MIT 9.520/6.860: Statistical Learning Theory and Applications **Fall 2022**

Problem Set 3

Student Name: Aniruddha Deshpande

Problem 1

Problem 1.1

For this first part, I don't use the re-parametrized form of the neural network and let:

$$f(W;x) = W_K\sigma(W_{K-1}\dot{.}.W_2\sigma(W_1x)...)$$

Now, the aim is to 1) Characterize what the term $\frac{\partial||W_k||^2}{\partial t}$ looks like and 2) Show that this rate of change of the squared Frobenius norm of all weight matrices W_k is positive once separability is achieved on the dataset $\mathcal{S}$, which will indicate that the norms of the weights keep increasing. Now, to do 1) see that $\forall k \in [K]$

$$\begin{aligned}||W_k||^2 &= \sum_{i,j}(W_k^{ij})^2 \\ \frac{1}{2}\frac{\partial||W_k||^2}{\partial t} &= \sum_{i,j} W_k^{ij}\frac{\partial W_k^{ij}}{\partial t}\end{aligned}$$

Now, using the gradient flow equation $\frac{\partial W_k^{ij}}{\partial t} = -\frac{\partial L(W)}{\partial W_k^{ij}}$ applied to the exponential loss $L(W) = \sum_{i=1}^n e^{-y_i f(W;x_i)}$ we get :

$$\begin{aligned}\frac{1}{2}\frac{\partial||W_k||^2}{\partial t} &= \sum_{i,j} W_k^{ij}(\sum_{n=1}^N e^{-y_n f(W;x_n)}\frac{\partial f(W;x_n)}{\partial W_k^{ij}}y_n) \\ &= \sum_{i,j}\sum_{n=1}^N W_k^{ij}\frac{\partial f(W;x_n)}{\partial W_k^{ij}}y_n e^{-y_n f(W;x_n)} \\ &= \sum_{n=1}^N y_n e^{-y_n f(W;x_n)}(\sum_{i,j} W_k^{ij}\frac{\partial f(W;x_n)}{\partial W_k^{ij}})\end{aligned}$$

Now, using the structural property of ReLU networks $\sum_{i,j} W_k^{ij}\frac{\partial f(W;x_n)}{\partial W_k^{ij}} = f(W;x_n)$, we get:

$$\frac{1}{2}\frac{\partial||W_k||^2}{\partial t} = \sum_{n=1}^{N} f(W;x_n)y_n e^{-y_n f(W;x_n)}$$

Now, a couple of things to note about this. The above term characterizes the rate of change of the Frobenius norms of the weights for layer k. However, we can see that this rate of change is independent of k! This means that the norms of all the weight matrices change at the same rate.

The second thing to note is that when complete separability on $\mathcal{S}$ is achieved ie. when we have $y_n f(W;x_n) > 0 \forall n \in [N]$, then the above term is always positive. This can be observed by looking at individual terms of the summation and noting that they are all positive. This completes the proof, and we can see that once complete separability is achieved , the rate of change of weight norms is always positive, meaning they keep increasing.

Problem 1.2

I will tackle this problem in 2 steps. 1) Characterize how the rate of change of the normalized margin $\mathcal{V}_1 = y_1 f(V;x_1)$ looks like, ie. characterise $\frac{\partial \mathcal{V}_1}{\partial t}$ and 2) Show that this quantity is always positive. First, we know that:

$$\mathcal{V}_1 = y_1 f(V_1, V_2 ... V_K; x_1)$$

Now, using the chain-rule for partial derivatives, we get:

$$\frac{\partial \mathcal{V}_1}{\partial t} = \sum_{k=1}^{K} (\frac{\partial \mathcal{V}_1}{\partial V_k})^T \frac{\partial V_k}{\partial t} \tag{3.1}$$

Note, that here the derivatives are with respect to matrices, so each derivative terms $\in \mathbb{R}^{d_{k-1} \times d_k}$. The inner product term is defined similarly as for vectors, and is just a sum of elementwise products. Now, to characterize each of the partial derivative terms. The first term is straightforward and is:

$$\frac{\partial \mathcal{V}_1}{\partial V_k} = y_1 \frac{\partial f(V;x_1)}{\partial V_k} \tag{3.2}$$

The second term is a little more work. I start by characterising it using the gradient flow equation $\frac{\partial V_k}{\partial t} = -\frac{\partial \mathcal{L}}{\partial V_k}$. To derive this, I will not use the full Loss function as defined in the question, but the approximated loss function obtained after a long time, where only the term with the largest margin dominates. That is the loss is given by:

$$\mathcal{L} = e^{-y_1 \rho f(V;x_1)} + \sum_{k=1}^{K} \lambda_k ||V_k||^2$$

$$\begin{aligned}\frac{\partial V_k}{\partial t} &= \dot{V}_k \\ &= e^{-\rho y_1 f(V;x_1)} \rho y_1 \frac{\partial f(V;x_1)}{\partial V_k} - 2\lambda_k V_k\end{aligned}$$

Now, I need to find what λ_k is. I will derive this using the same trick as used in the lecture, using $V_k^T \dot{V}_k = 0$. Pre-multiplying above term with V_k^T and setting it to 0 I get:

$$\begin{aligned}V_k^T \dot{V}_k &= e^{-\rho y_1 f(V;x_1)} \rho y_1 V_k^T \frac{\partial f(V;x_1)}{\partial V_k} - 2\lambda_k V_k^T V_k \\ &= e^{-\rho y_1 f(V;x_1)} \rho y_1 f(V;x_1) - 2\lambda_k \\ &= 0 \\ 2\lambda_k &= \rho y_1 f(V;x_1) e^{-\rho y_1 f(V;x_1)}\end{aligned}$$

Where, the second equality comes from the fact that $V_k^T \frac{\partial f(V;x_1)}{\partial V_k} = f(V;x_1)$ .Now, substituting this value of λ_k into the equation for $\dot{V}_k$ I get:

$$\begin{aligned}\dot{V}_k &= \rho y_1 \frac{\partial f(V;x_1)}{\partial V_k} e^{-\rho y_1 f(V;x_1)} - \rho y_1 f(V;x_1) e^{-\rho y_1 f(V;x_1)} V_k \\ &= \rho y_1 e^{-\rho y_1 f(V;x_1)} (\frac{\partial f(V;x_1)}{\partial V_k} - f(V;x_1) V_k)\end{aligned}$$

Now, combining this with 3.2 and substituting them into equation 3.1 , we get:

$$\begin{aligned}\frac{\partial \mathcal{V}_1}{\partial t} &= \sum_{k=1}^{K} y_1 (\frac{\partial f(V;x_1)}{\partial V_k})^T \rho y_1 e^{-\rho y_1 f(V;x_1)} (\frac{\partial f(V;x_1)}{\partial V_k} - f(V;x_1) V_k) \\ &= \sum_{k=1}^{K} \rho y_1^2 e^{-\rho y_1 f(V;x_1)} (\frac{\partial f(V;x_1)}{\partial V_k})^T (\frac{\partial f(V;x_1)}{\partial V_k} - f(V;x_1) V_k) \\ &= \sum_{k=1}^{K} C (\frac{\partial f(V;x_1)}{\partial V_k})^T (\frac{\partial f(V;x_1)}{\partial V_k} - f(V;x_1) V_k) \\ &= \sum_{k=1}^{K} C (||\frac{\partial f(V;x_1)}{\partial V_k}||^2 - f(V;x_1) (\frac{\partial f(V;x_1)}{\partial V_k})^T V_k) \\ &= \sum_{k=1}^{K} C (||\frac{\partial f(V;x_1)}{\partial V_k}||^2 - f^2(V;x_1))\end{aligned}$$

Now, to show that this term is always positive. I'm not fully sure how I can do that here!

Problem 2

Problem 2.1

To find the compositional kernel my approach will be to replace any inner product that appears in the kernel term with the inner product in the first feature space. See:

$$\begin{aligned} K(x,y) &= \langle \phi(x), \phi(y) \rangle \\ &= \langle x, y \rangle^d \end{aligned}$$

Now, consider the compositional kernel:

$$\begin{aligned} K^{(2)}(x,y) &= K(\phi(x), \phi(y)) \\ &= \langle \phi(\phi(x)), \phi(\phi(x)) \rangle \\ &= \langle \phi(x), \phi(y) \rangle^d \\ &= (\langle x, y \rangle^d)^d \\ &= \langle x, y \rangle^{d^2} \end{aligned}$$

Problem 2.2

I will use a similar approach for this question as in the previous part. Here, I will also replace distance terms that appear in the kernel. Distances (and norms) can also be written as inner products and replaced in the same way as I did in the previous part. Also for notation let $\mathcal{F}_0$ be the space in which the initial data points x and y lie in. Let $\mathcal{F}_1$ be the feature space that one mapping with the RBF kernel sends the data points into. Let, $d_{\mathcal{F}_i}(x,y)$ be the distance between the data points after mapping them to the feature space $\mathcal{F}_i$ for $i = 0, 1$.

$$\begin{aligned} K(x,y) &= exp(-\frac{||x-y||^2}{\gamma^2}) \\ &= exp(-\frac{d^2_{\mathcal{F}_0}(x,y)}{\gamma^2}) \end{aligned}$$

Now, we can write the compositional kernel as :-

$$K^{(2)}(x,y) = exp(-\frac{d^2_{\mathcal{F}_1}(x,y)}{\gamma^2})$$

To further simplify it, I will use the following:

$$\begin{aligned} d^2_{\mathcal{F}_1}(x,y) &= ||\phi(x)-\phi(y)||^2 \\ &= \langle\phi(x),\phi(x)\rangle + \langle\phi(y),\phi(y)\rangle - 2\langle\phi(x),\phi(y)\rangle \\ &= K(x,x)+K(y,y)-2K(x,y) \\ &= 2-2exp(-\frac{||x-y||^2}{\gamma^2}) \end{aligned}$$

Substituting this back I get:

$$K^{(2)}(x,y) = exp(-\frac{1}{\gamma^2}(2-2exp(-\frac{||x-y||^2}{\gamma^2})))$$

Now, to compare the kernels $K(x,y)$ and $K^{(2)}(x,y)$ for 1) $||x-y|| \to 0$ and 2) $||x-y|| \to \infty$. For 1), we get the values:

$$\begin{aligned} K(||x-y|| \to 0) &= 1 \\ K^{(2)}(||x-y|| \to 0) &= 1 \end{aligned}$$

For 2) we get:

$$\begin{aligned} K(||x-y|| \to \infty) &= 0 \\ K^{(2)}(||x-y|| \to \infty) &= exp(-\frac{2}{\gamma^2}) \end{aligned}$$

Problem 2.3

For this part I will use the following notation. Let $u_1, u_2, ...u_d \in \mathbb{R}^d$ be unit-length orthogonal basis vectors for $\mathbb{R}^d$. Then, we can write any $w \in \mathbb{R}^d$ as:

$$w = \sum_{i=1}^{d} w_i u_i$$

Now, given the assumption mentioned in the question we can write the data vectors x, y as :

$$
\begin{aligned}
x &= ||x||u_1 \\
y &= y_1u_1 + y_2u_2 \\
||y||^2 &= y_1^2 + y_2^2
\end{aligned}
$$

Now, to proceed to compute the given integral:

$$
k_n(x,y) = \int_{w_1...w_d} dw \frac{e^{-\frac{||w||^2}{2}}}{(2\pi)^{\frac{d}{2}}} \Theta(\langle w,x\rangle)\Theta(\langle w,y\rangle)\langle w,x\rangle^n\langle w,y\rangle^n
$$

However, given the forms of x and y , the terms $\Theta(\langle w,x\rangle), \Theta(\langle w,y\rangle), \langle w,x\rangle^n, \langle w,y\rangle^n$ depend only on w_1 and w_2 and not on any other components of w. So, we can re-write the integral as follows and marginalize over $w_3, w_4...w_d$. Note that the term inside the square brackets below is just the integral of the multi-dimensional Gaussian distribution over its entire support, and hence integrates to be equal to 1.

$$
\begin{aligned}
k_n(x,y) &= \int_{w_1,w_2} [\int_{w_3...w_d} (dw_3dw_4...dw_d) \frac{e^{-\frac{\sum_{j=3}^d w_j^2}{2}}}{(2\pi)^{\frac{d-2}{2}}}] \frac{e^{-\frac{w_1^2+w_2^2}{2}}}{(2\pi)} \Theta(\langle w,x\rangle)\Theta(\langle w,y\rangle)\langle w,x\rangle^n\langle w,y\rangle^n dw_1dw_2 \\
&= \frac{1}{2\pi}\int_{w_1,w_2} e^{-\frac{w_1^2+w_2^2}{2}} \Theta(\langle w,x\rangle)\Theta(\langle w,y\rangle)\langle w,x\rangle^n\langle w,y\rangle^n dw_1dw_2
\end{aligned}
$$

Now, to get the above integral in the form mentioned in the question I will multiply and divide by $||x||^n||y||^n$. This gives:

$$
k_n(x,y) = \frac{||x||^n||y||^n}{\pi}\frac{1}{2}\int_{w_1,w_2} e^{-\frac{w_1^2+w_2^2}{2}} \Theta(\langle w,x\rangle)\Theta(\langle w,y\rangle)\frac{\langle w,x\rangle^n}{||x||^n}\frac{\langle w,y\rangle^n}{||y||^n} dw_1dw_2
$$

By comparing with the term given in the question we now have:

$$
\begin{aligned}
J_n(\theta) &= \frac{1}{2}\int_{w_1,w_2} e^{-\frac{w_1^2+w_2^2}{2}} \Theta(\langle w,x\rangle)\Theta(\langle w,y\rangle)\frac{\langle w,x\rangle^n}{||x||^n}\frac{\langle w,y\rangle^n}{||y||^n} dw_1dw_2 \\
&= \frac{1}{2}\int_{w_1,w_2} e^{-\frac{w_1^2+w_2^2}{2}} \Theta(\langle w,x\rangle)\Theta(\langle w,y\rangle)(\frac{\langle w,x\rangle\langle w,y\rangle}{||x||||y||})^n dw_1dw_2
\end{aligned}
$$

Now, to further simplify this note the following equalities, that follow from the assumptions of the directions of x and y.

$$\begin{aligned}
\langle x, y\rangle &= ||x||y_1 \\
||x||y_2 &= ||x||\sqrt{||y||^2 - y_1^2} \\
&= \sqrt{||x||^2||y||^2 - \langle x, y\rangle^2} \\
\langle w, x\rangle\langle w, y\rangle &= ||x||w_1(y_1w_1 + y_2w_2) \\
&= ||x||y_1w_1^2 + ||x||y_2w_1w_2 \\
&= \langle x, y\rangle w_1^2 + \sqrt{||x||^2||y||^2 - \langle x, y\rangle^2}w_1w_2
\end{aligned}$$

Now, using this we get:

$$\begin{aligned}
\frac{\langle w, x\rangle\langle w, y\rangle}{||x||||y||} &= \frac{\langle x, y\rangle w_1^2 + \sqrt{||x||^2||y||^2 - \langle x, y\rangle^2}w_1w_2}{||x||||y||} \\
&= w_1^2\cos(\theta) + w_1w_2\sqrt{1 - \cos^2(\theta)} \\
&= w_1^2\cos(\theta) + w_1w_2\sin(\theta)
\end{aligned}$$

Substituting this back into the integral, we get:

$$J_n(\theta) = \frac{1}{2}\int_{w_1,w_2} e^{-\frac{w_1^2+w_2^2}{2}}\Theta(\langle w, x\rangle)\Theta(\langle w, y\rangle)(w_1^2\cos(\theta) + w_1w_2\sin(\theta))^n dw_1dw_2$$

Now, to simplify $\Theta(\langle w, x\rangle), \Theta(\langle w, y\rangle)$. Note, that since these are indicator functions, the way to simplify them would be to find the conditions on w_1, w_2 for which these equal to 1 and only compute the integral over the domain for which these conditions are satisfied. This looks like:

$$J_n(\theta) = \frac{1}{2}\int_{w_1,w_2|\langle w,x\rangle>0,\langle w,y\rangle>0} e^{-\frac{w_1^2+w_2^2}{2}}(w_1^2\cos(\theta) + w_1w_2\sin(\theta))^n dw_1dw_2$$

Now, to simplify these conditions:

$$\begin{aligned}
&\langle w, x\rangle > 0 \to ||x||w_1 > 0 \\
&\langle w, y\rangle > 0 \to ||y||\cos(\theta)w_1 + ||y||\sin(\theta)w_2 > 0
\end{aligned}$$

Now, since $||x||, ||y|| > 0$, we get:

$$J_n(\theta) = \frac{1}{2}\int_{w_1>0, w_1\cos(\theta)+w_2\sin(\theta)>0} e^{-\frac{w_1^2+w_2^2}{2}}(w_1^2\cos(\theta) + w_1w_2\sin(\theta))^n dw_1dw_2$$

Now, using the change of variables $u = w_1$ and $v = w_1\cos\theta + w_2\sin\theta$. Re-arranging this, we get :-

$$\begin{aligned} w_1 &= u \\ w_2 &= \frac{v - u\cos\theta}{\sin\theta} \end{aligned}$$

For change of variables, we also need to compute the Jacobian for this change of variables:

$$\begin{aligned} J(u,v) &= \begin{vmatrix} \frac{\partial w_1}{\partial u} & \frac{\partial w_1}{\partial v} \\ \frac{\partial w_2}{\partial u} & \frac{\partial w_2}{\partial v} \end{vmatrix} \\ &= \begin{vmatrix} 1 & 0 \\ -\frac{\cos\theta}{\sin\theta} & \frac{1}{\sin\theta} \end{vmatrix} \\ &= \frac{1}{\sin\theta} \end{aligned}$$

Doing the change of variables, we get:

$$\begin{aligned} J_n(\theta) &= \frac{1}{2}\int_{u>0,v>0} e^{-\frac{u^2 + \frac{v^2+u^2\cos^2\theta - 2uv\cos\theta}{\sin^2\theta}}{2}}(uv)^n J(u,v)dudv \\ &= \frac{1}{2}\int_{u>0,v>0} e^{-\frac{u^2+v^2-2uv\cos\theta}{2\sin^2\theta}}(uv)^n\frac{1}{\sin\theta}dudv \end{aligned}$$

Now, to do the second change of variables as suggested $u = r\cos(\frac{\psi}{2} + \frac{\pi}{4})$ and $v = r\sin(\frac{\psi}{2} + \frac{\pi}{4})$. Similar to the previous step, I will first compute the Jacobian $J(r, \psi)$ for this change of variables.

$$\begin{aligned}
J(r,\psi) &= \begin{vmatrix} \frac{\partial u}{\partial r} & \frac{\partial u}{\partial \psi} \\ \frac{\partial v}{\partial r} & \frac{\partial v}{\partial \psi} \end{vmatrix} \\
&= \begin{vmatrix} \cos(\frac{\psi}{2}+\frac{\pi}{4}) & -\frac{r}{2}\sin(\frac{\psi}{2}+\frac{\pi}{4}) \\ \sin(\frac{\psi}{2}+\frac{\pi}{4}) & \frac{r}{2}\cos(\frac{\psi}{2}+\frac{\pi}{4}) \end{vmatrix} \\
&= \frac{r}{2}(\cos^2\theta+\sin^2\theta) \\
&= \frac{r}{2}
\end{aligned}$$

Now, doing the change of variables, we get :

$$\begin{aligned}
J_n(\theta) &= \frac{1}{2\sin\theta}\int_{r>0,-\frac{\pi}{2}<\psi<\frac{\pi}{2}} e^{-\frac{r^2(1-2\sin(\frac{\psi}{2}+\frac{\pi}{4})\cos(\frac{\psi}{2}+\frac{\pi}{4})\cos\theta)}{2\sin^2\theta}}(r^2\sin(\frac{\psi}{2})\cos(\frac{\psi}{2}))^n\frac{r}{2}drd\psi \\
&= \frac{1}{2\sin\theta}\int_{r>0,-\frac{\pi}{2}<\psi<\frac{\pi}{2}} e^{-\frac{r^2(1-\sin(\psi+\frac{\pi}{2})\cos\theta)}{2\sin^2\theta}}(r^2\sin(\frac{\psi}{2})\cos(\frac{\psi}{2}))^n\frac{r}{2}drd\psi \\
&= \frac{1}{2\sin\theta}\int_{r>0,-\frac{\pi}{2}<\psi<\frac{\pi}{2}} e^{-\frac{r^2(1-\cos\psi\cos\theta)}{2\sin^2\theta}}(r^2\sin(\frac{\psi}{2})\cos(\frac{\psi}{2}))^n\frac{r}{2}drd\psi
\end{aligned}$$

Now, to further simplify this I will use the substitution $t = \frac{r^2(1-\cos\psi\cos\theta)}{2\sin^2\theta}$. Let ψ remain as is. The change of variables can be done similarly as before using the Jacobian and you get $2rdrd\psi = \frac{2\sin^2\theta}{1-\cos\psi\cos\theta}dtd\psi$

$$\begin{aligned}
J_n(\theta) &= \frac{1}{2\sin\theta}\int_{r>0,-\frac{\pi}{2}<\psi<\frac{\pi}{2}} e^{-\frac{r^2(1-\cos\psi\cos\theta)}{2\sin^2\theta}}(r^2\sin(\frac{\psi}{2})\cos(\frac{\psi}{2}))^n\frac{r}{2}drd\psi \\
&= \frac{1}{8\sin\theta}\int_{t>0,-\frac{\pi}{2}<\psi<\frac{\pi}{2}} e^{-t}r^{2n}\frac{1}{2^n}(2\sin(\frac{\psi}{2})\cos(\frac{\psi}{2}))^n 2rdrd\psi \\
&= \frac{1}{8\sin\theta}\int_{t>0,-\frac{\pi}{2}<\psi<\frac{\pi}{2}} e^{-t}t^n\frac{2^n\sin^{2n}\theta}{(1-\cos\psi\cos\phi)^n}\frac{1}{2^n}(\sin\psi)^n\frac{2\sin^2\theta}{1-\cos\psi\cos\theta}dtd\psi \\
&= \frac{\sin^{2n+1}(\theta)}{4}\int_{t>0,-\frac{\pi}{2}<\psi<\frac{\pi}{2}} e^{-t}t^n\frac{(\sin\psi)^n}{(1-\cos\psi\cos\phi)^n}\frac{1}{1-\cos\psi\cos\theta}dtd\psi \\
&= \frac{\sin^{2n+1}(\theta)}{4}\int_{t>0} e^{-t}t^n dt\int_{-\frac{\pi}{2}<\psi<\frac{\pi}{2}} \frac{(\sin\psi)^n}{(1-\cos\psi\cos\phi)^n}\frac{2}{1-\cos\psi\cos\theta}d\psi \\
&= \frac{n!\times\sin^{2n+1}(\theta)}{4}\int_{-\frac{\pi}{2}<\psi<\frac{\pi}{2}} \frac{(\sin\psi)^n}{(1-\cos\psi\cos\phi)^n}\frac{2}{1-\cos\psi\cos\theta}d\psi
\end{aligned}$$

Where, I get the last step using the identity $n! = \int_{t>0} e^{-t}t^n dt$. Now, substituting $n = 0$ I get:

$$J_n(\theta) = \frac{\sin\theta}{2}\int_{-\frac{\pi}{2}<\psi<\frac{\pi}{2}} \frac{1}{1-\cos\psi\cos\theta}d\psi$$

By breaking up the integral into the ranges $[-\frac{\pi}{2}, 0]$ and $(0, \frac{\pi}{2}]$ and doing a change in variables ($\psi = -\psi$) for the first range, we will get:

$$\begin{aligned} J_n(\theta) &= \frac{\sin\theta}{2}2\int_{0<\psi<\frac{\pi}{2}} \frac{1}{1-\cos\psi\cos\theta}d\psi \\ &= \sin\theta\int_{0<\psi<\frac{\pi}{2}} \frac{1}{1-\cos\psi\cos\theta}d\psi \end{aligned}$$

Now, using the identity given in the question to solve the integral. We have $a = 1$, $b = -\cos\theta$ and $\alpha = \frac{\pi}{2}$. Then, we get:

$$\begin{aligned} J_n(\theta) &= \sin\theta\frac{1}{\sqrt{1-\cos^2\theta}}\tan^{-1}(\frac{\sin\frac{\pi}{2}\sqrt{1-\cos^2\theta}}{-\cos\theta + 1\cos\frac{\pi}{2}}) \\ &= \sin\theta\frac{1}{\sin\theta}\tan^{-1}(\frac{\sin\theta}{-\cos\theta}) \\ &= \tan^{-1}(-\tan\theta) \\ &= \pi - \theta \end{aligned}$$

Problem 3

For this question, I will follow the proof used in the original Neural collapse paper (Ref. here). Theorem 1 in the paper proves exactly what is asked in this question. However, I will try to write out the proof in a more detailed manner to ensure (and communicate) complete understanding of the proof technique.

We are given that matrix $M \in \mathbb{R}^{p\times C}$ is the matrix where the centred class means $\mu_c - \mu_G \forall c \in [C]$ are the columns of the matrix. Now, I will first illustrate that NC1 and NC2 together imply NC3. To prove NC3, it will be helpful to characterize what the optimal linear classifier weights W looks like. Given the result in the question, we have :

$$W = \frac{1}{C}M^T\Sigma_T^{\dagger}$$

However, given NC1, we know that $\Sigma_W = 0$, the within class variance is zero. We also know that $\Sigma_T = \Sigma_W + \Sigma_B$. This can be easily verified from the definitions of the variance terms.

$$\begin{aligned}
\Sigma_T &= \frac{1}{NC}\sum_{c=1}^{C}\sum_{i=1}^{N}(h_{i,c}-\mu_G)(h_{i,c}-\mu_G)^T \\
&= \frac{1}{NC}\sum_{c=1}^{C}\sum_{i=1}^{N} h_{i,c}h_{i,c}^T + \mu_G\mu_G^T - 2\mu_G h_{i,c}^T \\
&= \frac{1}{NC}\sum_{c=1}^{C}\sum_{i=1}^{N} h_{i,c}h_{i,c}^T + \mu_G\mu_G^T - 2\mu_G h_{i,c}^T + 2\mu_c\mu_c^T - 2\mu_c\mu_c^T + 2\mu_c\mu_G^T - 2\mu_c\mu_c^T + 2\mu_c{}_{i,c}^T - 2\mu_c{}_{i,c}^T \\
&= \frac{1}{NC}\sum_{c=1}^{C}\sum_{i=1}^{N}(h_{i,c}h_{i,c}^T + \mu_c\mu_c^T - 2\mu_c h_{i,c}^T) + (\mu_c\mu_c^T + \mu_G\mu_G^T - 2\mu_c\mu_G^T) + \\
&2(\mu_c\mu_G^T - \mu_c\mu_c^T + \mu_c{}_{i,c}^T - \mu_G h_{i,c}^T) \\
&= \frac{1}{NC}\sum_{c=1}^{C}\sum_{i=1}^{N}(h_{i,c}-\mu_c)(h_{i,c}-\mu_c)^T + (\mu_c-\mu_G)(\mu_c-\mu_G)^T + 2(\mu_c\mu_G^T - \mu_c\mu_c^T + \mu_c h_{i,c}^T - \mu_G h_{i,c}^T) \\
&= \Sigma_W + \Sigma_B
\end{aligned}$$

Where, the last equality comes because $\sum_c \mu_c = \sum_c \mu_G$ by definition.

So, using this and $NC1$, we get $\Sigma_T = \Sigma_B$. Substituting this back into the equation for W, we get:

$$W = \frac{1}{C}M^T\Sigma_B^\dagger$$

Also, note that given the definition of matrix M and $\Sigma_B = \frac{1}{C}\sum_{c=1}^{C}\sum_{i=1}^{N}(\mu_c - \mu_G)(\mu_c - \mu_G)^T$, we can equivalently write $\Sigma_B = \frac{1}{C}MM^T$.

$$\begin{aligned}
W &= \frac{1}{C}M^T(\frac{1}{C}MM^T)^\dagger \\
&= M^T(M^T)^\dagger M^\dagger \\
&= M^\dagger
\end{aligned}$$

Now, all that remains to be done is to compute $M^\dagger$. By SVD we can write:

$$\begin{aligned}
M &= U\Sigma V^T \\
M^\dagger &= V\Sigma^{-1}U^T
\end{aligned}$$

However, given the structure of matrix M, it has $C-1$ non-zero singular values which are all equal. To see this:

$$\begin{aligned}\sigma(M) &= \sqrt{eig(M^TM)}\\ M^TM &= \frac{C}{C-1}I_{c\times c} - \frac{1}{C-1}\mathcal{I}_{C\times C}\end{aligned}$$

where, $I_{C\times C}$ is the identity matrix and $\mathcal{I}_{C\times C}$ is a matrix of all ones. We know that the eigenvalues of this matrix are just the difference between the eigenvalues of the 2 matrices. The identity matrix is full rank and has all C eigenvalues equal to $\frac{C}{C-1}$. The second matrix is rank 1 and has only one non-zero eigenvalue which is equal to $\frac{C}{C-1}$. Hence the difference of these 2 matrices has $C-1$non-zero eigenvalues which are all equal. This implies that M has $C-1$ non-zero equal singular values. This means that we can rewrite the inverse of the singular value matrix Σ as :

$$\Sigma^{-1} = \alpha\Sigma^T$$

Note, that his follows from the diagonal structure of the singular value matrix and the fact that all the non-zero singular values are equal. If they were different, this would not hold. So, we can then rewrite the pseudo-inverse in the following way:

$$\begin{aligned}M^\dagger &= V\Sigma^{-1}U^T\\ &= V\alpha\Sigma^TU^T\\ &= \alpha M^T\end{aligned}$$

Putting this together we get that:

$$W = \alpha M^T$$

This shows that the optimal linear classifier weight matrix W in the TPT phase is equal to M^T upto a constant factor, which is exactly what NC3 says.

From the above analysis we can also derive a term for the optimal bias term b. This is given by:

$$
\begin{aligned}
b &= \frac{1}{C}\nVdash_C - \frac{1}{C}M^T\Sigma_T^\dagger\mu_G \\
&= \frac{1}{C}\nVdash_C - \frac{1}{C}M^T\Sigma_B^\dagger\mu_G \\
&= \frac{1}{C}\nVdash_C - \frac{1}{C}M^T(\frac{1}{C}MM^T)^\dagger\mu_G \\
&= \frac{1}{C}\nVdash_C - M^\dagger\mu_G \\
&= \frac{1}{C}\nVdash_C - \alpha M^T\mu_G
\end{aligned}
$$

Now, to prove NC4. The classification rule for the network is given by:

$$
\begin{aligned}
&\arg\max_c \langle w_c, h\rangle + b_c \\
&= \arg\max_c \alpha\langle(\mu_c - \mu_G), h\rangle + (\frac{1}{C}\nVdash_C - \alpha M^T\mu_G)_c \\
&= \arg\max_c \alpha\langle(\mu_c - \mu_G), h\rangle + \frac{1}{C} - \alpha\langle(\mu_c - \mu_G), \mu_G\rangle \\
&= \arg\max_c \alpha\langle(\mu_c - \mu_G), (h - \mu_G)\rangle
\end{aligned}
$$

Now, the paper uses NC2 to say that the term $||\mu_c - \mu_G||^2$ is the same irrespective of the class c. This means that this term can be added to the maximization without changing what the maximizing class is, ie. it is essentially like adding a constant to the optimization problem. Given this as well as the fact that $||h - \mu_G||^2$ is independent of c, we get.

$$
\begin{aligned}
&\arg\max_c \langle(\mu_c - \mu_G), (h - \mu_G)\rangle \\
&= \arg\max_c 2\langle(\mu_c - \mu_G), (h - \mu_G)\rangle - ||\mu_c - \mu_G||^2 - ||h - \mu_G||^2 \\
&= \arg\min_c -2\langle(\mu_c - \mu_G), (h - \mu_G)\rangle + ||\mu_c - \mu_G||^2 + ||h - \mu_G||^2 \\
&= \arg\min_c ||(h - \mu_G) - (\mu_c - \mu_G)||^2 \\
&= \arg\min_c ||h - \mu_c||^2
\end{aligned}
$$

Which proves NC4.