APPENDIX A

We first show step by step how the normed gradient difference between two instances can be bounded by the last layer of the neural network.

Specifically, considering a T-layer perception, we define $\varphi^{(t)}(.)$ as the Lipschitz continuous activation function for layer t, and $\theta^{(t)}$ is the weight matrix for layer t. We simply define $p(l|x_i)$ as p_1 and define $p(l|x_j)$ as p_2. To compute the gradient of the loss function f with respect to the weights $\theta^{(T)}$ in the final layer, we use backpropagation as follows:

i).**Output of the neural network**:

- Let $o_i^{(T-1)}$ be the input to the final layer T.
- The output of the final layer before activation is $z_i^{(T)} = \theta^{(T)} \cdot o_i^{(T-1)}$.
- The activated output is $o_i^{(T)} = \varphi^{(T)}(z_i^{(T)})$.

ii). **Gradient with respect to** θ: By the chain rule, we have:

$$
\begin{aligned}
p_1\nabla f_i(\theta) &= p_1\frac{\mathrm{d}f_i}{\mathrm{d}\theta} = p_1\frac{\mathrm{d}f_i}{\mathrm{d}o_i^{(T)}}\cdot\frac{\mathrm{d}o_i^{(T)}}{\mathrm{d}z_i^{(T)}}\cdot\frac{\mathrm{d}z_i^{(T)}}{\mathrm{d}\theta}\\
&= p_1\frac{\mathrm{d}f_i}{\mathrm{d}o_i^{(T)}}\cdot\frac{\mathrm{d}o_i^{(T)}}{\mathrm{d}z_i^{(T)}}\cdot\frac{\mathrm{d}z_i^{(T)}}{\mathrm{d}z_i^{(T-1)}}\cdot\frac{\mathrm{d}z_i^{(T-1)}}{\mathrm{d}z_i^{(T-2)}}\cdot\ldots\cdot\frac{\mathrm{d}z_i^{(2)}}{\mathrm{d}z_i^{(1)}}\cdot\frac{\mathrm{d}z_i^{(1)}}{\mathrm{d}\theta}\\
&= p_1\nabla f_i^{(T)}(\theta)\cdot\varphi'^{(T)}(z_i^{(T)})\cdot[\theta^{(T)}\cdot\varphi'^{(T)}(z_i^{(T-1)})]\\
&\quad\cdot[\theta^{(T-1)}\cdot\varphi'^{(T-1)}(z_i^{(T-2)})]\cdot\ldots\cdot[\theta^{(2)}\cdot\varphi'^{(2)}(z_i^{(1)})]\cdot(o_i^{(0)})^\intercal\\
&= p_1\nabla f_i^{(T)}(\theta)\cdot\varphi'^{(T)}(z_i^{(T)})\cdot\Omega_i^{(T)}\cdot(o_i^{(0)})^\intercal
\end{aligned}
\tag{1}
$$

where $\Omega_i^{(T)}$ is denoted by $[\theta^{(T)}\cdot\varphi'^{(T)}(z_i^{(T-1)})]\cdot[\theta^{(T-1)}\cdot\varphi'^{(T-1)}(z_i^{(T-2)})]\cdot\ldots\cdot[\theta^{(2)}\cdot\varphi'^{(2)}(z_i^{(1)})]$.

iii). **Upper bound of approximation error**: The maximum distance between two gradients in the whole parameter space Θ could be expressed as:

$$
\begin{aligned}
&\|p_1\nabla f_i(\theta)-p_2\nabla f_j(\theta)\|\\
&=\|p_1\Omega_i^{(T)}\varphi'^{(T)}(z_i^{(T)})\nabla f_i^{(T)}(\theta)(o_i^{(0)})^\intercal-p_2\Omega_j^{(T)}\varphi'^{(T)}(z_j^{(T)})\nabla f_j^{(T)}(\theta)(o_j^{(0)})^\intercal\|\\
&=\|p_1\Omega_i^{(T)}\varphi'^{(T)}(z_i^{(T)})\nabla f_i^{(T)}(\theta)(o_i^{(0)})^\intercal-p_1\Omega_i^{(T)}\varphi'^{(T)}(z_j^{(T)})\nabla f_j^{(T)}(\theta)(o_i^{(0)})^\intercal\\
&+p_1\Omega_i^{(T)}\varphi'^{(T)}(z_j^{(T)})\nabla f_j^{(T)}(\theta)(o_i^{(0)})^\intercal-p_2\Omega_j^{(T)}\varphi'^{(T)}(z_j^{(T)})\nabla f_j^{(T)}(\theta)(o_j^{(0)})^\intercal\|\\
&\le\|p_1\Omega_i^{(T)}\|\cdot\|o_i^{(0)}\|\cdot\|\varphi'^{(T)}(z_i^{(T)})\nabla f_i^{(T)}(\theta)-\varphi'^{(T)}(z_j^{(T)})\nabla f_j^{(T)}(\theta)\|\\
&+\|\varphi'^{(T)}(z_j^{(T)})\nabla f_j^{(T)}(\theta)\|\cdot\|p_1\Omega_i^{(T)}o_i^{(0)}-p_2\Omega_j^{(T)}o_j^{(0)}\|\\
&\le\|p_1\Omega_i^{(T)}\|\cdot\|o_i^{(0)}\|\cdot\|\varphi'^{(T)}(z_i^{(T)})\nabla f_i^{(T)}(\theta)-\varphi'^{(T)}(z_j^{(T)})\nabla f_j^{(T)}(\theta)\|\\
&+\|\varphi'^{(T)}(z_j^{(T)})\nabla f_j^{(T)}(\theta)\|\cdot\|p_1\Omega_i^{(T)}o_i^{(0)}-p_2\Omega_i^{(T)}o_i^{(0)}\|\\
&+\|\varphi'^{(T)}(z_j^{(T)})\nabla f_j^{(T)}(\theta)\|\cdot\|p_2\Omega_i^{(T)}o_i^{(0)}-p_2\Omega_j^{(T)}o_j^{(0)}\|\\
&\le p_1\cdot\|\Omega_i^{(T)}\|\cdot\|o_i^{(0)}\|\cdot\|\varphi'^{(T)}(z_i^{(T)})\nabla f_i^{(T)}(\theta)-\varphi'^{(T)}(z_j^{(T)})\nabla f_j^{(T)}(\theta)\|\\
&+\|\varphi'^{(T)}(z_j^{(T)})\nabla f_j^{(T)}(\theta)\|\cdot(\|p_1-p_2\|\cdot\|\Omega_i^{(T)}o_i^{(0)}\|+p_2\cdot\|\Omega_i^{(T)}o_i^{(0)}-\Omega_j^{(T)}o_j^{(0)}\|)\\
&\le S_{ij}=p_1\cdot n_1\cdot\|\varphi'^{(T)}(z_i^{(T)})\nabla f_i^{(T)}(\theta)-\varphi'^{(T)}(z_j^{(T)})\nabla f_j^{(T)}(\theta)\|\\
&+\|\varphi'^{(T)}(z_j^{(T)})\nabla f_j^{(T)}(\theta)\|\cdot(\|p_1-p_2\|\cdot n_2+p_2\cdot n_3)
\end{aligned}
\tag{2}
$$

where $n_1=\max\limits_{T,i}(\|\Omega_i^{(T)}\|\cdot\|o_i^{(0)}\|)$, $n_2=\max\limits_{T,i}(\|\Omega_i^{(T)}o_i^{(0)}\|)$ and $n_3=\max\limits_{T,i,j}(\|\Omega_i^{(T)}o_i^{(0)}-\Omega_j^{(T)}o_j^{(0)}\|)$ are constants.

APPENDIX B

Next, we will show that solving Equation 7 is an NP-hard problem with submodular property. This allows us to design a greedy algorithm that effectively solves this problem with an approximate ratio.

THEOREM 1. *The problem of subset selection under uncertainty is NP-hard.*

PROOF. Consider a scenario where every instance in D is assigned to a hard label (with a probability of 1). Then the problem simplifies to $C=\arg\min\limits_{C\subseteq D}\sum_{i=1}^{N}\min_{c_j\in C}\|\nabla f_i(\theta)-\nabla f_j(\theta)\|, |C|\le K$. Naturally, the K-medoid problem [1] can be reduced to the special case. Therefore, our problem is also NP-hard.

THEOREM 2. *The problem of subset selection under uncertainty shows the submodular property.*

PROOF. We define the utility function as $B(C,\theta)=\sum_{i=1}^{N}\max_{c_j\in C}u_{ij}$, where $u_{ij}=1-$ normalized$(\max\limits_{\theta\in\Theta}\|p(l|x_i)\nabla f_i(\theta)-p(l|x_j)\nabla f_j(\theta)\|)$. The subset selection under uncertainty problem is equivalent to maximizing utility B. If B has the submodular property, for any $C\subseteq C^*\subseteq D$ and $o_i\in D\setminus C^*$, we have to prove (1)B is monotonous, *i.e.*, $B(C\cup\{o_i\},\theta)\ge B(C,\theta)$, and (2) B has the diminishing marginal returns property, *i.e.*, $B(C\cup\{o_i\},\theta)-B(C,\theta)\ge B(C^*\cup\{o_i\},\theta)-B(C^*,\theta)$. For simplicity, we use $B(o_i|C,\theta)$ to denote $B(C\cup\{o_i\},\theta)-B(C,\theta)$ in the following parts of the paper. The proof starts with considering u_{ij} to be known, which will be computed in Section **??**. In this situation, each instance in D will be assigned to an instance of the subset that maximizes the utility.

For (1), when o_i is added into C, if no instance in D will be assigned to o_i, then $B(C\cup\{o_i\},\theta)=B(C,\theta)$. If one or more instances in D are assigned to o_i, clearly $B(C\cup\{o_i\},\theta)>B(C,\theta)$. Hence, B is monotonous.

For (2), we can see that $B(C,\theta)$ is the sum of different terms w.r.t. different instances, and they are computed independently. Therefore, if there is only a single instance and the diminishing marginal returns property satisfies, then B has the property. Suppose that the instance is denoted by o_*. Given C, C^* and $\{o_i\}$, we prove the diminishing marginal returns for o_* in all possible three cases of $c_*=\arg\max\limits_{c_j\in C\cup(C^*\setminus C)\cup\{o_i\}}u_{*j}$.

[Case 1: $c_*\in C$**]** In this case, obviously, $B(C\cup\{o_i\},\theta)-B(C,\theta)=B(C^*\cup\{o_i\},\theta)-B(C^*,\theta)=0$ because o_* will not change its assignment when $C^*\setminus C$ and o_i are added.

[Case 2: $c_*\in C^*\setminus C$**]** Apparently, $B(C^*\cup\{o_i\},\theta)-B(C^*,\theta)=0$, which must be smaller than $B(C\cup\{o_i\},\theta)-B(C,\theta)$.

[Case 3: $c_*=o_i$**]** There are two cases here.

(1) If $\max\limits_{c_j\in C}u_{*j}\ge\max\limits_{c_j\in C^*\setminus C}u_{*j}$, $B(C\cup\{o_i\},\theta)-B(C,\theta)=B(C^*\cup\{o_i\},\theta)-B(C^*,\theta)>0$.

(2) If $\max\limits_{c_j\in C}u_{*j}<\max\limits_{c_j\in C^*\setminus C}u_{*j}$, $B(C\cup\{o_i\},\theta)-B(C,\theta)>B(C^*\cup\{o_i\},\theta)-B(C^*,\theta)>0$. The reason is that when $C^*\setminus C$ is added to C, o_* will change its assignment and thus the utility is increased. Afterwards, o_i is added, and the utility is further increased.

REFERENCES

[1] M. R. Garey and David S. Johnson. 1979. *Computers and Intractability: A Guide to the Theory of NP-Completeness.* W. H. Freeman.