1. $z_1 \cdots z_j \cdots z_c$ $\quad W_{hj}$

$y_1 \cdots y_h \cdots y_{n_H}$ $\quad W_{ih}$

$x_1 \cdots x_i \cdots x_d$

⇓

$$net_h = \sum_{i=1}^{d} W_{ih} x_i \qquad y_h = f(net_h)$$

$$net_j = \sum_{h=1}^{n_H} W_{hj} y_h \qquad z_j = Softmax(net_j)$$

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

$$\Delta W_{hj} = -\eta \frac{\partial J(w)}{\partial W_{hj}} = -\eta \sum_{k=1}^{c} \frac{\partial J(w)}{\partial z_k} \frac{\partial z_k}{\partial W_{hj}} = \eta \sum_{k=1}^{c} (t_k - z_k) \frac{\partial z_k}{\partial net_j} \frac{\partial net_j}{\partial W_{hj}}$$

$$\frac{\partial z_k}{\partial net_j} = \begin{cases} z_j - z_j^2 & k = j \\ -z_k z_j & k \neq j \end{cases} \qquad \frac{\partial net_j}{\partial W_{hj}} = y_h$$

$$\therefore \Delta W_{hj} = -\sum_{k \neq j}^{c} (t_k - z_k) z_k z_j y_h + (t_j - z_j)(z_j - z_j^2) y_h$$

$$\Delta W_{ih} = -\eta \frac{\partial J(w)}{\partial W_{ih}} = -\eta \sum_{k=1}^{c} \frac{\partial J(w)}{\partial z_k} \sum_{i=1}^{c} \frac{\partial z_k}{\partial net_i} \sum_{m=1}^{n_H} \frac{\partial net_i}{\partial y_m} \frac{\partial y_m}{\partial net_m} \frac{\partial net_m}{\partial W_{ih}}$$

$$= -\eta \sum_{k=1}^{c} \frac{\partial J(w)}{\partial z_k} \sum_{j=1}^{c} \frac{\partial z_k}{\partial net_j} \frac{\partial net_j}{\partial y_h} \frac{\partial y_h}{\partial net_h} \frac{\partial net_h}{\partial W_{ih}}$$

$$\frac{\partial J(w)}{\partial z_k} = -(t_k - z_k) \qquad \frac{\partial z_k}{\partial net_j} = \begin{cases} z_j - z_j^2 & k = j \\ -z_k z_j & k \neq j \end{cases}$$

$$\frac{\partial net_j}{\partial y_h} = W_{hj} \qquad \frac{\partial y_h}{\partial net_h} = y_h(1 - y_h) \qquad \frac{\partial net_h}{\partial W_{ih}} = x_i$$

$$\Delta W_{ih} = -\eta \sum_{k=1}^{c} (z_k - t_k) \left(\sum_{j \neq k}^{c} \left[z_k z_j W_{hj} y_h (1 - y_h) x_i \right] + (z_k - z_k^2) W_{hk} y_h (1 - y_h) x_i \right)$$

2. 反向传播的步骤

① 前向传播，样本从输入层逐层计算到输出层，得到预测标签，并与真实标签计算损失值。

② 反向计算第 N, N-1 ··· 1 层的误差项，即损失函数对相应隐层节点的偏导数。

③ 根据求导法则计算损失函数对各层参数的梯度，并对参数进行更新。

$$\Delta W_{hj} = \eta \delta_j y_h \qquad \delta_j = -\frac{\partial J(w)}{\partial net_j} = f'(net_j)(t_j - z_j)$$

δ_j 表示权重所联边的指向节点（输出节点）收集到的误差信号

y_h 表示权重所联边的起始节点的输出

$$\Delta w_{ih} = \eta \delta_h x_i \qquad \delta_h = -\frac{\partial J(w)}{\partial net_h} = f'(net_h) \sum_j w_{hj} \delta_j$$

δ_h 表示 w_{ih} 连接的边的指向节点（隐层节点 h）收集到的误差信号

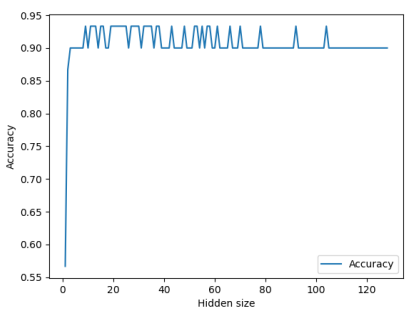
x_i 表示 w_{ih} 连接的边的起始节点（输入层节点 i）的输出（即样本的第 i 个分量）

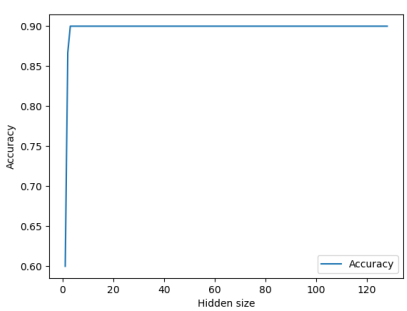
影响网络性能的因素

① 损失函数 ② 激活函数 ③ 隐层数 ④ 结点个数 ⑤ 初始权重 ⑥ 学习率 ⑦ 训练停止准则

3. (a) 隐含层节点数对训练精度的影响

单样本更新：　　批量更新：

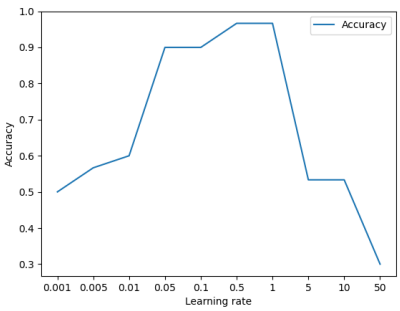




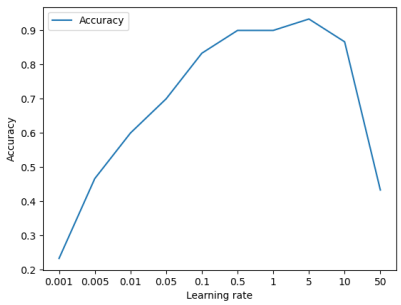
可以看出隐含层节点数较小时模型拟合能力较弱，分类的准确率较低，隐含层节点数增大时模型性能提升并趋于稳定，同时多批量更新表现出比单样本更新更加稳定。

(b) 梯度更新步长对训练的影响

单样本更新：



批量更新



可以看出梯度更新步长较小时参数更新较慢模型难以优化，更新步长较大时模型可能会越过全局最优点，收敛到局部最优点，同时单样本更新性能的波动比批量更新大，没有批量更新稳定。

(c) 目标函数随迭代步数增加变化曲线