第一题

根据题意，输入层有 d 个结点，隐藏层有 n_H 个结点，输出层有 c 个结点。设结点之间的连接权重为 w，在本题中以 z_j 表示第 j 个输出结点的值，y_h 表示第 h 个隐藏层结点的值，x_i 表示第 i 个输入结点的值。定义 h 和 j 之间的连接权重为 w_{hj}，i 和 h 之间的连接权重为 w_{ih}。以这三个点为例推导权重更新，即使用反向传播算法求对 w_{hj} 和 w_{ih} 的更新，其中用常数 η 表示学习率。用 net_j 表示输入结点 j 的加权和，则有，

$$\begin{aligned} net_h &= \sum_{i=1}^{d} w_{ih}x_i,\ \ y_h = f_1(net_h) \\ net_j &= \sum_{h=1}^{n_H} w_{hj}y_h,\ \ z_j = f_2(net_j) \end{aligned} \tag{1}$$

f_1 和 f_2 分别表示两个激活函数，根据题中定义，分别为 sigmoid 和 softmax 函数，对两者进行求导，

$$\begin{aligned} f_1'(s) &= \left(\frac{1}{1+e^{-s}}\right)' \\ &= \frac{e^{-s}}{(1+e^{-s})^2} \\ &= f_1(s)(1-f_1(s)) \end{aligned} \tag{2}$$

因为输出层各节点之间互相影响，对 z_j 求导分两种情况。当 $i=j$ 时

$$\begin{aligned} \frac{\partial z_j}{\partial s_i} &= \frac{\partial}{\partial s_i}\left(\frac{e^{s_j}}{\sum_k e^{s_k}}\right) \\ &= \frac{(e^{s_j})'\sum_k e^{s_k} - e^{s_j}(\sum_k e^{s_k})'}{(\sum_k e^{s_k})^2} \\ &= z_j(1-z_j) \end{aligned} \tag{3}$$

当 $i \neq j$ 时

$$\begin{aligned}\frac{\partial z_j}{\partial s_i} &= \frac{\partial}{\partial s_i}\left(\frac{e^{s_j}}{\sum_k e^{s_k}}\right) \\ &= \frac{\left(e^{s_j}\right)' \sum_k e^{s_k} - e^{s_j}\left(\sum_k e^{s_k}\right)'}{\left(\sum_k e^{s_k}\right)^2} \\ &= \frac{0 \cdot \sum_k e^{s_k} - e^{s_j}\left(\sum_k e^{s_k}\right)'}{\left(\sum_k e^{s_k}\right)^2} \\ &= \frac{-e^{s_j} e^{s_i}}{\left(\sum_k e^{s_k}\right)^2} \\ &= -z_j z_i\end{aligned} \tag{4}$$

根据题意，误差计算式为，

$$J(\mathbf{w}) = \frac{1}{2}\sum_{j=1}^{c}(t_j - z_j)^2 \tag{5}$$

用梯度下降法求隐含层到输出层的权重调节量，

$$\begin{aligned}\Delta w_{hj} &= -\eta \frac{\partial J}{\partial w_{hj}} \\ &= -\eta \sum_{p=1}^{c} \frac{\partial J}{\partial z_p} \frac{\partial z_p}{\partial net_j} \frac{\partial net_j}{\partial w_{hj}} \\ &= \eta \sum_{p=1}^{c} (t_p - z_p) \frac{\partial z_p}{\partial net_j} y_h\end{aligned} \tag{6}$$

其中根据公式 (1)、(3)、(4) 得，

$$\frac{\partial z_p}{\partial net_j} = \begin{cases} z_p(1 - z_p), & p = j \\ -z_p z_j, & p \neq j \end{cases} \tag{7}$$

$$\begin{aligned}y_h &= f_1(net_h) \\ &= f_1\left(\sum_{i=1}^{d} w_{ih} x_i\right) \\ &= \frac{1}{1 + e^{\sum_{i=1}^{d} w_{ih} x_i}}\end{aligned} \tag{8}$$

接下来求输入层到隐含层的权重调节量，

$$\begin{aligned}\Delta w_{ih} &= -\eta \frac{\partial J}{\partial w_{ih}} \\ &= -\eta \sum_{j=1}^{c} \sum_{p=1}^{c} \frac{\partial J}{\partial z_p} \frac{\partial z_p}{\partial net_j} \frac{\partial net_j}{\partial w_{ih}} \\ &= -\eta \sum_{j=1}^{c} \sum_{p=1}^{c} \frac{\partial J}{\partial z_p} \frac{\partial z_p}{\partial net_j} \frac{\partial net_j}{\partial y_h} \frac{\partial y_h}{\partial w_{ih}} \\ &= -\eta \sum_{j=1}^{c} \sum_{p=1}^{c} \frac{\partial J}{\partial z_p} \frac{\partial z_p}{\partial net_j} \frac{\partial net_j}{\partial y_h} \frac{\partial y_h}{\partial net_h} \frac{\partial net_h}{\partial w_{ih}} \\ &= \eta \sum_{j=1}^{c} \sum_{p=1}^{c} (t_p - z_p) \frac{\partial z_p}{\partial net_j} w_{hj} f_1^{'}(net_h) x_i\end{aligned} \tag{9}$$

$\frac{\partial z_p}{\partial net_j}$ 见公式 (7)，且根据公式 (1)、(3) 得，

$$f_1^{'}(net_h) = net_h(1 - net_h) \tag{10}$$

以上给出了权重根据输入 x, 输出 z, 目标 t, 原始权重 w 更新的公式。

第二题

隐藏层到输出层的权重更新公式为，

$$\begin{aligned}\Delta w_{hj} &= \eta \sum_{p=1}^{c} (t_p - z_p) \frac{\partial z_p}{\partial net_j} y_h \\ &= \eta \delta_j y_h\end{aligned} \tag{11}$$

其中，

$$\begin{aligned}\delta_j &= -\frac{\partial J}{\partial net_j} \\ &= \sum_{p=1}^{c} (t_p - z_p) \frac{\partial z_p}{\partial net_j}\end{aligned} \tag{12}$$

δ_j 是权重边指向结点（j 点）经过导数缩放后的输出误差，意为 j 点收集到的误差信号。所以，隐层到输出层的误差更新正比于权重边起始结点输出与权重边指向结点收集到误差信号的乘积。

输入层到隐藏层的权重更新公式为，

$$\begin{aligned}\Delta w_{ih} &= \eta \sum_{j=1}^{c}\sum_{p=1}^{c}(t_p - z_p)\frac{\partial z_p}{\partial net_j} w_{hj} f_1^{'}(net_h) x_i \\ &= \eta \sum_{j=1}^{c} \delta_j w_{hj} f_1^{'}(net_h) x_i \\ &= \eta \left(f_1^{'}(net_h) \sum_{j=1}^{c} \delta_j w_{hj} \right) x_i \\ &= \eta \delta_h x_i\end{aligned} \tag{13}$$

其中，

$$\begin{aligned}\delta_h &= -\frac{\partial J}{\partial net_h} \\ &= f_1^{'}(net_h) \sum_{j=1}^{c} \delta_j w_{hj}\end{aligned} \tag{14}$$

δ_h 是权重边指向结点（h 点）收集到的误差信号。是由上一层（输出层）收集误差的加权和（$\sum_{j=1}^{c} \delta_j w_{hj}$）经过这一层激活函数导数缩放后得到。所以，输入层到隐藏层的误差更新同样正比于权重边起始结点输出与权重边指向结点收集到误差信号的乘积。

第三题

程序详见附件中的 BP.py 文件。接下来研究这几个问题，

(1) **隐含层不同结点数目对训练精度的影响**

对于单样本方式，设置最大迭代次数为 $100,000$ 次，对于批量更新算法，设置最大 batch 为 $3,333$，取梯度更新步长为 0.1。认为参数更新率小于 10^{-5} 时模型收敛，停止迭代。选取隐含层结点个数分别为 1, 2, 4, 8, 16, 32, 64, 128, 256, 512 进行测试，结果如表 1、表 2 所示。

可以看到隐含层结点个数并不是越多越好，所以应该选取合适的隐含层结点个数来进行训练。

表 1: 不同隐含层结点个数对训练误差函数的影响

结点数 / 算法	1	2	4	8	16
单样本更新	0.27	0.24	0.12	0.047	0.051
批量更新	0.27	0.23	0.08	0.029	0.0013
结点数 / 算法	32	64	128	256	512
单样本更新	0.016	0.053	0.029	0.27	0.26
批量更新	0.00087	0.00061	0.017	0.067	0.13

表 2: 不同隐含层结点个数对训练步数的影响

结点数 / 算法	1	2	4	8	16
单样本更新	max	max	27487	59867	8468
批量更新	max	max	max	max	max
结点数 / 算法	32	64	128	256	512
单样本更新	17994	4795	10058	140	109
批量更新	max	max	max	max	max

注意：表 2 中 max 指训练步数达到最大，对于单样本更新算法，该值为 $100,000$，对于批量更新算法，该值为 $3,333$。

(2) 不同的梯度更新步长对训练的影响

设定与上一问题相同的最大迭代次数，隐含层神经元个数为 8，选取梯度更新步长分别为 0.001, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1, 10 进行测试，结果见表 3、表 4。可以看到随着步长的增加，学习速度变快，但是到一定程度时会出现“反复跳跃”的现象。

这是因为学习速率太大容易导致目标函数波动较大从而难以找到最优，而学习速率设置太小，每次权重的更新太小，则会导致收敛过慢耗时太长。

表 3: 不同梯度更新步长对训练误差函数的影响

	0.001	0.1	0.2	0.3	0.4	0.5
单样本更新	0.19	0.047	0.051	0.088	0.14	0.11
批量更新	0.17	0.029	0.017	0.018	0.045	0.029
	0.6	0.7	0.8	0.9	1	10
单样本更新	0.080	0.10	0.089	0.12	0.052	0.57
批量更新	0.049	0.00010	0.095	0.10	0.095	0.92

表 4: 不同梯度更新步长对训练步数的影响

结点数 / 算法	0.001	0.1	0.2	0.3	0.4	0.5
单样本更新	max	59867	30579	10440	16561	2594
批量更新	max	max	max	max	max	max
结点数 / 算法	0.6	0.7	0.8	0.9	1	10
单样本更新	max	20276	9127	36414	6407	51
批量更新	max	max	max	max	max	56

(3) 目标函数随着迭代步数增加的变化曲线

设置隐含层结点个数为 8，梯度更新步长为 0.001，绘制目标函数随着迭代步数增加的变化曲线如图 1、图 2 所示。可以看到，随着迭代次数的增加，损失函数均逐步下降。需要注意的是，如果出现随着迭代次数的增加，损失函数的值一会上升，一会下降的情况，说明梯度更新步长设置较大。

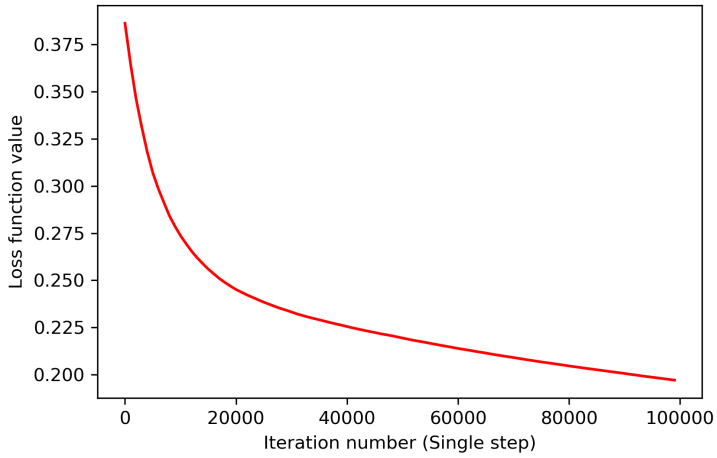

图 1: 单样本更新算法目标函数变化示意图

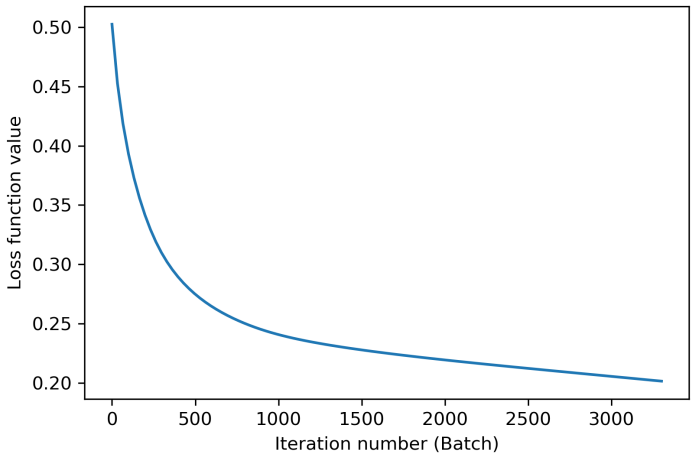

图 2: 批量更新算法目标函数变化示意图