定。

考察对凸函数F的优化过程,F有全局唯一最优点O,$\nabla F(O) = 0, \nabla^2 F(O) >= 0$,在$O$的邻域内有$A$,设由$O$指向$A$的向量为$a$,梯度下降学习率设为$\delta$,那么有：

$$A* = A - \delta \nabla F(A)$$

$$\begin{aligned}\nabla F(A) &= \nabla F(O) + H_F(O) \cdot a + O(|a|) \\ &= H_F(O) \cdot a + O(|a|)\end{aligned}$$

更新公式可转化为:

$$\begin{aligned}&A* - O = A - O - \delta \nabla F(A) \\ &\Rightarrow a* = a - \delta H_F(O) \cdot a + \delta O(|a|) \\ &\Rightarrow a* = (I - \delta H_F(O)) \cdot a + \delta O(|a|)\end{aligned}$$

随着更新的进行，向量a的模值趋近于0，$|a|$的低阶无穷小对整个式子影响越来越小，因此只需考察下面的式子：

$$a* = (I - \delta H_F(O)) \cdot a$$

这个式子说明，每次梯度下降更新前后的位置向量按矩阵$(I - \delta H_F(O))$变换。因此只需考察这个矩阵的性质，希望任意向量能由这个矩阵变换得到模值更小的向量。设$H_F(O)$的特征值为：$0 < \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$,那么$(I - \delta H_F(O))$的特征值为：$\sigma_i = 1 - \delta\lambda_i$。

为了保证更新向着收敛于极值点的方向进行，希望矩阵$(I - \delta H_F(O))$在所有特征向量方向上的变换系数绝对值都小于1：

$$\max_i |1 - \delta\lambda_i| < 1$$

考虑最大特征值λ_n:

$$1 - \delta\lambda_n > -1 \Rightarrow \delta < \frac{2}{\lambda_n}$$

不妨令$\delta = \frac{c}{\lambda_n}(0 < c < 2)$，那么有：

$$\max_i |1 - \delta\lambda_i| = \max_i |1 - \frac{\lambda_i}{\lambda_n}| \leq \max_i |1 - \frac{\lambda_1}{\lambda_n}| = \max_i |1 - \frac{1}{k(H_F(O))}|$$

这个式子的意义由以下两点说明：

1. 学习率δ满足$0 < \delta < \frac{2}{\lambda_n}$时，优化总向着极值点的方向进行(被更新的向量在各个特征方向上的放缩系数都小于1而大于-1)。

2. 当学习率满足上面的条件时，每次被更新的向量在各特征方向上的最大放缩系数由极值点黑塞阵的条件数决定。条件值越小，每次被更新向量在各特征方向上的变换系数整体越小，向量收敛至0的速度越快。

以上所有论述的结论是：在用梯度下降算法训练网络时，网络最优点附近黑塞阵的条件数越小，网络收敛的越快。

3 考察一个简单的概率模型

样本x来自空间$X=\{x_1,x_2,...,x_n\}$,样本类别$C=\{1,2,...,C\}$,模型参数$\Lambda=\{\lambda_1,\lambda_2,\cdots,\lambda_C\}$,样本$x$属于类别$c$的概率由下式确定：

$$P_\Lambda(c|x)=\frac{e^{\lambda_c^T\cdot x}}{\sum\limits_{c'\in C}e^{\lambda_{c'}^T\cdot x}}$$

优化目标为对数似然函数：

$$\begin{aligned}L&=-\tfrac{1}{N}\sum_{n=1}^{N}\ln P_\Lambda(c_n|x_n)\\&=-\tfrac{1}{N}\sum_{n=1}^{N}\ln\frac{e^{\lambda_{c_n}^T\cdot x_n}}{\sum\limits_{c'\in C}e^{\lambda_{c'}^T\cdot x_n}}\\&=-\tfrac{1}{N}\sum_{n=1}^{N}(\lambda_{c_n}^T\cdot x_n-\ln\sum_{c'\in C}e^{\lambda_{c'}^T\cdot x_n})\end{aligned}$$

3.1 计算目标的黑塞阵

求L对所有参数的一阶偏导：

$$\begin{aligned}\frac{\partial L}{\partial\lambda_{c,j}}&=-\tfrac{1}{N}\sum_{n=1}^{N}(\delta(c,c_n)\cdot x_{n,j}-\frac{e^{\lambda_c^T\cdot x_n}\cdot x_{n,j}}{\sum\limits_{c'\in C}e^{\lambda_{c'}^T\cdot x_n}})\\&=\tfrac{1}{N}\sum_{n=1}^{N}(P_\Lambda(c|x_n)-\delta(c,c_n))x_{n,j}\end{aligned}$$

其中：

$$\delta(c,c_n)=\begin{cases}0 & c\neq c_n\\1 & c=c_n\end{cases}$$

求L的黑塞阵中的任意元素：

$$\begin{aligned}
\frac{\partial L^2}{\partial \lambda_{c,j}\partial \lambda_{\bar{c},\bar{j}}} &= \frac{\partial \frac{1}{N}\sum_{n=1}^{N}\left(\frac{e^{\lambda_c^T\cdot x_n}}{\sum_{c'\in C} e^{\lambda_{c'}^T\cdot x_n}}-\delta(c,c_n)\right)x_{n,j}}{\partial \lambda_{\bar{c},\bar{j}}} \\
&= \frac{1}{N}\sum_{n=1}^{N}\frac{\frac{\partial e^{\lambda_c^T\cdot x_n}}{\partial \lambda_{\bar{c},\bar{j}}}\cdot\sum_{c'\in C} e^{\lambda_{c'}^T\cdot x_n}-e^{\lambda_c^T\cdot x_n}\cdot\frac{\partial \sum_{c'\in C} e^{\lambda_{c'}^T\cdot x_n}}{\partial \lambda_{\bar{c},\bar{j}}}}{\left(\sum_{c'\in C} e^{\lambda_{c'}^T\cdot x_n}\right)^2}\cdot x_{n,j} \\
&= \frac{1}{N}\sum_{n=1}^{N}\frac{\delta(c,\bar{c})e^{\lambda_c^T\cdot x_n}\cdot x_{n,\bar{j}}\cdot\sum_{c'\in C} e^{\lambda_{c'}^T\cdot x_n}-e^{\lambda_c^T\cdot x_n}\cdot e^{\lambda_{\bar{c}}^T\cdot x_n}\cdot x_{n,\bar{j}}}{\left(\sum_{c'\in C} e^{\lambda_{c'}^T\cdot x_n}\right)^2}\cdot x_{n,j} \\
&= \frac{1}{N}\sum_{n=1}^{N}\left\{\frac{\delta(c,\bar{c})e^{\lambda_c^T\cdot x_n}}{\sum_{c'\in C} e^{\lambda_{c'}^T\cdot x_n}}-\frac{e^{\lambda_c^T\cdot x_n}\cdot e^{\lambda_{\bar{c}}^T\cdot x_n}}{\left(\sum_{c'\in C} e^{\lambda_{c'}^T\cdot x_n}\right)^2}\right\}\cdot x_{n,j}\cdot x_{n,\bar{j}} \\
&= \frac{1}{N}\sum_{n=1}^{N}\left\{\delta(c,\bar{c})P_\Lambda(c|x_n)-P_\Lambda(c|x_n)P_\Lambda(\bar{c}|x_n)\right\}\cdot x_{n,j}\cdot x_{n,\bar{j}}
\end{aligned}$$

3.2 考虑特定点的黑塞阵

为了说明某些操作对目标函数黑塞阵的影响，不妨考察多元函数上某个特定点的黑塞阵在这种操作前后的变化。如果这个特定点具有一般性，那么对多元函数上所有的点来说，采取某些操作对它们的影响应该是相同的。这里的目标函数是关于参数矩阵Λ的多元函数，为了计算方便，考察目标函数在零点Λ_0的黑塞阵，即所有参数均取0时的黑塞阵，此时有：

$$P_{\Lambda_0}(c|x)=\frac{e^{\lambda_c^T\cdot x}}{\sum_{c'\in C} e^{\lambda_{c'}^T\cdot x}}=\frac{e^0}{\sum_{c'\in C} e^0}=\frac{1}{C}$$

目标函数在Λ_0处的黑塞阵元素：

$$\begin{aligned}
\frac{\partial L^2}{\partial \lambda_{c,j}\partial \lambda_{\bar{c},\bar{j}}} &= \frac{1}{N}\sum_{n=1}^{N}\left\{\delta(c,\bar{c})P_{\Lambda_0}(c|x_n)-P_{\Lambda_0}(c|x_n)P_{\Lambda_0}(\bar{c}|x_n)\right\}\cdot x_{n,j}\cdot x_{n,\bar{j}} \\
&= \frac{1}{N}\sum_{n=1}^{N}\left\{\delta(c,\bar{c})C^{-1}-C^{-2}\right\}\cdot x_{n,j}\cdot x_{n,\bar{j}} \\
&= C^{-1}\left\{\delta(c,\bar{c})-C^{-1}\right\}\frac{1}{N}\sum_{n=1}^{N} x_{n,j}\cdot x_{n,\bar{j}}
\end{aligned}$$

目标函数的黑塞阵是$C\times J$大小的方阵，注意到只有$\delta(c,\bar{c})$只有在$c=\bar{c}$时取1，其余时候取0，考虑用$C\times C$大小的方阵来表示这些$\delta(c,\bar{c})$的值，而当$(c,\bar{c})$确定时，每一个$(j,\bar{j})$对应着一项$\frac{1}{N}\sum_{n=1}^{N} x_{n,j}\cdot x_{n,\bar{j}}$,考虑用$J\times J$大小的方阵来表示这些项，那么两个方阵的克罗内克积恰好可以表示整个黑塞阵：

$$S=C^{-1}(I_C-C^{-1}\cdot 1_C)$$

$$X=\frac{1}{N}\sum_{n=1}^{N} x_n\cdot x_n^T$$

$$H_L(\Lambda_0)=S\otimes X$$

这可能很难理解，我在下面展开部分矩阵以辅助说明上面三个式子：

$$S=\frac{1}{C}\left\{\begin{pmatrix}1&0&\cdots&0\\0&1&\cdots&0\\\vdots&\vdots&\ddots&\vdots\\0&0&\cdots&1\end{pmatrix}-\frac{1}{C}\begin{pmatrix}1&1&\cdots&1\\1&1&\cdots&1\\\vdots&\vdots&\ddots&\vdots\\1&1&\cdots&1\end{pmatrix}\right\}=\begin{pmatrix}\frac{1}{C}-\frac{1}{C^2}&-\frac{1}{C^2}&\cdots&-\frac{1}{C^2}\\-\frac{1}{C^2}&\frac{1}{C}-\frac{1}{C^2}&\cdots&-\frac{1}{C^2}\\\vdots&\vdots&\ddots&\vdots\\-\frac{1}{C^2}&-\frac{1}{C^2}&\cdots&\frac{1}{C}-\frac{1}{C^2}\end{pmatrix}$$

$$X=\frac{1}{N}\sum_{n=1}^{N}x_n\cdot x_n^T=\begin{pmatrix}\frac{1}{N}\sum_{n=1}^{N}x_{n,1}\cdot x_{n,1}&\frac{1}{N}\sum_{n=1}^{N}x_{n,1}\cdot x_{n,2}&\cdots&\frac{1}{N}\sum_{n=1}^{N}x_{n,1}\cdot x_{n,J}\\\frac{1}{N}\sum_{n=1}^{N}x_{n,2}\cdot x_{n,1}&\frac{1}{N}\sum_{n=1}^{N}x_{n,2}\cdot x_{n,2}&\cdots&\frac{1}{N}\sum_{n=1}^{N}x_{n,2}\cdot x_{n,J}\\\vdots&\vdots&\ddots&\vdots\\\frac{1}{N}\sum_{n=1}^{N}x_{n,J}\cdot x_{n,1}&\frac{1}{N}\sum_{n=1}^{N}x_{n,J}\cdot x_{n,2}&\cdots&\frac{1}{N}\sum_{n=1}^{N}x_{n,J}\cdot x_{n,J}\end{pmatrix}$$

$$H_L(\Lambda_0)=S\otimes X=\begin{pmatrix}(\frac{1}{C}-\frac{1}{C^2})X&(-\frac{1}{C^2})X&\cdots&(-\frac{1}{C^2})X\\(-\frac{1}{C^2})X&(\frac{1}{C}-\frac{1}{C^2})X&\cdots&(-\frac{1}{C^2})X\\\vdots&\vdots&\ddots&\vdots\\(-\frac{1}{C^2})X&(-\frac{1}{C^2})X&\cdots&(\frac{1}{C}-\frac{1}{C^2})X\end{pmatrix}$$

可以这样理解这个黑塞阵，$(c,\bar{c})$确定黑塞阵中块坐标，$(j,\bar{j})$确定了块内的元素坐标，实际上这由这两个坐标确定的元素与把这两组坐标代入黑塞阵元素计算式得到的值总是相等的。为了计算黑塞阵的条件值，先计算S和X的特征值，S的对角线减去C^{-1}后变成秩为1的矩阵，说明C^{-1}是S的$C-1$重特征值，而$|S|=0$(各行减去第一行，再用各列加在第一列)，说明0也是S的特征值，记作：

$$\lambda_1(S)=0,\lambda_2(S)=C^{-1}$$

记X的特征值为序列：$0\leq\lambda_1(X)\leq\lambda_2(X)\leq\cdots\leq\lambda_J(X)$,克罗内克积的特征值是两被积矩阵特征值集合的笛卡儿积，记作：

$$\lambda(H_L(\Lambda_0))=\{0,\frac{\lambda_1(X)}{C},\frac{\lambda_2(X)}{C},\cdots,\frac{\lambda_i(X)}{C}\}$$

黑塞阵的条件数为：

$$\kappa(H_L(\Lambda_0))=\frac{\lambda_i(X)}{\lambda_1(X)}=\kappa(X)$$

在零点处，目标函数黑塞阵的条件数就是样本协方差和矩阵的条件数。降低这个值就能使以零点附近的最优点为优化目标的梯度下降收敛的更快。

3.3 样本特征的相关性对条件数的影响

对样本x_i的每一维特征，设其均值为μ_i,均方差为σ_i^2,那么有：

$$\begin{aligned}\sigma_i^2&=\tfrac{1}{N}\sum_{n=1}^{N}(x_i-\mu_i)^2\\&=\tfrac{1}{N}\sum_{n=1}^{N}x_i^2-\tfrac{1}{N}\sum_{n=1}^{N}2x_i\mu_i+\mu_i^2\\&=\tfrac{1}{N}\sum_{n=1}^{N}x_i^2-\mu_i^2\end{aligned}$$

属性间的协方差表示为：

$$\begin{aligned}\operatorname{cov}(x_i, x_j) &= E(x_i \cdot x_j) - E(x_i)E(x_j) \\ &= \frac{1}{N}\sum_{n=1}^{N} x_{n,i} \cdot x_{n,j} - \mu_i \cdot \mu_j\end{aligned}$$

矩阵X有运算：

$$X - \mu \cdot \mu^T = \begin{pmatrix} \frac{1}{N}\sum_{n=1}^{N} x_{n,1} \cdot x_{n,1} & \frac{1}{N}\sum_{n=1}^{N} x_{n,1} \cdot x_{n,2} & \cdots & \frac{1}{N}\sum_{n=1}^{N} x_{n,1} \cdot x_{n,J} \\ \frac{1}{N}\sum_{n=1}^{N} x_{n,2} \cdot x_{n,1} & \frac{1}{N}\sum_{n=1}^{N} x_{n,2} \cdot x_{n,2} & \cdots & \frac{1}{N}\sum_{n=1}^{N} x_{n,2} \cdot x_{n,J} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{N}\sum_{n=1}^{N} x_{n,J} \cdot x_{n,1} & \frac{1}{N}\sum_{n=1}^{N} x_{n,J} \cdot x_{n,2} & \cdots & \frac{1}{N}\sum_{n=1}^{N} x_{n,J} \cdot x_{n,J} \end{pmatrix} - \begin{pmatrix} \mu_1 \cdot \mu_1 & \mu_1 \cdot \mu_2 & \cdots & \mu_1 \cdot \mu_J \\ \mu_2 \cdot \mu_1 & \mu_2 \cdot \mu_2 & \cdots & \mu_2 \cdot \mu_J \\ \vdots & \vdots & \ddots & \vdots \\ \mu_J \cdot \mu_1 & \mu_J \cdot \mu_2 & \cdots & \mu_J \cdot \mu_J \end{pmatrix}$$

$$= \begin{pmatrix} \sigma_1^2 & \operatorname{cov}(x_1, x_2) & \cdots & \operatorname{cov}(x_1, x_J) \\ \operatorname{cov}(x_2, x_1) & \sigma_2^2 & \cdots & \operatorname{cov}(x_2, x_J) \\ \vdots & \vdots & \ddots & \vdots \\ \operatorname{cov}(x_J, x_1) & \operatorname{cov}(x_J, x_2) & \cdots & \sigma_J^2 \end{pmatrix} = A$$

由Weyl不等式(Weyl’ s inequalities [3])收紧X特征值的上下界。Weyl被用来界定收扰动影响的对称阵的特征值范围。考察矩阵$W = H+P$,W, H, P分别有特征值$\{\lambda_1 \le \lambda_2 \le \cdots \le \lambda_n\}$,$\{h_1 \le h_2 \le \cdots \le h_n\}$，$\{p_1 \le p_2 \le \cdots \le p_n\}$,它们之间有不等式关系：

$$h_i + p_1 \le \lambda_i \le h_i + p_n$$

即受扰动前后的特征值差异介于扰动阵最大和最小特征值之间。我认为这个结论的证明过程已经超出了计科的讨论范围，因此这里直接使用它而不再深究。

在这里，X可以看作矩阵A受到$\mu \cdot \mu^T$扰动的结果。即：

$$X = A + \mu \cdot \mu^T$$

$\mu \cdot \mu^T$的性质很好，它是向量的协方差矩阵，秩为1，这说明0是它的$J-1$重特征值，而由$Tr(\mu \cdot \mu^T) = \mu^T \cdot \mu = ||\mu||_2^2$ 可知，$||\mu||_2^2$是它剩下的单重特征值。因此有：

$$h_i \le \lambda_i \le h_i + ||\mu||_2^2$$

先考察样本的所有特征互不相关的情况，此时$cov(x_i, x_j) = 0$,A是对角阵，不妨假设这些这些均方差有：

$$\sigma_1^2 \le \sigma_2^2 \le \cdots \le \sigma_J^2$$

由Weyl不等式有：

$$\sigma_1^2 \le \lambda_i \le \sigma_J^2 + ||\mu||_2^2$$

据此可以给X的条件数定界：

$$1 \le \kappa(X) \le \frac{\sigma_J^2 + ||\mu||_2^2}{\sigma_1^2}$$

再考察样本特征互相关的情况，此时A不在再是对角阵，但A可以看作上述对角阵的受到去对角协方差阵扰动的结果：

$$X = diag(A) + \mu \cdot \mu^T + \begin{pmatrix} 0 & \operatorname{cov}(x_1, x_2) & \cdots & \operatorname{cov}(x_1, x_J) \\ \operatorname{cov}(x_2, x_1) & 0 & \cdots & \operatorname{cov}(x_2, x_J) \\ \vdots & \vdots & \ddots & \vdots \\ \operatorname{cov}(x_J, x_1) & \operatorname{cov}(x_J, x_2) & \cdots & 0 \end{pmatrix}$$

由于去对角协方差阵的特征值是不易得的，直接套用Weyl不等式似乎不可行。这里借助Gersgorin discs([1] Theorem 6.1.1)来收紧X的特征值范围,这个不等式的证明这里同样不做讨论:

令$R_i = \sum_{j,j\neq i} |\mathrm{cov}(x_i, x_j)|$，对受去对角协方差阵扰动前后的$X$的特征值$\lambda_i$和$\lambda_i^*$有:

$$\lambda_i - R_i \leq \lambda_i^* \leq \lambda_i + R_i$$

带入之前由Weyl不等式确定的边界：

$$\sigma_1^2 - R_1 \leq \lambda_J^* \leq \sigma_J^2 + ||\mu||_2^2 + R_J$$

对样本特征相关的情况，矩阵X的条件数范围：

$$1 \leq \kappa(X) \leq \frac{\sigma_J^2 + ||\mu||_2^2 + R_J}{\sigma_1^2 - R_1}$$

对规范化的样本，所有的特征自相关系数为1，均值为0，上述范围可以收紧到：

$$1 \leq \kappa(X) \leq \frac{1 + R_J}{1 - R_1}$$

这个式子的意义是更小的R_j和R_1能限制更小的矩阵X的条件数，而由$R_i = \sum_{j,j\neq i} |\mathrm{cov}(x_i, x_j)|$可知，$R_i$刻画的是样本各特征之间的相关程度，在所有样本特征互相无关时，R_i才能取到最小值0，相应地规范化样本下的矩阵X条件数可以取到最小值1，由第二节的结论，特征去相关后的样本能让目标函数在梯度下降优化中在零点附近收敛的更快。

至此这篇文章的主要结论已经论证完毕：使用特征间去相关的样本能加速用梯度下降训练的神经网络在零点附近的收敛速度。

4 一些后话

上面两节存疑的地方在于以下几点：

1. 2.3中的讨论忽略了$|a|$的高阶无穷小，但在a模值较大，即当前点离最优化点很远时，考虑这一项会使2.3中的结论变得很松，当前点离最优点越近，这种近似就越有效。这也是为何后文一直强调在‘极值点附近的’收敛速度，但2.3中没有定量地分析‘附近’与‘收敛速度’间的关系。

2. 3.2中为了简化计算用考察零点附近的黑塞阵来替代极考察值点附近的黑塞阵，但在优化任务中极值点离零点是近是远是无法事先知道的。实际上这套理论的结果跟实际测试符合的很好，我认为但无法论证这是因为两点：一是在零点成立的这些性质可能在同一凸域下的其它点也成立，这就包括了优化的极值点。二是在实际的优化任务中，总有手段使优化极值点靠近零点，比如添加L_1正则项使最优权重矩阵稀疏，或添加L_2正则项使最优权重矩阵在各个特征方向上收缩从而更靠近零点。

3. 最后的结论没有反应特征相关性与条件数之间严格的线性关系，只是通过特征的相关性确定了条件数的上界。这导致的后果是，能据此确定特征去相关的样本一定比特征相关的样本更好，但无法得出特征相关性更弱的样本是比特征相关性更强的样本更好。

4. 有关样本特征去相关，请搜索关键词‘白化’，‘高斯’，‘去相关’，这里不再赘述。

参考文献

[1] R. Horn and C. Johnson. *Matrix Analysis*. Cambridge University Press, 2005.

[2] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *Computer Science*, 2015.

[3] Lambert M. Surhone, Miriam T. Timpledon, and Susan F. Marseken. *Weyl's Inequality*. Betascript Publishing, 2010.

[4] Simon Wiesler and Hermann Ney. A convergence analysis of log-linear training. *Advances in Neural Information Processing Systems*, pages 657–665, 2011.