DOI: 10.7641/CTA.2014.30600

多层感知器模型互反奇异性区域学习动态的理论分析

郭伟立[1], 魏海坤[1†], 赵军圣[1,2], 张侃健[1]

(1. 东南大学 自动化学院 复杂工程系统测量与控制教育部重点实验室, 江苏 南京 210096;

2. 聊城大学 数学科学学院, 山东 聊城 252059)

摘要: 多层感知器神经网络(MLPs)的学习过程经常发生一些奇异性行为, 容易陷入平坦区, 这都和MLPs的参数空间中存在的奇异性区域有直接关系. 当MLPs的两个隐节点的权值接近互反时, 置换对称性会导致学习困难. 对MLPs的互反奇异性区域附近的学习动态进行分析. 本文首先得到了平均学习方程的解析表达式, 然后给出了互反奇异性区域附近的理论学习轨迹, 并通过数值方法得到了其附近的实际学习轨迹. 通过仿真实验, 分别观察了MLPs的平均学习动态, 批处理学习动态和在线学习动态, 并进行了比较分析.

关键词: 多层感知器; 神经网络; 学习动态; 奇异性; 互反

中图分类号: TP273　　**文献标识码**: A

Theoretical analysis of learning dynamics near the opposite singularities in multilayer perceptrons

GUO Wei-li[1], WEI Hai-kun[1†], ZHAO Jun-sheng[1,2], ZHANG Kan-jian[1]

(1. Key Laboratory of Measurement and Control of CSE, Ministry of Education, School of Automation, Southeast University, Nanjing Jiangsu 210096, China;

2. School of Mathematics Science, Liaocheng University, Liaocheng Shandong 252059, China)

Abstract: Owing to the existence of singularities in the parameter space, multilayer perceptrons (MLPs) may behave extremely slowly in learning or even be trapped in plateaus. When weights of two hidden units are nearly mutually opposite, the learning process will encounter difficulties because of the permutation symmetry. We investigate the learning dynamics of MLPs near opposite singularities, and derive the analytical expressions for averaged learning equations. Then, we obtain the theoretical learning trajectories near the opposite singularities. Furthermore, real learning trajectories near the opposite singularities are also calculated by using numerical methods. In simulations, we study the averaged learning dynamics, the batch mode learning dynamics and the online learning dynamics, respectively.

Key words: multilayer perceptrons; neural networks; learning dynamics; singularity; opposite

1　引言(Introduction)

笔者称学习机器是正则的: 如果其Fisher信息阵是可逆矩阵, 反之笔者称学习机器是奇异的. 研究表明包括前馈神经网络在内的几乎所有的学习机器都是奇异的[1]. 在分层系统的奇异性区域中, 由于Fisher信息阵退化, 对梯度下降法而言负梯度方向不再是下降速度最快的方向, 学习过程中Cramér-Rao定理不再成立[2–4]. 此时学习过程就会发生一些奇异性行为, 如容易陷入局部极小点, 有时学习速度变得很慢, 发生平坦区现象(如图1)[5–10]. 受奇异性区域的影响, 赤池信息准则(AIC)、贝叶斯信息准则(BIC)、最小描述长度(MDL)等模型选择准则也不再有效. 文献[11–12]研究了奇异性区域在Bayes推断中的影响, 并提出了一种应用更广泛BIC(WBIC)准则.

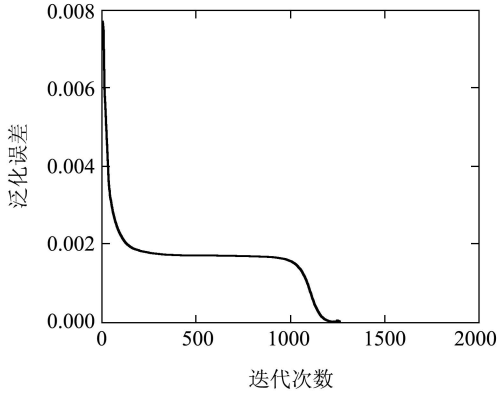


图 1 学习过程发生平坦区现象

Fig. 1 Plateau phenomenon occurred in the learning process

收稿日期: 2013−06−18; 录用日期: 2013−10−10.

†通信作者. E-mail: hkwei@seu.edu.cn.

基金项目: 国家自然科学基金重大项目资助项目(11190015); 高等学校博士学科点专项科研基金资助项目(20100092110020); 国家自然科学基金资助项目(61374006).

对典型的前馈神经网络

$$f(\boldsymbol{x},\boldsymbol{\theta})=\sum_{i=1}^{k}w_i\phi(\boldsymbol{x},\boldsymbol{J}_i),\tag{1}$$

其中: $\boldsymbol{J}_i$为输入层到隐层的权值, w_i为隐层到输出层的权值. 当w_i(或$\|\boldsymbol{J}_i\|$) $=0$时, $w_i\phi(\boldsymbol{x},\boldsymbol{J}_i)=0$, 此时无论$\|\boldsymbol{J}_i\|$(或$w_i$)的取值为何, 对$f(\boldsymbol{x},\boldsymbol{\theta})$都不会造成影响, 则在神经网络的学习过程中就无法辨识$\|\boldsymbol{J}_i\|$(或w_i). 另外, 当神经网络的参数$\boldsymbol{J}_i=\pm\boldsymbol{J}_j$时, $w_i\phi(\boldsymbol{x},\boldsymbol{J}_i)+w_j\phi(\boldsymbol{x},\boldsymbol{J}_j)=(w_i\pm w_j)\phi(\boldsymbol{x},\boldsymbol{J}_i)$, 这两个隐节点可以看作权值为$\boldsymbol{J}=\boldsymbol{J}_i=\pm\boldsymbol{J}_j$和$w=w_i\pm w_j$的一个隐节点, 此时系统只能辨识出两个输出权值的和(或差)$w_i\pm w_j$, 但不能分别分辨出w_i和w_j. 所以前馈神经网络中存在两类奇异性区域[13–14]:

$$1)\qquad w_i\|\boldsymbol{J}_i\|=\boldsymbol{0},\tag{2}$$

$$2)\qquad \boldsymbol{J}_i=\pm\boldsymbol{J}_j.\tag{3}$$

多层感知器神经网络(multilayer perceptrons, MLPs)和径向基函数(radial basis function, RBF)网络都是典型的前馈神经网络, 在实际中都获得了广泛的应用. 研究表明RBF神经网络中只存在重合奇异性($\boldsymbol{J}_i=\boldsymbol{J}_j$)和消去奇异性($w_i=0$)这两类奇异性区域[2]. 文献[15]对分层网络的重合奇异性区域($\boldsymbol{J}_i=\boldsymbol{J}_j$)附近的学习动态做了通用的数学分析, 得到了重合奇异性区域附近的通用学习轨迹. 使用文献[15]中的方法, 文献[16]得到了RBF网络的平均学习方程, 分析了重合奇异性区域附近的学习动态, 并提出了奇异性区域附近平坦区现象发生的机制. 文献[17]讨论了toy模型中重合奇异性区域附近的学习动态, 该toy模型由两个隐节点的学生模型学习一个隐节点的教师模型, 并且所有参数的维数设定为1. 以上研究成果都是以重合奇异性区域为主要研究对象. 而当MLPs的激活函数是双极性函数, 即$\phi(\boldsymbol{x},-\boldsymbol{J})=-\phi(\boldsymbol{x},\boldsymbol{J})$时, 根据前面的分析, MLPs的参数空间还存在着互反奇异性区域$\mathcal{R}=\{\boldsymbol{\theta}|\boldsymbol{J}_i=-\boldsymbol{J}_j\}$.

在本文中, 主要考虑MLPs中参数空间的互反奇异性区域:

$$\mathcal{R}=\{\boldsymbol{\theta}|\boldsymbol{J}_i=-\boldsymbol{J}_j\}.\tag{4}$$

首先求得MLPs的平均学习方程的解析表达式. 应用坐标变换, 得到互反奇异性区域附近的渐近学习轨迹, 并得到实际的学习轨迹. 在仿真实验中分别研究MLPs的平均学习动态、批处理学习动态和在线学习动态, 并对仿真结果进行分析.

2　问题描述(Problem statement)

文献[16]中对RBF网络的研究结果表明研究具有两个隐节点的模型已经足够, 所以本文主要研究具有两个隐节点的MLPs:

$$f(\boldsymbol{x},\boldsymbol{\theta})=w_1\phi(\boldsymbol{x},\boldsymbol{J}_1)+w_2\phi(\boldsymbol{x},\boldsymbol{J}_2),\tag{5}$$

其中: $\phi(\cdot)$是双极性的激活函数, $\phi(\boldsymbol{x},\boldsymbol{J}_i)=\phi(\boldsymbol{J}_i^{\mathrm{T}}\boldsymbol{x})$ $(i=1,2)$是第i个隐节点的输出, $\boldsymbol{\theta}=\{w_1,w_2,\boldsymbol{J}_1,\boldsymbol{J}_2\}$表示所有的系统参数.

假设教师模型也表示为具有两个隐节点的MLPs:

$$\begin{aligned}y=&f_0(\boldsymbol{x})=f(\boldsymbol{x},\boldsymbol{\theta}_0)+\varepsilon=\\&v_1\phi(\boldsymbol{x},\boldsymbol{t}_1)+v_2\phi(\boldsymbol{x},\boldsymbol{t}_2)+\varepsilon,\end{aligned}\tag{6}$$

其中: ε是服从0均值高斯分布的附加噪声, 与训练样本$\boldsymbol{x}$不相关, $\boldsymbol{\theta}_0=\{v_1,v_2,\boldsymbol{t}_1,\boldsymbol{t}_2\}$表示教师的所有参数. 当教师函数不能由MLPs表示时, $f(\boldsymbol{x},\boldsymbol{\theta}_0)$假定为MLPs表示的最佳逼近.

不失一般性, 假设输入$\boldsymbol{x}$服从均值为$\boldsymbol{0}$和方差为单位阵$\boldsymbol{I}_n$的高斯分布:

$$q(\boldsymbol{x})=(\sqrt{2\pi})^{-n}\exp\left(-\frac{\|\boldsymbol{x}\|^2}{2}\right).\tag{7}$$

本文定义损失函数为

$$l(y,\boldsymbol{x},\boldsymbol{\theta})=\frac{1}{2}(y-f(\boldsymbol{x},\boldsymbol{\theta}))^2.\tag{8}$$

对在线学习情形, 一次训练一个样本, 则在t时刻, 若训练样本为$(y_t,\boldsymbol{x}_t)$, 参数$\boldsymbol{\theta}_t$的迭代公式为

$$\boldsymbol{\theta}_{t+1}=\boldsymbol{\theta}_t-\eta\frac{\partial l(y_t,\boldsymbol{x}_t,\boldsymbol{\theta}_t)}{\partial\boldsymbol{\theta}_t},\tag{9}$$

其中η是学习率.

$\dfrac{\partial l(y_t,\boldsymbol{x}_t,\boldsymbol{\theta}_t)}{\partial\boldsymbol{\theta}_t}$可表示为

$$\frac{\partial l(y_t,\boldsymbol{x}_t,\boldsymbol{\theta}_t)}{\partial\boldsymbol{\theta}_t}=\left\langle\frac{\partial l(y_t,\boldsymbol{x}_t,\boldsymbol{\theta}_t)}{\partial\boldsymbol{\theta}_t}\right\rangle+\delta_t,\tag{10}$$

其中: $\langle\cdot\rangle$表示期望, 相应的教师分布函数$p_0(y,\boldsymbol{x})$为

$$p_0(y,\boldsymbol{x})=q(\boldsymbol{x})\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}(y-f_0(\boldsymbol{x}))^2\right),\tag{11}$$

δ_t为依赖于$\boldsymbol{\theta}_t$, 均值为$\boldsymbol{0}$的随机向量.

而对批处理学习情形, t时刻参数$\boldsymbol{\theta}_t$的迭代公式为

$$\boldsymbol{\theta}_{t+1}=\boldsymbol{\theta}_t-\eta\frac{1}{N}\sum_{i=1}^{N}\frac{\partial l(y_i,\boldsymbol{x}_i,\boldsymbol{\theta}_t)}{\partial\boldsymbol{\theta}_t}.\tag{12}$$

综合式(9)和式(12), 本文研究如下的平均方程:

$$\boldsymbol{\theta}_{t+1}=\boldsymbol{\theta}_t-\eta\left\langle\frac{\partial l(y_t,\boldsymbol{x}_t,\boldsymbol{\theta}_t)}{\partial\boldsymbol{\theta}_t}\right\rangle.\tag{13}$$

当输入输出$(\boldsymbol{x}_t,y_t)$遍历时, 式(12)可写为连续时间的形式, 即

$$\dot{\boldsymbol{\theta}}=-\eta\left\langle\frac{\partial l(y,\boldsymbol{x},\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right\rangle=-\eta\frac{\partial L(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}},\tag{14}$$

其中$L(\boldsymbol{\theta})=\langle l(y,\boldsymbol{x},\boldsymbol{\theta})\rangle$.

显然, 当N比较大时, 式(12)可以近似等价于式(13). 所以式(14)不仅可以反映在线学习的特性, 也可以反映批处理学习的特性. 故在本文主要研究平均学

习方程(14). 平均损失函数$L(\boldsymbol{\theta})$也是教师函数和学生函数之间的泛化误差.

3 平均学习方程的解析表达式(Analytical expression of averaged learning equations)

由于平均学习方程(14)既能反映在线学习的特性, 也能反映批处理学习的特性, 对研究MLPs的学习动态具有重要的作用, 所以给出平均学习方程的解析表达式具有重要的意义. 双曲正切函数$\tanh(x)=\dfrac{\mathrm{e}^{x}-\mathrm{e}^{-x}}{\mathrm{e}^{x}+\mathrm{e}^{-x}}$是一种双极性的激活函数, 在MLPs中具有广泛的应用. 然而其积分难以求解使得求取平均学习方程非常困难, 所以本文选取误差函数$\phi(x)=\sqrt{\dfrac{2}{\pi}}\cdot\int_{0}^{x}\exp(-\dfrac{t^{2}}{2})\mathrm{d}t$作为MLPs的激活函数. 误差函数也是一种双极性的S型函数, 具有易积分的特点. 选取误差函数作为激活函数能够大大简化计算, 得到更丰富的理论分析成果. 基于此, 本文可得到平均学习方程的解析表达式.

定理 1 若选取误差函数作为MLPs的激活函数, 则平均学习方程(14)的解析表达式为

$$\dot{\boldsymbol{J}}_i=\eta w_i(\sum_{j=1}^{2}v_jI_2(\boldsymbol{t}_j,\boldsymbol{J}_i)-\sum_{j=1}^{2}w_jI_2(\boldsymbol{J}_j,\boldsymbol{J}_i)),\tag{15}$$

$$\dot{w}_i=\eta(\sum_{j=1}^{2}v_jI_1(\boldsymbol{t}_j,\boldsymbol{J}_i)-\sum_{j=1}^{2}w_jI_1(\boldsymbol{J}_j,\boldsymbol{J}_i)),\tag{16}$$

其中: $i=1,2$,

$$I_1(\boldsymbol{t},\boldsymbol{J})=\frac{2}{\pi}\arcsin\frac{\boldsymbol{t}^{\mathrm{T}}\boldsymbol{J}}{\sqrt{1+\|\boldsymbol{t}\|^2}\sqrt{1+\|\boldsymbol{J}\|^2}},\tag{17}$$

$$I_2(\boldsymbol{t},\boldsymbol{J})=\frac{2}{\pi}\sqrt{\det(\boldsymbol{B}(\boldsymbol{t},\boldsymbol{J})^{-1})}\boldsymbol{A}^{-1}\boldsymbol{t},\tag{18}$$

$$\boldsymbol{A}=\boldsymbol{I}_n+\boldsymbol{J}\boldsymbol{J}^{\mathrm{T}},\tag{19}$$

$$\boldsymbol{B}(\boldsymbol{t},\boldsymbol{J})=\boldsymbol{A}+\boldsymbol{t}\boldsymbol{t}^{\mathrm{T}}=\boldsymbol{I}_n+\boldsymbol{J}\boldsymbol{J}^{\mathrm{T}}+\boldsymbol{t}\boldsymbol{t}^{\mathrm{T}},\tag{20}$$

$$\boldsymbol{A}^{-1}=(\boldsymbol{I}_n+\boldsymbol{J}\boldsymbol{J}^{\mathrm{T}})^{-1}=\boldsymbol{I}_n-\frac{\boldsymbol{J}\boldsymbol{J}^{\mathrm{T}}}{1+\|\boldsymbol{J}\|^2},\tag{21}$$

$$\boldsymbol{B}(\boldsymbol{t},\boldsymbol{J})^{-1}=(\boldsymbol{I}_n-\frac{\boldsymbol{A}^{-1}\boldsymbol{t}\boldsymbol{t}^{\mathrm{T}}}{1+\boldsymbol{t}^{\mathrm{T}}\boldsymbol{A}^{-1}\boldsymbol{t}})\boldsymbol{A}^{-1},\tag{22}$$

$$\det(\boldsymbol{B}(\boldsymbol{t},\boldsymbol{J})^{-1})=\frac{1}{(1+\|\boldsymbol{t}\|^2)(1+\|\boldsymbol{J}\|^2)-(\boldsymbol{t}^{\mathrm{T}}\boldsymbol{J})^2}.\tag{23}$$

证 证明过程见附录A.

由定理1的结果, 本文还可以得出泛化误差$L(\boldsymbol{\theta})$的解析表达式为

$$L(\boldsymbol{\theta})=\frac{1}{2}\sum_{i,j}v_iv_jI_1(\boldsymbol{t}_i,\boldsymbol{t}_j)-\sum_{i,j}v_iw_jI_1(\boldsymbol{t}_i,\boldsymbol{J}_j)+\frac{1}{2}\sum_{i,j}w_iw_jI_1(\boldsymbol{J}_i,\boldsymbol{J}_j).\tag{24}$$

4 互反奇异性区域附近的解析学习轨迹(Analytical trajectories near opposite singularity)

4.1 坐标变换(Coordinate transformation)

为了方便进行理论分析, 本文首先引入与文献[15]研究重合奇异性相类似的坐标变换:

$$\boldsymbol{u}=\boldsymbol{J}_1+\boldsymbol{J}_2,\tag{25}$$

$$\boldsymbol{v}=\frac{w_2\boldsymbol{J}_2+w_1\boldsymbol{J}_1}{w_2-w_1},\tag{26}$$

$$w=w_2-w_1,\tag{27}$$

$$z=\frac{w_2+w_1}{w_2-w_1},\tag{28}$$

则原坐标可表示为

$$\boldsymbol{J}_1=\frac{1}{2}(1+z)\boldsymbol{u}-\boldsymbol{v},\tag{29}$$

$$\boldsymbol{J}_2=\frac{1}{2}(1-z)\boldsymbol{u}+\boldsymbol{v},\tag{30}$$

$$w_1=\frac{1}{2}(z-1)w,\tag{31}$$

$$w_2=\frac{1}{2}(z+1)w.\tag{32}$$

在新的坐标系统中模型参数为$\boldsymbol{\xi}=\{\boldsymbol{v},w,\boldsymbol{u},z\}$, 互反奇异性区域可表示为$\mathcal{R}=\{\boldsymbol{\xi}|\boldsymbol{u}=\boldsymbol{0}\}$. 显然$z=1$ $(z=-1)$表示$w_1=0(w_2=0)$, 即文献[15]所定义的消去奇异性区域.

此时学生模型可表示为

$$f(\boldsymbol{x},\boldsymbol{\xi})=\frac{1}{2}(z-1)w\phi(\boldsymbol{x},\frac{1}{2}(1+z)\boldsymbol{u}-\boldsymbol{v})+\frac{1}{2}(z+1)w\phi(\boldsymbol{x},\frac{1}{2}(1-z)\boldsymbol{u}+\boldsymbol{v}).\tag{33}$$

对$f(\boldsymbol{x},\boldsymbol{\xi})$在$\boldsymbol{u}=\boldsymbol{0}$处做泰勒展开:

$$f(\boldsymbol{x},\boldsymbol{\xi})=w\phi(\boldsymbol{x},\boldsymbol{v})+\frac{1}{8}w(1-z^2)\boldsymbol{u}^{\mathrm{T}}\frac{\partial^2\phi(\boldsymbol{x},\boldsymbol{v})}{\partial\boldsymbol{v}\partial\boldsymbol{v}^{\mathrm{T}}}\boldsymbol{u}+O(\boldsymbol{u}^3).\tag{34}$$

此时学习方程(14)可重新表示为

$$\dot{\boldsymbol{\xi}}=-\eta\boldsymbol{T}\boldsymbol{T}^{\mathrm{T}}\langle\frac{\partial l(y,\boldsymbol{x},\boldsymbol{\xi})}{\partial\boldsymbol{\xi}}\rangle,\tag{35}$$

其中

$$\boldsymbol{T}=\frac{\partial\boldsymbol{\xi}}{\partial\boldsymbol{\theta}^{\mathrm{T}}}=\begin{bmatrix}\dfrac{z-1}{2} & \dfrac{z+1}{2w}\boldsymbol{u} & \dfrac{z+1}{2} & \dfrac{1-z}{2w}\boldsymbol{u}\\ 0 & -1 & 0 & 1\\ 1 & 0 & 1 & 0\\ 0 & \dfrac{z+1}{w} & 0 & \dfrac{1-z}{w}\end{bmatrix}\tag{36}$$

为坐标变换的Jacobi矩阵.

由式(35)可得

$$\dot{\boldsymbol{v}}=\frac{z^2+1}{2}l_{\boldsymbol{v}}+\frac{z^2+1}{2w^2}\boldsymbol{u}\boldsymbol{u}^{\mathrm{T}}l_{\boldsymbol{v}}-$$

$$\frac{z}{w}\boldsymbol{u}l_w + zl_{\boldsymbol{u}} + \frac{z^2+1}{w^2}\boldsymbol{u}l_z, \tag{37}$$

$$\dot{w} = -\frac{z}{w}\boldsymbol{u}^{\mathrm{T}}l_{\boldsymbol{v}} + 2l_w - \frac{2z}{w}l_z, \tag{38}$$

$$\dot{\boldsymbol{u}} = zl_{\boldsymbol{v}} + 2l_{\boldsymbol{u}}, \tag{39}$$

$$\dot{z} = \frac{z^2+1}{w^2}\boldsymbol{u}^{\mathrm{T}}l_{\boldsymbol{v}} - \frac{2z}{w}l_w + \frac{2(z^2+1)}{w^2}l_z, \tag{40}$$

其中:

$$l_{\boldsymbol{v}} = w\langle e(y,\boldsymbol{x},\boldsymbol{\xi})\frac{\partial\phi(\boldsymbol{x},\boldsymbol{v})}{\partial\boldsymbol{v}}\rangle + O(\boldsymbol{u}^2), \tag{41}$$

$$l_w = \langle e(y,\boldsymbol{x},\boldsymbol{\xi})\phi(\boldsymbol{x},\boldsymbol{v})\rangle + \frac{1}{8}(1-z^2)\times \langle e(y,\boldsymbol{x},\boldsymbol{\xi})\boldsymbol{u}^{\mathrm{T}}\frac{\partial^2\phi(\boldsymbol{x},\boldsymbol{v})}{\partial\boldsymbol{v}\partial\boldsymbol{v}^{\mathrm{T}}}\boldsymbol{u}\rangle + O(\boldsymbol{u}^3), \tag{42}$$

$$l_{\boldsymbol{u}} = \frac{1}{4}w(1-z^2)\langle e(y,\boldsymbol{x},\boldsymbol{\xi})\frac{\partial^2\phi(\boldsymbol{x},\boldsymbol{v})}{\partial\boldsymbol{v}\partial\boldsymbol{v}^{\mathrm{T}}}\boldsymbol{u}\rangle + O(\boldsymbol{u}^2), \tag{43}$$

$$l_z = -\frac{1}{4}wz\langle e(y,\boldsymbol{x},\boldsymbol{\xi})\boldsymbol{u}^{\mathrm{T}}\frac{\partial^2\phi(\boldsymbol{x},\boldsymbol{v})}{\partial\boldsymbol{v}\partial\boldsymbol{v}^{\mathrm{T}}}\boldsymbol{u}\rangle + O(\boldsymbol{u}^3) \tag{44}$$

是$L(\boldsymbol{\xi})$关于$\boldsymbol{\xi}$的负梯度, $e(y,\boldsymbol{x},\boldsymbol{\xi}) = f_0(\boldsymbol{x}) - f(\boldsymbol{x},\boldsymbol{\xi})$是误差.

4.2 互反奇异性区域附近的渐近学习轨迹 (Asymptotical learning trajectories near opposite singularity)

当学习轨迹到达互反奇异性区域, 即$\boldsymbol{u}=\boldsymbol{0}$附近时, 学习过程可看作为一个隐节点的学生模型去逼近两个隐节点的教师模型. 此时教师模型和学生模型可表示为如下的形式:

$$f(\boldsymbol{x},\boldsymbol{\theta_0}) = v_1\phi(\boldsymbol{x},\boldsymbol{t}_1) + v_2\phi(\boldsymbol{x},\boldsymbol{t}_2), \tag{45}$$

$$f(\boldsymbol{x},\boldsymbol{\theta}^*) = w^*\phi(\boldsymbol{x},\boldsymbol{v}^*). \tag{46}$$

$(\boldsymbol{v}^*, w^*)$表示教师模型的最佳逼近. 此时互反奇异性区域就可表示为

$$\mathcal{R}^* = \{\boldsymbol{\xi}|\boldsymbol{v}=\boldsymbol{v}^*,\ w=w^*,\ \boldsymbol{u}=\boldsymbol{0},\ z\in\mathbb{R}\}. \tag{47}$$

在$\mathcal{R}^*$附近的学习动态可由$l_{\boldsymbol{v}}, l_w, l_{\boldsymbol{u}}$和$l_z$中$\boldsymbol{u}$的高阶项所得到, 则式(41)–(44)变为

$$l_{\boldsymbol{v}}(\boldsymbol{\xi}^*) = O(\boldsymbol{u}^2), \tag{48}$$

$$l_w(\boldsymbol{\xi}^*) = \frac{1}{2}\frac{1-z^2}{w^{*2}}\boldsymbol{u}^{\mathrm{T}}H(\boldsymbol{v}^*,w^*)\boldsymbol{u} + O(\boldsymbol{u}^3), \tag{49}$$

$$l_{\boldsymbol{u}}(\boldsymbol{\xi}^*) = (1-z^2)H(\boldsymbol{v}^*,w^*)\boldsymbol{u} + O(\boldsymbol{u}^2), \tag{50}$$

$$l_z(\boldsymbol{\xi}^*) = -z\boldsymbol{u}^{\mathrm{T}}H(\boldsymbol{v}^*,w^*)\boldsymbol{u} + O(\boldsymbol{u}^3), \tag{51}$$

其中

$$H(\boldsymbol{v}^*,w^*) = \frac{1}{4}w^*\langle e(y,\boldsymbol{x},\boldsymbol{\xi})\frac{\partial^2\phi(\boldsymbol{x},\boldsymbol{v})}{\partial\boldsymbol{v}\partial\boldsymbol{v}^{\mathrm{T}}}\rangle|_{\boldsymbol{\xi}=\boldsymbol{\xi}^*}. \tag{52}$$

此时$l_{\boldsymbol{u}}(\boldsymbol{\xi}^*)$是$\boldsymbol{u}$的同阶无穷小量, $l_{\boldsymbol{v}}(\boldsymbol{\xi}^*), l_w(\boldsymbol{\xi}^*)$和$l_z(\boldsymbol{\xi}^*)$都是$\boldsymbol{u}^2$的同阶无穷小量. 本文只关注$\boldsymbol{u}$和$z$的变化, 则对$\mathcal{R}^*$附近的学习动态, 式(39)和式(40)可重新表示为

$$\dot{\boldsymbol{u}} = 2(1-z^2)H(\boldsymbol{v}^*,w^*)\boldsymbol{u}, \tag{53}$$

$$\dot{z} = -\frac{z(1-z^2)}{w^{*2}}\boldsymbol{u}^{\mathrm{T}}H(\boldsymbol{v}^*,w^*)\boldsymbol{u} - \frac{2z(z^2+1)}{w^{*2}}\boldsymbol{u}^{\mathrm{T}}H(\boldsymbol{v}^*,w^*)\boldsymbol{u}. \tag{54}$$

定理 2　由式(53)和式(54), 可得到$\mathcal{R}^*$附近的平均学习方程的轨迹为

$$h(\boldsymbol{u}) = \frac{1}{2}\boldsymbol{u}^{\mathrm{T}}\boldsymbol{u} = \frac{2w^{*2}}{3}\log\frac{(z^2+3)^2}{|z|} + C, \tag{55}$$

其中$h(\boldsymbol{u}) = \frac{1}{2}\boldsymbol{u}^{\mathrm{T}}\boldsymbol{u}$, C为由初始模型参数$(u^{(0)}, z^{(0)})$决定的常数. 轨迹如图2所示.

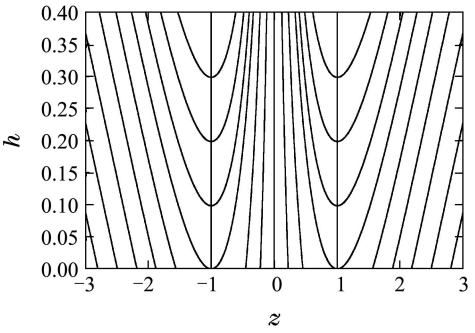


图 2 $\mathcal{R}^*$附近的理论学习轨迹

Fig. 2 Theoretical learning trajectories near $\mathcal{R}^*$

证　令$h(\boldsymbol{u}) = \frac{1}{2}\boldsymbol{u}^{\mathrm{T}}\boldsymbol{u}$, 由式(53)和式(54)可得

$$\dot{h} = \boldsymbol{u}^{\mathrm{T}}\dot{\boldsymbol{u}} = \frac{2w^{*2}(z^2-1)}{z(z^2+3)}\dot{z}. \tag{56}$$

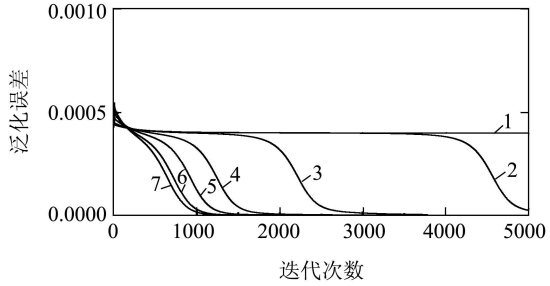
通过解微分方程(56), 可得到$\mathcal{R}^*$附近的平均学习方程的轨迹为式(55)中的形式.

4.3 实际平均学习轨迹(Real averaged learning trajectories)

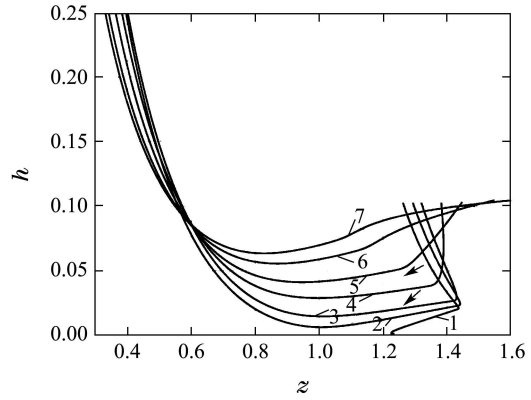
在给出互反奇异性区域附近的理论学习轨迹后, 笔者现在研究MLPs在学习过程中互反奇异性区域附近实际的学习轨迹. 由于平均学习方程(14)是常微分方程, 所以给定教师函数的权值和学生函数的初始权值后, 通过解微分方程就可以得出学生函数各个参数的学习轨迹, 进一步就可以得到h和z的学习轨迹. 本文取教师函数的权值为$\boldsymbol{t}_1 = (0.2, -0.5)^{\mathrm{T}}$, $w_1 = 0.5$, $\boldsymbol{t}_2 = (0.7, 0.4)^{\mathrm{T}}$, $w_2 = 0.3$. 给定学生函数的初始权值, 可以得出$h\sim z$的轨迹. 图3表示由不同的学生函数初始权值得出的泛化误差轨迹和相应的$h\sim z$轨迹.

对比图2和图3(b), 可以看出图3(b)和图2的右半部分比较相似, 而随着h越来越大时, 两图之间的差别越来越大. 这是因为理论学习轨迹是通过在$\boldsymbol{u}=\boldsymbol{0}$, 即$h=0$处做泰勒展开求得, 所以在$h$比较小时能保证足够的精度, 而当$h$越来越大时, 误差也会越来越大. 由

图3(b)的曲线1可以看出在训练之后h接近为0, 即两个隐节点接近互反, 此时由图3(a)中的曲线1可以看出泛化误差基本不变. 图3(b)中的曲线2–7表明学习轨迹穿越$z=1$即消去奇异性区域, 对应图3(a)中的曲线可以看出误差曲线发生了明显的平坦区现象, 最终到达最佳逼近.



(a) 泛化误差轨迹



(b) $h\sim z$轨迹

图 3 $\mathcal{R}^*$附近的实际学习轨迹

Fig. 3 Real learning trajectories near $\mathcal{R}^*$

5 仿真例子(Simulation examples)

本节对平均学习动态(14)、批处理学习动态(12)和在线学习动态(9)进行比较. 通过解微分方程(15)和式(16), 可以用数值方法求得MLPs的平均学习动态. 批处理学习动态和在线学习动态通过仿真实验得出. 选取如下的教师函数:

$$f_0(\boldsymbol{x})=v_1\phi(\boldsymbol{x},\boldsymbol{t}_1)+v_2\phi(\boldsymbol{x},\boldsymbol{t}_2),\qquad(57)$$

其中: $w_1=0.5,\boldsymbol{t}_1=(0.2,-0.5)^{\mathrm{T}},w_2=0.3,\boldsymbol{t}_2=(0.7,0.4)^{\mathrm{T}}$.

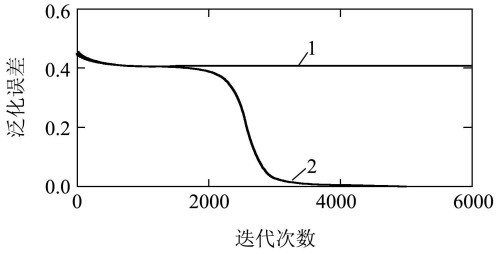
5.1 平均学习动态(Averaged learning dynamics)

本文首先研究MLPs的平均学习动态. 除了泛化误差这个重要的指标外, 由于h接近0时可表示两个隐节点接近互反, 所以h的轨迹也具有重要的意义. 表1和表2为学生函数的初始权值和通过解平均学习方程得到的最终权值.

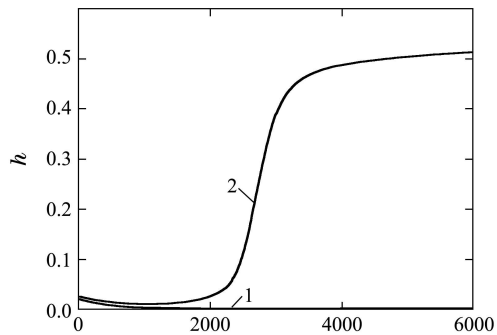
图4中的曲线1和曲线2表示当学生函数初始权值分别取表1和表2中的值时泛化误差$L(\boldsymbol{\theta})$和h的学习轨迹. 通过图4(b)中的曲线1, 可以看到h接近于0, 即学生函数的两个隐节点接近互反, 此时学生参数进入互反奇异性区域. 由图4(a)中的曲线1可以看出泛化误差基本不变, 学习过程受到互反奇异性区域的严重影响. 由表1的最终权值可以看出两个隐节点并没有完全互反, 但此时

$$\frac{\partial L(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}=10^{-4}\times(0.1655,0.0095,-0.1066,-0.1057,0.0501,0.1223)^{\mathrm{T}}$$

已经很小, 说明即使经过更长的运算, 两个隐节点已经基本停止迭代, 学生参数也很难离开互反奇异性区域, 学生函数难以达到全局最优.



(a) 泛化误差轨迹



(b) h的轨迹

图 4 平均学习动态

Fig. 4 Averaged learning dynamics

表2中的初始权值仅是对表1中的初始权值做一个小的改动, 由图4(a)中的曲线2可以看出$L(\boldsymbol{\theta})$有一段时间基本不变, 然后突然快速下降, 发生了平坦区现象. 由表2的最终权值可以看出虽然$\boldsymbol{J}_2=(-0.1811,0.4943)^{\mathrm{T}}$, $w_2=-0.5179$与教师参数$\boldsymbol{t}_1=(0.2,-0.5)^{\mathrm{T}}$, $w_1=0.5$符号不同, 但由于激活函数为双极性函数, $w_2\phi(\boldsymbol{x},\boldsymbol{J}_2)=-v_1\phi(\boldsymbol{x},-\boldsymbol{t}_1)=v_1\phi(\boldsymbol{x},\boldsymbol{t}_1)$, 即学生函数已经到达了全局最优点, 实现了最佳逼近.

表 1 学习过程陷入互反奇异性区域

Table 1 The learning process trapped in the opposite singularities

	初始$\boldsymbol{J}$	初始w	最终$\boldsymbol{J}$	最终w
隐节点1	(0.40, 0.10)	−0.12	(0.2901, −0.1457)	−0.1152
隐节点2	(−0.50, 0.35)	−1.08	(−0.3309, 0.1667)	−0.9129

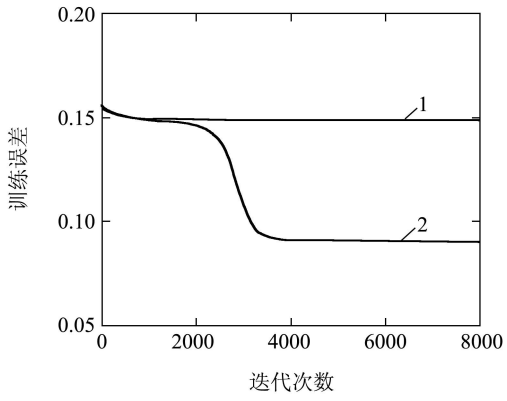
表 2 学习过程受奇异性影响发生平坦区现象

Table 2 Plateau occurred in the learning process affected by singularities

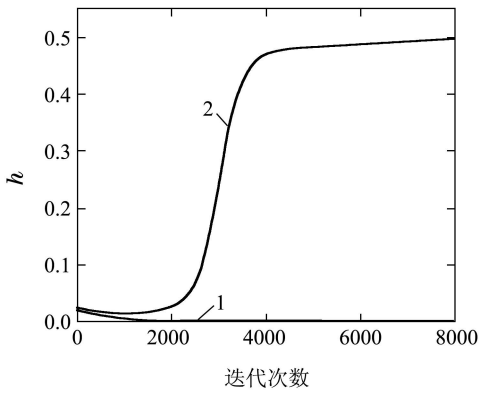
	初始$\boldsymbol{J}$	初始w	最终$\boldsymbol{J}$	最终w
隐节点1	(0.40, 0.10)	−0.12	(0.6672, 0.3947)	0.3199
隐节点2	(−0.50, 0.35)	−0.90	(−0.1811, 0.4943)	−0.5179

5.2 批处理学习动态(Batch mode learning dynamics)

下面本文研究批处理学习中MLPs的学习动态. 由教师函数(43)生成200个样本, 样本输入服从标准高斯分布, 噪声ε服从均值为0和方差为0.05的高斯分布. 本文使用梯度下降法进行训练, 学习率$\eta=0.005$, 迭代次数为8000. 在批处理学习中本文用训练误差代替泛化误差. 图5(a)和图5(b)分别表示训练误差和h的轨迹, 其中曲线1为学生参数取表1中的初始值, 曲线2为学生参数取表2中的初始值.



(a) 训练误差轨迹



(b) h的轨迹

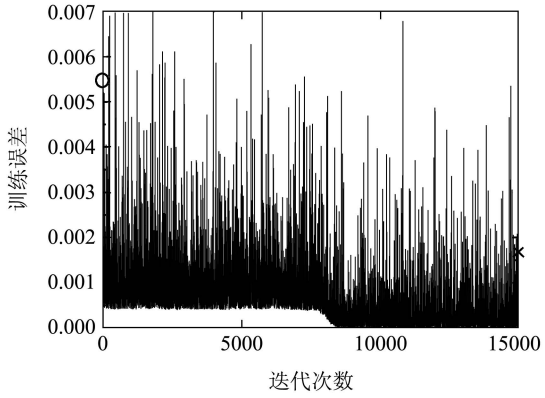
图 5 批处理学习动态

Fig. 5 Batch mode dynamics of learning

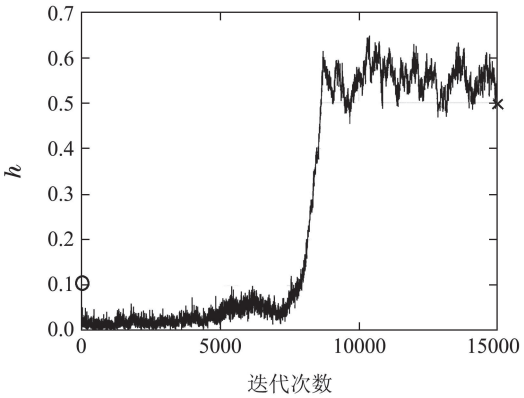
由于样本数和噪声的影响, 在批处理学习中训练误差达不到0. 而平均学习方程在计算时对教师分布函数取期望, 所以取期望后系统的平均学习动态就避免了样本数和噪声的影响. 对比平均学习动态和批处理学习动态的泛化误差轨迹和h的学习轨迹, 可以发现两者的轨迹十分相似, 而且训练结束后由平均学习方程训练得到的模型最终权值和由批处理模式训练得到的模型最终权值基本相同, 因此本文可以使用平均学习方程来研究MLPs的批处理学习动态. 使用平均学习方程可避免样本和噪声的影响, 并且学习速度更快, 训练结果也更稳定.

5.3 在线学习动态(Online learning dynamics)

现在本文研究在线学习情形下MLPs的学习动态. 批处理学习模式为所有样本同时训练, 而在线学习模式为一次训练一个样本. 样本的生成方式与批处理学习模式中相同. 笔者生成15000个样本, 每个样本训练200次, 学习率$\eta=0.05$. 图6(a)和图6(b)分别表示训练误差和h的轨迹, 学生参数初始值取表1中的初始值. “○”和“×”分别表示在线学习过程中的初始状态和最终状态.



(a) 训练误差轨迹



(b) h的轨迹

图 6 在线学习动态

Fig. 6 Online dynamics of learning

由图6的曲线可以看出由于样本的随机性和噪声的存在, 训练过程一直波动. 对比图4和图6的曲线可以看出在线学习过程中的学习曲线是沿着平均学习过程的学习曲线波动. 由图6(b)可以看出在训练开始后h迅速下降, 接近于0, 学习过程在重合奇异性区域附近波动, 学习过程受到奇异性区域的影响. 由图

4和图5的仿真结果可以看出平均学习模式和批处理学习模式中学习过程受到重合奇异性区域影响后会最终陷入重合奇异性区域, 学生模型到达局部最优点. 这是由于平均学习动态计算过程中经过对教师函数取期望后不再受样本输入随机性和附加噪声的影响, 而批处理模式中一次学习多个样本, 也会减少这种随机性的影响, 所以平均学习动态和批处理学习动态会陷入互反奇异性区域. 而在线学习模式的情形不同, 由于一次只训练一个样本, 样本的随机性和附加噪声的影响使得学习过程在重合奇异性区域附近随机游走, 由图6可以看出学习过程最终离开重合奇异性区域从而达到最优值, 发生了明显的平坦区现象.

6 结论(Conclusions)

多层感知器神经网络(MLPs)是神经网络中应用最广泛的前馈神经网络之一, 然而在用BP算法训练时经常出现学习速度很慢, 容易陷入极小点等奇异行为. 经研究发现, 这些奇异行为都是由参数空间中存在的奇异性区域导致的. 在这些奇异性区域, Fisher信息阵退化, 导致学习困难, 学习过程变得很慢. 当MLPs的两个隐节点权值接近互反时, Fisher信息阵接近奇异, 学习过程受到很大影响, 因此很有必要研究互反奇异性区域附近的学习动态. 为了克服理论分析中传统激活函数$\tan(x)$积分困难, 本文选取误差函数作为MLPs的激活函数, 并且得到了MLPs的平均学习方程的解析表达式. 引入坐标变换和进行泰勒展开, 得到了MLPs在互反奇异性附近的理论学习轨迹, 并通过使用平均学习方程得到了实际的学习轨迹, 进行了对比分析. 在仿真实验中本文分别对MLPs的平均学习动态、批处理学习动态和在线学习动态进行了比较分析. 仿真结果表明可以使用平均学习方程来研究MLPs的批处理学习动态.

参考文献(References):

[1] WATANABE S. Almost all learning machines are singular [C] //*Proceedings of IEEE Symposium on Foundations of Computational Intelligence*. Honolulu, HI: IEEE, 2007: 383 – 388.

[2] FUKUMIZU K. A regularity condition of the information matrix of a multilayer perceptron network [J]. *Neural Networks*, 1996, 9(5): 871 – 879.

[3] FUKUMIZU K, AMARI S. Local minima and plateaus in hierarchical structure of multilayer perceptrons [J]. *Neural Networks*, 2000, 13(3): 317 – 327.

[4] AMARI S, PARK H, OZEKI Z. Singularities affect dynamics of learning in neuromanifolds [J]. *Neural Computation*, 2006, 18(5): 1007 – 1065.

[5] SAAD D, SOLLA A. Exact solution for online learning in multilayer neural networks [J]. *Physical Review Letters*, 1995, 74(21): 4337 – 4340.

[6] FREEMAN J A S, SAAD D. Dynamics of on-line learning in radial basis function networks [J]. *Phisical Review E*, 1997, 56(1): 907 – 918.

[7] AMARI S, NAGAOKA H. *Information Geometry* [M]. New York: AMS and Oxford University Press, 2000.

[8] PARK H, INOUE M, OKADA M. Online learning dynamics of multilayer perceptrons with unidentifiable parameters [J]. *Journal of Physics A: Mathematical and General*, 2003, 36(47): 11753 – 11764.

[9] AMARI S, OZEKI T, COUSSEAU F, et al. Dynamics of learning in hierarchical models–singularity and milnor attractor [C] //*Advances in Cognitive Neurodynamics (II)*. Hangzhou: Springer Netherlands, 2011: 3 – 9.

[10] 魏海坤, 李奇, 宋文忠. 梯度算法下RBF网的参数变化动态 [J]. 控制理论与应用, 2007, 24(3): 356 – 360.
(WEI Haikun, LI Qi, SONG Wenzhong. Gradient learning dynamics of radial basis function networks [J]. *Control Theory & Applications*, 2007, 24(3): 356 – 360.)

[11] WATANABE S. Algebraic geometrical methods for hierarchical learning machines [J]. *Neural Networks*, 2001, 14(8): 1049 – 1060.

[12] WATANABE S. A widely applicable bayesian information criterion [J]. *Journal of Machine Learning Research*, 2013, 14(3): 867 – 897.

[13] PARK H, OZEKI T. Singularity and slow convergence of the EM algorithm for gaussian mixtures [J]. *Neural Process Letters*, 2009, 29(1): 45 – 59.

[14] FREEMAN J A S, SAAD D. On-line learning in radial basis function networks [J]. *Neural Computation*, 1997, 9(7): 1601 – 1622.

[15] WEI H, ZHANG J, COUSSEAU F, et al. Dynamics of learning near singularities in layered networks [J]. *Neural Computation*, 2008, 20(3): 813 – 843.

[16] WEI H, AMARI S. Dynamics of learning near singularities in radial basis function networks [J]. *Neural Networks*, 2008, 21(7): 989 – 1005.

[17] COUSSEAU F, OZEKI T, AMARI S. Dynamics of learning in multilayer perceptrons near singularities [J]. *IEEE Transactions on Neural Networks*, 2008, 19(8): 1313 – 1328.

附录 A(Appendix A)

由式(6)可得

$$y - f_0(\boldsymbol{x}) = \varepsilon \sim \mathcal{N}(0,1), \tag{A1}$$

则有

$$\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}\exp(-\frac{1}{2}(y-f_0(\boldsymbol{x}))^2)\mathrm{d}y = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}\exp(-\frac{\varepsilon^2}{2})\mathrm{d}\varepsilon = 1. \tag{A2}$$

$I_1(\boldsymbol{s},\boldsymbol{v})$和$I_2(\boldsymbol{s},\boldsymbol{v})$可重新表示为

$$\begin{aligned} I_1(\boldsymbol{s},\boldsymbol{v}) =\ & (2\pi)^{-\frac{n}{2}}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\phi(\boldsymbol{s}^{\mathrm{T}}\boldsymbol{x})\phi(\boldsymbol{v}^{\mathrm{T}}\boldsymbol{x})\mathrm{e}^{(-\frac{1}{2}\|\boldsymbol{x}\|^2)}\cdot \\ & \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}(y-f_0(\boldsymbol{x}))^2\right)\mathrm{d}y\mathrm{d}\boldsymbol{x} = \\ & (2\pi)^{-\frac{n}{2}}\int_{-\infty}^{\infty}\phi(\boldsymbol{s}^{\mathrm{T}}\boldsymbol{x})\phi(\boldsymbol{v}^{\mathrm{T}}\boldsymbol{x})\mathrm{e}^{(-\frac{1}{2}\|\boldsymbol{x}\|^2)}\mathrm{d}\boldsymbol{x}, \end{aligned} \tag{A3}$$

$$\begin{aligned} I_2(\boldsymbol{s},\boldsymbol{v}) =\ & (2\pi)^{-\frac{n}{2}}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\phi(\boldsymbol{s}^{\mathrm{T}}\boldsymbol{x})\frac{\partial\phi(\boldsymbol{v}^{\mathrm{T}}\boldsymbol{x})}{\partial\boldsymbol{v}}\mathrm{e}^{(-\frac{1}{2}\|\boldsymbol{x}\|^2)}\cdot \\ & \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}(y-f_0(\boldsymbol{x}))^2\right)\mathrm{d}y\mathrm{d}\boldsymbol{x} = \\ & (2\pi)^{-\frac{n}{2}}\int_{-\infty}^{\infty}\phi(\boldsymbol{s}^{\mathrm{T}}\boldsymbol{x})\frac{\partial\phi(\boldsymbol{v}^{\mathrm{T}}\boldsymbol{x})}{\partial\boldsymbol{v}}\mathrm{e}^{(-\frac{1}{2}\|\boldsymbol{x}\|^2)}\mathrm{d}\boldsymbol{x}. \end{aligned} \tag{A4}$$

由莱布尼兹积分法则可得

$$I_2(\boldsymbol{s},\boldsymbol{v}) = \frac{\partial I_1(\boldsymbol{s},\boldsymbol{v})}{\partial\boldsymbol{v}}, \tag{A5}$$

$$I_2(\boldsymbol{s},\boldsymbol{v}) =$$
$$(2\pi)^{-\frac{n}{2}}\int_{-\infty}^{\infty}\phi(\boldsymbol{s}^{\mathrm{T}}\boldsymbol{x})\frac{\partial\phi(\boldsymbol{v}^{\mathrm{T}}\boldsymbol{x})}{\partial\boldsymbol{v}}\mathrm{e}^{(-\frac{1}{2}\|\boldsymbol{x}\|^2)}\mathrm{d}\boldsymbol{x} =$$
$$(2\pi)^{-\frac{n}{2}}\frac{2}{\pi}\int_{-\infty}^{\infty}\phi(\boldsymbol{s}^{\mathrm{T}}\boldsymbol{x})\boldsymbol{x}\mathrm{e}^{(-\frac{1}{2}(\boldsymbol{v}^{\mathrm{T}}\boldsymbol{x})^2)}\mathrm{e}^{(-\frac{1}{2}\|\boldsymbol{x}\|^2)}\mathrm{d}\boldsymbol{x} =$$
$$(2\pi)^{-\frac{n}{2}}\frac{2}{\pi}\int_{-\infty}^{\infty}\boldsymbol{x}\phi(\boldsymbol{s}^{\mathrm{T}}\boldsymbol{x})\mathrm{e}^{(-\frac{1}{2}(\|\boldsymbol{x}\|^2+(\boldsymbol{v}^{\mathrm{T}}\boldsymbol{x})^2))}\mathrm{d}\boldsymbol{x} =$$
$$\frac{2}{\pi}\sqrt{\det(\boldsymbol{B}^{-1})}\boldsymbol{A}^{-1}\boldsymbol{s}, \tag{A6}$$

其中:

$$\boldsymbol{A} = \boldsymbol{I}_n + \boldsymbol{v}\boldsymbol{v}^{\mathrm{T}}, \tag{A7}$$
$$\boldsymbol{B} = \boldsymbol{A} + \boldsymbol{s}\boldsymbol{s}^{\mathrm{T}}. \tag{A8}$$

根据Sherman-Morrison公式, 可得

$$\boldsymbol{A}^{-1} = (\boldsymbol{I}_n + \boldsymbol{v}\boldsymbol{v}^{\mathrm{T}})^{-1} = \boldsymbol{I}_n - \frac{\boldsymbol{v}\boldsymbol{v}^{\mathrm{T}}}{1+\|\boldsymbol{v}\|^2}, \tag{A9}$$

$$\boldsymbol{B}^{-1} = (\boldsymbol{A}+\boldsymbol{s}\boldsymbol{s}^{\mathrm{T}})^{-1} = (\boldsymbol{I}_n - \frac{\boldsymbol{A}^{-1}\boldsymbol{s}\boldsymbol{s}^{\mathrm{T}}}{1+\boldsymbol{s}^{\mathrm{T}}\boldsymbol{A}^{-1}\boldsymbol{s}})\boldsymbol{A}^{-1}. \tag{A10}$$

使用Sylvester行列式定理可得

$$\det(\boldsymbol{A}^{-1}) = 1 - \frac{\|\boldsymbol{v}\|^2}{1+\|\boldsymbol{v}\|^2} = \frac{1}{1+\|\boldsymbol{v}\|^2}, \tag{A11}$$

$$\det(\boldsymbol{B}^{-1}) = \frac{1}{(1+\|\boldsymbol{s}\|^2)(1+\|\boldsymbol{v}\|^2)-(\boldsymbol{s}^{\mathrm{T}}\boldsymbol{v})^2}. \tag{A12}$$

由式(A5), I_1可通过I_2关于$\boldsymbol{v}$积分求得

$$I_1(\boldsymbol{s},\boldsymbol{v}) = \int_{-\infty}^{\boldsymbol{v}} I_2(\boldsymbol{s},\boldsymbol{v})\mathrm{d}\boldsymbol{v} =$$
$$\frac{2}{\pi}\int_{-\infty}^{\boldsymbol{v}}\frac{\boldsymbol{A}^{-1}\boldsymbol{s}}{\sqrt{(1+\|\boldsymbol{s}\|^2)(1+\|\boldsymbol{v}\|^2)-(\boldsymbol{s}^{\mathrm{T}}\boldsymbol{v})^2}}\mathrm{d}\boldsymbol{v} =$$
$$\frac{2}{\pi}(\arcsin\frac{\boldsymbol{s}^{\mathrm{T}}\boldsymbol{v}}{\sqrt{1+\|\boldsymbol{s}\|^2}\sqrt{1+\|\boldsymbol{v}\|^2}}+C), \tag{A13}$$

其中C是一个常数.

由式(A3)可知$I_1(\boldsymbol{0},\boldsymbol{0}) = 0$, 故$C = 0$. 由此可得

$$I_1(\boldsymbol{s},\boldsymbol{v}) = \frac{2}{\pi}\arcsin\frac{\boldsymbol{s}^{\mathrm{T}}\boldsymbol{v}}{\sqrt{1+\|\boldsymbol{s}\|^2}\sqrt{1+\|\boldsymbol{v}\|^2}}. \tag{A14}$$

作者简介:

郭伟立　(1987–), 男, 博士研究生, 研究方向为前馈神经网络奇异学习动态, E-mail: weiliguo@seu.edu.cn;

魏海坤　(1971–), 男, 博士, 教授, 研究方向为流程工业综合自动化系统设计、实时系统设计、神经网络学习理论、模式识别和人工智能等, E-mail: hkwei@seu.edu.cn;

赵军圣　(1980–), 男, 博士研究生, 讲师, 研究方向为非线性系统奇异学习动态, E-mail: zhaojunshshao@163.com;

张侃健　(1972–), 男, 博士, 教授, 研究方向为非线性系统的鲁棒控制、优化控制等, E-mail: kjzhang@seu.edu.cn.