Appendices

6.1 Proof of Theorem 4.1

We first decompose the population Hessian as:

$$\widehat{H}(\Theta) = H(\Theta^*) + (H(\Theta) - H(\Theta^*)) + \left(\widehat{H}(\Theta) - H(\Theta)\right). \tag{6.1.1}$$

Using Weyl's inequality, we have:

$$\begin{aligned} \lambda_{\min}(\widehat{H}(\Theta)) &\geq \lambda_{\min}(H(\Theta^*)) - \|H(\Theta) - H(\Theta^*)\|_2 - \|\widehat{H}(\Theta) - H(\Theta)\|_2, \\ \lambda_{\max}(\widehat{H}(\Theta)) &\leq \lambda_{\max}(H(\Theta^*)) + \|H(\Theta) - H(\Theta^*)\|_2 + \|\widehat{H}(\Theta) - H(\Theta)\|_2. \end{aligned} \tag{6.1.2}$$

Theorem now follows by combining Lemma 6.1, Lemma 6.4, and Lemma 6.5.

Below we present the above mentioned lemmas that are critical to the proof.

Lemma 6.1. *Consider setting of Theorem 4.1. Let $H(\Theta^*)$ be the Hessian of* (4.0.2)*. Then,*

$$C_1 I \preceq H(\Theta^*) \preceq C_2 I,$$

where C_1, C_2 are global constants.

Proof. The Hessian evaluated at W^*, V^*, a^*, b^* is positive semi-definite as it can be written as an outer product form $H(\Theta^*) = \mathbb{E}_{x\sim D} QQ^T$, where

$$Q = \begin{pmatrix} \phi(x_i^T w_1^* + b_1^*)x_1 \\ \vdots \\ \phi(x_i^T w_i^* + b_i^*) \\ \vdots \\ (x_1^T v_1^* + a_1^*)\phi'(x_1^T w_1^* + b_1^*)x_1 \\ \vdots \\ (x_K^T v_K^* + a_K^*)\phi'(x_K^T w_K^* + b_K^*)x_K \end{pmatrix} \tag{6.1.3}$$

To prove positive definiteness of $H(\Theta^*)$ we invoke the Schur-Product theorem on Hadamard product of matrices(see Theorem 5.2.1 in Horn and Johnson (1991)).

Theorem 6.2 (Theorem 5.2.1 of Horn and Johnson (1991))**.** *If A and B are positive semi-definite, then so is $A \circ B$. If, in addition, B is positive definite and A has no diagonal entry equal to 0, then $A \circ B$ is positive definite.*

We also need Theorem 5.3.4 Horn and Johnson (1991) to obtain a lower bound of eigenvalues of Hadamard product of matrices.

Theorem 6.3 (Theorem 5.3.4 of Horn and Johnson (1991))**.** *Let A and B be two $n \times n$ positive definite matrices. It follows that any eigenvalue of the Hadamard product $A \circ B$ satisfies:*

$$\lambda(A \circ B) \geq [\min_{i\in[n]} a_{ii}]\lambda_{min}(B)$$

We first note that for two random vectors U and Z the Hadamard product $\mathbb{E}[UU^T \circ ZZ^T]$ can be lower bounded with respect to the PSD cone as follows:

$$\mathbb{E}[UU^T \circ ZZ^T] \succeq \mathbb{E}[UU^T] \circ \mathbb{E}[Z]\mathbb{E}[Z^T]$$

To show this we note that:

$$\mathbb{E}[(UU^T - \mathbb{E}[U]\mathbb{E}[U^T]) \circ (ZZ^T - \mathbb{E}[Z]\mathbb{E}[Z^T])] \succeq 0$$

This means that:

$$\mathbb{E}[UU^T \circ ZZ^T] \succeq \mathbb{E}[UU^T] \circ \mathbb{E}[Z]\mathbb{E}[Z^T] + \mathbb{E}[U]\mathbb{E}[U^T] \circ (\mathbb{E}[ZZ^T] - \mathbb{E}[Z]\mathbb{E}[Z^T])] \succeq \mathbb{E}[UU^T] \circ \mathbb{E}[Z]\mathbb{E}[Z^T]$$

where the last inequality follows because the second term in the 2nd expression is positive semi-definite being a Hadamard product of two PSD matrices.

Consequently, it follows from the Schur Product theorem that if $\mathbb{E}[UU^T] \succ 0$, then $\mathbb{E}[UU^T \circ ZZ^T] \succ 0$. Furthermore, we can lower bound the eigenvalues of $\mathbb{E}[UU^T \circ ZZ^T]$ by considering the smallest eigenvalues of $\mathbb{E}[UU^T] \circ \mathbb{E}[Z]\mathbb{E}[Z^T]$.

We let $b_i = a_i = 1$ for simplicity. To invoke Theorem 6.2 we set:

$$U = \begin{pmatrix} x_1 \\ \vdots \\ x_K \\ (x_1^T v_1^* + 1)x_1 \\ \vdots \\ (x_K^T v_K^* + 1)x_K \end{pmatrix} \quad Z = \begin{pmatrix} \phi(x_1^T w_1^* + 1) \\ \vdots \\ \phi(x_K^T w_K^* + 1) \\ \phi'(x_1^T w_1^* + 1) \\ \vdots \\ \phi'(x_K^T w_K^* + 1) \end{pmatrix} \tag{6.1.4}$$

and note that $H(\Theta^*) = \mathbb{E}[UU^T \circ ZZ^T]$. Consequently, we are left to establish that $\mathbb{E}[UU^T]$ is definite. Furthermore, from Theorem 6.3, we know that so long as $\mathbb{E}[Z_i^2]$ is bounded away from zero, we can also characterize the smallest eigenvalue of $H(\Theta^*)$. In particular,

$$\lambda_{\min}(H(\Theta^*)) \geq \min\left\{\min_i \mathbb{E}[\phi_i(x_i^T w_i^* + 1)]^2, \min_i \mathbb{E}[\phi_i'(x_i^T w_i^* + 1)]^2\right\} \lambda_{\min}(\mathbb{E}[UU^T]) \tag{6.1.5}$$

Note that $\min_i \mathbb{E}[\phi_i(x_i^T w_i^* + 1)]^2, \min_i \mathbb{E}[\phi_i'(x_i^T w_i^* + 1)]^2$ is a constant dependent only on ϕ and $\|w_i^*\|$.

Now, we will show that $\lambda_{\min}(\mathbb{E}[UU^T]) \geq 1/\sqrt{3}$. To prove this result we first express the minimim eigenvalue as follows:

$$\lambda_{min}(\mathbb{E}[UU^T]) = \min_{\Omega} \mathbb{E}_{x\sim D}\left[(c^T, e^T)U\right]^2 \tag{6.1.6}$$

$$= \min_{\Omega} \mathbb{E}_{x\sim D}\Big(\sum_{k=1}^K x_k^T c_k + (x_k^T v_k^* + 1)x_k^T e_k\Big)^2 \tag{6.1.7}$$

$$= \min_{\Omega} \mathbb{E}_{x\sim D}\Big(\sum_{k=1}^K x_k^T (c_k + e_k) + (x_k^T v_k^*)x_k^T e_k\Big)^2 \tag{6.1.8}$$

$$= \min_{\Omega} \mathbb{E}_{x\sim D}\Big(\sum_{k=1}^K x_k^T (c_k + e_k)\Big)^2 + \mathbb{E}_{x\sim D}\Big(\sum_{k=1}^K (x_k^T v_k^*)x_k^T e_k\Big)^2 \tag{6.1.9}$$

$$+ 2\mathbb{E}_{x\sim D}\Big(\sum_{k=1}^K x_k^T (c_k + e_k)(x_k^T v_k^*)x_k^T e_k\Big) \tag{6.1.10}$$

$$= \min_{\Omega} \mathbb{E}_{x\sim D}\Big(\sum_{k=1}^K x_k^T (c_k + e_k)\Big)^2 + \mathbb{E}_{x\sim D}\Big(\sum_{k=1}^K (x_k^T v_k^*)x_k^T e_k\Big)^2 \tag{6.1.11}$$

where Ω denotes the set $\{c, e : \sum_{k=1}^K \|c_k\|^2 + \|e_k\|^2 = 1\}$. The last equality follows by Isserlis' theorem for zero-mean Gaussian random variables, which states that odd-order moments for Gaussian random variables are zero. In particular Isserlis' Theorem asserts that, for Jointly Gaussian random variables $X_1, X_2, X_3, \ldots, X_{2n}$:

$$\mathbb{E}[\, X_1 X_2 \cdots X_{2n}\,] = \sum \prod \mathbb{E}[\, X_i X_j\,] = \sum \prod \mathbf{Cov}(\, X_i, X_j\,), \tag{6.1.12}$$

$$\mathbb{E}[\, X_1 X_2 \cdots X_{2n-1}\,] = 0, \tag{6.1.13}$$

where the notation $\sum\prod$ means summing over all distinct ways of partitioning $X_1, \ldots, X_{2n}$ into pairs X_i, X_j and each summand is the product of the n pairs

We substitute $d_k = c_k + e_k$ and $e_k = d_k - c_k$ and obtain:

$$\mathbb{E}_{x\sim D}\Big(\sum_{k=1}^{K} x_k^T(d_k)\Big)^2 + \mathbb{E}_{x\sim D}\Big(\sum_{k=1}^{K}(x_k^T v_k^*)x_k^T(d_k - c_k)\Big)^2 \geq \|d\|_2^2\sigma_{\min}^2 + \mathbb{E}_{x\sim D}\Big(\sum_{k=1}^{K}(x_k^T v_k^*)x_k^T(d_k - c_k)\Big)^2, \quad (6.1.14)$$

where d, c are such that $\sum_{k=1}^{K}\|d_k - c_k\|^2 + \|c_k\|^2 = 1$.

Denote $f_k = c_k - d_k$ for notational simplicity. Expanding we obtain:

$$\mathbb{E}_{x\sim D}\Big(\sum_{k=1}^{K}(x_k^T v_k^*)x_k^T(f_k)\Big)^2 = \sum_{k=1}^{K}\mathbb{E}_{x\sim D}((v_k^*)^T x_k)^2((f_k)^T x_k)^2 + \sum_{j\neq k}((v_k^*)^T(f_k))^2((v_j^*)^T(f_j))^2. \quad (6.1.15)$$

Using independence of x_j, x_k for $j \neq k$, by using Isserlis' and Stein's Theorem, and by using Cauchy-Schwartz inequality, we have:

$$\mathbb{E}_{x\sim D}\Big(\sum_{k=1}^{K}(x_k^T v_k^*)x_k^T(f_k)\Big)^2 = 2\sum_{k=1}^{K}\|v_k^*\|^2\|f_k\|^2 + \left(\sum_k (v_k^*)^T f_k\right)^2. \quad (6.1.16)$$

Using $\|v_k^*\| \geq 1, \forall k$, we have:

$$R = \mathbb{E}_{x\sim D}\Big(\sum_{k=1}^{K} x_k^T(d_k)\Big)^2 + \mathbb{E}_{x\sim D}\Big(\sum_{k=1}^{K}(x_k^T v_k^*)x_k^T(f_k)\Big)^2 \geq \sum_{k=1}^{K}\|d_k\|^2 + 2\sum_{k=1}^{K}\|d_k - c_k\|^2$$
$$\geq 2 + \sum_{k=1}^{K}\|d_k\|^2 - 2\sum_{k=1}^{K}\|c_k\|^2, \quad (6.1.17)$$

where second inequality follows from $\sum_{k=1}^{K}\|d_k - c_k\|^2 + \|c_k\|^2 = 1$. Using triangle inequality, we have:

$$\sqrt{\sum_{k=1}^{K}\|d_k\|^2} \geq \left|\sqrt{1 - \sum_{k=1}^{K}\|c_k\|^2} - \sqrt{\sum_{k=1}^{K}\|c_k\|^2}\right|.$$

Combining the above observation with (6.1.17) and denoting $\alpha = \sqrt{\sum_{k=1}^{K}\|c_k\|^2} \leq 1$, we have:

$$R \geq 2 + (\sqrt{1-\alpha^2} - \alpha)^2 - 2\alpha^2 = 3 - 2\alpha^2 - 2\alpha\sqrt{1-\alpha^2}$$
$$= 2 - 2\alpha^2 + (\sqrt{1-\alpha^2} - \alpha)^2 \geq 1/\sqrt{3}, \quad (6.1.18)$$

where the last inequality follows by considering two cases $|\alpha| \leq \sqrt{2/3}$ and $\alpha \geq \sqrt{2/3}$.

Lower bound on Hessian's eigenvalues now follows by combining 6.1.5, 6.1.11, and (6.1.18). A similar argument leads to the desired upper bound on Hessian's eigenvalue $\lambda_{\max}(H(\Theta^*))$. □

Lemma 6.4 (Smoothness of Hessian near optimum)**.** *Consider setting of Theorem 4.1. Then the following holds:*

$$\|H(\Theta) - H(\Theta^*)\|_2 \leq \sqrt{K}\cdot C_1(\|V - V^*\|_F + \|W - W^*\|_F),$$

where C_1 is a global constant.

Proof. Let $\Delta = \nabla^2 L(W,V) - \nabla^2 L(W^*,V^*)$ be the difference in Hessian. We study the different blocks of Δ separately. Consider:

$$\Delta_{v_i,v_j} = \mathbb{E}_{x\sim D}\left[\left(\phi_i\phi_j - \phi_i^*\phi_j^*\right)x_i x_j^T\right],$$

where we use the short hand notation $\phi_i = \phi(x_i^T w_i + 1)$ and $\phi_i^* = \phi(x_i^T w_i^* + 1)$.

We can bound the norm as follows,

$$\begin{aligned}
\|\Delta_{v_i,v_j}\| &= \|\mathbb{E}_{x\sim D}\left[\left(\phi_i\phi_j - \phi_i^*\phi_j^*\right)x_i x_j^T\right]\| \\
&\le \max_{\|\alpha\|=1,\|\beta\|=1} \mathbb{E}_{x\sim D}\left[|\phi_i\phi_j - \phi_i^*\phi_j^*||x_i^T\alpha||x_j^T\beta|\right] \\
&\le \max_{\|\alpha\|=1,\|\beta\|=1} \mathbb{E}_{x\sim D}\left[|\phi_i\phi_j - \phi_i^*\phi_j^*||x_i^T\alpha||x_j^T\beta|\right] \\
&\le \max_{\|\alpha\|=1,\|\beta\|=1} \mathbb{E}_{x\sim D}\left[\left(|\phi_i - \phi_i^*||\phi_j| + |\phi_i^*||\phi_j - \phi_j^*|\right)|x_i^T\alpha||x_j^T\beta|\right] \\
&\le \max_{\|\alpha\|=1,\|\beta\|=1} \mathbb{E}_{x\sim D}\Big[\Big(L_0|x_i^T(w_i - w_i^*)|B|x_j^T w_j|^q \\
&\qquad\qquad + B|x_i^T w_i^*|^q L_0|x_j^T(w_j - w_j^*)|\Big)|x_i^T\alpha||x_j^T\beta|\Big] \\
&\le \max_{\|\alpha\|=1,\|\beta\|=1} L_0 B\left(\|w_i - w_i^*\|\|w_j\|^q + \|w_j - w_j^*\|\|w_i^*\|^q\right)\|\alpha\|\|\beta\|. \\
&= L_0 B\left(\|w_i - w_i^*\|\|w_j\|^q + \|w_j - w_j^*\|\|w_i^*\|^q\right), \qquad (6.1.19)
\end{aligned}$$

where we used triangle inequality and assumption on ϕ given in the Lemma statement.

Next consider Δ_{w_i,w_j} for $i \neq j$:

$$\Delta_{w_i,w_j} = \mathbb{E}_{x\sim D}\left[\left(x_i^T v_i x_j^T v_j \phi_i'\phi_j' - x_i^T v_i^* x_j^T v_j^* \phi_i'^*\phi_j'^*\right)x_i x_j^T\right],$$

Similarly, we can bound the norm of the above quantity as follows,

$$\begin{aligned}
\|\Delta_{w_i,w_j}\| &= \|\mathbb{E}_{x\sim D}\left[\left(x_i^T v_i x_j^T v_j \phi_i'\phi_j' - x_i^T v_i^* x_j^T v_j^* \phi_i'^*\phi_j'^*\right)x_i x_j^T\right]\| \\
&= \max_{\|\alpha\|=1,\|\beta\|=1} \mathbb{E}_{x\sim D}\left[|x_i^T v_i x_j^T v_j \phi_i'\phi_j' - x_i^T v_i^* x_j^T v_j^* \phi_i'^*\phi_j'^*||x_i^T\alpha||x_j^T\beta|\right] \\
&\le \max_{\|\alpha\|=1,\|\beta\|=1} \mathbb{E}_{x\sim D}\Big[\Big(|x_i^T v_i \phi_i'(\phi_j' x_j^T v_j - \phi_j'^* x_j^T v_j^*)| \\
&\qquad\qquad + |\phi_j'^* x_j^T v_j^*(x_i^T v_i\phi_i' - x_i^T v_i^*\phi_i'^*)|\Big)|x_i^T\alpha||x_j^T\beta|\Big] \\
&\le \max_{\|\alpha\|=1,\|\beta\|=1} \mathbb{E}_{x\sim D}\Big[\Big(|x_i^T v_i||\phi_i'|\left(|\phi_j' - \phi_j'^*||x_j^T v_j| + |\phi_j'^*||x_j^T(v_j - v_j^*)|\right) \\
&+ |\phi_j'^*||x_j^T v_j^*|\left(|\phi_i' - \phi_i'^*||x_i^T v_i| + |\phi_i'^*||x_i^T(v_i - v_i^*)|\right)\Big)|x_i^T\alpha||x_j^T\beta|\Big] \\
&\le \max_{\|\alpha\|=1,\|\beta\|=1} \mathbb{E}_{x\sim D}\Big[\Big(|x_i^T v_i|L_1|x_i^T w_i|^p\left(L_2|x_j^T(w_j - w_j^*)||x_j^T v_j|\right. \\
&\quad \left. + L_1|x_j^T w_j^*|^p|x_j^T(v_j - v_j^*)|\right) + L_1|x_j^T w_j^*|^p|x_j^T v_j^*|\left(L_2|x_i^T(w_i - w_i^*)||x_i^T v_i|\right. \\
&\qquad\qquad \left. + L_1|x_i^T w_i^*|^p|x_i^T(v_i - v_i^*)|\right)\Big)|x_i^T\alpha||x_j^T\beta|\Big] \\
&\le \|v_i\|L_1\|w_j^*\|^p\left(L_2\|w_j - w_j^*\|\|v_j\| + L_1\|w_j^*\|^p\|v_j - v_j^*\|\right) \\
&+ L_1\|w_j^*\|^p\|v_j^*\|\left(L_2\|w_i - w_i^*\|\|v_i\| + L_1\|w_i^*\|^p\|v_i - v_i^*\|\right). \qquad (6.1.20)
\end{aligned}$$

When $i = j$, consider

$$\Delta_{w_i,w_i} = \mathbb{E}_{x\sim D}\left[\left(\sum_{k=1}^K x_k^T v_k\phi_k - y\right)x_i^T v_i\phi_i'' x_i x_i^T\right],$$

We can bound the norm as follows,

$$\begin{aligned}
\|\Delta_{w_i,w_i}\| &= \max_{\|\alpha\|=1} \mathbb{E}_{x\sim D}\left[|\sum_{k=1}^{K} x_k^T v_k\phi_k - x_k^T v_k^*\phi_k^*||x_i^T v_i||\phi_i''|(x_i^T\alpha)^2\right] \\
&\leq \max_{\|\alpha\|=1} \mathbb{E}_{x\sim D}\left[\left(\sum_{k=1}^{K}|\phi_k||x_k^T(v_k-v_k^*)| + |\phi_k-\phi_k^*||x_k^T v_k^*|\right)|x_i^T v_i||\phi_i''|(x_i^T\alpha)^2\right] \\
&\leq \max_{\|\alpha\|=1} \mathbb{E}_{x\sim D}\Bigg[\Bigg(\sum_{k=1}^{K} B|x_k^T w_k|^q\|x_k^T(v_k-v_k^*)\| \\
&\qquad + L_0|x_k^T(w_k-w_k^*)||x_k^T v_k^*|\Bigg)|x_i^T v_i|L_2(x_i^T\alpha)^2\Bigg] \\
&\leq \left(\sum_{k=1}^{K} B\|w_k\|^q\|v_k-v_k^*\| + L_0\|w_k-w_k^*\|\|v_k^*\|\right)\|v_i\|L_2. \qquad (6.1.21)
\end{aligned}$$

Next we consider:

$$\Delta_{w_i,v_j} = \mathbb{E}_{x\sim D}\left[\left(x_i^T v_i\phi_i'\phi_j - x_i^T v_i^*\phi_i'^*\phi_j^*\right)x_i x_j^T\right],$$

where $i\neq j$. The norm of the above quantity can be bound as follows,

$$\begin{aligned}
\|\Delta_{w_i,v_j}\| &= \max_{\|\alpha\|=1,\|\beta\|=1} \mathbb{E}_{x\sim D}\left[|x_i^T v_i\phi_i'(\phi_j-\phi_j^*) + \phi_j^*\left(x_i^T(v_i-v_i^*)\phi_i' + x_i^T v_i^*(\phi_i'-\phi_i'^*)\right)||x_i^T\alpha||x_j^T\beta|\right] \\
&\leq \max_{\|\alpha\|=1,\|\beta\|=1} \mathbb{E}_{x\sim D}\Big[\Big(|x_i^T v_i|L_1|x_i^T w_i|^p L_0|x_j^T(w_j-w_j^*)| \\
&+ B|x_j^T w_j^*|^q\left(|x_i^T(v_i-v_i^*)L_1|x_i^T w_i|^p + |x_i^T v_i^*|L_2|x_i^T(w_i-w_i^*)\right)\Big)|x_i^T\alpha||x_j^T\beta|\Big] \\
&\leq L_0L_1\|v_i\|\|w_i\|^p\|w_j-w_j^*\| + B\|w_j^*\|^q\left(L_1\|w_i\|^p\|v_i-v_i^*\| + L_2\|v_i^*\|\|w_i-w_i^*\|\right). \qquad (6.1.22)
\end{aligned}$$

Now, consider:

$$\Delta_{w_i,v_i} = \mathbb{E}_{x\sim D}\left[\left(\sum_{k=1}^{K} x_k^T v_k\phi_k - y\right)\phi_i' x_i x_i^T\right],$$

We can bound the norm of the above quantity as follows,

$$\begin{aligned}
\|\Delta_{w_i,v_i}\| &= \max_{\|\alpha\|=1} \mathbb{E}_{x\sim D}\left[|\sum_{k=1}^{K} x_k^T v_k\phi_k - x_k^T v_k^*\phi_k^*||\phi_i'|(x_i^T\alpha)^2\right] \\
&= \max_{\|\alpha\|=1} \mathbb{E}_{x\sim D}\left[|\sum_{k=1}^{K} x_k^T(v_k-v_k^*)\phi_k + x_k^T v_k^*(\phi_k-\phi_k^*)||\phi_i'|(x_i^T\alpha)^2\right] \\
&\leq \max_{\|\alpha\|=1} \mathbb{E}_{x\sim D}\left[\sum_{k=1}^{K}\left(|x_k^T(v_k-v_k^*)|B|x_k^T w_k|^q + |x_k^T v_k^*|L_0|x_k^T(w_k-w_k^*)|\right)L_1|x_i^T w_i|^p(x_i^T\alpha)^2\right] \\
&\leq \left(\sum_{k=1}^{K} B\|w_k\|^q\|v_k-v_k^*\| + L_0\|v_k^*\|\|w_k-w_k^*\|\right)L_1\|w_i\|^p. \qquad (6.1.23)
\end{aligned}$$

Lemma now follows by using Gershgorin's theorem along with (6.1.19), (6.1.20), (6.1.21), (6.1.22), and (6.1.23), and the assumption that $\|w_i\|\leq 2$ and $\|v_i\|\leq 2$.

That is,

$$\begin{aligned}
\|\Delta\| &\leq \max\{\max_i\|\Delta_{w_i,w_i}\| + \sum_{j\neq i}\|\Delta_{w_i,w_j}\| + \|\Delta_{w_i,v_i}\| + \sum_{j\neq i}\|\Delta_{w_i,v_j}\|, \qquad (6.1.24)\\
&\qquad \max_i\|\Delta_{v_i,v_i}\| + \sum_{j\neq i}\|\Delta_{v_i,v_j}\| + \|\Delta_{w_i,v_i}\| + \sum_{j\neq i}\|\Delta_{w_i,v_j}\|\} \qquad (6.1.25)\\
&\leq \sqrt{K}C_1(\|W-W^*\|_F + \|V-V^*\|_F), \qquad (6.1.26)
\end{aligned}$$

where $C_1>0$ is a global constant. □

Lemma 6.5 (Concentration of Hessian)**.** *Consider setting of Theorem 4.1. Then, the empirical Hessian of* (4.0.2) *satisfies the following (w.p.* $\geq 1 - 10K^2/n^{10}$*):*

$$\|\widehat{H}(\Theta) - H(\Theta)\|_2 \leq K \cdot C\frac{(\log n)^C \cdot d}{\sqrt{n}},$$

where $C > 0$ *is a global constant.*

Proof. Let $\Delta = H(\Theta) - \frac{1}{n}\sum_{s=1}^{n} H^s(\Theta)$, where $H^s(\Theta)$ is the Hessian of s-th data point. Then Δ can be written as a block matrix with various blocks Δ_{v_i,v_i}, Δ_{v_i,v_j}, Δ_{w_i,w_i}, Δ_{w_i,w_j}, and Δ_{w_i,v_i} and Δ_{w_i,v_j} for all i, j s.t. $i \neq j$.

We first consider Δ_{v_i,v_j} for all i, j, which is given by:

$$\Delta_{v_i,v_j} = \frac{1}{n}\sum_{s=1}^{n} \phi_i^s \phi_j^s (x_i^s)(x_i^s)^T - \mathbb{E}_{x\sim D}[\phi_i \phi_j x_i x_j^T], \tag{6.1.27}$$

where we use the shorthand notation $\phi_i^s = \phi(w_i^T x_i^s + 1)$, $\phi_i = \phi(w_i^T x_i + 1)$. Since w_i is fixed wrt data points, with probability $\geq 1 - K/n^{100}$, $\mathbf{w}_i^T \mathbf{x}_i^s \leq C \log n, \forall s, i$ and some constant $C > 0$. Similarly, $\|x_i^s\|^2 \leq d + C\sqrt{d \log n}$ with probability at least $1 - 1/n^{100}$ and a constant $C > 0$. Now, using the requirement $\phi(z)$ along with standard Matrix Chernoff bound, we have (w.p. $\geq 1 - 1/n^{100}$):

$$\|\Delta_{v_i,v_i}\|_2 \leq \frac{B^2(\log n)^{q+1} \cdot d}{\sqrt{n}}.$$

Using a similar argument for all remaining blocks, we get (w.p. $\geq 1 - 10K^2/n^{100}$):

$$\|\Delta\|_2 \leq K \cdot C\frac{(\log n)^C \cdot d}{\sqrt{n}},$$

where $C > 0$ is a global constants. □

6.2 Proof of Lemma 1

We briefly sketch the proof here. It follows along similar lines as Rago et al. (1996); Appadwedula et al. (2008). For simplicity of exposition we consider two devices, a and b and binary classification. Let $g_a(x)$, $g_b(x)$ be gating functions that take binary values corresponding to whether or not devices transmit. Assume that both classes are equiprobable. Let x_a, x_b denote device feature realizations corresponding to the two devices for an instance $x = [x_a, x_b]$ and $p_a(x_a), q_a(x_a)$ the class-conditional likelihoods for device a under the two classes. Similarly, $p_b(x_b), q_b(x_b)$ for device b. Due to conditional independence, the joint class conditional probability is the product of the class-conditional marginals.

We let $\psi_{ab}(x_a, x_b), \psi_a(x_a), \psi_b(x_b), \psi_0$ denote the binary classification rule when both devices transmit, only device a transmits, only device b transmits, and no device transmits respectively. They take a binary value with zero denoting the first class and one denoting the second class. Note that these cases are mutually exclusive. For this reason, the error probability, P_e can be written sum of four terms and the fusion rules can be derived in a straightforward manner. Error occurs when the fusion rule classifies as class one while the true class is zero and

vice versa. So we have,

$$\begin{aligned}
P_e = &\int\int g_a(x_a)g_b(x_b)\psi_{ab}(x_a,x_b)p_a(x_a)p_b(x_b)dx_a dx_b + \int\int g_a(x_a)g_b(x_b)(1-\psi_{ab}(x))q_a(x_a)q_b(x_b)dx_a dx_b \\
&+ \int\int g_a(x_a)(1-g_b(x_b))\psi_a(x_a)p_a(x_a)p_b(x_b)dx_a dx_b \\
&+ \int\int g_a(x_a)(1-g_b(x_b))(1-\psi_a(x_a))q_a(x_a)q_b(x_b)dx_a dx_b \\
&+ \int\int (1-g_a(x_a))g_b(x_b))\psi_b(x_b)p_a(x_a)p_b(x_b)dx_a dx_b \\
&+ \int\int (1-g_a(x_a))g_b(x_b)(1-\psi_b(x_b))q_a(x_a)q_b(x_b)dx_a dx_b \\
&+ \psi_0 \int\int (1-g_a(x_a))(1-g_b(x_b))p_a(x_a)p_b(x_b)dx_a dx_b \\
&+ (1-\psi_0)\int\int (1-g_a(x_a))(1-g_b(x_b))q_a(x_a)q_b(x_b)dx_a dx_b
\end{aligned}$$

Let us first consider the last two terms corresponding to both devices being inactive.

Case 0: Clearly, the optimal fusion rule for given gatings, g_a, g_b is to select the minimum of the two terms, namely,

$$\int\int (1-g_a(x_a))(1-g_b(x_b))p_a(x_a)p_b(x_b)dx_a dx_b \underset{\psi_0=1}{\overset{\psi_0=0}{\gtrless}} \int\int (1-g_a(x_a))(1-g_b(x_b))q_a(x_a)q_b(x_b)dx_a dx_b$$

This can be further simplified notice the conditional independence of the two features. Consequently, we obtain:

$$\log\frac{\int(1-g_a(x_a))p_a(x_a)dx_a}{\int(1-g_a(x_a))q_a(x_a)dx_a} + \log\frac{\int(1-g_b(x_b))p_b(x_b)dx_b}{\int(1-g_b(x_b))q_b(x_b)dx_b} \underset{\psi_0=1}{\overset{\psi_0=0}{\gtrless}} 0 \tag{6.2.1}$$

In a similar fashion we can see that the optimal fusion rule in other cases are:

Case 1: $g_a(x_a) = g_b(x_b) = 1$

$$p_a(x_a)p_b(x_b) \underset{\psi_{ab}=1}{\overset{\psi_{ab}=0}{\gtrless}} q_a(x_a)q_b(x_b)$$

which, is precisely the likelihood ratio test, namely,

$$\ell_a(x_a) + \ell_b(x_b) \underset{\psi_{ab}=1}{\overset{\psi_{ab}=0}{\gtrless}} 0$$

where, $\ell_a(x_a) = \log\frac{p_a(x_a)}{q_a(x_a)}$, $\ell_b(x_b) = \log\frac{p_b(x_b)}{q_b(x_b)}$

Case 2: $g_a(x_a) = 1 - g_b(x_b) = 1$ where device 2 is inactive. In this case the central processor receives measurement x_a:

$$p_a(x_a)\int(1-g_b(x_b))p_b(x_b)dx_b \underset{\psi_a=1}{\overset{\psi_a=0}{\gtrless}} q_a(x_a)\int(1-g_b(x_b))q_b(x_b)dx_b$$

Again, this is just a likelihood ratio test, namely,

$$\ell_a(x_a) - \log\frac{\int(1-g_b(x_b))p_b(x_b)dx_b}{\int(1-g_b(x_b))q_b(x_b)dx_b} \underset{\psi_a=1}{\overset{\psi_a=0}{\gtrless}} 0$$

Case 3: $g_b(x_b) = 1 - g_a(x_a) = 1$ where device 1 is now inactive. In this case the central processor receives measurement x_b. This case is similar to case 2.

Note that we can combine the four cases and write the fusion rule compactly as follows:

$$f(x_a, x_b) = \ell_a(x_a)g_a(x_a) + \ell_b(x_b)g_b(x_b) + \eta_a(1 - g_a(x_a)) + \eta_b(1 - g_b(x_b)) \underset{1}{\overset{0}{\gtrless}} 0$$

where η_a, η_b are the first and second terms in Eq. 6.2.1 and are constants independent of x_a and x_b. The reason follows from linearity of the likelihood ratio test under conditional independence.

Consequently, we are left to show that $g_a(x_a)$ and $g_b(x_b)$ are also functions of the local likelihood ratios. To do this we again examine our objective function 6.1.2. It is composed of the error probability plus a linear activation penalty $(g_a + g_b)$. For a fixed gating function g_b, the fusion rules described above we see that, for a fixed x_a, the expressions are a affine-linear function in $g_a(x_a)p_a(x_a)$, $g_a(x_a)q_a(x_a)$ and $\lambda g_a(x_a)(p_a(x_a) + q_a(x_a)$.

In a similar fashion as before the problem can be analyzed in an analogous manner to the cases already considered above. The optimal gating function can therefore be expressed in terms of the likelihood ratios.