1. (1) $\theta_{MAP}=\arg\max\limits_{\theta} p(\theta|x,y)=\arg\max\limits_{\theta}\frac{p(y|x,\theta)p(\theta|x)}{p(y|x)}=\arg\max\limits_{\theta} p(y|x,\theta)p(\theta)$

(2) 设维度为n，则 $p(\theta)=\frac{1}{(2\pi\eta^2)^{n/2}}e^{-\frac{1}{2\eta^2}\theta^T I_n\theta}=\frac{1}{(2\pi\eta^2)^{n/2}}e^{-\frac{1}{2\eta^2}\|\theta\|_2^2}$

$\Rightarrow \log p(\theta)=-\frac{n}{2}\log 2\pi\eta^2-\frac{1}{2\eta^2}\|\theta\|_2^2 \propto -\frac{1}{2\eta^2}\|\theta\|_2^2$

此时 $\theta_{MAP}=\arg\max\limits_{\theta} p(y|x,\theta)p(\theta)=\arg\max\limits_{\theta}(\log p(y|x,\theta)+\log p(\theta))$

$=\arg\max\limits_{\theta}(\log p(y|x,\theta)-\frac{1}{2\eta^2}\|\theta\|_2^2)=\arg\min\limits_{\theta}(-\log p(y|x,\theta)+\frac{1}{2\eta^2}\|\theta\|_2^2)$

即为所求. 比较知此处 $\lambda=\frac{1}{2\eta^2}$.

(3) 设 $X\in\mathbb{F}^{m\times n}$, $\theta_{MAP}=\arg\max\limits_{\theta} p(\theta|X,\vec{y})=\arg\max\limits_{\theta}\frac{p(\vec{y}|X,\theta)p(\theta)}{p(\vec{y}|X)}=\arg\max\limits_{\theta} p(\vec{y}|X,\theta)p(\theta)$.

由 $\varepsilon^{(i)}\sim N(0,\sigma^2)$, 有 $y^{(i)}\sim N(\theta^Tx^{(i)},\sigma^2)$, $p(\vec{y}^{(i)}|x^{(i)},\theta)=\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2\sigma^2}(y^{(i)}-\theta^Tx^{(i)})^2}$

$\Rightarrow p(\vec{y}|X,\theta)=\prod\limits_{i=1}^{m}\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2\sigma^2}(y^{(i)}-\theta^Tx^{(i)})^2}=\frac{1}{(2\pi\sigma^2)^{m/2}}e^{-\frac{1}{2\sigma^2}\sum\limits_{i=1}^{m}(y^{(i)}-\theta^Tx^{(i)})^2}=\frac{1}{(2\pi\sigma^2)^{m/2}}e^{-\frac{1}{2\sigma^2}\|\vec{y}-X\theta\|_2^2}$

由 $\theta\sim N(0,\eta^2 I_n)$ 有 $p(\theta)=\frac{1}{(2\pi\eta^2)^{n/2}}e^{-\frac{1}{2\eta^2}\|\theta\|_2^2}$

$\Rightarrow \theta_{MAP}=\arg\max\limits_{\theta}\frac{1}{(2\pi\sigma^2)^{m/2}}e^{-\frac{1}{2\sigma^2}\|\vec{y}-X\theta\|_2^2}\cdot\frac{1}{(2\pi\eta^2)^{n/2}}e^{-\frac{1}{2\eta^2}\|\theta\|_2^2}$

$=\arg\max\limits_{\theta}(-\frac{m}{2}\log 2\pi\sigma^2-\frac{1}{2\sigma^2}\|\vec{y}-X\theta\|_2^2-\frac{n}{2}\log(2\pi\eta^2)-\frac{1}{2\eta^2}\|\theta\|_2^2)$

$=\arg\min\limits_{\theta}(\frac{1}{2\sigma^2}\|\vec{y}-X\theta\|_2^2-\frac{1}{2\eta^2}\|\theta\|_2^2)$.

令 $L(\theta)=\frac{1}{2\sigma^2}\|\vec{y}-X\theta\|_2^2-\frac{1}{2\eta^2}\|\theta\|_2^2$, $\nabla L(\theta)=\frac{2}{\sigma^2}X^T(X\theta-\vec{y})+\frac{2}{\eta^2}\theta$.

令 $\nabla L(\theta)=0$, 有 $X^TX\theta+\frac{\sigma^2}{\eta^2}\theta-X^T\vec{y}=0\Rightarrow(X^TX+\frac{\sigma^2}{\eta^2}I_n)\theta=X^T\vec{y}$

$\Rightarrow$ 闭式解为 $\theta_{MAP}=(X^TX+\frac{\sigma^2}{\eta^2}I_n)^{-1}X^T\vec{y}$

(4) $\theta_i\sim L(0,b)\Rightarrow p(\theta_i)=\frac{1}{2b}e^{-\frac{1}{b}|\theta_i|}\Rightarrow p(\theta)=\prod\limits_{i=1}^{n}p(\theta_i)=\frac{1}{(2b)^n}e^{-\frac{1}{b}\sum\limits_{i=1}^{n}|\theta_i|}=\frac{1}{(2b)^n}e^{-\frac{1}{b}\|\theta\|_1}$

与(3)相似，可得到 $\theta_{MAP}=\arg\max\limits_{\theta}\frac{1}{(2\pi\sigma^2)^{m/2}}e^{-\frac{1}{2\sigma^2}\|\vec{y}-X\theta\|_2^2}\cdot\frac{1}{(2b)^n}e^{-\frac{1}{b}\|\theta\|_1}$

$=\arg\min\limits_{\theta}(\frac{1}{2\sigma^2}\|X\theta-\vec{y}\|_2^2+\frac{1}{b}\|\theta\|_1)$

$=\arg\min\limits_{\theta}(\|X\theta-\vec{y}\|_2^2+\frac{2\sigma^2}{b}\|\theta\|_1)$

即为所求. 比较有 $\gamma=\frac{2\sigma^2}{b}$.

2. (1) 对正例, $\theta^Tx:=z>0\Rightarrow\lim\limits_{c\to+\infty}p(y=1|x)=\lim\limits_{c\to+\infty}\frac{1}{1+e^{-c\theta^Tx}}=\lim\limits_{c\to+\infty}\frac{1}{1+e^{-cz}}=1$

对负例, $\theta^Tx:=\hat{z}<0\Rightarrow\lim\limits_{c\to+\infty}p(y=1|x)=\lim\limits_{c\to+\infty}\frac{1}{1+e^{-c\theta^Tx}}=\lim\limits_{c\to+\infty}\frac{1}{1+e^{-c\hat{z}}}=0$

(2) $L(c\theta)=-\sum\limits_{i=1}^{m}(y^{(i)}\log\frac{1}{1+e^{-c\theta^Tx}}+(1-y^{(i)})\log(1-\frac{1}{1+e^{-c\theta^Tx}}))$

$=\sum\limits_{i=1}^{m}(y^{(i)}\log(1+e^{-c\theta^Tx})+(1-y^{(i)})\log\frac{1+e^{-c\theta^Tx}}{e^{-c\theta^Tx}})$

对每一个i，若 $y^{(i)}=1\Leftrightarrow\theta^Tx>0$, $\log(1+e^{-c\theta^Tx})\to\log 1=0\Rightarrow$ 求和中此项为0

若 $y^{(i)}=0\Leftrightarrow\theta^Tx<0$, $\log\frac{1+e^{-c\theta^Tx}}{e^{-c\theta^Tx}}\to\log 1=0\Rightarrow$ 求和中此项为0

$\Rightarrow\lim\limits_{c\to+\infty}L(c\theta)=0\Rightarrow L(c\theta)$ 随 c 增大无限下降至0.

(3) 由于数据不可分，即任意超平面均不能完全分开这两个类，故无论如何调整θ，总有部分样本的预测无法达到其真实标签的极限，故无法同时优化所有样本的预测概率. 因此存在一个有限的最小损失值. 由于有下界时必有下确界，因此存在 $L_{min}\in\mathbb{R}$ s.t. $\forall\varepsilon,\exists\theta\in\mathbb{R}^d$ s.t. $L_{min}\le L(\theta)\le L_{min}+\varepsilon$, $\forall\theta\in\mathbb{R}^d$, $L_{min}\le L(\theta)$. 特别地，取 $\varepsilon=\frac{1}{n}$, 记 θ_n 符合条件，则取 $\theta=\theta_n$ 并令 $n\to+\infty$. 即有损失函数能够收敛至 L_{min}.

(4) ① 不能. 不论学习率高低，θ均会发散至无穷远处，无法达到收敛点.

② 能. 由于 $\sum\limits_{n=1}^{+\infty}\frac{1}{n^2}<+\infty$, 此时快速减小的学习率使得算法更稳定，可能表现出更好的收敛性

③ 不能. 缩放不改变数据集的可分性，仍然可能发散.

④ 能. L^2 正则化可以限制θ的模长防止其无限增大，使优化过程稳定并收敛至有限值.

⑤ 有可能. 添加 Gauss 噪声可能导致数据集不再可分，此时 $L(\theta)$ 有某一下界，此时可能收敛.

3. (1) $a_1=\frac{e^{z_1}}{e^{z_1}+e^{z_2}+e^{z_3}}=0.0900$ $a_2=0.2447$ $a_3=0.6653\Rightarrow a^L=\begin{bmatrix}0.0900\\0.2447\\0.6653\end{bmatrix}$

(2) $NLL(a,y)=-y_3\log a_3=1.6094$

(3) $NLL(a^L,y)=-\sum\limits_{j=1}^{n}y_j\log a_j^L=-\sum\limits_{j=1}^{n}y_j(\log\frac{e^{z_j^L}}{\sum\limits_{k=1}^{n}e^{z_k^L}})=-\sum\limits_{j=1}^{n}y_j(z_j^L-\log\sum\limits_{k=1}^{n}e^{z_k^L})$

$=-\sum\limits_{j=1}^{n}y_jz_j^L+\sum\limits_{j=1}^{n}y_j\log(\sum\limits_{k=1}^{n}e^{z_k^L})$

由 y 是 onehot 编码，故 $\exists!j^*$ s.t. $y_{j^*}=1$, 其余 $y_j=0\Leftrightarrow y_j=\delta_{jj^*}$

故 $NLL(a^L,y)=-z_{j^*}^L+\log(\sum\limits_{k=1}^{n}e^{z_k^L})$. 首先有 $\frac{\partial NLL(a^L,y)}{\partial w_{kj}^L}=\sum\limits_{l=1}^{n}\frac{\partial NLL(a^L,y)}{\partial z_l^L}\cdot\frac{\partial z_l^L}{\partial w_{kj}^L}$

由于 $z_j^L=\sum\limits_k w_{kj}^Lx_k+w_{0j}$. 有 $\frac{\partial z_l^L}{\partial w_{kj}^L}=\frac{\partial}{\partial w_{kj}}\sum\limits_k w_{lj}^Lx_k=x_k\delta_{jl}$

$\frac{\partial NLL(a^L,y)}{\partial z_l^L}=-\delta_{j^*l}+\frac{e^{z_l^L}}{\sum\limits_{k=1}^{n}e^{z_k^L}}=-\delta_{j^*l}+a_l^L$

故 $\frac{\partial NLL(a^L,y)}{\partial z_l^L}=\sum\limits_{l=1}^{N}(-\delta_{j^*l}+a_l^L)x_k\delta_{jl}=(a_j^L-\delta_{jj^*})x_k=x_k(a_j^L-y_j)$

因此 $\nabla_{W^L}NLL(a^L,y)=(x_k(a_j^L-y_j))_{m\times n}=x(a^L-y)^T$.

此处 $W^L=\begin{bmatrix}1&-1&-2\\-1&2&1\end{bmatrix}$, $W_0^L=[1,1,1]$, $x=[1,1]^T$, 则 $z=(W^L)^Tx+(W_0^L)^T=[0,1,-1]^T$

$a^L=SM(z^L)=[0.2447,0.6652,0.0900]^T$. 又 $y=[0,1,0]$

有 $\nabla_{W^L}NLL(a^L,y)=x(a^L-y)^T=\begin{bmatrix}0.2447&-0.3348&0.0900\\0.2447&-0.3348&0.0900\end{bmatrix}$

(4) 由 $a^L=[0.2447,0.6652,0.0900]$, 类别1对应第2个输出，故概率为66.52%

(5) $W_{NEW}^L=W^L-\alpha\nabla_{W^L}NLL(a^L,y)=\begin{bmatrix}1&-1&-2\\-1&2&1\end{bmatrix}-0.5\begin{bmatrix}0.2447&-0.3348&0.0900\\0.2447&-0.3348&0.0900\end{bmatrix}$

$=\begin{bmatrix}0.8776&-0.8326&-2.0450\\-1.1224&2.1674&0.9550\end{bmatrix}$

(6) $W_1^L=\begin{bmatrix}0.8776&-0.8326&-2.0450\\-1.1224&2.1674&0.9550\end{bmatrix}\Rightarrow z_1^L=(W_1^L)^Tx+W_0^L=[-0.2447,1.3348,-1.0900]^T$

$a_1^L=SM(z_1^L)=[0.1592,0.7725,0.0684]\Rightarrow$ 概率为77.25%

4. (1) $\begin{bmatrix}z_1^1\\z_2^1\\z_3^1\\z_4^1\end{bmatrix}=\begin{bmatrix}1&0\\0&1\\-1&0\\0&-1\end{bmatrix}\begin{bmatrix}3\\14\end{bmatrix}+\begin{bmatrix}-1\\-1\\-1\\-1\end{bmatrix}=\begin{bmatrix}2\\13\\-4\\-15\end{bmatrix}$, $\begin{bmatrix}a_1^1\\a_2^1\\a_3^1\\a_4^1\end{bmatrix}=ReLU\begin{bmatrix}2\\13\\-4\\-15\end{bmatrix}=\begin{bmatrix}2\\13\\0\\0\end{bmatrix}$

$\begin{bmatrix}z_1^2\\z_2^2\end{bmatrix}=\begin{bmatrix}1&1&1&1\\-1&-1&-1&-1\end{bmatrix}\begin{bmatrix}2\\13\\0\\0\end{bmatrix}+\begin{bmatrix}0\\2\end{bmatrix}=\begin{bmatrix}15\\-13\end{bmatrix}$, $\begin{bmatrix}a_1^2\\a_2^2\end{bmatrix}=SM\begin{bmatrix}15\\-13\end{bmatrix}=\begin{bmatrix}1.0000\\6.9144\times10^{-13}\end{bmatrix}$

可视为 $\begin{bmatrix}a_1^2\\a_2^2\end{bmatrix}=\begin{bmatrix}1\\0\end{bmatrix}$ 为最终输出，$(f'(z_1^1),f'(z_2^1),f'(z_3^1),f'(z_4^1))=(2,13,0,0)$

(2) 对 $x^{(1)}=\begin{bmatrix}0.5\\0.5\end{bmatrix}$, $\begin{bmatrix}{}^1z_1^{(1)}\\{}^1z_2^{(1)}\\{}^1z_3^{(1)}\\{}^1z_4^{(1)}\end{bmatrix}=\begin{bmatrix}0.5\\0.5\\-1.5\\-1.5\end{bmatrix}$, $\begin{bmatrix}{}^1a_1^{(1)}\\{}^1a_2^{(1)}\\{}^1a_3^{(1)}\\{}^1a_4^{(1)}\end{bmatrix}=\begin{bmatrix}0\\0\\0\\0\end{bmatrix}$ 对 $x^{(2)}=\begin{bmatrix}0\\2\end{bmatrix}$, $\begin{bmatrix}{}^1z_1^{(2)}\\{}^1z_2^{(2)}\\{}^1z_3^{(2)}\\{}^1z_4^{(2)}\end{bmatrix}=\begin{bmatrix}-1\\1\\-1\\-3\end{bmatrix}$,

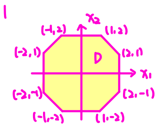
$\begin{bmatrix}{}^1a_1^{(2)}\\{}^1a_2^{(2)}\\{}^1a_3^{(2)}\\{}^1a_4^{(2)}\end{bmatrix}=\begin{bmatrix}0\\1\\0\\0\end{bmatrix}$ 对 $x^{(3)}=\begin{bmatrix}-3\\0.5\end{bmatrix}$, $\begin{bmatrix}{}^1z_1^{(3)}\\{}^1z_2^{(3)}\\{}^1z_3^{(3)}\\{}^1z_4^{(3)}\end{bmatrix}=\begin{bmatrix}-4\\-0.5\\2\\-1.5\end{bmatrix}$, $\begin{bmatrix}{}^1a_1^{(3)}\\{}^1a_2^{(3)}\\{}^1a_3^{(3)}\\{}^1a_4^{(3)}\end{bmatrix}=\begin{bmatrix}0\\0\\2\\0\end{bmatrix}$, 结果矩阵为 $\begin{bmatrix}0&0&0\\0&1&0\\0&0&2\\0&0&0\end{bmatrix}$

对分类为第二类，即 $a_1^2<a_2^2$，需要 $z_1^2<z_2^2\Leftrightarrow a_1^1+a_2^1+a_3^1+a_4^1<-a_1^1-a_2^1-a_3^1-a_4^1+2$

$\Leftrightarrow a_1^1+a_2^1+a_3^1+a_4^1<1\Leftrightarrow\max\{z_1^1,0\}+\max\{z_2^1,0\}+\max\{z_3^1,0\}+\max\{z_4^1,0\}<1$

$\Leftrightarrow\max\{x_1-1,0\}+\max\{x_2-1,0\}+\max\{-x_1-1,0\}+\max\{-x_2-1,0\}<1$

由于表达式较复杂，故给出图像：(所有线均为直线)



对 $(x_1,x_2)\in D$, 分类结果为第二类

对 $(x_1,x_2)\in D^c$, 分类结果为第一类

(3) 对 Case 1, 由上述分析，此处 $a_1^1+a_2^1+a_3^1+a_4^1=0$, 输出为 $[0.1192,0.8807]^T$ 为第二类

对 Case 2, $a_1^1+a_2^1+a_3^1+a_4^1=1$, 输出为 $[0.5,0.5]^T$ 无法分类

对 Case 3, $a_1^1+a_2^1+a_3^1+a_4^1=3$, 输出为 $[0.9820,0.0180]^T$ 为第一类.

(4) $\frac{\partial Loss}{\partial z_k^2}=a_k^2-y_k\Rightarrow\frac{\partial Loss}{\partial z_1^2}=1$, $\frac{\partial Loss}{\partial z_2^2}=-1\Rightarrow\frac{\partial Loss}{\partial z^2}=\begin{bmatrix}1\\-1\end{bmatrix}$. $\frac{\partial Loss}{\partial W_0^2}=\frac{\partial Loss}{\partial z^2}=\begin{bmatrix}1\\-1\end{bmatrix}$,

$\frac{\partial Loss}{\partial W^2}=a^1(\frac{\partial Loss}{\partial z^2})^T=\begin{bmatrix}2\\13\\0\\0\end{bmatrix}[1\ -1]=\begin{bmatrix}2&-2\\13&-13\\0&0\\0&0\end{bmatrix}$ $\frac{\partial Loss}{\partial a^1}=W^2\frac{\partial Loss}{\partial z^2}=\begin{bmatrix}1&-1\\1&-1\\1&-1\\1&-1\end{bmatrix}\begin{bmatrix}1\\-1\end{bmatrix}=\begin{bmatrix}2\\2\\2\\2\end{bmatrix}$

$\frac{\partial Loss}{\partial z_j^1}=\frac{\partial Loss}{\partial a_j^1}\cdot ReLU'(z_j^1)=\frac{\partial Loss}{\partial a_j^1}\cdot\chi_{(0,+\infty)}(z_j^1)\Rightarrow\frac{\partial Loss}{\partial z_1^1}=2$, $\frac{\partial Loss}{\partial z_2^1}=2$. $\frac{\partial Loss}{\partial z_3^1}=0$. $\frac{\partial Loss}{\partial z_4^1}=0$

$\Rightarrow\frac{\partial Loss}{\partial z^1}=\begin{bmatrix}2\\2\\0\\0\end{bmatrix}$. $\frac{\partial Loss}{\partial W_0^1}=\frac{\partial Loss}{\partial z^1}=\begin{bmatrix}2\\2\\0\\0\end{bmatrix}$, $\frac{\partial Loss}{\partial W^1}=x(\frac{\partial Loss}{\partial z^1})^T=\begin{bmatrix}3\\14\end{bmatrix}[2\ 2\ 0\ 0]=\begin{bmatrix}6&6&0&0\\28&28&0&0\end{bmatrix}$

$\Rightarrow\widehat{W}^2=W^2-\eta\frac{\partial Loss}{\partial W^2}=\begin{bmatrix}0.8&-0.8\\-0.3&0.3\\1&-1\\1&-1\end{bmatrix}$, $\widehat{W}_0^2=W_0^2-\eta\frac{\partial Loss}{\partial W_0^2}=\begin{bmatrix}-0.1\\2.1\end{bmatrix}$

$\widehat{W}^1=W^1-\eta\frac{\partial Loss}{\partial W^1}=\begin{bmatrix}0.4&-0.6&-1&0\\-2.8&-1.8&0&-1\end{bmatrix}$, $\widehat{W}_0^1=W_0^1-\eta\frac{\partial Loss}{\partial W_0^1}=\begin{bmatrix}-1.2\\-1.2\\-1\\-1\end{bmatrix}$

以上即为更新后权重.