Proceedings of Machine Learning Research xx:1–8, 2020 ACML 2020

Supplementary File: Robust Deep Ordinal Regression under Label Noise

Bhanu Garg BHANUGARG05@GMAIL.COM and **Naresh Manwani** NARESH.MANWANI@IIIT.AC.IN
International Institute of Information Technology, Hyderabad, India

Editors: Sinno Jialin Pan and Masashi Sugiyama

1. Proof of Theorem 1

1.1. Rank consistency proof for $\tilde{l}_{CE}$

We need to show that $b_1 \geq b_2 \geq \ldots \geq b_{K-1}$ at the optimal solution. Let $\mathbf{b} = [b_1, b_2, .., b_{K-1}]^T$, and $\mathbf{b}^*$ be the optimal value of $\mathbf{b}$. Let $(\mathbf{x}_i, \tilde{y}_i),\ i = 1 \ldots N$ be the training set. Let for some j suppose $b_j < b_{j+1}$. Then we show that by replacing b_j with b_{j+1} or replacing b_{j+1} with b_j can further decrease the loss $\mathbf{\tilde{L}_{CE}} = \mathbf{N}^{-1}\mathbf{L}_{CE}$, where $\mathbf{\tilde{L}_{CE}} = \begin{bmatrix} \tilde{l}_{CE}(g(\mathbf{x}), \mathbf{b}, 1) \\ \vdots \\ \tilde{l}_{CE}(g(\mathbf{x}), \mathbf{b}, j+1) \\ \vdots \\ \tilde{l}_{CE}(g(\mathbf{x}), \mathbf{b}, K) \end{bmatrix}$ and $\mathbf{L}_{CE} = \begin{bmatrix} l_{CE}(g(\mathbf{x}), \mathbf{b}, 1) \\ \vdots \\ l_{CE}(g(\mathbf{x}), \mathbf{b}, j+1) \\ \vdots \\ l_{CE}(g(\mathbf{x}), \mathbf{b}, K) \end{bmatrix}$. We see that the change in $\mathbf{\tilde{L}_{CE}}$ depends on $\mathbf{L}_{CE}$ as follows.

$$\Delta\mathbf{\tilde{L}_{CE}} = \mathbf{N^{-1}}\Delta\mathbf{L}_{CE} = \mathbf{N}^{-1} \begin{bmatrix} \Delta l_{CE}(g(\mathbf{x}), \mathbf{b}, 1) \\ \vdots \\ \Delta l_{CE}(g(\mathbf{x}), \mathbf{b}, j+1) \\ \vdots \\ \Delta l_{CE}(g(\mathbf{x}), \mathbf{b}, K) \end{bmatrix}$$

We now have to find the change $\Delta l_{CE}(g(\mathbf{x}_i), \mathbf{b}, k)$ for every $i \in [N]$ and every $k \in [K-1]$. In order to do that, we first consider the following three partitions of the training set.

$$\begin{aligned} A_1 &= \{\mathbf{x}_i : y_i < j+1 \implies z^j_{y_i} = z^{j+1}_{y_i} = 0\} \\ A_2 &= \{\mathbf{x}_i : y_i > j+1 \implies z^j_{y_i} = z^{j+1}_{y_i} = 1\} \\ A_3 &= \{\mathbf{x}_i : y_i = j+1 \implies z^j_{y_i} = 1, z^{j+1}_{y_i} = 0\} \end{aligned}$$

The above three sets are mutually exclusive and exhaustive, i.e., $A_1 \cup A_2 \cup A_3 = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$. Let $h_j(\mathbf{x}) = \sigma(g(\mathbf{x}) + b_j)$. Now, we first find the change $\Delta l_{CE}(g(\mathbf{x}_i), \mathbf{b}, k)$ for every $k \in [K-1]$ in these sets individually.

1. **Change in l_{CE} for $\mathbf{x}_i \in A_1$:** The change in l_{CE} when replacing b_j with b_{j+1} is,
$$\Delta^a l_{CE}(g(\mathbf{x}_i), \mathbf{b}, y_i) = \log(1 - h_j(\mathbf{x}_i)) - \log(1 - h_{j+1}(\mathbf{x}_i)).$$
The change in l_{CE} when replacing b_{j+1} with b_j is,
$$\Delta^b l_{CE}(g(\mathbf{x}_i), \mathbf{b}, y_i) = \log(1 - h_{j+1}(\mathbf{x}_i)) - \log(1 - h_j(\mathbf{x}_i)).$$
The total change in loss l_{CE} after swapping b_j and b_{j+1} is $\Delta l_{CE}(g(\mathbf{x}), \mathbf{b}, y_i) = (\Delta^a + \Delta^b) l_{CE}(g(\mathbf{x}), \mathbf{b}, y_i) = 0$.

2. **Change in l_{CE} for A_2:** The change in l_{CE} when replacing b_j with b_{j+1} is
$$\Delta^a l_{CE}(g(\mathbf{x}), \mathbf{b}, y_i) = \log(h_j(\mathbf{x})) - \log(h_{j+1}(\mathbf{x})).$$
The change in l_{CE} replacing b_{j+1} with b_j
$$\Delta^b l_{CE}(g(\mathbf{x}), \mathbf{b}, y_i) = \log(h_{j+1}(\mathbf{x})) - \log(h_j(\mathbf{x})).$$
The total change in loss L_{CE} after swapping b_j and b_{j+1} is $(\Delta^a + \Delta^b) l_{CE}(g(\mathbf{x}), \mathbf{b}, y_i) = 0$.

3. **Change in l_{CE} for A_3:** The change in l_{CE} when replacing b_j with b_{j+1} is
$$\Delta^a l_{CE}(g(\mathbf{x}), \mathbf{b}, y_i) = \log(h_j(\mathbf{x})) - \log(h_{j+1}(\mathbf{x})).$$
The change in l_{CE} replacing b_{j+1} with b_j
$$\Delta^b l_{CE}(g(\mathbf{x}), \mathbf{b}, y_i) = -\log(1 - h_j(\mathbf{x})) - \log(1 - h_{j+1}(\mathbf{x})).$$
The total change in loss l_{CE} after swapping b_j and b_{j+1} and given that $b_j \geq b_{j+1}$ is
$$\begin{aligned}(\Delta^a + \Delta^b) l_{CE}(g(\mathbf{x}), \mathbf{b}, y_i) &= \log(h_j(\mathbf{x})) - \log(h_{j+1}(\mathbf{x})) \\ &\quad - (\log(1 - h_{j+1}(\mathbf{x})) - \log(1 - h_j(\mathbf{x})) < 0\end{aligned}$$

Hence
$$(\Delta^a + \Delta^b) l_{CE}(g(\mathbf{x}), \mathbf{b}, y_i) = \begin{cases} \delta, & \text{if } y_i = j+1 \\ 0, & \text{if } y_i \neq j+1 \end{cases}$$
for some $\delta < 0$. Now consider the equations
$$(\Delta^a + \Delta^b)\tilde{\mathbf{L}}_{\mathbf{CE}} = \mathbf{N}^{-1} \begin{bmatrix} (\Delta^a + \Delta^b) l_{CE}(g(\mathbf{x}), \mathbf{b}, 1) \\ \vdots \\ (\Delta^a + \Delta^b) l_{CE}(g(\mathbf{x}), \mathbf{b}, j+1) \\ \vdots \\ (\Delta^a + \Delta^b) l_{CE}(g(\mathbf{x}), \mathbf{b}, K) \end{bmatrix}$$
$$\Rightarrow \begin{bmatrix} (\Delta^a + \Delta^b) \tilde{l}_{CE}(g(\mathbf{x}), \mathbf{b}, 1) \\ \vdots \\ (\Delta^a + \Delta^b) \tilde{l}_{CE}(g(\mathbf{x}), \mathbf{b}, j+1) \\ \vdots \\ (\Delta^a + \Delta^b) \tilde{l}_{CE}(g(\mathbf{x}), \mathbf{b}, K) \end{bmatrix} = \mathbf{N^{-1}} \begin{bmatrix} 0 \\ \vdots \\ \delta \\ \vdots \\ 0 \end{bmatrix}$$

The change in loss $\tilde{l}_{CE}$ is as follows.

$$\begin{aligned}(\Delta^a + \Delta^b)R_\rho &= (\Delta^a + \Delta^b)\mathbb{E}_{\tilde{y}}[\tilde{l}_{CE}(g(\mathbf{x}), \mathbf{b}, \tilde{y})] = \mathbb{E}_{\tilde{y}}[(\Delta^a + \Delta^b)\tilde{l}_{CE}(g(\mathbf{x}), \mathbf{b}, \tilde{y})] \\ &= \mathbb{E}_{\tilde{y}}[\mathbf{N}^{-1}_{(\tilde{y},i+1)}\delta] = \delta\mathbb{E}_{\tilde{y}}[\mathbf{N}^{-1}_{(\tilde{y},i+1)}] = \delta\sum_{k=1}^{K} P(\tilde{y} = k)\mathbf{N}^{-1}_{(k,i+1)} \\ &= \delta\sum_{k=1}^{K}\mathbf{N}^{-1}_{(k,i+1)}\sum_{j=1}^{K} P(y = j)P(\tilde{y} = k|y = j) \\ &= \delta\sum_{j=1}^{K} P(y = j)\sum_{k=1}^{K}\eta_{(j,k)}\mathbf{N}^{-1}_{(k,i+1)} = \delta\sum_{j=1}^{K} P(y = j)\mathbb{I}_{\{j=i+1\}} = \delta P(y = i + 1) \le 0\end{aligned}$$

That means by swapping b_j and b_{j+1}, we can further reduce the total loss $\tilde{L}_{CE}$, which is a contradiction to the assumption that $\mathbf{b}$ is the optimal solution under $\tilde{L}_{CE}$. This completes the proof that $\tilde{l}_{CE}$ is also rank consistent.

1.2. Rank consistency proof for $\tilde{l}_{IMC}$

We need to show that $b_1 \ge b_2 \ge \ldots \ge b_{K-1}$ at the optimal solution. We use a similar methodology as Theorem 1 Section 1.1 to prove this. Let $\mathbf{b} = [b_1, b_2, .., b_{K-1}]^T$, and $\mathbf{b}^*$ be the optimal value of $\mathbf{b}$.

Let for some j suppose $b_j < b_{j+1}$. Then we show that by replacing b_j with b_{j+1} or replacing b_{j+1} with b_j can further decrease the loss $\tilde{\mathbf{L}} = \mathbf{N}^{-1}\mathbf{L}$. Consider the following sets.

$$\begin{aligned}A_1 &= \{i : y_i < j + 1 \implies z^j_{y_i} = z^{j+1}_{y_i} = -1\} \\ A_2 &= \{i : y_i > j + 1 \implies z^j_{y_i} = z^{j+1}_{y_i} = +1\} \\ A_3 &= \{i : y_i = j + 1 \implies z^j_{y_i} = -1, z^{j+1}_{y_i} = +1\}\end{aligned}$$

The above three sets are mutually exclusive and exhaustive, i.e., $A_1 \cup A_2 \cup A_3 = \{1, 2, .., N\}$.

1. **Change in l_{IMC} for A_1:** The change in l_{IMC} when replacing b_j with b_{j+1} is

   $$\Delta^a l_{IMC}(f(\mathbf{x}), y_i) = \max(0, -1(g(\mathbf{x}_i) + b_{j+1}) + 1) - \max(0, -1(g(\mathbf{x}_i) + b_j) + 1)$$

   The change in l_{IMC} when replacing b_{j+1} with b_j

   $$\Delta^b l_{IMC}(f(\mathbf{x}), y_i) = \max(0, -1(g(\mathbf{x}_i) + b_j) + 1) - \max(0, -1(g(\mathbf{x}_i) + b_{j+1}) + 1)$$

   The total change in loss L_{IMC} after swapping b_j and b_{j+1} is $(\Delta^a + \Delta^b)l_{IMC}(f(\mathbf{x}), y_i) = 0$

2. **Change in l_{IMC} for A_2:** The change in l_{IMC} when replacing b_j with b_{j+1} is

   $$\Delta^a l_{IMC}(f(\mathbf{x}), y_i) = \max(0, +1(g(\mathbf{x}_i) + b_{j+1}) + 1) - \max(0, +1(g(\mathbf{x}_i) + b_j) + 1)$$

   The change in l_{IMC} replacing b_{j+1} with b_j

   $$\Delta^b l_{IMC}(f(\mathbf{x}), y_i) = \max(0, +1(g(\mathbf{x}_i) + b_j) + 1) - \max(0, +1(g(\mathbf{x}_i) + b_{j+1}) + 1)$$

   The total change in loss L_{IMC} after swapping b_j and b_{j+1} is $(\Delta^a + \Delta^b)l_{IMC}(f(\mathbf{x}), y_i) = 0$

3. **Change in l_{IMC} for A_3:** The change in l_{IMC} when replacing b_j with b_{j+1} is

$$\begin{aligned}\Delta^a l_{IMC}(f(\mathbf{x}), y_i) &= \max(0, -1(g(\mathbf{x}_i) + b_{j+1}) + 1) - \max(0, -1(g(\mathbf{x}_i) + b_j) + 1) \\ &= \max(0, -b_{j+1} - g(\mathbf{x}_i) + 1) - \max(0, -b_j - g(\mathbf{x}_i) + 1 + 1) \leq 0\end{aligned}$$

The change in l_{IMC} replacing b_{j+1} with b_j

$$\begin{aligned}\Delta^b l_{IMC}(f(\mathbf{x}), y_i) &= \max(0, +1(g(\mathbf{x}_i) + b_j) + 1) - \max(0, +1(g(\mathbf{x}_i) + b_{j+1}) + 1) \\ &= \max(0, g(\mathbf{x}_i) + b_j + 1) - \max(0, g(\mathbf{x}_i) + b_{j+1} + 1) \leq 0\end{aligned}$$

Now suppose $\Delta^a l_{IMC}(f(\mathbf{x}), y_i) = 0$. Since $b_j < b_{j+1}$ we have

$$g(\mathbf{x}_i) + b_j \geq 1 \quad \text{and} \quad g(\mathbf{x}_i) + b_{j+1} > 1 \tag{1}$$

From 1, we have in $\Delta^b l_{IMC}(f(\mathbf{x}), y_i)$,

$$\begin{aligned}\Delta^b l_{IMC}(f(\mathbf{x}), y_i) &= \max(0, g(\mathbf{x}_i) + b_j + 1) - \max(0, g(\mathbf{x}_i) + b_{j+1} + 1) \\ &= b_{j+1} - b_j < 0\end{aligned}$$

Similarly, if $\Delta^b l_{IMC}(f(\mathbf{x}), y_i) = 0$, we will have $\Delta^a l_{IMC}(f(\mathbf{x}), y_i) < 0$. The total change in loss l_{IMC} after swapping b_j and b_{j+1} and given that $b_j < b_{j+1}$ is

$$(\Delta^a + \Delta^b) l_{IMC}(f(\mathbf{x}), y_i) < 0$$

Hence

$$(\Delta^a + \Delta^b) l_{IMC}(f(\mathbf{x}), y_i) = \begin{cases} \delta, & \text{if } y_i = j+1 \\ 0, & \text{if } y_i \neq j+1 \end{cases}$$

for some $\delta < 0$. Now using similar arguments as Theorem-1, Section 1.2 we get that $\tilde{l}_{IMC}$ is rank consistent too.

2. Proof of Theorem 2

We are given that $\mathbb{E}_{\tilde{y}}[b_i^t - b_{i+1}^t] \geq 0,\ i \in [K-1]$. Let at the t^{th} iteration example $(\mathbf{x}^t, \tilde{y}^t)$ is being presented to the network. Loss $\tilde{l}_{CE}$ corresponding to $(\mathbf{x}^t, \tilde{y}^t)$ is as follows.

$$\begin{aligned}\tilde{l}_{CE}(g(\mathbf{x}^t), \mathbf{b}, \tilde{y}^t) &= \sum_{j=1}^{K} \mathbf{N}^{-1}_{(\tilde{y}^t, j)} l_{CE}(g(\mathbf{x}^t), \mathbf{b}, j) \\ &= -\sum_{j=1}^{K} \mathbf{N}^{-1}_{(\tilde{y}^t, j)} \sum_{i=1}^{K-1} \left(\log h_i(\mathbf{x}^t)^{z_i^j} + \log(1 - h_i(\mathbf{x}^t))^{(1 - z_i^j)} \right)\end{aligned}$$

For every $j = 1 \ldots K-1$, z_i^j are defined as follows. $z_i^j = 1,\ \forall i < j$ and $z_i^j = 0,\ \forall i \geq j$. The update equation using SGD requires to compute the partial derivative of the parameters with respect to the loss function $\tilde{l}_{CE}$. We see the following.

$$\begin{aligned}
\frac{\partial \tilde{l}_{CE}(g(\mathbf{x}^t),\mathbf{b},\tilde{y}^t)}{\partial b_i} &= -\sum_{j=1}^{K} \mathbf{N}^{-1}_{(\tilde{y}^t,j)} \Big[z_i^j \frac{\partial \log(h_i(\mathbf{x}^t))}{\partial b_i} + (1-z_i^j)\frac{\partial \log(1-h_i(\mathbf{x}^t))}{\partial b_i} \Big] \\
&= -\sum_{j=1}^{K} \mathbf{N}^{-1}_{(\tilde{y}^t,j)} \left(\frac{z_i^j}{h_i(\mathbf{x}^t)} - \frac{1-z_i^j}{1-h_i(\mathbf{x}^t)} \right) \frac{\partial h_i(\mathbf{x}^t)}{\partial b_i} \\
&= -\sum_{j=1}^{K} \mathbf{N}^{-1}_{(\tilde{y}^t,j)} \Big(z_i^j (1-h_i(\mathbf{x}^t)) - (1-z_i^j) h_i(\mathbf{x}^t) \Big)
\end{aligned}$$

The update equations for thresholds $b_1,\ldots,b_{K-1}$ using SGD are as follows. Let α be the learning rate.

$$\begin{aligned}
b_i^{t+1} &= b_i^t - \alpha \frac{\partial \tilde{l}_{CE}(g^t(\mathbf{x}^t),\mathbf{b}^t,\tilde{y}^t)}{\partial b_i} \\
&= b_i^t + \alpha \sum_{j=1}^{K} \mathbf{N}^{-1}_{(\tilde{y}^t,j)} \Big(z_i^j (1-\sigma(g^t(\mathbf{x}^t)+b_i^t) - (1-z_i^j)\sigma(g^t(\mathbf{x}^t)+b_i^t) \Big)
\end{aligned}$$

Using the above equation, we compute the following.

$$\begin{aligned}
b_i^{t+1} - b_{i+1}^{t+1} &= b_i^t - b_{i+1}^t + \alpha \sum_{j=1}^{K} \mathbf{N}^{-1}_{(\tilde{y}^t,j)} \Big(z_i^j (1-h_i^t(\mathbf{x}^t)) - (1-z_i^j) h_i^t(\mathbf{x}^t) - z_{i+1}^j (1-h_{i+1}^t(\mathbf{x}^t)) \\
&\quad + (1-z_{i+1}^j) h_{i+1}^t(\mathbf{x}^t) \Big) \\
&= b_i^t - b_{i+1}^t + \alpha \sum_{j=1}^{K} \mathbf{N}^{-1}_{(\tilde{y}^t,j)} \Big[z_i^j - h_i^t(\mathbf{x}^t) - z_{i+1}^j + h_{i+1}^t(\mathbf{x}^t) \Big]
\end{aligned}$$

For every $j \in \{1,\ldots,K\}$, there can be three possibilities as follows. (a) $z_i^j = z_{i+1}^j = 0$, (b) $z_i^j = z_{i+1}^j = 1$ and (c) $z_i^j = 1$, $z_{i+1}^j = 0$. Thus, we can rewrite $b_i^{t+1} - b_{i+1}^{t+1}$ as follows.

$$\begin{aligned}
b_i^{t+1} - b_{i+1}^{t+1} &= b_i^t - b_{i+1}^t + \alpha \sum_{z_i^j = z_{i+1}^j} \mathbf{N}^{-1}_{(\tilde{y}^t,j)} \Big[h_{i+1}^t(\mathbf{x}^t) - h_i^t(\mathbf{x}^t) \Big] + \alpha \sum_{z_i^j=1, z_{i+1}^j=0} \mathbf{N}^{-1}_{(\tilde{y}^t,j)} \Big[1 + h_{i+1}^t(\mathbf{x}^t) - h_i^t(\mathbf{x}^t) \Big] \\
&= b_i^t - b_{i+1}^t + \alpha \sum_{j=1}^{K} \mathbf{N}^{-1}_{(\tilde{y}^t,j)} \Big[h_{i+1}^t(\mathbf{x}^t) - h_i^t(\mathbf{x}^t) \Big] + \alpha \sum_{z_i^j=1, z_{i+1}^j=0} \mathbf{N}^{-1}_{(\tilde{y}^t,j)}
\end{aligned}$$

Using properties of noise matrix, we know that $\sum_{j=1}^{K} \mathbf{N}^{-1}_{(\tilde{y}^t,j)} = 1$. Thus,

$$b_i^{t+1} - b_{i+1}^{t+1} = b_i^t - b_{i+1}^t - \alpha \Big[h_i^t(\mathbf{x}^t) - h_{i+1}^t(\mathbf{x}^t) \Big] + \alpha \sum_{z_i^j=1, z_{i+1}^j=0} \mathbf{N}^{-1}_{(\tilde{y}^t,j)}$$

The only possibility for $z_i^j = 1, z_{i+1}^j = 0$ is $j = i + 1$. Thus,

$$b_i^{t+1} - b_{i+1}^{t+1} = b_i^t - b_{i+1}^t - \alpha\Big[h_i^t(\mathbf{x}^t) - h_{i+1}^t(\mathbf{x}^t)\Big] + \alpha\mathbf{N}^{-1}_{(\tilde{y}^t,i+1)}.$$

Since $\mathbf{N}^{-1}_{(\tilde{y}^t,i+1)}$ updates depend on $\tilde{y}^t$, we take the expectation on both sides with respect to $\tilde{y}$, we get the following.

$$\mathbb{E}_{\tilde{y}}[b_i^{t+1} - b_{i+1}^{t+1}] \geq \mathbb{E}_{\tilde{y}}\Big[b_i^t - b_{i+1}^t - \alpha\big(h_i^t(\mathbf{x}^t) - h_{i+1}^t(\mathbf{x}^t)\big)\Big] + \alpha\mathbb{E}_{\tilde{y}}[\mathbf{N}^{-1}_{(\tilde{y}^t,i+1)}]$$

We know that, $h_i^t(\mathbf{x}^t) = \sigma(g^t(\mathbf{x}^t) + b_i^t)$. Also, $b_i^t \geq b_{i+1}^t$. Using the Mean-Value Theorem, $\exists\theta \in (b_{i+1}^t, b_i^t)$ such that

$$\begin{aligned}\frac{h_i^t(\mathbf{x}^t) - h_{i+1}^t(\mathbf{x}^t)}{b_i^t - b_{i+1}^t} &= \frac{\partial\sigma(g^t(\mathbf{x}^t) + b)}{\partial b}_{|\theta} \\ &= \sigma(g^t(\mathbf{x}^t) + \theta)(1 - \sigma(g^t(\mathbf{x}^t) + \theta)).\end{aligned}$$

We know that $0 < \sigma(g^t(\mathbf{x}^t) + \theta)(1 - \sigma(g^t(\mathbf{x}^t) + \theta)) \leq 0.25,\ \forall\theta \in \mathbb{R}$. Using this, we get,

$$\begin{aligned}b_i^t - b_{i+1}^t - \alpha(\sigma(g^t(\mathbf{x}^t) + b_i^t) - \sigma(g^t(\mathbf{x}^t) + b_{i+1}^t)) &= (1 - \alpha\frac{\partial\sigma(g^t(\mathbf{x}^t) + b')}{\partial b})(b_i^t - b_{i+1}^t) \\ &\geq (1 - 0.25\alpha)(b_i^t - b_{i+1}^t) \geq 0\end{aligned}$$

where the last inequality holds when $\alpha \leq 4$. Thus for $b_i^t \geq b_{i+1}^t$, we get

$$b_i^t - b_{i+1}^t - \alpha\Big[h_i(\mathbf{x}^t) - h_{i+1}(\mathbf{x}^t)\Big] \geq 0,\ \forall\alpha \leq 4. \tag{2}$$

We know that

$$\mathbb{E}_{\tilde{y}}[b_i^{t+1} - b_{i+1}^{t+1}] \geq \mathbb{E}_{\tilde{y}}\Big[b_i^t - b_{i+1}^t - \alpha\big(h_i^t(\mathbf{x}^t) - h_{i+1}^t(\mathbf{x}^t)\big)\Big] + \alpha\mathbb{E}_{\tilde{y}}[\mathbf{N}^{-1}_{(\tilde{y}^t,i+1)}].$$

Now, we using the result in eq.(2), we get the following.

$$\begin{aligned}\mathbb{E}_{\tilde{y}}[b_i^{t+1} - b_{i+1}^{t+1}] &\geq \alpha\mathbb{E}_{\tilde{y}}[\mathbf{N}^{-1}_{(\tilde{y}^t,i+1)}] = \alpha\sum_{k=1}^{K} P(\tilde{y} = k)\mathbf{N}^{-1}_{(k,i+1)} \\ &= \alpha\sum_{k=1}^{K}\mathbf{N}^{-1}_{(k,i+1)}\sum_{j=1}^{K} P(y = j)P(\tilde{y} = k|y = j) = \alpha\sum_{j=1}^{K} P(y = j)\sum_{k=1}^{K}\eta_{(j,k)}\mathbf{N}^{-1}_{(k,i+1)} \\ &= \alpha\sum_{j=1}^{K} P(y = j)\mathbb{I}_{\{j=i+1\}} = \alpha P(y = i + 1) \geq 0.\end{aligned}$$

Thus, we have shown that $\mathbb{E}_{\tilde{y}}[b_i^{t+1} - b_{i+1}^{t+1}] \geq 0$. This completes our proof that SGD gives the optimal solution maintaining rank consistency.

3. Proof of Theorem 3

Let at t^{th} iteration example $(\mathbf{x}^t, \tilde{y}^t)$ is being presented to the network. Loss $\tilde{l}_{IMC}$ corresponding to $(\mathbf{x}^t, \tilde{y}^t)$ is described as follows.

$$\tilde{l}_{IMC}(g(\mathbf{x}^t), \mathbf{b}, \tilde{y}^t) = \sum_{j=1}^{K} \mathbf{N}^{-1}_{(\tilde{y}^t,j)} \sum_{i=1}^{K-1} \left[0, 1 - z_i^j \left(g(\mathbf{x}^t) + b_i\right)\right]_+$$

Where $z_i^j = 1,\ \forall i < j$ and $z_i^j = -1,\ \forall i \geq j$. We first find the sub-gradient of $\tilde{l}_{IMC}$ w.r.t b_i.

$$\frac{\partial \tilde{l}_{IMC}(g(\mathbf{x}^t), \mathbf{b}, \tilde{y}^t)}{\partial b_i} = -\sum_{j=1}^{K} \mathbf{N}^{-1}_{(\tilde{y}^t,j)} z_i^j \mathbb{I}[z_i^j(g(\mathbf{x}^t) + b_i) < 1]$$

Hence the SGD based update equation for b_i (with step size α) is as follows.

$$\begin{aligned}
b_i^{t+1} &= b_i^t + \alpha \sum_{j=1}^{K} \mathbf{N}^{-1}_{(\tilde{y}^t,j)} z_i^j \mathbb{I}[z_i^j(g^t(\mathbf{x}^t) + b_i^t) < 1] \\
&= b_i^t + \alpha \sum_{j \leq i} \mathbf{N}^{-1}_{(\tilde{y}^t,j)} z_i^j \mathbb{I}[z_i^j(g^t(\mathbf{x}^t) + b_i^t) < 1] + \alpha \sum_{j > i} \mathbf{N}^{-1}_{(\tilde{y}^t,j)} z_i^j \mathbb{I}[z_i^j(g^t(\mathbf{x}^t) + b_i^t) < 1] \\
&= b_i^t - \alpha \sum_{j \leq i} \mathbf{N}^{-1}_{(\tilde{y}^t,j)} \mathbb{I}[g^t(\mathbf{x}^t) + b_i^t > -1] + \alpha \sum_{j > i} \mathbf{N}^{-1}_{(\tilde{y}^t,j)} \mathbb{I}[g^t(\mathbf{x}^t) + b_i^t < 1]
\end{aligned}$$

Where we used the definition of z_i^j. Now, we take the expectation with respect to $\tilde{y}^t$ on both size, and using the fact that $\mathbb{E}_{\tilde{y}^t}[\mathbf{N}^{-1}_{(\tilde{y}^t,j)}] = P(y = j)$, we get the following.

$$\begin{aligned}
\mathbb{E}_{\tilde{y}^t}[b_i^{t+1} - b_i^t] &= -\alpha \sum_{j \leq i} \mathbb{E}_{\tilde{y}^t}[\mathbf{N}^{-1}_{(\tilde{y}^t,j)}] \mathbb{I}[g^t(\mathbf{x}^t) + b_i^t > -1] + \alpha \sum_{j > i} \mathbb{E}_{\tilde{y}^t}[\mathbf{N}^{-1}_{(\tilde{y}^t,j)}] \mathbb{I}[g^t(\mathbf{x}^t) + b_i^t < 1] \\
&= -\alpha \sum_{j \leq i} P(y = j) \mathbb{I}[g^t(\mathbf{x}^t) + b_i^t > -1] + \alpha \sum_{j > i} P(y = j) \mathbb{I}[g^t(\mathbf{x}^t) + b_i^t < 1] \\
&= -\alpha P(y \leq i) \mathbb{I}[g^t(\mathbf{x}^t) + b_i^t > -1] + \alpha P(y > i) \mathbb{I}[g^t(\mathbf{x}^t) + b_i^t < 1]
\end{aligned}$$

Using this, we now compute the following.

$$\begin{aligned}
\mathbb{E}_{\tilde{y}^t}[b_i^{t+1} - b_{i+1}^{t+1} - b_i^t + b_{i+1}^t] &= \mathbb{E}_{\tilde{y}^t}[b_i^{t+1} - b_i^t] - \mathbb{E}_{\tilde{y}^t}[b_i^{t+1} - b_i^t] \\
&= -\alpha P(y \leq i) \mathbb{I}[g^t(\mathbf{x}^t) + b_i^t > -1] + \alpha P(y > i) \mathbb{I}[g^t(\mathbf{x}^t) + b_i^t < 1] \\
&\quad + \alpha P(y \leq i+1) \mathbb{I}[g^t(\mathbf{x}^t) + b_i^t > -1] - \alpha P(y > i+1) \mathbb{I}[g^t(\mathbf{x}^t) + b_i^t < 1] \\
&= \alpha[P(y \leq i+1) - P(y \leq i)] \mathbb{I}[g^t(\mathbf{x}^t) + b_i^t > -1] + \alpha[P(y > i) - P(y > i+1)] \mathbb{I}[g^t(\mathbf{x}^t) + b_i^t < 1]
\end{aligned}$$

But, we know that

$$\begin{aligned}
P(y > i) - P(y > i+1) \geq 0,\ \forall i \in [K-1] \\
P(y \leq i+1) - P(y \leq i) \geq 0,\ \forall i \in [K-1]
\end{aligned}$$

and $\mathbb{I}[.] \in \{0,1\}$. Thus,

$$\begin{aligned}&\mathbb{E}_{\tilde{y}}[(b_i^{t+1} - b_{i+1}^{t+1}) - (b_i^t - b_{i+1}^t)] \geq 0\\ \Rightarrow &\mathbb{E}_{\tilde{y}}[b_i^{t+1} - b_{i+1}^{t+1}] \geq \mathbb{E}_{\tilde{y}}[b_i^t - b_{i+1}^t] = b_i^t - b_{i+1}^t \geq 0\end{aligned}$$

This completes the proof.

4. Generalisation bounds

Using unbiased estimator, we have

$$\begin{aligned}\tilde{l}(g(\mathbf{x}), \mathbf{b}, y) &= \sum_{j=1}^{K} \mathbf{N}_{(y,j)}^{-1} l(g(\mathbf{x}), \mathbf{b}, j) = \sum_{j=1}^{K} \mathbf{N}_{(y,j)}^{-1} \sum_{i=1}^{K-1} l^i(g(\mathbf{x}), \mathbf{b}, z_i^j)\\ &= \sum_{i=1}^{K-1} (\sum_{j=1}^{K} \mathbf{N}_{(y,j)}^{-1} l^i(g(\mathbf{x}), \mathbf{b}, z_i^j)) = \sum_{i=1}^{K-1} \tilde{l}^i(g(\mathbf{x}), \mathbf{b}, i)\end{aligned}$$

For any i, if l^i is $L-$Lipschitz, then $\tilde{l}^i$ is $\tilde{L} = (\sum_{j=1}^{K} |\mathbf{N}_{(y,j)}^{-1}|)L \leq ML$ Lipschitz constant, where $M = \max_y \sum_{j=1}^{K} |\mathbf{N}_{(y,j)}^{-1}|$. Using Lipschitz composition property of basic Rademacher generalisation bounds on i^{th} binary classifier, with probability atleast $1 - \delta$

$$R_{\tilde{l}^i, D_\rho}(f^i) \leq \hat{R}_{\tilde{l}^i, S}(f^i) + 2ML\mathfrak{R}(\mathcal{F}) + \sqrt{\frac{log(1/\delta)}{2n}} \tag{3}$$

where f^i is the i^{th} binary classifier. Adding the maximal deviations between expected risk and empirical risk for all the $K-1$ classifiers,

$$R_{\tilde{l}, D_\rho}(f) \leq \hat{R}_{\tilde{l}, S}(f) + 2ML(K-1)\mathfrak{R}(\mathcal{F}) + (K-1)\sqrt{\frac{log(1/\delta)}{2n}} \tag{4}$$

which if true for any f. Let $\hat{f} \leftarrow \arg\min_{f \in \mathcal{F}} \hat{R}_{\tilde{l}, S}(f)$ and $f^* \leftarrow \arg\min_{f \in \mathcal{F}} R_{l,D}(f)$. Following Theorem 3 from Natarajan et al. (2013),

$$\begin{aligned}R_{l,D}(\hat{f}) - R_{l,D}(f^*) &= R_{\tilde{l}, D_\rho}(\hat{f}) - R_{\tilde{l}, D_\rho}(f^*)\\ &= \hat{R}_{\tilde{l}, S}(\hat{f}) - \hat{R}_{\tilde{l}, S}(f^*) + (R_{\tilde{l}, D_\rho}(\hat{f}) - \hat{R}_{\tilde{l}, S}(\hat{f})) + (\hat{R}_{\tilde{l}, S}(f^*) - R_{\tilde{l}, D_\rho}(f^*))\\ &\leq 2 \max_{f \in \mathcal{F}} |R_{\tilde{l}, D_\rho}(f) - \hat{R}_{\tilde{l}, S}(f)| \qquad (5)\end{aligned}$$

From 4 and 5, we get,

$$R_{l,D}(\hat{f}) \leq R_{l,D}(f^*) + 4ML(K-1)\mathfrak{R}(\mathcal{F}) + 2(K-1)\sqrt{\frac{log(1/\delta)}{2n}}$$

References

Nagarajan Natarajan, Inderjit S. Dhillon, Pradeep Ravikumar, and Ambuj Tewari. Learning with noisy labels. In *NIPS*, pages 1196–1204, 2013.