A ADDITIONAL PROOFS

Proof of Lemma 8. We denote $h^{(t)} := h_{\mathbf{u}^{(t)},W^{(t)},b}$. For every j we have:

$$\frac{\partial}{\partial \mathbf{u}^{(j)}} L_{f,\mathcal{D}}(h^{(t)}) = \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}}\left[\frac{\partial}{\partial \mathbf{u}^{(j)}}\ell(h^{(t)}(\mathbf{x}), f(\mathbf{x}))\right] = \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}}\left[\sigma(W^{(t)}x_{j\dots j+k-1})\ell'(h^{(t)}(\mathbf{x}), f(\mathbf{x}))\right]$$

Therefore:

$$\left\|\frac{\partial}{\partial \mathbf{u}^{(j)}} L_{f,\mathcal{D}}(h^{(t)})\right\| \leq \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}}\left[\left\|\sigma(W^{(t)}x_{j\dots j+k-1})\right\|\right] \leq c\sqrt{q}$$

And from the updates of gradient-descent we have:

$$\left\|\mathbf{u}^{(t,j)}\right\| = \left\|\eta\sum_{t=1}^{T}\frac{\partial}{\partial \mathbf{u}^{(j)}} L_{f,\mathcal{D}}(h^{(t)})\right\| \leq \eta\sum_{t=1}^{T}\left\|\frac{\partial}{\partial \mathbf{u}^{(j)}} L_{f,\mathcal{D}}(h^{(t)})\right\| \leq c\eta T\sqrt{q}$$

Now, we have that:

$$\frac{\partial}{\partial W} L_{f,\mathcal{D}}(h^{(t)}) = \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}}\left[\sum_{j=1}^{n-k}\mathbf{u}^{(t,j)}\mathbf{x}_{j\dots j+k-1}^{\top}\sigma'(W^{(t)}\mathbf{x}_{j\dots j+k-1}+b)\ell'(h^{(t)}(\mathbf{x}), f(\mathbf{x}))\right]$$

And so:

$$\left\|\frac{\partial}{\partial W} L_{f,\mathcal{D}}(h^{(t)})\right\| \leq \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}}\left[\sum_{j=1}^{n-k}\left\|\mathbf{u}^{(t,j)}\right\|\|\mathbf{x}_{j\dots j+k-1}\|\right] \leq c(n-k)\eta T\sqrt{q}\sqrt{k}$$

Again, by the updates of gradient-descent:

$$\left\|W^{(T)}-W^{(0)}\right\| = \left\|\eta\sum_{t=1}^{T}\frac{\partial}{\partial W} L_{f,\mathcal{D}}(h^{(t)})\right\| \leq \eta\sum_{t=1}^{T}\left\|\frac{\partial}{\partial W} L_{f,\mathcal{D}}(h^{(t)})\right\| \leq c\eta^2T^2n\sqrt{qk}$$

□

Proof of Lemma 9.

$$\begin{aligned}
&\left|L_{f,\mathcal{D}}(h_{\mathbf{u}^*,W^{(T)},b}) - L_{f,\mathcal{D}}(h_{\mathbf{u}^*,W^{(0)},b})\right| \\
&= \left|\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}}\left[\ell(h_{\mathbf{u}^*,W^{(T)},b}(\mathbf{x}), f(\mathbf{x}))\right] - \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}}\left[\ell(h_{\mathbf{u}^*,W^{(0)},b}(\mathbf{x}), f(\mathbf{x}))\right]\right| \\
&\leq \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}}\left[\left|\ell(h_{\mathbf{u}^*,W^{(T)},b}(\mathbf{x}), f(\mathbf{x})) - \ell(h_{\mathbf{u}^*,W^{(0)},b}(\mathbf{x}), f(\mathbf{x}))\right|\right] \\
&\leq \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}}\left[\left|h_{\mathbf{u}^*,W^{(T)},b}(\mathbf{x}) - h_{\mathbf{u}^*,W^{(0)},b}(\mathbf{x})\right|\right] \\
&= \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}}\left[\left|\sum_{j=1}^{n-k}\left\langle\mathbf{u}^{*(j)}, \sigma(W^{(T)}\mathbf{x}_{j\dots j+k-1}+\mathbf{b}) - \sigma(W^{(0)}\mathbf{x}_{j\dots j+k-1}+\mathbf{b})\right\rangle\right|\right] \\
&\leq \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}}\left[\sum_{j=1}^{n-k}\left\|\mathbf{u}^{*(j)}\right\|\left\|\sigma(W^{(T)}\mathbf{x}_{j\dots j+k-1}+\mathbf{b}) - \sigma(W^{(0)}\mathbf{x}_{j\dots j+k-1}+\mathbf{b})\right\|\right] \\
&\leq \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}}\left[\sum_{j=1}^{n-k}\left\|\mathbf{u}^{*(j)}\right\|\left\|W^{(T)}-W^{(0)}\right\|\|\mathbf{x}_{j\dots j+k-1}\|\right] \leq c\eta^2T^2nk\sqrt{q}\sum_{j=1}^{n-k}\left\|\mathbf{u}^{*(j)}\right\|
\end{aligned}$$

□

Proof. (second part of Theorem 12)

For some π, observe that:

$$\begin{aligned}\frac{\partial}{\partial u_i} L_{\chi_{\pi(I)},\mathcal{D}}(h_{\mathbf{u},\mathbf{w},\mathbf{b}}) &= -\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}}\left[\sigma\left(\left\langle \mathbf{w}^{(i)},\mathbf{x}\right\rangle + b_i\right)\chi_I(\pi(\mathbf{x}))\right] \\ &= -\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}}\left[\sigma\left(\left\langle \pi(\mathbf{w}^{(i)}),\mathbf{x}\right\rangle + b_i\right)\chi_I(\mathbf{x})\right] = \frac{\partial}{\partial u_i} L_{\chi_I,\mathcal{D}}(\pi(h_{\mathbf{u},\mathbf{w},\mathbf{b}}))\end{aligned}$$

Let S be a maximal set of permutations such that for every $\pi_1 \neq \pi_2 \in S$ we have $\pi_1(I) \neq \pi_2(I)$, and note that $|S| = \binom{n}{k}$. Let $g_i(\mathbf{x}) = \sigma\left(\left\langle \mathbf{w}^{(i)},\mathbf{x}\right\rangle + b_i\right)$ and note that $\|g_i\|_{\mathcal{D}}^2 \leq c^2$. Therefore:

$$\begin{aligned}\sum_{\pi\in S_j}\left(\frac{\partial}{\partial u_i} L_{\chi_I,\mathcal{D}}(\pi(h_{\mathbf{u},\mathbf{w},\mathbf{b}}))\right)^2 &= \sum_{\pi\in S}\left(\frac{\partial}{\partial u_i} L_{\chi_{\pi(I)},\mathcal{D}}(h_{\mathbf{u},\mathbf{w},\mathbf{b}})\right)^2 \\ &= \sum_{\pi\in S}\left\langle g_i,\chi_{\pi(I)}\right\rangle_{\mathcal{D}}^2 \leq \sum_{I'\subseteq[n]}\left\langle g_i,\chi_{I'}\right\rangle_{\mathcal{D}}^2 = \|g_i\|_{\mathcal{D}}^2 \leq c^2\end{aligned}$$

And therefore:

$$\mathop{\mathbb{E}}_{\mathbf{w}\sim\mathcal{W}}\left\|\frac{\partial}{\partial \mathbf{u}} L_{\chi_I,\mathcal{D}}(h_{\mathbf{u},\mathbf{w},\mathbf{b}})\right\|_2^2 = \mathop{\mathbb{E}}_{\mathbf{w}\sim\mathcal{W}}\mathop{\mathbb{E}}_{\pi\sim S}\left\|\frac{\partial}{\partial \mathbf{u}} L_{\chi_I,\mathcal{D}}(\pi(h_{\mathbf{u},\mathbf{w},\mathbf{b}}))\right\|_2^2 \leq c^2 q\binom{n}{k}^{-1}$$

□