第27卷第3期 数学研究与评论 Vol.27, No.3
2007年8月 JOURNAL OF MATHEMATICAL RESEARCH AND EXPOSITION Aug., 2007

文章编号: 1000-341X(2007)03-0643-11 **文献标识码**: **A**

带惩罚项与随机输入的 BP 神经网络在线梯度学习算法的收敛性

鲁慧芳[1,2], 吴微[1], 李正学[1]
(1. 大连理工大学应用数学系, 辽宁 大连 116024; 2. 山东交通学院数理系, 山东 济南 250023)
(E-mail: wuweiw@dlut.edu.cn)

摘 要: 本文对三层 BP 神经网络中带有惩罚项的在线梯度学习算法的收敛性问题进行了研究. 在网络训练每一轮开始执行之前, 对训练样本随机进行重排, 以使网络学习更容易跳出局部极小. 文中给出了误差函数的单调性定理以及该算法的弱收敛和强收敛性定理.

关键词: BP 神经网络; 在线梯度法; 收敛性; 惩罚项; 随机输入.
MSC(2000): 92B20; 68T05
中图分类: O24

1 引 言

作为一种有效的学习算法, 在线梯度法已广泛地应用于工程计算等领域. 文献 [3],[5] 和 [10] 分别对带随机输入的三层 BP 网络和带惩罚项无隐层的 BP 网络中的情况进行了细致的研究. 本文考虑带有一个隐层且结构为 p-n-1 的网络. 给定训练样本集 $\{\xi^j, O^j\}_{j=1}^J \subset R^p \times R$, 设 $g: R \to R$ 和 $f: R \to R$ 分别为隐单元和输出单元上的活化函数. 令输入层到隐层的权矩阵为 $V = (v_{ij})_{n\times p}$, 记 $v_i = (v_{i1}, v_{i2}, \dots, v_{ip})^{\mathrm{T}}$ $(1 \le i \le n)$, 隐层到输出层的权向量为 $w = (w_1, w_2, \dots, w_n)^{\mathrm{T}}$.

本文的研究是在 Euclidean 空间中进行的. 假设 K 是任一正整数, 对于 $x = (x_1, \dots, x_K)^{\mathrm{T}}$, $y = (y_1, \dots, y_K)^{\mathrm{T}} \in R^K$, 定义 $x \cdot y = \sum_{k=1}^K x_k y_k$, $\|x\| = (x \cdot x)^{1/2}$. 为了表法简洁, 我们还引入记号

$$G(x) = (g(x_1), g(x_2), \dots, g(x_n))^{\mathrm{T}}, \ \ \forall x = (x_1, \dots, x_n) \in R^n, \tag{1.1}$$

$$\varphi_j = G(V\xi^j), \ \ 1 \le j \le J, \tag{1.2}$$

$$\varphi_j^k = G(V^k\xi^j), \ \ k = 0, 1, \dots, \ 1 \le j \le J. \tag{1.3}$$

对于给定的输入样本 $\xi \in R^p$, 网络的实际输出值为

$$\zeta = f(w \cdot G(V\xi)).$$

对每个训练样本 j, 定义平方误差函数

$$\widetilde{E}_j(w \cdot \varphi_j) = \frac{1}{2}\left(O^j - \zeta^j\right)^2 = \frac{1}{2}\left[O^j - f(w \cdot \varphi_j)\right]^2. \tag{1.4}$$

收稿日期: 2005-09-28; **接受日期**: 2006-02-28
基金项目: 国家自然科学基金 (10471017).

相应的总误差函数可表为

$$\widetilde{E}(w,V)=\sum_{j=1}^{J}\widetilde{E}_j(w\cdot\varphi_j). \tag{1.5}$$

引入惩罚项，则 (1.4) 和 (1.5) 变为

$$E_j(w\cdot\varphi_j)=\frac{1}{2}\left[(O^j-f(w\cdot\varphi_j))^2+\lambda(w\cdot\varphi_j)^2\ \right], \tag{1.6}$$

$$E(w,V)=\sum_{j=1}^{J}E_j(w\cdot\varphi_j), \tag{1.7}$$

其中 λ 是非负的实数.

网络训练的目的是为了求得权值 w^* 和 V^*, 使

$$E(w^*,V^*)=\min E(w,V). \tag{1.8}$$

为此，人们常常采用简单有效的在线梯度法. 我们用一种特殊的随机方式选择 ξ^j, 即在每一轮训练执行之前，首先对训练样本 $\xi^1,\dots,\xi^J$ 随机排序，然后把得到的 $\xi^{m1},\dots,\xi^{mJ}$ 顺序输入网络. 对任意给定的初始权值 w^0 和 V^0, 按下列公式迭代地修改权向量

$$w^{mJ+j}=w^{mJ+j-1}+\triangle_j^m w^{mJ+j-1},\ m=0,1,\dots,\ 1\le j\le J, \tag{1.9}$$

$$v_i^{mJ+j}=v_i^{mJ+j-1}+\triangle_j^m v_i^{mJ+j-1},\ m=0,1,\dots,\ 1\le j\le J,\ 1\le i\le n, \tag{1.10}$$

其中

$$\triangle_j^m w=-\eta_m\frac{\partial E_{mj}\left(w\cdot\varphi_{mj}\right)}{\partial w}=-\eta_m E'_{mj}(w\cdot\varphi_{mj})\varphi_{mj}, \tag{1.11}$$

$$\triangle_j^m v_i=-\eta_m\frac{\partial E_{mj}\left(w\cdot\varphi_{mj}\right)}{\partial v_i}=-\eta_m E'_{mj}(w\cdot\varphi_{mj})w_i g'(v_i\cdot\xi^{mj})\xi^{mj}, \tag{1.12}$$

这里 η_m 是第 m 轮的学习步长.

引入惩罚项的目的是提高网络的推广能力，并且控制权值的大小 [5,10]. 而采用随机方式输入样本能使网络更易跳出局部极小 [2]. 本文同时使用了这两种学习方式. 由于在每一轮网络学习中，各样本恰好被使用一次，从而在很大程度上保留了一般梯度法的输入特点，为最后得到确定性收敛结果提供了重要的前提条件.

2 预备定理

我们首先给出文中所需要的几个假设条件，其中 C_0,C_1 等表示某一正的常数，且同一记号在不同的地方可以表示不同的值.

(A1) $|g(t)|,|g'(t)|,|g''(t)|\le C_0,\ t\in R.$

(A2) $|f(t)|,|f'(t)|,|f''(t)|\le C_0\ ,\ t\in R.$

(A3) $\|\,w^{mJ+j}\|\le C_0,\ \ m=0,1,\dots,\ 1\le j\le J.$

(A4) $\|\,v_i^{mJ+j}\|\le C_0,\ \ m=0,1,\dots,\ 1\le i\le n,\ 1\le j\le J.$

(A5) 学习步长序列 $\{\eta_m\}$ 的迭代公式为

$$\frac{1}{\eta_{m+1}} = \frac{1}{\eta_m} + \beta, \quad m = 0, 1, \ldots, \tag{2.1}$$

其中正常数 β 的值将在后面的证明中给出.

根据上述条件，我们容易得到如下的七个引理. 其中引理 2.1–2.4 的证明可在文献 [3] 及其引文中找到，引理 2.5 和 2.6 容易证得，而引理 2.7 是文献 [1] 中引理 3.5.10 的直接推论.

引理 2.1 设 $X = (x_1, \ldots, x_K)^{\mathrm{T}} \in R^K$, 则

$$\|X\| \le \sum_{i=1}^{K} |x_i|, \tag{2.2}$$

$$\Big(\sum_{i=1}^{K} |x_i|\Big)^2 \le K \sum_{i=1}^{K} |x_i|^2. \tag{2.3}$$

引理 2.2 学习步长序列 $\{\eta_m\}_{m=1}^{\infty}$ 具有下列性质

$$0 < \eta_m < \eta_{m-1} \le \eta_0; \tag{2.4}$$

$$\eta_m < \frac{\delta}{m},\ \delta > 0. \tag{2.5}$$

引理 2.3 假设级数 $\sum_{n=1}^{\infty} \frac{a_n^2}{n}$ $(a_n > 0)$ 收敛. 如果存在常数 $\mu > 0$, 使得

$$|a_{n+1} - a_n| < \frac{\mu}{n},$$

那么

$$\lim_{n\to\infty} a_n = 0.$$

引理 2.4 令 $v_{id}^m = v_i^{(m+1)J} - v_i^{mJ}$, $m = 0, 1, \ldots,$ $1 \le i \le n$. 若条件 (A1) 成立，那么

$$\|G(X)\| \le C_1,\ \forall X \in R^n, \tag{2.6}$$

$$\big\|\varphi_j^{mJ+s} - \varphi_j^{mJ}\big\| \le C_1 \sum_{i=1}^{n} \big\|v_i^{mJ+s} - v_i^{mJ}\big\|, \tag{2.7}$$

$$\begin{aligned}&\big\|G((v_1^{mJ}, \ldots, v_i^{mJ}, v_{i+1}^{(m+1)J}, \ldots, v_n^{(m+1)J})^{\mathrm{T}}\xi^j) - \varphi_j^{mJ}\big\| \le C_1 \sum_{k=1}^{n} \|v_{kd}^m\|,\\ &\quad m = 0, 1, \ldots, 1 \le i \le n-1, 1 \le s \le J.\end{aligned} \tag{2.8}$$

引理 2.5 若条件 (A1), (A2) 成立，序列 $\{w^{mJ+j}\}$ 和 $\{v_i^{mJ+j}\}$ 由学习算法 (1.9) 和 (1.10) 生成，则

$$\big|E_j''(t)\big| \le C_1, \quad 1 \le j \le J, \quad t \in R, \tag{2.9}$$

$$|E_j'(t_j)| \le C_1,\ 1 \le j \le J, \tag{2.10}$$

其中，t_j 是介于 $w^{mJ} \cdot \varphi_{mj}^{mJ}$ 和 $w^{mJ+j-1} \cdot \varphi_{mj}^{mJ+j-1}$ 之间的任一实数，$m = 0, 1, \ldots$.

引理 2.6 令 $w_d^m = w^{(m+1)J} - w^{mJ}$, $m=0,1,\ldots$. 若条件 (A1), (A2) 成立, 序列 $\{w^{mJ+j}\}$ 和 $\{v_i^{mJ+j}\}$ 由学习算法 (1.9) 及 (1.10) 生成, 则有

$$\left\| E_{ww}(w^{mJ}, V^{mJ}) \right\| \le C_1, \tag{2.11}$$

$$\left\| E_{wv_i}(w^{mJ}, V^{mJ}) \right\| \le C_1,\ 1 \le i \le n. \tag{2.12}$$

此外, 对任意的 $0 \le t \le 1$, 令

$$\begin{aligned}
&\tilde{w} = w^{mJ} + t w_d^m,\\
&V_1 = (v_1^{mJ} + t v_{1d}^m,\ v_2^{(m+1)J},\ \ldots,\ v_n^{(m+1)J})^{\mathrm{T}},\\
&V_i = (v_1^{mJ},\ \ldots,\ v_{i-1}^{mJ}, v_i^{mJ} + t v_{id}^m,\ v_{i+1}^{(m+1)J},\ \ldots,\ v_n^{(m+1)J})^{\mathrm{T}},\ 2 \le i \le n-1,\\
&V_n = (v_1^{mJ},\ \ldots,\ v_{n-1}^{mJ},\ v_n^{mJ} + t v_{nd}^m)^{\mathrm{T}},
\end{aligned}$$

则有

$$\left\| E_{ww}(\tilde{w}, V^{(m+1)J}) \right\| \le C_1, \tag{2.13}$$

$$\left\| E_{v_i v_i}(w^{mJ}, V_i) \right\| \le C_1,\ 1 \le i \le n. \tag{2.14}$$

引理 2.7 设函数 $h: R^K \longrightarrow R$ 在紧集 D 上连续可微, 并且 h 在 D 中的临界点集 $\Omega = \{x \in D | \nabla h(x) = 0\}$ 是有限的. 若序列 $\{x^m\}_{m=0}^{\infty} \subset D$ 满足

$$\lim_{m\to\infty} \|x^{m+1} - x^m\| = 0, \quad \lim_{m\to\infty} \|\nabla h(x^m)\| = 0,$$

则存在 $x^* \in \Omega$, 使得

$$\lim_{m\to\infty} x^m = x^*.$$

3 关于误差函数的几个定理

为了简化后面的证明, 对 $k=1,\ 2$, 记

$$\Big\| \sum_{j=1}^{J} \Delta_j^m w^{mJ} \Big\|^k = \sigma_{k,\,1}^m, \quad \sum_{i=1}^{n} \Big\| \sum_{j=1}^{J} \Delta_j^m v_i^{mJ} \Big\|^k = \sigma_{k,\,2}^m,$$

$$\sum_{j=1}^{J} \|\Delta_j^m w^{mJ}\|^k = \sigma_{k,\,3}^m, \quad \sum_{i=1}^{n} \sum_{j=1}^{J} \|\Delta_j^m v_i^{mJ}\|^k = \sigma_{k,\,4}^m.$$

此外, 我们还引入

$$R^{m,\,j} = \triangle_j^m w^{mJ+j-1} - \triangle_j^m w^{mJ},\ m=0,1,\ldots,\ 1 \le j \le J, \tag{3.1}$$

$$r_i^{m,\,j} = \triangle_j^m v_i^{mJ+j-1} - \triangle_j^m v_i^{mJ},\ m=0,1,\ldots,\ 1 \le j \le J,\ 1 \le i \le n. \tag{3.2}$$

显然

$$R^{m,\,1} = 0, \tag{3.3}$$

$$r_i^{m,1}=0. \tag{3.4}$$

定理 3.1 设条件 (A1)–(A3) 成立，序列 $\{w^{mJ+j}\}$ 和 $\{v_i^{mJ+j}\}$ 由学习算法 (1.9) 及 (1.10) 生成，则对于 $m=0,1,\ldots,\ j=1,\ldots,J,\ i=1,\cdots,n$,

$$w^{mJ+j}=w^{mJ}+\sum_{k=1}^{j}(\triangle_k^m w^{mJ}+R^{m,k}), \tag{3.5}$$

$$v_i^{mJ+j}=v_i^{mJ}+\sum_{k=1}^{j}(\triangle_k^m v_i^{mJ}+r_i^{m,k}). \tag{3.6}$$

进而有

$$\max\Big\{\sum_{j=1}^{J}\|R^{m,j}\|,\ \sum_{i=1}^{n}\sum_{j=1}^{J}\|r_i^{m,j}\|\Big\}\le C_2\eta_m(\sigma_{1,3}^m+\sigma_{1,4}^m). \tag{3.7}$$

证明 (3.5) 与 (3.6) 式可由 (1.9) 和 (1.10) 式直接证得.

根据 (1.11) 和 (3.1) 式，得

$$\begin{aligned}R^{m,j}=&\triangle_j^m w^{mJ+j-1}-\triangle_j^m w^{mJ}\\=&\eta_m[E'_{mj}(w^{mJ}\cdot\varphi_{mj}^{mJ})\varphi_{mj}^{mJ}-E'_{mj}(w^{mJ+j-1}\cdot\varphi_{mj}^{mJ+j-1})\varphi_{mj}^{mJ+j-1}]\\=&\eta_m E'_{mj}(w^{mJ}\cdot\varphi_{mj}^{mJ})(\varphi_{mj}^{mJ}-\varphi_{mj}^{mJ+j-1})+\\&\eta_m[E'_{mj}(w^{mJ}\cdot\varphi_{mj}^{mJ})-E'_{mj}(w^{mJ+j-1}\cdot\varphi_{mj}^{mJ+j-1})]\varphi_{mj}^{mJ+j-1}\\=&\eta_m E'_{mj}(w^{mJ}\cdot\varphi_{mj}^{mJ})(\varphi_{mj}^{mJ}-\varphi_{mj}^{mJ+j-1})+\\&\eta_m E''_{mj}(t'_j)[w^{mJ}\cdot(\varphi_{mj}^{mJ}-\varphi_{mj}^{mJ+j-1})+(w^{mJ}-w^{mJ+j-1})\cdot\varphi_{mj}^{mJ+j-1}]\varphi_{mj}^{mJ+j-1},\end{aligned}$$

其中 t'_j 介于 $w^{mJ}\cdot\varphi_{mj}^{mJ}$ 和 $w^{mJ+j-1}\cdot\varphi_{mj}^{mJ+j-1}$ 之间.

由 (A3), (2.6), (2.7), (2.9) 和 (2.10) 式，推出

$$\|R^{m,j}\|\le C'_2\eta_m\big(\|w^{mJ+j-1}-w^{mJ}\|+\sum_{i=1}^{n}\|v_i^{mJ+j-1}-v_i^{mJ}\|\big), \tag{3.8}$$

其中 $C'_2=C_1^2(1+C_1+C_0C_1)$.

同理,

$$\sum_{i=1}^{n}\|r_i^{m,j}\|\le C''_2\eta_m\big(\|w^{mJ+j-1}-w^{mJ}\|+\sum_{i=1}^{n}\|v_i^{mJ+j-1}-v_i^{mJ}\|\big), \tag{3.9}$$

其中， $C''_2=C_1(C_0C_1+C_0n+C_1n+C_0\max_{1\le j\le J}|\xi^j|)\max_{1\le j\le J}|\xi^j|$.

下面，我们利用数学归纳法证明：对任意 $k=1,\ldots,J$, 存在正常数 $C_{2,k}$ 使得

$$\max\Big\{\|R^{m,k}\|,\ \sum_{i=1}^{n}\|r_i^{m,k}\|\Big\}\le C_{2,k}\eta_m\sum_{s=1}^{k-1}\big(\|\triangle_s^m w^{mJ}\|+\sum_{i=1}^{n}\|\triangle_s^m v_i^{mJ}\|\big). \tag{3.10}$$

由 (3.3), (3.4) 式易知， $k=1$ 时存在 $C_{2,1}$ 使得 (3.10) 式成立. 设 j 是大于 1 的任意整数，且 $k<j$ 时，存在 $C_{2,k}$ 使得 (3.10) 式成立. 下面我们证明 $k=j$ 时，存在 $C_{2,j}$ 使得 (3.10) 式也成立.

令 $C_2''' = \max_{1\le k\le j-1} C_{2,k}$. 由 (3.5), (3.6) 和 (3.8) 式知

$$
\begin{aligned}
\|R^{m,j}\| \le & C_2'\eta_m\big(\|w^{mJ+j-1}-w^{mJ}\|+\sum_{i=1}^{n}\|v_i^{mJ+j-1}-v_i^{mJ}\|\big)\\
= & C_2'\eta_m\big(\|\sum_{k=1}^{j-1}(\triangle_k^m w^{mJ}+R^{m,k})\|+\sum_{i=1}^{n}\|\sum_{k=1}^{j-1}(\triangle_k^m v_i^{mJ}+r_i^{m,k})\|\big)\\
\le & C_2'\eta_m\big[\sum_{k=1}^{j-1}\|\triangle_k^m w^{mJ}\|+\sum_{k=1}^{j-1}C_2'''\eta_m\sum_{s=1}^{k-1}(\|\triangle_s^m w^{mJ}\|+\sum_{i=1}^{n}\|\triangle_s^m v_i^{mJ}\|)+\\
& \sum_{i=1}^{n}\sum_{k=1}^{j-1}\|\triangle_k^m v_i^{mJ}\|+\sum_{k=1}^{j-1}C_2'''\eta_m\sum_{s=1}^{k-1}(\|\triangle_s^m w^{mJ}\|+\sum_{i=1}^{n}\|\triangle_s^m v_i^{mJ}\|)\big]\\
\le & C_2'\eta_m\sum_{k=1}^{j-1}\big(\|\triangle_k^m w^{mJ}\|+\sum_{i=1}^{n}\|\triangle_k^m v_i^{mJ}\|\big)+\\
& 2C_2'C_2'''\eta_m^2\sum_{k=1}^{j-1}\sum_{s=1}^{j-1}\big(\|\triangle_s^m w^{mJ}\|+\sum_{i=1}^{n}\|\triangle_s^m v_i^{mJ}\|\big)\\
\le & C_2'(1+2C_2'''J\eta_0)\eta_m\sum_{k=1}^{j-1}\big(\|\triangle_k^m w^{mJ}\|+\sum_{i=1}^{n}\|\triangle_k^m v_i^{mJ}\|\big).
\end{aligned}
$$

同理,

$$\sum_{i=1}^{n}\|r_i^{m,j}\|\le C_2''(1+2C_2'''J\eta_0)\eta_m\sum_{k=1}^{j-1}\big(\|\triangle_k^m w^{mJ}\|+\sum_{i=1}^{n}\|\triangle_k^m v_i^{mJ}\|\big).$$

即 $k=j$ 时，存在 $C_{2,j}=(C_2'+C_2'')(1+2C_2'''J\eta_0)$, 使得 (3.10) 式成立.

又令 $C_2''''=\max_{1\le k\le j} C_{2,k}$, 则对于任意 k, 有

$$\max\big\{\|R^{m,k}\|,\sum_{i=1}^{n}\|r_i^{m,k}\|\big\}\le C_2''''\eta_m\sum_{s=1}^{k-1}\big(\|\triangle_s^m w^{mJ}\|+\sum_{i=1}^{n}\|\triangle_s^m v_i^{mJ}\|\big).$$

从而

$$
\begin{aligned}
\sum_{j=1}^{J}\|R^{m,j}\| &\le C_2''''\eta_m\sum_{j=1}^{J}\sum_{k=1}^{j-1}\big(\|\triangle_k^m w^{mJ}\|+\sum_{i=1}^{n}\|\triangle_k^m v_i^{mJ}\|\big)\\
&\le C_2''''\eta_m\sum_{j=1}^{J}\sum_{k=1}^{J}\big(\|\triangle_k^m w^{mJ}\|+\sum_{i=1}^{n}\|\triangle_k^m v_i^{mJ}\|\big)\\
&\le C_2''''J\eta_m(\sigma_{1,3}^m+\sigma_{1,4}^m).
\end{aligned}
$$

同理,

$$\sum_{j=1}^{J}\sum_{i=1}^{n}\|r_i^{m,j}\|\le C_2''''J\eta_m(\sigma_{1,3}^m+\sigma_{1,4}^m).$$

取 $C_2=C_2''''J$, 至此定理得证. □

定理 3.2 设条件 (A1)–(A3) 成立，序列 $\{w^{mJ+j}\}$ 和 $\{v_i^{mJ+j}\}$ 由学习算法 (1.9) 及 (1.10) 式生成，则对于 $m=0,1,\dots$，存在一个与 m 无关的非负常数 C_3 使得

$$E(w^{(m+1)J},V^{(m+1)J})-E(w^{mJ},V^{mJ})\le-\frac{1}{\eta_m}(\sigma_{2,1}^m+\sigma_{2,2}^m)+C_3(\sigma_{2,3}^m+\sigma_{2,4}^m). \tag{3.11}$$

证明 令

$$\begin{aligned}
&V_i^{mJ}=(v_1^{mJ},\dots,v_i^{mJ},v_{i+1}^{(m+1)J},\dots,v_n^{(m+1)J})^{\mathrm{T}},\ 1\le i\le n-1,\\
&V_n^{mJ}=(v_1^{mJ},\dots,v_n^{mJ})^{\mathrm{T}}=V^{mJ},\\
&\tilde V_1=(v_1^{mJ}+t_1v_{1d}^m,v_2^{(m+1)J},\dots,v_n^{(m+1)J})^{\mathrm{T}},\\
&\tilde V_i=(v_1^{mJ},\dots,v_{i-1}^{mJ},v_i^{mJ}+t_iv_{id}^m,v_{i+1}^{(m+1)J},\dots,v_n^{(m+1)J})^{\mathrm{T}},\ 2\le i\le n-1,\\
&\tilde V_n=(v_1^{mJ},\dots,v_{n-1}^{mJ},v_n^{mJ}+t_nv_{nd}^m)^{\mathrm{T}},
\end{aligned}$$

其中 t_i $(1\le i\le n)$ 介于 0 和 1 之间. 根据泰勒展式可得

$$\begin{aligned}
&E(w^{(m+1)J},V^{(m+1)J})-E(w^{mJ},V^{mJ})\\
&\quad=E(w^{(m+1)J},v_1^{(m+1)J},\dots,v_n^{(m+1)J})-E(w^{mJ},v_1^{mJ},\dots,v_n^{mJ})\\
&\quad=E(w^{(m+1)J},v_1^{(m+1)J},\dots,v_n^{(m+1)J})-E(w^{mJ},v_1^{(m+1)J},\dots,v_n^{(m+1)J})+\\
&\qquad E(w^{mJ},v_1^{(m+1)J},\dots,v_n^{(m+1)J})-E(w^{mJ},v_1^{mJ},v_2^{(m+1)J},\dots,v_n^{(m+1)J})+\cdots+\\
&\qquad E(w^{mJ},v_1^{mJ},\dots,v_{n-1}^{mJ},v_n^{(m+1)J})-E(w^{mJ},v_1^{mJ},\dots,v_n^{mJ})\\
&\quad=(w_d^m)^{\mathrm{T}}E_w(w^{mJ},V^{(m+1)J})+\frac12(w_d^m)^{\mathrm{T}}E_{ww}(w^{mJ}+t^*w_d^m,V^{(m+1)J})w_d^m+\\
&\qquad\sum_{i=1}^n\big[(v_{id}^m)^{\mathrm{T}}E_{v_i}(w^{mJ},V_i^{mJ})+\frac12(v_{id}^m)^{\mathrm{T}}E_{v_iv_i}(w^{mJ},\tilde V_i)v_{id}^m\big],
\end{aligned} \tag{3.12}$$

其中 $0\le t^*\le1$. 而

$$\begin{aligned}
&E_w(w^{mJ},V^{(m+1)J})\\
&\quad=\sum_{j=1}^J E_j'(w^{mJ}\cdot\varphi_j^{(m+1)J})\varphi_j^{(m+1)J}\\
&\quad=\sum_{j=1}^J\big[E_j'(w^{mJ}\cdot\varphi_j^{mJ})+(E_j'(w^{mJ}\cdot\varphi_j^{(m+1)J})-E_j'(w^{mJ}\cdot\varphi_j^{mJ}))\big]\varphi_j^{(m+1)J}\\
&\quad=\sum_{j=1}^J\big[E_j'(w^{mJ}\cdot\varphi_j^{mJ})\varphi_j^{mJ}+E_j'(w^{mJ}\cdot\varphi_j^{mJ})(\varphi_j^{(m+1)J}-\varphi_j^{mJ})\big]+\\
&\qquad\sum_{j=1}^J E_j''(t_j'')[w^{mJ}\cdot(\varphi_j^{(m+1)J}-\varphi_j^{mJ})]\varphi_j^{(m+1)J}\\
&\quad=E_w(w^{mJ},V^{mJ})+\sum_{j=1}^J E_j'(w^{mJ}\cdot\varphi_j^{mJ})(\varphi_j^{(m+1)J}-\varphi_j^{mJ})+
\end{aligned}$$

$$\sum_{j=1}^{J} E_j''(t_j'')\big[w^{mJ}\cdot(\varphi_j^{(m+1)J}-\varphi_j^{mJ})\big]\varphi_j^{(m+1)J},$$

其中 t_j'' 介于 $w^{mJ}\cdot\varphi_j^{(m+1)J}$ 和 $w^{mJ}\cdot\varphi_j^{mJ}$ 之间.

利用 (A3), (1.6), (2.6), (2.17) 和 (2.9)–(2.11) 式，有

$$\begin{aligned}
&(w_d^m)^{\mathrm{T}}E_w(w^{mJ},V^{(m+1)J})\\
&\quad\le (w_d^m)^{\mathrm{T}}E_w(w^{mJ},V^{mJ})+C_1J(1+C_0C_1)\|\varphi_j^{(m+1)J}-\varphi_j^{mJ}\|\|w_d^m\|\\
&\quad\le (w_d^m)^{\mathrm{T}}E_w(w^{mJ},V^{mJ})+C_1^2J(1+C_0C_1)\sum_{i=1}^{n}\|v_{id}^m\|\|w_d^m\|\\
&\quad\le (w_d^m)^{\mathrm{T}}E_w(w^{mJ},V^{mJ})+C_3'(\sum_{i=1}^{n}\|v_{id}^m\|^2+\|w_d^m\|^2),
\end{aligned}$$

其中 $C_3=\frac{1}{2}C_1^2n(1+C_0C_1)$. 同理，由 (A1), (A3) 和 (2.9) 式可知

$$(v_{id}^m)^{\mathrm{T}}E_{v_i}(w^{mJ},V_i^{mJ})\le (v_{id}^m)^{\mathrm{T}}E_{v_i}(w^{mJ},V^{mJ})+C_3''\sum_{k=1}^{n}\|v_{kd}^m\|^2,\quad 1\le i\le n.$$

由 (1.11),(1.12) 和 (3.5)–(3.7) 式，得

$$\begin{aligned}
&(w_d^m)^{\mathrm{T}}E_w(w^{mJ},V^{mJ})+\sum_{i=1}^{n}(v_{id}^m)^{\mathrm{T}}E_{v_i}(w^{mJ},V^{mJ})\\
&\quad=(\sum_{j=1}^{J}\triangle_j^m w^{mJ}+\sum_{j=1}^{J}R^{m,j})^{\mathrm{T}}(-\frac{1}{\eta_m}\sum_{j=1}^{J}\triangle_j^m w^{mJ})+\\
&\qquad\sum_{i=1}^{n}(\sum_{j=1}^{J}\triangle_j^m v_i^{mJ}+\sum_{j=1}^{J}r_i^{m,j})^{\mathrm{T}}(-\frac{1}{\eta_m}\sum_{j=1}^{J}\triangle_j^m v_i^{mJ})\\
&\quad=-\frac{1}{\eta_m}\|\sum_{j=1}^{J}\triangle_j^m w^{mJ}\|^2-\frac{1}{\eta_m}\sum_{i=1}^{n}\|\sum_{j=1}^{J}\triangle_j^m v_i^{mJ}\|^2-\\
&\qquad\frac{1}{\eta_m}(\sum_{j=1}^{J}R^{m,j})^{\mathrm{T}}(\sum_{j=1}^{J}\triangle_j^m w^{mJ})-\frac{1}{\eta_m}\sum_{i=1}^{n}(\sum_{j=1}^{J}r_i^{m,j})^{\mathrm{T}}(\sum_{j=1}^{J}\triangle_j^m v_i^{mJ})\\
&\quad\le-\frac{1}{\eta_m}(\sigma_{2,1}^m+\sigma_{2,2}^m)+\frac{1}{\eta_m}\sum_{j=1}^{J}\|R^{m,j}\|\sum_{j=1}^{J}\|\triangle_j^m w^{mJ}\|+\frac{1}{\eta_m}\sum_{i=1}^{n}\sum_{j=1}^{J}\|r_i^{m,j}\|\sum_{j=1}^{J}\|\triangle_j^m v_i^{mJ}\|\\
&\quad\le-\frac{1}{\eta_m}(\sigma_{2,1}^m+\sigma_{2,2}^m)+\frac{1}{\eta_m}C_2\eta_m(\sigma_{1,3}^m+\sigma_{1,4}^m)\sigma_{1,3}^m+\frac{1}{\eta_m}C_2\eta_m(\sigma_{1,3}^m+\sigma_{1,4}^m)\sigma_{1,4}^m\\
&\quad\le-\frac{1}{\eta_m}(\sigma_{2,1}^m+\sigma_{2,2}^m)+C_2[(\sigma_{1,3}^m)^2+(\sigma_{1,4}^m)^2]\\
&\quad\le-\frac{1}{\eta_m}(\sigma_{2,1}^m+\sigma_{2,2}^m)+C_2nJ(\sigma_{2,3}^m+\sigma_{2,4}^m).
\end{aligned}\tag{3.13}$$

又因为

$$\|w_d^m\|=\|\sum_{j=1}^{J}\triangle_j^m w^{mJ}+\sum_{j=1}^{J}R^{m,j}\|\le\sigma_{1,1}^m+\sum_{j=1}^{J}\|R^{m,j}\|,$$

根据 (3.7) 式

$$\|w_d^m\|^2 \le C_3''(\sigma_{2,3}^m + \sigma_{2,4}^m). \tag{3.14}$$

同理,

$$\sum_{i=1}^n \|v_{id}^m\|^2 \le C_3'''(\sigma_{2,3}^m + \sigma_{2,4}^m). \tag{3.15}$$

结合 (2.13), (2.14), (3.12)–(3.15) 式得

$$\begin{aligned} & E(w^{(m+1)J}, V^{(m+1)J}) - E(w^{mJ}, V^{mJ}) \\ & \le -\frac{1}{\eta_m}(\sigma_{2,1}^m + \sigma_{2,2}^m) + C_2 nJ(\sigma_{2,3}^m + \sigma_{2,4}^m) + \frac{1}{2}C_1(\sum_{i=1}^n \|v_{id}^m\|^2 + \|w_d^m\|^2) \\ & \le -\frac{1}{\eta_m}(\sigma_{2,1}^m + \sigma_{2,2}^m) + C_3(\sigma_{2,3}^m + \sigma_{2,4}^m), \end{aligned}$$

其中 $C_3 = C_2 nJ + \frac{1}{2}C_1(C_3'' + C_3''')$. □

下面两个定理的证明可参见文献 [3], 此略.

定理 3.3　设条件 (A1)–(A3) 成立，序列 $\{w^{mJ+j}\}$ 和 $\{v_i^{mJ+j}\}$ 由学习算法 (1.9) 及 (1.10) 式生成. 适当选取常数 $\rho > 0$, 则存在 $\beta > 0$, 当 $0 < \eta_0 < \min\{\frac{1}{\beta_0}, \rho\}$ 且 $\beta_0 < \beta \le \frac{1}{\eta_0}$ 时，对任意整数 $m > 0$, 有

$$\frac{1}{\eta_m}(\sigma_{2,1}^m + \sigma_{2,2}^m) \ge C_3(\sigma_{2,3}^m + \sigma_{2,4}^m). \tag{3.16}$$

定理 3.4　设条件 (A1)–(A3) 成立，序列 $\{w^{mJ+j}\}$ 和 $\{v_i^{mJ+j}\}$ 由学习算法 (1.9) 及 (1.10) 式生成. 若 η_0, β 按定理 3.3 中的方式选取，则

$$\sum_{m=1}^\infty \frac{1}{m}\|E_w(w^{mJ}, V^{mJ})\|^2 < \infty, \tag{3.17}$$

$$\sum_{m=1}^\infty \frac{1}{m}\|E_{v_i}(w^{mJ}, V^{mJ})\|^2 < \infty, \quad 1 \le i \le n. \tag{3.18}$$

4　主要结果

定理 4.1　若定理 3.3 中的条件均成立，则

$$E(w^{(m+1)J}, V^{(m+1)J}) \le E(w^{mJ}, V^{mJ}), \quad m = 0, 1, \ldots. \tag{4.1}$$

定理 4.2　若定理 3.4 中的条件均成立，则

$$\lim_{m\to\infty} \|E_w(w^{mJ+j}, V^{mJ+j})\| = 0, \quad 1 \le j \le J, \tag{4.2}$$

$$\lim_{m\to\infty} \|E_{v_i}(w^{mJ+j}, V^{mJ+j})\| = 0, \quad 1 \le j \le J, \quad 1 \le i \le n. \tag{4.3}$$

证明　利用 (A1), (A3), (1.9)–(1.12), (2.5), (2.9) 和 (2.10) 式，得

$$\| w_d^m\| < \frac{C}{m}, \quad m = 1, 2, \ldots, \tag{4.4}$$

$$\| v_{id}^m \| < \frac{C}{m}, \ \ m = 1, 2, \ldots, 1 \leq i \leq n. \tag{4.5}$$

结合 (2.7) 式有

$$\begin{aligned}
&\|E_w(w^{(m+1)J}, V^{(m+1)J}) - E_w(w^{mJ}, V^{mJ})\| \\
&\leq \sum_{j=1}^{J} \|E_j'(w^{(m+1)J} \cdot \varphi_j^{(m+1)J})\varphi_j^{(m+1)J} - E_j'(w^{mJ} \cdot \varphi_j^{mJ})\varphi_j^{mJ}\| \\
&\leq \sum_{j=1}^{J} \|E_j'(w^{(m+1)J} \cdot \varphi_j^{(m+1)J})(\varphi_j^{(m+1)J} - \varphi_j^{mJ}) + \\
&\quad (E_j'(w^{(m+1)J} \cdot \varphi_j^{(m+1)J}) - E_j'(w^{mJ} \cdot \varphi_j^{mJ}))\varphi_j^{mJ}\| \\
&\leq C\|\varphi_j^{(m+1)J} - \varphi_j^{mJ}\| + C\sum_{j=1}^{J} \|E_j''(t_j^*)[w^{(m+1)J} \cdot (\varphi_j^{(m+1)J} - \varphi_j^{mJ}) + w_d^m \varphi_j^{mJ}]\| \\
&\leq C\sum_{i=1}^{n} \|v_{id}^m\| + C\|w_d^m\| \\
&\leq \frac{C}{m},
\end{aligned} \tag{4.6}$$

其中 t_j^* 介于 $w^{(m+1)J} \cdot \varphi_j^{(m+1)J}$ 和 $w^{mJ} \cdot \varphi_j^{mJ}$ 之间.

由 (3.15), (4.6) 式和引理 2.3, 得

$$\lim_{m\to\infty} \|E_w(w^{mJ}, V^{mJ})\| = 0.$$

仿照 (4.6) 式的推导，有

$$\|E_w(w^{mJ+j}, V^{mJ+j})\| \leq \|E_w(w^{mJ}, V^{mJ})\| + \frac{C}{m}, \ m = 1, 2, \ldots, \ 1 \leq j \leq J - 1.$$

从而

$$\lim_{m\to\infty} \|E_w(w^{mJ+j}, V^{mJ+j})\| = 0, \ m = 1, 2, \ldots, \ 1 \leq j \leq J.$$

(4.3) 式同理可证. □

令 $W = (w^{\mathrm{T}}, v_1^{\mathrm{T}}, \ldots, v_n^{\mathrm{T}})^{\mathrm{T}}$, 并设定理 3.3 中的所有条件及 (A4) 均成立，则 $W \in D$, 其中 $D \subset R^{(p+1)n}$ 是某一紧集. 此时有

定理 4.3 设定理 3.3 中所有条件及 (A4) 均成立. 若 $D_0 = \{W \in D | \nabla E(W) = 0\}$ 包含有限个点，则存在一点 $W^* \in D_0$ 使得

$$\lim_{k\to\infty} W^k = W^*.$$

证明 由 (A1)–(A3), (1.9)–(1.12) 和 (2.5) 式，有

$$\max\{\|w^{mJ+j} - w^{mJ+j-1}\|, \ \|v_i^{mJ+j} - v_i^{mJ+j-1}\|\} \leq \frac{C}{m}, \ \ m = 1, 2, \ldots,$$

故

$$\lim_{k\to\infty} \|W^{k+1} - W^k\| = \|((w^{k+1} - w^k)^{\mathrm{T}}, (v_1^{k+1} - v_1^k)^{\mathrm{T}}, \ldots, (v_n^{k+1} - v_n^k)^{\mathrm{T}})\|^{\mathrm{T}} = 0. \tag{4.7}$$

根据 (4.2) 和 (4.3) 式，我们得到

$$\lim_{k\to\infty}\|\nabla E(W^k)\| = 0. \tag{4.8}$$

根据 (4.7), (4.8) 式和引理 2.7, 存在 $W^* \in D_0$, 使得

$$\lim_{k\to\infty} W^k = W^*.$$

参考文献:

[1] 袁亚湘，孙文瑜．最优化理论与方法 [M]. 北京：科学出版社，1997.
YUAN Ya-xiang, SUN Wen-yu. *Optimization Theory and Methods* [M]. Beijing: Science Press, 1997. (in Chinese)

[2] 徐秉铮，张百灵，韦岗．神经网络理论与应用 [M]. 广州：华南理工大学出版社，1994.
XU Bing-zheng, ZHANG Bai-ling, WEI Gang. *Theory and Application of Neural Networks* [M]. Guangzhou: South China University of Technology Press, 1994. (in Chinese)

[3] LI Zheng-xue, WU Wei, FENG Guo-rui. et al. *Convergence of an online gradient method for BP neural networks with stochastic inputs* [J]. Lecture Notes in Computer Science, 2005, **3610**: 720–729.

[4] LESHNO M, LIN V Y, PINKER A, SCHOCHEN S. *Multilayer feedforward networks with a nonpolynomial activation function can approximation any function* [J]. Neural Network, 1993, **6**: 861–867.

[5] KONG Jun, WU Wei. *Online gradient methods with a punishing term for neural networks* [J]. Northeast. Math. J., 2001, **17**(3): 371–378.

[6] SOLLICH P, BARBER D. *Online learning from finite training sets and robustness to input bias* [J]. Neural Computation, 1998, **10**: 2201–2217.

[7] LIANG Yan-chun. *Successive approximation training algorithm for feedforward neural network* [J]. Neuroconputing, 2002, **42**: 311–322.

[8] LI Zheng-xue, WU Wei, TIAN Yu-long. *Convergence of an online gradient method for feedforward neural networks with stochastic inputs* [J]. J. Comput. Appl. Math., 2004, **163**(1): 165–176.

[9] OH S H. *Improving the error backpropagation algorithm with a modified error function* [J]. IEEE Transaction on Neural Netork, 1997, **8**: 799–803.

[10] SHAO Hong-mei, WU Wei, LI Feng. *Convergence of online gradient method with a penalty term for feedforward neural networks with stochastic inputs* [J]. Numer. Math. J. Chinese Univ., 2005, **14**(1): 87–96.

Convergence of Online Gradient Method with a Penalty Term for BP Neural Network with Stochastic Inputs

LU Hui-fang[1,2], *WU Wei*[1], *LI Zheng-xue*[1]

(1. Department of Mathematics, Dalian University of Technology, Liaoning 116024, China;
2. Department of Mathematics and Physics, Shandong Jiaotong University, Shandong 250023, China)

Abstract: In this paper, we present and discuss an online gradient method with a penalty term for three-layer BP neural networks. The input training examples are reset stochastically before the performance of each batch so that the learning is easy to jump off from local minima. The monotonicity and the convergence of deterministic nature are proved.

Key words: BP neural networks; online gradient method; convergence; penalty term; stochastic inputs.