Bishop Questions

David Ruhe

August 4, 2020

4.1

If the convex hulls intersect, then $\exists \boldsymbol{z} : \boldsymbol{z} = \sum_n \alpha_n \boldsymbol{x}_n = \sum_n \beta_n \boldsymbol{y}_n$

$$\hat{w}^T \boldsymbol{z} + w_0 = \hat{w}^T \sum_n \alpha_n \boldsymbol{x}_n + w_0 = \hat{w}^T \sum_n \alpha_n \boldsymbol{x}_n + \sum_n \alpha_n w_0 = \sum_n \alpha_n \hat{w}^T x_n + w_0 > 0 \tag{1}$$

$$\hat{w}^T \boldsymbol{z} + w_0 = \hat{w}^T \sum_n \beta_n \boldsymbol{y}_n + w_0 = \hat{w}^T \sum_n \beta \boldsymbol{y}_n + \sum_n \beta w_0 = \sum_n \beta \hat{w}^T y_n + w_0 < 0 \tag{2}$$

This is a contradiction.

4.2

I've found that if you add the bias (a row of 1's) to X explicitly, then $\dagger X^T X$ will be a matrix of which all the rows and columns sum to 1. I'm not sure how this is proven. It does mean that $a^T y(x) + b = a^T T^T X^{\dagger T} x + b = -\boldsymbol{b} h^{\dagger T} x + b = -b + b = 0$, as required.

4.3

4.4

$$\mathcal{L} = \boldsymbol{w}^T(\boldsymbol{m}_2 - \boldsymbol{m}_1) + \lambda(\boldsymbol{w}^T \boldsymbol{w} - 1) \tag{3}$$

$$\nabla_{\boldsymbol{w}} \mathcal{L} = (\boldsymbol{m}_2 - \boldsymbol{m}_1) + 2\lambda \boldsymbol{w} = 0 \tag{4}$$

$$\boldsymbol{w} = -\frac{1}{2\lambda}(\boldsymbol{m}_2 - \boldsymbol{m}_1) \propto (\boldsymbol{m}_2 - \boldsymbol{m}_1) \tag{5}$$

4.5

It's quite straightforward if we fill in the given equations. Numerator:

$$
\begin{aligned}
(m_2 - m_1)^2 &= (\boldsymbol{w}^T\boldsymbol{m}_2 - \boldsymbol{w}^T\boldsymbol{m}_1)^2 && (6)\\
&= \boldsymbol{w}^T\boldsymbol{m}_2\boldsymbol{w}^T\boldsymbol{m}_2 - \boldsymbol{w}^T\boldsymbol{m}_2\boldsymbol{w}^T\boldsymbol{m}_1 - \boldsymbol{w}^T\boldsymbol{m}_1\boldsymbol{w}^T\boldsymbol{m}_2 + \boldsymbol{w}^T\boldsymbol{m}_1\boldsymbol{w}^T\boldsymbol{m}_1 && (7)\\
&= \boldsymbol{w}^T(\boldsymbol{m}_2\boldsymbol{m}_2^T - \boldsymbol{m}_2\boldsymbol{m}_1^T - \boldsymbol{m}_1\boldsymbol{m}_2^T + \boldsymbol{m}_1\boldsymbol{m}_1^T)\boldsymbol{w} && (8)\\
&= \boldsymbol{w}^T(\boldsymbol{m}_2 - \boldsymbol{m}_1)(\boldsymbol{m}_2 - \boldsymbol{m}_1)^T\boldsymbol{w} && (9)
\end{aligned}
$$

The denominator uses exactly the same approach. I'll show for s_1^2.

$$
\begin{aligned}
s_1^2 &= \sum_{n\in C_1} (\boldsymbol{w}^T\boldsymbol{x}_n - \boldsymbol{w}^T\boldsymbol{m}_1)^2 && (10)\\
&= \sum_{n\in C_1} (\boldsymbol{w}^T\boldsymbol{x}_n\boldsymbol{w}^T\boldsymbol{x}_n - \boldsymbol{w}^T\boldsymbol{x}_n\boldsymbol{w}^T\boldsymbol{m}_1 - \boldsymbol{w}^T\boldsymbol{m}_1\boldsymbol{w}^T\boldsymbol{x}_n + \boldsymbol{w}^T\boldsymbol{m}_1\boldsymbol{w}^T\boldsymbol{m}_1) && (11)\\
&= \sum_{n\in C_1} \boldsymbol{w}^T(\boldsymbol{x}_n\boldsymbol{x}_n^T - \boldsymbol{x}_n\boldsymbol{m}_1^T - \boldsymbol{m}_1\boldsymbol{x}_n^T + \boldsymbol{m}_1\boldsymbol{m}_1^T)\boldsymbol{w} && (12)\\
&= \sum_{n\in C_1} \boldsymbol{w}^T(\boldsymbol{x}_n - \boldsymbol{m}_1)(\boldsymbol{x}_n - \boldsymbol{m}_1)^T\boldsymbol{w} && (13)
\end{aligned}
$$

4.6

$$
\begin{aligned}
\sum_{n=1}^{N}(\boldsymbol{w}^T\boldsymbol{x}_n + w_0 - t_n)\boldsymbol{x}_n &= 0 && (14)\\
= \sum_{n=1}^{N}(\boldsymbol{w}^T\boldsymbol{x}_n - \boldsymbol{w}^T\boldsymbol{m} - t_n)\boldsymbol{x}_n & && (15)
\end{aligned}
$$

$$
\boldsymbol{w}^T\sum_{n=1}^{N}(\boldsymbol{x}_n - \boldsymbol{m})\boldsymbol{x}_n = \sum_{n=1}^{N} t_n\boldsymbol{x}_n \qquad (16)
$$

This gives us the three terms that we have to solve for.

$$
\begin{aligned}
\sum_{n=1}^{N} t_n\boldsymbol{x}_n &= \sum_{n\in C_1}^{N_1} t_n\boldsymbol{x}_n - \sum_{n\in C_2}^{N_2} t_n\boldsymbol{x}_n && (17)\\
&= \sum_{n\in C_1}^{N_1} \frac{N}{N_1}\boldsymbol{x}_n - \sum_{n\in C_2}^{N_2} \frac{N}{N_2}\boldsymbol{x}_n && (18)\\
&= N\left(\sum_{n\in C_1}^{N_1} \frac{1}{N_1}\boldsymbol{x}_n - \sum_{n\in C_2}^{N_2} \frac{1}{N_2}\boldsymbol{x}_n\right) && (19)\\
&= N(\boldsymbol{m}_1 - \boldsymbol{m}_2) && (20)
\end{aligned}
$$

$$
\begin{aligned}
-\boldsymbol{w}^T\boldsymbol{m}\sum_{n=1}^{N}\boldsymbol{x}_n &= -\frac{1}{N}\boldsymbol{w}^T(N_1\boldsymbol{m}_1+N_2\boldsymbol{m}_2)(N_1\boldsymbol{m}_1+N_2\boldsymbol{m}_2) && (21)\\
&= -\frac{1}{N}(N_1\boldsymbol{m}_1+N_2\boldsymbol{m}_2)(N_1\boldsymbol{m}_1+N_2\boldsymbol{m}_2)^T\boldsymbol{w} && (22)\\
&= -\frac{1}{N}(N_1^2\boldsymbol{m}_1\boldsymbol{m}_1^T+N_1N_2\boldsymbol{m}_1\boldsymbol{m}_2^T+N_2N_1\boldsymbol{m}_2\boldsymbol{m}_1^T+N_2^2\boldsymbol{m}_2\boldsymbol{m}_2^T)\boldsymbol{w} && (23)\\
&= -\frac{1}{N}\left((N-N_2)N_1\boldsymbol{m}_1\boldsymbol{m}_1^T+N_1N_2\boldsymbol{m}_1\boldsymbol{m}_2^T+N_2N_1\boldsymbol{m}_2\boldsymbol{m}_1^T+(N-N_1)N_2\boldsymbol{m}_2\boldsymbol{m}_2^T\right)\boldsymbol{w} && (24)\\
&= \left(-N_1\boldsymbol{m}_1\boldsymbol{m}_1^T+\frac{N_1N_2}{N}\boldsymbol{m}_1\boldsymbol{m}_1^T-\frac{N_1N_2}{N}\boldsymbol{m}_1\boldsymbol{m}_2^T-\frac{N_2N_1}{N}\boldsymbol{m}_2\boldsymbol{m}_1^T-N_2\boldsymbol{m}_2\boldsymbol{m}_2^T+\frac{N_1N_2}{N}\boldsymbol{m}_2\boldsymbol{m}_2^T\right)\boldsymbol{w} && (25)\\
&= \left(-N_1\boldsymbol{m}_1\boldsymbol{m}_1^T-N_2\boldsymbol{m}_2\boldsymbol{m}_2^T+\frac{N_1N_2}{N}(\boldsymbol{m}_2-\boldsymbol{m}_1)(\boldsymbol{m}_2-\boldsymbol{m}_1)^T\right)\boldsymbol{w} && (26)\\
&= \left(-N_1\boldsymbol{m}_1\boldsymbol{m}_1^T-N_2\boldsymbol{m}_2\boldsymbol{m}_2^T+\frac{N_1N_2}{N}\boldsymbol{S}_B\right)\boldsymbol{w} && (27)\\
& && (28)
\end{aligned}
$$

We add the remaining terms to the final term:

$$
\begin{aligned}
&\left(-N_1\boldsymbol{m}_1\boldsymbol{m}_1^T-N_2\boldsymbol{m}_2\boldsymbol{m}_2^T+\sum_{n=1}^{N}\boldsymbol{x}_n\boldsymbol{x}_n^T\right)\boldsymbol{w} && (29)\\
&= \left(N_1\boldsymbol{m}_1\boldsymbol{m}_1^T-2N_1\boldsymbol{m}_1\boldsymbol{m}_1^T+N_2\boldsymbol{m}_2\boldsymbol{m}_2^T-2N_2\boldsymbol{m}_2\boldsymbol{m}_2^T+\sum_{n=1}^{N}\boldsymbol{x}_n\boldsymbol{x}_n^T\right)\boldsymbol{w} && (30)\\
&= \left(N_1\boldsymbol{m}_1\boldsymbol{m}_1^T-\sum_{n\in C_1}\boldsymbol{x}_n\boldsymbol{m}_1^T-\boldsymbol{m}_1\sum_{n\in C_1}\boldsymbol{x}_1^T+...+\sum_{n=1}^{N}\boldsymbol{x}_n\boldsymbol{x}_n^T\right)\boldsymbol{w} && (31)\\
&= \left(\sum_{n\in C_1}\boldsymbol{m}_1\boldsymbol{m}_1^T-\boldsymbol{x}_n\boldsymbol{m}_1^T-\boldsymbol{m}_1\boldsymbol{x}_n^T+\boldsymbol{x}_n\boldsymbol{x}_n^T+...\right)\boldsymbol{w} && (32)
\end{aligned}
$$

Which gives the product we need. "..." denotes symmetric steps but for the class 2.

4.7

$$\sigma(-a)=\frac{1}{e^x+1}=\frac{e^{-x}}{1+e^{-x}}=\frac{e^{-x}+1}{e^{-x}+1}-\frac{1}{e^{-x}+1} \tag{33}$$

$$y=\frac{1}{1+e^{-x}} \tag{34}$$

$$e^{-x}=\frac{1-y}{y} \tag{35}$$

$$y=e^x(1-y) \tag{36}$$

$$e^x = y/(1-y) \tag{37}$$
$$x = \ln\left[y/(1-y)\right] \tag{38}$$

4.8

$$p(C_1|\boldsymbol{x}) = \sigma(a) \tag{39}$$

So we have to show: $a = \boldsymbol{w}^T\boldsymbol{x} + w_0$

$$a = \ln\frac{p(\boldsymbol{x}|C_1)}{p(\boldsymbol{x}|C_2)} + \ln\frac{p(C_1)}{p(C_2)} \tag{40}$$

$$\begin{aligned}
\ln\frac{p(\boldsymbol{x}|C_1)}{p(\boldsymbol{x}|C_2)} &= \ln\left[\frac{1}{(2\pi)^{D/2}}\frac{1}{|\boldsymbol{\Sigma}|^{1/2}}\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_1)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1)\right)\right] && (41)\\
&\quad - \ln\left[\frac{1}{(2\pi)^{D/2}}\frac{1}{|\boldsymbol{\Sigma}|^{1/2}}\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_2)\right)\right] && (42)\\
&= -\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_1)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1) + \frac{1}{2}(\boldsymbol{x}-\mu_2)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_2) && (43)\\
&= \frac{1}{2}\left[2\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_1^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 - 2\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2 + \boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right] && (44)\\
& && (45)
\end{aligned}$$

From which the result easily follows.

4.9

Hint 1: Using a Lagrange multiplier, make sure $\sum_{j=1}^{k}\pi_j = 1$ before optimizing.

$$\begin{aligned}
\ln p(\boldsymbol{X}|\boldsymbol{T}) &= \sum_{n=1}^{N}\ln\prod_{j=1}^{K}(\pi_j\mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Sigma})_j^t) && (46)\\
&= \sum_{n=1}^{N}\sum_{j=1}^{K}t_j\ln(\pi_j\mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k\boldsymbol{\Sigma})) && (47)
\end{aligned}$$

Now we optimize with constraint $\sum_{j=1}^{K}\pi_j = 1$.

$$\mathcal{L}(\boldsymbol{\pi},\lambda) = \ln(p(\boldsymbol{X},\boldsymbol{T})) - \lambda(\sum_{j=1}^{K}\pi_j - 1) \tag{48}$$

$$\frac{\partial\mathcal{L}}{\partial\pi_j} = \sum_{n=1}^{N}\frac{t_j}{\pi_j} - \lambda = 0 \tag{49}$$

$$\lambda = \sum_{n=1}^{N}\frac{t_j}{\pi_j} = N\frac{t_j}{\pi_j} = \frac{N_j}{\pi_j} \tag{50}$$

$$\pi_j = \frac{N_j}{\lambda} \tag{51}$$

$$\frac{\partial \mathcal{L}}{\partial \lambda} = \sum_{j=1}^{K} \pi_j = 1 \tag{52}$$

Plugging this into 380.

$$\sum_{j=1}^{K} \frac{N_j}{\lambda} = 1 \iff \lambda = N \tag{53}$$

Gives us the desired result.

4.10

The likelihood of the model model for the entire dataset can be described as

$$\prod_{n=1}^{N} \prod_{k=1}^{K} \mathcal{N}(\boldsymbol{\phi}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma})^{t_{nk}} \tag{54}$$

Computing the log-derivative:

$$\begin{aligned}
&\sum_{n=1}^{N} \sum_{k=1}^{K} t_{nk} \frac{d}{d\mu_k} \log \mathcal{N}(\boldsymbol{\phi}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}) && (55)\\
&= \sum_{n=1}^{N} \sum_{k=1}^{K} t_{nk} \frac{d}{d\mu_k} [-\frac{1}{2}(\boldsymbol{\phi} - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\phi} - \boldsymbol{\mu}_k)] && (56)\\
&= \sum_{n=1}^{N} \sum_{k=1}^{K} t_{nk} t_{nk} \boldsymbol{\Sigma}^{-1}](\boldsymbol{\phi} - \boldsymbol{\mu}_k) = 0 && (57)\\
&\Rightarrow \sum_{n=1}^{N} \sum_{k=1}^{K} t_{nk}(\boldsymbol{\phi}_n - \boldsymbol{\mu}_k) = 0 && (58)\\
&\Rightarrow \sum_{n=1}^{N} t_{nk}(\boldsymbol{\phi}_n - \boldsymbol{\mu}_k) = 0 \qquad \text{Since } t_k \text{ is 1-of-K} && (59)\\
&\Rightarrow \sum_{n=1}^{N} t_{nk} \boldsymbol{\phi}_n = \sum_{n=1}^{N} t_{nk} \boldsymbol{\mu}_k && (60)\\
&\Rightarrow \sum_{n=1}^{N} t_{nk} \boldsymbol{\phi}_n = \boldsymbol{\mu}_k \sum_{n=1}^{N} t_{nk} = N \boldsymbol{\mu}_k && (61)\\
& && (62)
\end{aligned}$$

The process for Σ is highlly similar.

$$\sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk}\frac{d}{d\mu_k}\log\mathcal{N}(\boldsymbol{\mu}|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}) \tag{63}$$

$$= \sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk}\frac{d}{d\mu_k}\left[\log|\boldsymbol{\Sigma}|^{-1/2} - \frac{1}{2}(\boldsymbol{\phi}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\phi}-\boldsymbol{\mu}_k\right] \tag{64}$$

$$= \sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk}\left(-1/2\boldsymbol{\Sigma} - 1/2(\boldsymbol{\phi}-\boldsymbol{\mu}_k)(\boldsymbol{\phi}-\boldsymbol{\mu}_k)^T)\right) = 0 \tag{65}$$

$$\Rightarrow \boldsymbol{\Sigma}\sum_{n=1} t_{nk} = N_k\sum_{n=1}^{N} S_k \qquad \Rightarrow \boldsymbol{\Sigma}N = N_k\sum_{n=1}^{N} S_k \tag{66}$$

4.11

If you realise that the binary case can be seen as a $L = 2$ case it is a natural extension of the example in the book:

$$p(\phi|C_k) = \prod_{i=1}^{D}\prod_{j=1}^{L}\mu_{ijk}^{\phi_{ij}} \tag{67}$$

$$\log p(\phi, C_k) = \log p(\Phi|C_k) + \log p(C_k) = \sum_{i=1}^{D}\sum_{j=1}^{L}\phi_{ij}\log\mu_{ijk} + \log p(C_k) \tag{68}$$

4.12

$$\frac{d\sigma}{da} = (1+e^{-a})^{-2}e^{-a} \tag{69}$$

$$= \frac{1}{1+e^{-a}}\frac{1}{1+e^{-a}}e^{-a} \tag{70}$$

$$= \sigma(a)\frac{e^{-a}}{1+e^{-a}} \tag{71}$$

$$= \sigma(a)\left[\frac{1+e^{-a}}{1+e^{-a}} - \frac{1}{1+e^{-a}}\right] \tag{72}$$

4.13

$$
\begin{aligned}
\nabla_{\boldsymbol{w}} E(\boldsymbol{w}) &= -\sum_{n=1}^{N} \frac{t_n}{y_n} \nabla_{\boldsymbol{w}} \sigma(\boldsymbol{w}^T \boldsymbol{\phi}_n) - \frac{1-t_n}{1-y_n} \nabla_{\boldsymbol{w}} \sigma(\boldsymbol{w}^T \boldsymbol{\phi}_n) && (73)\\
&= -\sum_{n=1}^{N} \frac{t_n}{y_n} \sigma(\boldsymbol{w}^T \boldsymbol{\phi}_n)(1-\sigma(\boldsymbol{w}^T \boldsymbol{\phi}_n))\boldsymbol{\phi}_n - \frac{1-t_n}{1-y_n} \sigma(\boldsymbol{w}^T \boldsymbol{\phi}_n)(1-\sigma(\boldsymbol{w}^T \boldsymbol{\phi}_n))\boldsymbol{\phi}_n && (74)\\
&= -\sum_{n=1}^{N} t_n(1-y_n)\boldsymbol{\phi}_n - (1-t_n)y_n\boldsymbol{\phi}_n && (75)\\
&= -\sum_{n=1}^{N} (t_n - y_n)\boldsymbol{\phi}_n && (76)
\end{aligned}
$$

4.14

Hint 1: approach it with an argument, using that we have a perfect decision boundary at $\boldsymbol{w}^T\boldsymbol{\phi} = 0$.

We know that if C_1 is labelled with $t_{C_1} = 1$ and C_2 is labelled with $t_{C_2} = 0$ then we want $p(C_1|\phi) = \sigma(\boldsymbol{w}^T\phi) > 0.5$ and $p(C_2|\phi) = \sigma(\boldsymbol{w}^T\phi) < 0.5$ which happens if the decision boundary perfectly separates them at $\boldsymbol{w}^T\phi = 0$. Now the binary cross entropy will be minimal as $p(C_1|\phi) \to 1$ which happens when $\boldsymbol{w} \to \infty$. And vice versa.

4.15

For H to be positive semidefinite we need to prove $u^T H u \geq 0$.

$$
u^T H u = \sum_{n=1^N} y_n(1-y_n) u^T \phi \phi^T u \qquad \text{We know that } y_n(1_y n) \text{ will be positive} \qquad (77)
$$

$$
\begin{aligned}
&= \sum_{n=1}^{N} y_n(1-y_n) \sum_{i}^{d} \sum_{j}^{d} u_i \phi_i \phi_j u_j && (78)\\
&= \sum_{n=1}^{N} (y_n(1-y_n)) \left(\sum_{i}^{d} u_i^2 \phi_i^2 + \sum_{i=1}^{d} \sum_{j=1, j\neq i}^{d} u_i \phi_i u_j \phi_j \right) && (79)\\
&= \sum_{n=1}^{N} (y_n(1-y_n)) \left(\sum_{i}^{d} u_i \phi_i \right)^2 && (80)
\end{aligned}
$$

And so it's a sum of positive terms, which will be positive.

By definition, since the Hessian is the second derivative of the error function is positive, therefore the function is convex (error in book).

4.16

$$p(\boldsymbol{t}, \boldsymbol{w}) = \prod_{n=1}^{N} y_n^{\pi}[1 - y_n]^{1-t_n} \tag{81}$$

$$\ln p = \sum_{n=1}^{N} \pi_n \ln y_n + (1 - \pi_n) \ln(1 - y_n) \tag{82}$$

4.17

$$p(C_k|\boldsymbol{\phi}) = y_k = \frac{\exp a_k}{\sum_{j=1} \exp a_j} \tag{83}$$

$$\begin{aligned}
\frac{\partial y_k}{\partial a_j} &= -\exp a_k(\sum_j \exp(a_j)^{-2} \exp(a_j) && (84)\\
&= \begin{cases} y_k(0 - y_j) & j \neq k \\ y_k(1 - y_j) & j = k \end{cases} && (85)\\
&= y_k(I_{kj} - y_j) && (86)
\end{aligned}$$

4.18

$$\nabla_{\boldsymbol{w}_j} E(\boldsymbol{W}) = -\sum_{n=1}^{N} \sum_{k=1}^{K} t_{nk} \nabla_{\boldsymbol{w}_j} \ln y_n k \tag{87}$$

$$\nabla_{\boldsymbol{w}_j} \ln(y_{nk}) = -(I_{kj} - y_j)\boldsymbol{\phi}_n \tag{88}$$

$$\begin{aligned}
\nabla_{\boldsymbol{w}_j} E(\boldsymbol{W}) &= -\sum_{n=1}^{N} \sum_{k=1}^{K} t_{nk}(I_{kj} - y_j n)\boldsymbol{\phi}_n && (89)\\
&= \boldsymbol{\phi} \sum_{n=1}^{N} \sum_{k=1}^{K} t_{nk} y_{jn} \boldsymbol{\phi} - t_{nk} I_{kj} && (90)\\
&= \boldsymbol{\phi} \sum_{n=1}^{N} t_{nj} - y_{jn}\boldsymbol{\phi} \underbrace{\sum_{k=1}^{K} t_{nk}}_{=1} && (91)\\
&= \sum_{n=1}^{N} \boldsymbol{\phi}(y_{jn} - t_{nj}) && (92)
\end{aligned}$$

4.19

Hint 1: Use binary cross netropy and 4.114 as the activation function. Use the fundamental theorem of calculus.

$$p(\boldsymbol{t}, \boldsymbol{w}) = \sum_{n=1}^{N} [t_n \ln y_n + (1 - t_n \ln(1 - y_n)] \tag{93}$$

$$\nabla_{\boldsymbol{w}} \sum_{n=1}^{N} [t_n \ln \Phi(a) + (1 - t_n) \ln(1 - \Phi(a))] = \sum_{n=1}^{N} (\frac{t_n}{\Phi(a)} - \frac{1 - t_n}{1 - \Phi(a)}) \Phi(a) \boldsymbol{\phi}_n \tag{94}$$

4.20

$$u^T H u = \sum_i \sum_j u_i H_{ij} u_j = \sum_i \sum_j u_i \left[\sum_{j=1}^{N} y_n (1 - y_n) \phi_n \phi_n^T \right]_{ij} u_j \tag{95}$$

with $H_{ij} = \sum_{n=1}^{N} y_n (1 - y_n) \phi_{ni} \phi_{nj}$

4.21

$$\begin{aligned}
\Phi(a) &= \int_0^a \mathcal{N}(0, 1) d\theta & (96) \\
&= \int_0^a 1/\sqrt{2\pi} \exp(-\frac{1}{2}\theta^2) d\theta & (97) \\
&= \frac{1}{2} + \int_{-\infty}^{a} \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}\theta^2) d\theta & (98) \\
&= \frac{1}{2} + \frac{1}{\sqrt{2\pi}} \frac{\sqrt{\pi}}{2} \int_{-\infty}^{a} \frac{2}{\sqrt{\pi}} \exp(-\frac{1}{2}\theta^2) d\theta & (99) \\
&= \frac{1}{2} \left(1 + \frac{1}{\sqrt{2}} \mathrm{erf}(a) \right) & (100)
\end{aligned}$$

4.22

$$\begin{aligned}
\ln p(D) &= \ln \left[f(z_0) \frac{(2\pi)^{M/2}}{|A|^{1/2}} \right] & (101) \\
&= \ln f(z_0) = \frac{M}{2} \ln 2\pi - \frac{1}{2} \ln |A| & (102)
\end{aligned}$$

z_0 is the location of the $\boldsymbol{\theta}_{MAP}$ estimate so

$$\ln f(\boldsymbol{\theta}) \Big| z_0 = \ln f(\boldsymbol{\theta}_{MAP}) = \ln p(D|\boldsymbol{\theta}_{MAP}) + \ln p(\boldsymbol{\theta}_{MAP}) \tag{103}$$

4.23

The second term vanishes. Then we're left with $\ln|H| = \ln\left|\sum_{n=1}^{N} H_n\right| = \ln\left|N\frac{1}{N}\sum_{n=1}^{N} H_n\right| = \ln|N + \ln 1/N \sum_N H_n| = \ln N^M \left|1/N\sum_{n=1}^{N} h_n\right| = M\ln N + \ln\left|\frac{1}{N}\sum_{n=1}^{N} H_n\right|$.

We can drop the last term since it does not grow with N. We have used the property $|cA| = c^M\,|A|$ where $A \in R^{M\times M}$.

4.24

4.25

$$\frac{d}{da}\sigma(a) = \sigma(a)(1-\sigma(a)) \tag{104}$$

$$\begin{aligned}
\frac{d}{da}\Phi(\lambda a) &= \frac{d\Phi(\lambda a)}{d\lambda a}\frac{d\lambda a}{da} && (105)\\
&= \lambda\frac{1}{2\sqrt{2}}\frac{\text{derf}(\lambda a)}{d\lambda a} && (106)\\
&= \frac{\lambda}{2\sqrt{2}}\frac{2}{\sqrt{\pi}}\int_0^{\lambda a}\frac{d}{d\lambda a}\exp(\frac{-(\lambda a)^2}{2})d(\lambda a) && (107)\\
&= \frac{\lambda}{\sqrt{2\pi}}\exp(\frac{-(\lambda a)^2}{2}) && (108)
\end{aligned}$$

Where we used the fundamental theorem of calculus.

$$\begin{aligned}
\frac{d}{da}\sigma(a)\bigg|_{a=0} &= \frac{d}{da}\Phi(\lambda a)\bigg|_{a=0} && (109)\\
\frac{1}{4} &= \frac{\lambda}{\sqrt{2\pi}} && (110)\\
\frac{1}{4}\sqrt{2\pi} &= \lambda && (111)\\
\lambda^2 &= \frac{2}{16}\pi && (112)
\end{aligned}$$

4.26