CDAN: Cost Dependant Deep Abstention Network

No Author Given

No Institute Given

A Proofs

The bounded multiclass abstention loss is given as,

$$\begin{aligned} L^{BMA}(\mathbf{h}(\mathbf{x}_i), \boldsymbol{\rho}, y) = &\frac{d}{\mu}\Big[[\mu - h_y(\mathbf{x}_i) + \max_{j\neq y} h_j(\mathbf{x}_i) + \rho_y]_+ \\ &- [-\mu^2 - h_y(\mathbf{x}_i) + \max_{j\neq y} h_j(\mathbf{x}_i) + \rho_y]_+\Big] \\ &+ \frac{1-d}{\mu}\Big[[\mu - h_y(\mathbf{x}_i) + \max_{j\neq y} h_j(\mathbf{x}_i) - \rho_y]_+ \\ &- [-\mu^2 - h_y(\mathbf{x}_i) + \max_{j\neq y} h_j(\mathbf{x}_i) - \rho_y]_+\Big]. \end{aligned}$$

where $\mathbf{h}(\mathbf{x}) = [h_1(\mathbf{x}),\ h_2(\mathbf{x}),\ \dots\ , h_k(\mathbf{x})] \in \mathbb{R}^k$ and y is the ground truth label. Let $\hat{y} = \arg\max_{r\in[k]} h_r(\mathbf{x}_i)$. Then, the corresponding classifier is given as:

$$f_d^{BMA}(\mathbf{h}(\mathbf{x}_i), \boldsymbol{\rho}(.)) = \begin{cases} h_{\hat{y}}(\mathbf{x}_i), & h_{\hat{y}}(\mathbf{x}) - \max\limits_{y'\neq\hat{y}} h_{y'}(\mathbf{x}_i) > \rho_{\hat{y}} \\ \text{reject}, & h_{\hat{y}}(\mathbf{x}) - \max\limits_{y'\neq\hat{y}} h_{y'}(\mathbf{x}_i) \leq \rho_{\hat{y}} \end{cases}$$

Define sets $\mathcal{H}_i,\ i = 1\dots k+1$ as follows.

$$\mathcal{H}_i = \{\mathbf{h} \in \mathbb{R}^k \mid f_d^{BMA}(\mathbf{h}, \boldsymbol{\rho}) = i\}$$

$$\begin{aligned} \mathcal{H}_y^{\boldsymbol{\rho}} &= \left\{\mathbf{h} \in \mathbb{R}^k : h_y > \max_{j\in k} h_j + \rho_y \quad \forall j \neq y\right\}; y \in [k] \\ \mathcal{H}_{k+1}^{\boldsymbol{\rho}} &= \left\{\mathbf{h} \in \mathbb{R}^k : h_{\hat{y}} < h_{y'} + \rho_y\right\} \end{aligned} \tag{1}$$

where $k+1$ represents the rejection option and $\hat{y} = \arg\max_{i\in k} h(\mathbf{x}_i)$ and $y' = \arg\max_{i\in k, i\neq y} h(\mathbf{x}_i)$.

Now, $\forall u \in [k]$ and $\forall \mathbf{p} \in \Delta_k$, where Δ_k represents the probability simplex i.e. $\sum_i^k p_i = 1$ and $p_i \geq 0$.

$$\begin{aligned} \mathbf{p}^\top L^{BMA}(\mathbf{e}_u) &= 2(1 - p_u) \\ \mathbf{p}^\top L^{BMA}(\mathbf{0}) &= 1 \end{aligned} \tag{2}$$

where $\mathbf{e_u}$ is the vector in $\mathbb{R}^k$ and 1 in the u^{th} position and 0 everywhere else and $\mathbf{0}$ is a vector which is 0 everywhere.

We also define for y, the ground truth and $\tilde{y} \in [1, .., k] - \{y\}$,

$$\begin{aligned}
L^{BMA}(h(x), \boldsymbol{\rho}, y) \geq \frac{d}{\mu}\Big[&[\mu - (h_{\hat{y}}(x_i) - \max_{y' \in k, y' \neq \hat{y}} h_{y'}(\mathbf{x}_i)) + \rho_{\hat{y}}]_+ \\
&- [-\mu^2 - (h_{\hat{y}}(x_i) - \max_{y' \in k, y' \neq \hat{y}} h_{y'}(\mathbf{x}_i)) + \rho_{\hat{y}}]_+\Big] \\
&+ \frac{1-d}{\mu}\Big[[\mu - (h_{\hat{y}}(x_i) - \max_{y' \in k, y' \neq \hat{y}} h_{y'}(\mathbf{x}_i)) - \rho_{\hat{y}}]_+ \\
&- [-\mu^2 - (h_{\hat{y}}(x_i) - \max_{y' \in k, y' \neq \hat{y}} h_{y'}(\mathbf{x}_i)) - \rho_{\hat{y}}]_+\Big] \\
L^{BMA}(h(x), \boldsymbol{\rho}, \tilde{y}) \geq \frac{d}{\mu}\Big[&[\mu - (\max_{y' \in k, y' \neq \hat{y}} h_{y'}(\mathbf{x}_i) - h_{\hat{y}}(x_i)) + \rho_{y'}]_+ \\
&- [-\mu^2 - (\max_{y' \in k, y' \neq \hat{y}} h_{y'}(\mathbf{x}_i) - h_{\hat{y}}(x_i)) + \rho_{y'}]_+\Big] \\
&+ \frac{1-d}{\mu}\Big[[\mu - (\max_{y' \in k, y' \neq \hat{y}} h_{y'}(\mathbf{x}_i) - h_{\hat{y}}(x_i)) - \rho_{y'}]_+ \\
&- [-\mu^2 - (\max_{y' \in k, y' \neq \hat{y}} h_{y'}(\mathbf{x}_i) - h_{\hat{y}}(x_i)) - \rho_{y'}]_+\Big]
\end{aligned}$$

Theorem 1. *Let $k \in \mathbb{N}$ and $\rho_i \in [0, 1] \quad \forall i \in k$. Then for all $\mathbf{h} \in H$, $\rho_{\hat{y}} \geq \frac{1}{2}\mu(1+\mu)$ and $d \leq 0.5$:*

$$(1+\mu)(R_d(h, \rho) - R_d(h_d^*, \rho_d^*)) \leq \left(R_d^{BMA}(h, \rho) - R_d^{BMA}(h_d^*, \rho_d^*)\right)$$

Proof. To establish the theorem we need to prove,

$$(1+\mu)(p^T L_d(f_{\boldsymbol{\rho}}^{BMA}(\boldsymbol{h})) - \min_t p^T L_d(t)) \leq p^T L^{BMA}(\boldsymbol{h}) - \inf_{h^* \in \mathbb{R}^k} p^T L^{BMA}(\mathbf{h}^*) \tag{3}$$

□

The risk associated with BMA loss can be computed as expected loss with respect to all the classes,

$$\begin{aligned}
R_{p_y}^{BMA}(\boldsymbol{h}(x), \boldsymbol{\rho}) &= \sum_{i=1}^{k} L^{BMA}(\boldsymbol{h}(x), \boldsymbol{\rho}, i) \\
R_{p_y}^{BMA}(\boldsymbol{h}(x), \rho) &\geq p_y L^{BMA}(\boldsymbol{h}, \rho, y) + (1 - p_y) L^{BMA}(\boldsymbol{h}, \rho, \tilde{y})
\end{aligned}$$

where $p_y = P(Y = y | X = x)$. The $R_{p_y}^{BMA}(\boldsymbol{h}(x), \boldsymbol{\rho})$ can take different values in different cases based on $h(x)$ values. The different regions are $0 \leq \rho_{\hat{y}} - \mu$,$\rho_{\hat{y}} - \mu \leq \rho_{\hat{y}} - \mu^2$,$\rho_{\hat{y}} - \mu^2 \leq \rho_{\hat{y}} + \mu^2$, $\rho_{\hat{y}} + \mu^2 \leq \rho_{\hat{y}} + \mu$ or greater than $\rho_{\hat{y}} + \mu$.

We discuss the various lower bounds for $L^{BMA}(\mathbf{h}, \boldsymbol{\rho}, y)$ in various regions where $z = h_y - h_{y'}$, we use L_y^{BMA} as an abbreviation for $L^{BMA}(\mathbf{h}, \boldsymbol{\rho}, y)$

$$L_y^{BMA} \geq \begin{cases} (1+\mu) & \text{if } z \leq -\rho_{\hat{y}} - \mu \\ \\ (1+\mu) & \text{if } -\rho_{\hat{y}} - \mu \leq z \leq -\rho_{\hat{y}} - \mu^2 \\ \\ d(1+\mu) + (\mu - z - \rho_{\hat{y}})\tfrac{1-d}{\mu} & \text{if } -\rho_{\hat{y}} - \mu^2 \leq z \leq -\rho_{\hat{y}} + \mu^2 \\ \\ d(1+\mu) + (\mu - z - \rho_{\hat{y}})\tfrac{1-d}{\mu} & \text{if } -\rho_{\hat{y}} + \mu^2 \leq z \leq -\rho_{\hat{y}} + \mu \\ \\ (1+\mu)d & \text{if } \rho_{\hat{y}} - \mu \leq z \leq 0 \\ (1+\mu)d & \text{if } 0 \leq z \leq \rho_{\hat{y}} - \mu \\ (1+\mu)d & \text{if } \rho_{\hat{y}} - \mu \leq z \leq \rho_{\hat{y}} - \mu^2 \\ (\rho_{\hat{y}} + \mu - z)\dfrac{d}{\mu} & \text{if } \rho_{\hat{y}} - \mu^2 \leq z \leq \rho_{\hat{y}} + \mu^2 \\ (\rho_{\hat{y}} + \mu - z)\dfrac{d}{\mu} & \text{if } \rho_{\hat{y}} + \mu^2 \leq z \leq \rho_{\hat{y}} + \mu \\ 0 & \text{if } z \geq \rho_{\hat{y}} + \mu \end{cases}$$

and similarly, the various values $L^{BMA}(\mathbf{h}, \boldsymbol{\rho}, \tilde{y})$ can take in various regions where $z' = h_{y'} - h_y$, we use $L_{\tilde{y}}^{BMA}$ as an abbreviation for $L^{BMA}(\mathbf{h}, \boldsymbol{\rho}, \tilde{y})$

$$L_{\tilde{y}}^{BMA} \geq \begin{cases} (1+\mu) & \text{if } z \leq -\rho_{y'} - \mu \\ \\ (1+\mu) & \text{if } -\rho_{y'} - \mu \leq z \leq -\rho_{y'} - \mu^2 \\ \\ d(1+\mu) + (\mu - z - \rho_{y'})\tfrac{1-d}{\mu} & \text{if } -\rho_{y'} - \mu^2 \leq z \leq -\rho_{y'} + \mu^2 \\ \\ d(1+\mu) + (\mu - z - \rho_{y'})\tfrac{1-d}{\mu} & \text{if } -\rho_{y'} + \mu^2 \leq z \leq -\rho_{y'} + \mu \\ \\ (1+\mu)d & \text{if } \rho_{y'} - \mu \leq z \leq 0 \\ (1+\mu)d & \text{if } 0 \leq z' \leq \rho_{y'} - \mu \\ (1+\mu)d & \text{if } \rho_{y'} - \mu \leq z' \leq \rho_{y'} - \mu^2 \\ (\rho_{y'} + \mu - z')\dfrac{d}{\mu} & \text{if } \rho_{y'} - \mu^2 \leq z' \leq \rho_{y'} + \mu^2 \\ (\rho_{y'} + \mu - z')\dfrac{d}{\mu} & \text{if } \rho_{y'} + \mu^2 \leq z' \leq \rho_{y'} + \mu \\ 0 & \text{if } z' \geq \rho_{y'} + \mu \end{cases}$$

We can also establish two properties,

$$p^\top L^{BMA}(\mathbf{e_u}) = (1 - p_u)(1 + \mu)$$
$$p^\top L^{BMA}(\mathbf{0}) = d(1 + \mu)$$

$\mathbf{e_y}$ is a vector in R^k with 1 in the y^{th} position and 0 elsewhere and 0 represents the 0 vector in R^k. The equality and inequality follow from following conditions. should be observed $\forall i$.

Case1 : $p_y \geq 0.5$ for some $y \in [k]$

where $y \in arg\max_t p^T L_d(t)$

Case 1a: $h \in \mathcal{H}_y^{\boldsymbol{\rho}}$

$$p^T L_d(y) - p^T L_d(y) = 0 \tag{4}$$

RHS of equation 3 is 0 hence the equation it's trivial.

Case 1b. $\boldsymbol{h} \in \mathcal{H}_{k+1}^{\boldsymbol{\rho}}$ and $z = h_{\hat{y}} - h_{y'} \leq \rho_{\hat{y}}$, $q = \sum_{i \in \arg\max_i h_i} p_i$. We can look at 3 cases individually based on the value of q under the condition Case 1b.

LHS.

A. $q = p_y$

$$\begin{aligned} p^T L^{BMA}(\boldsymbol{h}) - & p^T L^{BMA}(\mathbf{e_y}) \\ & \geq p_{\hat{y}} L^{BMA}(\mathbf{h}, \boldsymbol{\rho}, y) + (1 - p_{\hat{y}}) L^{BMA}(\mathbf{h}, \boldsymbol{\rho}, \tilde{y}) - L^{BMA}(\mathbf{e_y}) \end{aligned}$$

We use the values of $L^{BMA}(\mathbf{h}, \boldsymbol{\rho}, y)$ values for the condition when $z = h_{\hat{y}} - h_{y'} \leq \rho_{\hat{y}}$ and $L^{BMA}(\mathbf{h}, \boldsymbol{\rho}, \tilde{y})$ value when $z' = h_{y'} - h_{\hat{y}} \geq -\rho_{\hat{y}}$. Though $L^{BMA}(\mathbf{h}, \boldsymbol{\rho}, \tilde{y})$ isn't defined for $z' \geq -\rho_{\hat{y}}$, the minimum we can achieve is $d(1 + \mu)$ irrespective of the relationship of $\rho_{\hat{y}}$ and $\rho_{y'}$. Note that $L^{BMA}(\mathbf{h}, \boldsymbol{\rho}, \tilde{y})$ is only defined for $z' \geq 0$ and achieves minimum value of $d(1 + \mu)$ at $0 \leq z' \leq \rho_{y'} - \mu$.

$$\begin{aligned} p^T L^{BMA}(\boldsymbol{h}) - & p^T L^{BMA}(\mathbf{e_y}) \\ & \geq p_y d(1 + \mu) + (1 - p_y) d(1 + \mu) - L^{BMA}(\mathbf{e_y}) \\ & = p_y d(1 + \mu) + (1 - p_y) d(1 + \mu) - (1 - p_y)(1 + \mu) \\ & = (p_y d + d(1 - p_y) - (1 - p_y))(1 + \mu) \\ & = (1 + \mu)(p_y + d - 1) \end{aligned}$$

LHS.

B. $q < p_y$

$$\begin{aligned} p^T L^{BMA}(\mathbf{h}) - & p^T L^{BMA}(e_y) \\ & \geq q L^{BMA}(\mathbf{h}, \boldsymbol{\rho}, y) + (1 - q) L^{BMA}(\mathbf{h}, \boldsymbol{\rho}, \tilde{y}) - L^{BMA}(e_y) \\ & \geq q d(1 + \mu) + (1 - q) d(1 + \mu) - L^{BMA}(\mathbf{e_y}) \\ & = q d(1 + \mu) + (1 - q) d(1 + \mu) - (1 - p_y)(1 + \mu) \\ & = (q d + d(1 - q) - (1 - p_y))(1 + \mu) \\ & = (1 + \mu)(p_y + d - 1) \end{aligned}$$

LHS.
C. $q > p_{\hat{y}}$

$$\begin{aligned} p^T L^{BMA}(\boldsymbol{h}) - & p^T L^{BMA}(e_y) \\ & \geq q L^{BMA}(\mathbf{h}, \boldsymbol{\rho}, y) + (1-q) L^{BMA}(\mathbf{h}, \boldsymbol{\rho}, \tilde{y}) - L^{BMA}(e_y) \\ & \geq qd(1+\mu) + (1-q)d(1+\mu) - L^{BMA}(\mathbf{e_y}) \\ & = qd(1+\mu) + (1-q)d(1+\mu) - (1-p_y)(1+\mu) \\ & = (qd + d(1-q) - (1-p_y))(1+\mu) \\ & = (1+\mu)(p_y + d - 1) \end{aligned}$$

RHS.

$$\begin{aligned} p^T L_d(f_{\boldsymbol{\rho}}^{BMA}(\boldsymbol{h})) - \min_t p^T L_d(t) & = p^T L_d(k+1) - p^T L_d(\hat{y}) \\ & = d - (1 - p_{\hat{y}}) \end{aligned}$$

Thus,

$$\begin{aligned} p^T L^{BMA}(\boldsymbol{h}) - \inf_{h^* \in \mathbb{R}^k} & p^T L^{BMA}(\mathbf{h}^*) \geq \\ & (1+\mu)(p^T L_d(f_{\boldsymbol{\rho}}^{BMA}(\boldsymbol{h})) - \min_t p^T L_d(t)) \end{aligned} \tag{5}$$

Case 1c. $\boldsymbol{h} \in R^k - (\mathcal{H}_{\hat{y}}^{\boldsymbol{\rho}} \cup \mathcal{H}_{k+1}^{\boldsymbol{\rho}})$ and $h_{y'} - h_{\hat{y}} \geq \rho_{y'}$
LHS.

$$\begin{aligned} & p^T L^{BMA}(\mathbf{h}) - p^T L^{BMA}(e_y) \\ & \quad = p_{\tilde{y}} L^{BMA}(\mathbf{h}, \boldsymbol{\rho}, \tilde{y}) + (1 - p_{\tilde{y}}) L^{BMA}(\mathbf{h}, \boldsymbol{\rho}, y) - L^{BMA}(e_y) \\ & \quad \geq p_{\tilde{y}}(\mathbf{0}) + (1 - p_{\tilde{y}})(1+\mu) - (1-p_y)(1+\mu) \\ & \quad = (p_y - p_{\tilde{y}})(1+\mu) \end{aligned}$$

For $z = h_{\hat{y}} - h_{y'} < 0$, we would be incorrectly predicting an incorrect class. So in such a case, loss would be $1+\mu$ with respect to z.

RHS.

$$p^T L_d(f_{\boldsymbol{\rho}}^{BMA}(\boldsymbol{h})) - \min_t p^T L_d(t) = (1 - p_{\tilde{y}}) - (1 - p_{\hat{y}}) = p_{\hat{y}} - p_{\tilde{y}}$$

$$\begin{aligned} p^T L^{BMA}(\mathbf{h}) - \inf_{h^* \in \mathbb{R}^k} & p^T L^{BMA}(\mathbf{h}^*) \geq \\ & (1+\mu)(p^T L_d(f_{\boldsymbol{\rho}}^{BMA}(\boldsymbol{h})) - \min_t p^T L_d(t)) \end{aligned} \tag{6}$$

Case 2: $p_{\tilde{y}} < 0.5 \qquad \forall \tilde{y} \in [k]$
such that, $k+1 \in arg\min_t p^T l_t$
Case 2a: $\boldsymbol{h} \in \mathcal{H}_{k+1}^{\boldsymbol{\rho}}$ trivial since RHS = 0

$$p^T L_d(k+1) - p^T L_d(k+1) = 0$$

Case 2b: $\boldsymbol{h} \in R^k - (\mathcal{H}^{\boldsymbol{\rho}}_{k+1})$ and $h_{\hat{y}} \geq h_{y'} + \rho_{\hat{y}}$

LHS.

$$\begin{aligned} & p^T L^{BMA}(\mathbf{h}) - p^T L^{BMA}(\mathbf{0}) \\ & = p_y L^{BMA}(\mathbf{h}, \boldsymbol{\rho}, y) + (1 - p_{\hat{y}}) L^{BMA}(\mathbf{h}, \boldsymbol{\rho}, \tilde{y}) - L^{BMA}(0) \\ & \geq p_y(0) + (1 - p_y)(1 + \mu) - d(1 + \mu) \\ & = (1 - d - p_y)(1 + \mu) \end{aligned}$$

RHS.

$$p^T L_d(f^{BMA}_{\boldsymbol{\rho}}(\boldsymbol{h})) - \min_t p^T L_d(t) = (1 - p_y) - d$$

$$\begin{aligned} p^T L^{BMA}(\mathbf{h}) - & \inf_{h^* \in \mathbb{R}^k} p^T L^{BMA}(\mathbf{h}^*) \geq \\ & (1 + \mu)(p^T L_d(f^{BMA}_{\boldsymbol{\rho}}(\boldsymbol{h})) - \min_t p^T L_d(t)) \end{aligned} \tag{7}$$

Thus, from equations eqn. 5, 6 and 7 we establish the theorem in eqn. 3

Theorem 2. *Let $H \subseteq \mathbb{R}^{\mathcal{X} \times \mathcal{Y}}$ be a hypothesis set with $\mathcal{Y} = \{1, \ldots, k\}$ and $\|\boldsymbol{\rho}\| \leq \tilde{\rho}$ i.e. $\{\rho_i \leq \tilde{\rho} : \forall \rho_i \in \boldsymbol{\rho}\}$. Then, for any $n \geq 1$,$q \geq 1$, $1 \leq p < \infty$, and any set $S = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$,$\delta > 0$, with probability at least $1 - \delta$, the following holds for all $h \in H$:*

$$\begin{aligned} R_{BMA}(h) \leq \widehat{R}_{BMA}(h) + & 2(1 + \mu)\sqrt{\frac{\log \frac{4}{\delta}}{2m}} + (1 + \mu)\sqrt{\frac{\log \frac{2}{\delta}}{2m}} \\ & + k\frac{\tilde{\rho}}{\sqrt{m}} + 2\left(\frac{1 - d}{\mu}\right) k^2 \beta \left(2U^{\left[\frac{1}{p^*} - \frac{1}{q}\right]_+}\right)^{(n-1)} \sqrt{\frac{\min\{p^*, 4\log(2D)\} \max_i \|\mathbf{x}_i\|^2_{p^*}}{m}} \end{aligned}$$

Proof. Let $\widetilde{H}$ be the family of hypotheses mapping $\mathcal{X} \times \mathcal{Y}$ to $\mathbb{R}$ defined by $\widetilde{H} = \{z = (x, y) \mapsto \phi_h(\mathbf{x}, y) : h \in H\}$ where $\phi_h(\mathbf{x}, y)$ is the margin function. Here $\phi_h(x, y)$ is defined as

$$\phi_h(\mathbf{x}, y) = h_y(\mathbf{x}) - \max_{y' \in k, y' \neq y} h_{y'}(\mathbf{x}_i) - \rho_y$$

where $y = \arg\max_{r \in k} h_r(x_i)$.

Consider the family of functions $\widetilde{\mathcal{H}} = \left\{L^{BMA}(c) : c \in \widetilde{H}\right\}$ derived from H, which can take values in [0,1+μ] . By lemma 4 , with probability at least $1 - \delta$, for all $h \in H$,

$$\begin{aligned} R^{BMA}(h, \boldsymbol{\rho}) \leq R^{BMA}(h, \boldsymbol{\rho}) + & 2\mathbb{L}\mathfrak{R}_m(\tilde{H}) \\ & + 2(b - a)\sqrt{\frac{\ln\left(\frac{4}{\delta}\right)}{2m}} + (b - a)\sqrt{\frac{ln\left(\frac{2}{\delta}\right)}{m}} \end{aligned}$$

We also use the Talagrand's contraction lemma without absolute values :reference-2011 paper: that is $\mathfrak{R}_m\left(L^{BMA}(\widetilde{H})\right) \leq \mathbb{L}\mathfrak{R}_m(\tilde{H})$ using the $\mathbb{L}$-Lipschitzness of L^{BMA}.

Here, $\mathfrak{R}_m(\widetilde{H})$ can be upper bounded as follows:

$$\begin{aligned}\mathfrak{R}_m(\widetilde{H}) &= \frac{1}{m}\mathop{\mathrm{E}}_{S,\sigma}\left[\sup_{h\in H}\left|\sum_{i=1}^{m}\sigma_i\phi_{\mathbf{h}}\left(\mathbf{x}_i, y_i\right)\right|\right]\\ &= \frac{1}{m}\mathop{\mathrm{E}}_{S,\sigma}\left[\sup_{h\in H}\left|\sum_{i=1}^{m}\sum_{y\in\mathcal{Y}}\sigma_i\phi_{\mathbf{h}}\left(\mathbf{x}_i, y\right)1_{y=y_i}\right|\right]\end{aligned}$$

Using Subadditivity of supremum,

$$\begin{aligned}\mathfrak{R}_m(\widetilde{H}) &\leq \frac{1}{m}\mathop{\mathrm{E}}_{S,\sigma}\left[\sup_{h\in H}\sum_{y\in\mathcal{Y}}\left|\sum_{i=1}^{m}\sigma_i\phi_{\mathbf{h}}\left(\mathbf{x}_i, y\right)1_{y=y_i}\right|\right]\\ &= \frac{1}{m}\sum_{y\in\mathcal{Y}}\mathop{\mathrm{E}}_{S,\sigma}\left[\left|\sup_{h\in H}\sum_{i=1}^{m}\sigma_i\phi_{\mathbf{h}}\left(\mathbf{x}_i, y\right)1_{y=y_i}\right|\right]\\ &= \frac{1}{m}\sum_{y\in\mathcal{Y}}\mathop{\mathrm{E}}_{S,\sigma}\left[\left|\sup_{h\in H}\sum_{i=1}^{m}\sigma_i\phi_{\mathbf{h}}\left(\mathbf{x}_i, y\right)\left(\frac{2\left(1_{y=y_i}\right)-1}{2}+\frac{1}{2}\right)\right|\right]\end{aligned}$$

We define, $\epsilon_i = 2(1_{y=y_i}) - 1$ and get,

$$\begin{aligned}\mathfrak{R}_m(\widetilde{H}) &\leq \frac{1}{2m}\sum_{y\in\mathcal{Y}}\mathop{\mathrm{E}}_{S,\sigma}\left[\left|\sup_{h\in H}\sum_{i=1}^{m}\sigma_i\epsilon_i\phi_{\mathbf{h}}\left(\mathbf{x}_i, y\right)\right|\right]\\ &\qquad + \frac{1}{2m}\sum_{y\in\mathcal{Y}}\mathop{\mathrm{E}}_{S,\sigma}\left[\left|\sup_{h\in H}\sum_{i=1}^{m}\sigma_i\phi_{\mathbf{h}}\left(\mathbf{x}_i, y\right)\right|\right]\\ &= \frac{1}{m}\sum_{y\in\mathcal{Y}}\mathop{\mathrm{E}}_{S,\sigma}\left[\sup_{h\in H}\left|\sum_{i=1}^{m}\sigma_i\phi_{\mathbf{h}}\left(\mathbf{x}_i, y\right)\right|\right]\end{aligned}$$

where by definition $\epsilon_i \in \{-1, +1\}$ and we use the fact that σ_i and $\sigma_i\epsilon_i$ have the same distribution. We define $\Psi(H)$ for any hypotheses H for the mapping $\mathcal{X} \times \mathcal{Y}$ to $\mathbb{R}$ as $\Psi(H) = \{\mathbf{x} \mapsto h_y(\mathbf{x}) : y \in \mathcal{Y}, h \in H\}$.

Then we define, $\Psi(H)^{(k-1)} = \{\max\{h_1, \ldots, h_{k-1}\} : h_i \in \Psi(H), i \in [1, k-1]\}$. Now, rewriting $\phi_{\mathbf{h}}\left(\mathbf{x}_i, y\right)$ explicitly and using again the sub-additivity of sup. We follow this with observing that $-\sigma_i$ and σ_i are distributed in the same way.

$$\begin{aligned}
\mathfrak{R}_m(\widetilde{H}) \leq \frac{1}{m} \sum_{y \in \mathcal{Y}} \underset{S,\sigma}{\mathrm{E}} \Bigg[& \sup_{h \in H} \sup_{\rho \in \tilde{\rho}} \\
& \left| \sum_{i=1}^{m} \sigma_i \left(h_y\left(x_i\right) - \max_{y' \neq y} h_{y'}\left(x_i\right) - \rho_y \right) \right| \Bigg] \\
\leq \sum_{y \in \mathcal{Y}} \Bigg[& \frac{1}{m} \underset{S,\sigma}{\mathrm{E}} \left[\sup_{\mathbf{h} \in H} \left| \sum_{i=1}^{m} \sigma_i h_y\left(\mathbf{x}_i\right) \right| \right] \\
& + \frac{1}{m} \underset{S,\sigma}{\mathrm{E}} \left[\sup_{\mathbf{h} \in H} \left| \sum_{i=1}^{m} -\sigma_i \max_{y' \neq y} h_{y'}\left(\mathbf{x}_i\right) \right| \right] \\
& + \frac{1}{m} \underset{S,\sigma}{\mathrm{E}} \left[\sup_{\mathbf{h} \in H} \sup_{\rho \in \tilde{\rho}} \left| \sum_{i=1}^{m} -\sigma_i \rho_y \right| \right] \Bigg] \\
= \sum_{y \in \mathcal{Y}} \Bigg[& \frac{1}{m} \underset{S,\sigma}{\mathrm{E}} \left[\sup_{\mathbf{h} \in H} \left| \sum_{i=1}^{m} \sigma_i h_y\left(\mathbf{x}_i\right) \right| \right] \\
& + \frac{1}{m} \underset{S,\sigma}{\mathrm{E}} \left[\sup_{\mathbf{h} \in H} \left| \sum_{i=1}^{m} \sigma_i \max_{y' \neq y} h_{y'}\left(\mathbf{x}_i\right) \right| \right] \\
& + \frac{1}{m} \underset{S,\sigma}{\mathrm{E}} \left[\sup_{\mathbf{h} \in H} \sup_{\rho \in \tilde{\rho}} \left| \sum_{i=1}^{m} \sigma_i \rho_y \right| \right] \Bigg]
\end{aligned}$$

We now use lemma 3 to get,

$$\begin{aligned}
\mathfrak{R}_m(\widetilde{H}) &\leq \sum_{y \in \mathcal{Y}} \Bigg[\frac{1}{m} \underset{S,\sigma}{\mathrm{E}} \left[\sup_{\mathbf{h} \in \Psi(H)} \left| \sum_{i=1}^{m} \sigma_i h\left(\mathbf{x}_i\right) \right| \right] \\
&\quad + \frac{1}{m} \underset{S,\sigma}{\mathrm{E}} \left[\sup_{h \in \Psi(H)^{(k-1)}} \left| \sum_{i=1}^{m} \sigma_i h\left(x_i\right) \right| \right] + \frac{1}{m} \tilde{\rho} \sqrt{m}(1) \Bigg] \\
&\leq k \left[\frac{k}{m} \underset{S,\sigma}{\mathrm{E}} \left[\sup_{h \in \Psi(H)} \left| \sum_{i=1}^{m} \sigma_i h\left(x_i\right) \right| \right] + \frac{\tilde{\rho}}{\sqrt{m}} \right] \\
&= k^2 \mathfrak{R}_m\left(\Psi(H)\right) + k \frac{\tilde{\rho}}{\sqrt{m}}
\end{aligned}$$

In case of neural networks, $\Psi(H)$ would be

$$\mathcal{N}^{n,U}_{\beta_{p,q} \leq \beta}(\mathbf{x}) = \mathbf{w}^T \left(\Phi\left(W_{n-1} \Phi\left(W_{n-2}\left(\ldots \Phi\left(W_1 \mathbf{x} \right) \right) \right) \right) \right)$$

where n is the no of layers, U is the width of the hidden layer, Φ is ReLU activation function and $\beta_{p,q}(W) = \prod_{k=1}^{n} \|W_k\|_{p,q} \leq \beta$ that is product of weights in each layer which is upper bounded by β.

The upper bound on rademachar complexity for the neural network $\mathfrak{R}_m\left(\mathcal{N}^{n,U}_{\beta_{p,q}\leq\beta}(\mathbf{x})\right)$ is presented in Lemma 7. We use the result of Lemma 7, to obtain,

$$\mathfrak{R}_m(\widetilde{H}) \leq k\frac{\tilde{\rho}}{\sqrt{m}} + k^2\beta\left(2U^{\left[\frac{1}{p^*}-\frac{1}{q}\right]_+}\right)^{(d-1)} \sqrt{\frac{\min\left\{p^*, 4\log(2D)\right\}\max_i \|x_i\|^2_{p^*}}{m}}$$

The lipschitz constant can be found out by looking at the maximum gradient of the loss function for different intervals presented in equation of the margin function. Thus we get lipschitz constant as,

$$\mathbb{L} = \max\left\{\frac{d}{\mu}, \frac{1-d}{\mu}\right\} = \frac{1-d}{\mu}$$

Now we use the result of lemma 4 to get the generalisation bounds on CDAN with input independent ρ. We get,

$$R^{BMA}(h,\rho) \leq \widehat{R}^{BMA}(h,\rho) + 2(1+\mu)\sqrt{\frac{\log\frac{4}{\delta}}{2m}} + (1+\mu)\sqrt{\frac{\log\frac{2}{\delta}}{2m}} + k\frac{\tilde{\rho}}{\sqrt{m}} + 2\left(\frac{1-d}{\mu}\right)k^2\beta\left(2U^{\left[\frac{1}{p^*}-\frac{1}{q}\right]_+}\right)^{(n-1)}\left(\sqrt{\frac{\min\left\{p^*, 4\log(2D)\right\}\max_i \|\mathbf{x}_i\|^2_{p^*}}{m}}\right)$$

where $b - a = 1 + \mu$, since $L_{BMA} \in \{0, 1+\mu\}$. □

Lemma 3. *Let $\mathcal{F}_1, \ldots, \mathcal{F}_l$ be l hypothesis sets in $\mathbb{R}^{\mathcal{X}}$, $l \geq 1$, and let $\mathcal{G} = \{\max\{h_1, \ldots, h_l\} : h_i \in \mathcal{F}_i, i \in [1,l]\}$. Then, for any sample S of size m, the empirical Rademacher complexity of $\mathcal{G}$ can be upper bounded as follows:*

$$\widehat{\mathfrak{R}}_S(\mathcal{G}) \leq \sum_{j=1}^{l} \widehat{\mathfrak{R}}_S\left(\mathcal{F}_j\right)$$

Proof. Let $S = (x_1, \ldots, x_m)$ be a sample of size m. We first prove the result in the case $l = 2$. By definition of the max operator, for any $h_1 \in \mathcal{F}_1$ and $h_2 \in \mathcal{F}_2$,

$$\max\{h_1, h_2\} = \frac{1}{2}\left[h_1 + h_2 + |h_1 - h_2|\right]$$

Thus we can write

$$\begin{aligned}\widehat{\mathfrak{R}}_S(\mathcal{G}) &= \frac{1}{m}\underset{\boldsymbol{\sigma}}{\mathrm{E}}\left[\sup_{\substack{h_1\in\mathcal{F}_1\\h_2\in\mathcal{F}_2}}\sum_{i=1}^{m}\sigma_i\max\left\{h_1\left(x_i\right), h_2\left(x_i\right)\right\}\right]\\ &= \frac{1}{2m}\underset{\boldsymbol{\sigma}}{\mathrm{E}}\left[\sup_{\substack{h_1\in\mathcal{F}_1\\h_2\in\mathcal{F}_2}}\sum_{i=1}^{m}\sigma_i\left(h_1\left(x_i\right) + h_2\left(x_i\right) + \left|\left(h_1 - h_2\right)\left(x_i\right)\right|\right)\right]\\ &\leq \frac{1}{2}\widehat{\mathfrak{R}}_S\left(\mathcal{F}_1\right) + \frac{1}{2}\widehat{\mathfrak{R}}_S\left(\mathcal{F}_2\right)\end{aligned}$$

$$+\frac{1}{2m}\underset{\boldsymbol{\sigma}}{\mathrm{E}}\left[\sup_{\substack{h_1\in\mathcal{F}_1\\h_2\in\mathcal{F}_2}}\sum_{i=1}^{m}\sigma_i\left|\left(h_1-h_2\right)\left(x_i\right)\right|\right] \tag{8}$$

Now using the fact that $x \mapsto |x|$ is 1-Lipschitz, by Talagrand's contraction lemma, the last term can be bounded as follows

$$\frac{1}{2m}\underset{\boldsymbol{\sigma}}{\mathrm{E}}\left[\sup_{h_1\in\mathcal{F}_1}\sum_{i=1}^{m}\sigma_i\left|\left(h_1-h_2\right)\left(x_i\right)\right|\right]$$

$$\leq \frac{1}{2m}\underset{\boldsymbol{\sigma}}{\mathrm{E}}\left[\sup_{\substack{h_1\in\mathcal{F}_1\\h_2\in\mathcal{F}_2}}\sum_{i=1}^{m}\sigma_i\left(h_1-h_2\right)\left(x_i\right)\right]$$

Using the sub-additivity of supremum we get,

$$\leq \frac{1}{2}\widehat{\mathfrak{R}}_S\left(\mathcal{F}_1\right)+\frac{1}{2m}\underset{\boldsymbol{\sigma}}{\mathrm{E}}\left[\sup_{h_2\in\mathcal{F}_2}\sum_{i=1}^{m}-\sigma_i h_2\left(x_i\right)\right]$$

We also use the fact that σ_i and $-\sigma_i$ have the same distribution for any $i \in [1, m]$ to get,

$$=\frac{1}{2}\widehat{\mathfrak{R}}_S\left(\mathcal{F}_1\right)+\frac{1}{2}\widehat{\mathfrak{R}}_S\left(\mathcal{F}_2\right) \tag{9}$$

Combining eq. 8 and 9 yields $\widehat{\mathfrak{R}}_S(\mathcal{G}) \leq \widehat{\mathfrak{R}}_S\left(\mathcal{F}_1\right)+\widehat{\mathfrak{R}}_S\left(\mathcal{F}_2\right)$. The general case is an extension of the case with $l = 2$ using $\max\{h_1,\ldots,h_l\} = \max\{h_1, \max\{h_2,\ldots,h_l\}\}$ that leads to a recurrence. □

Lemma 4. *Let* $\mathrm{Y} \subseteq \mathbb{R}$, *and* let $\mathcal{F} \subseteq [a,b]^{\mathcal{X}}$ *for some* $a \leq b$. *Let* $\ell : \mathrm{Y} \times [a,b] \to [0,B]$ *be such that* $\ell(y,\hat{y})$ *is* $\mathbb{L}$ *-Lipschitz in its second argument for some* $\mathbb{L} > 0$. *Let* P *be any probability distribution on* $\mathcal{X} \times \mathrm{Y}$, *with marginal* μ *on* $\mathcal{X}$. *If* f *is selected from* $\mathcal{F}$, *then for any* $0 < \delta \leq 1$, *with probability at least* $1-\delta$ *(over* $S \sim D^m$*)*

$$R_{BMA}(f,\rho) \leq \hat{R}_{BMA}(f,\rho) + 2\mathbb{L}\mathfrak{R}_m(\mathcal{F}) + \\ 2(b-a)\sqrt{\frac{\ln\left(\frac{4}{\delta}\right)}{2m}} + (b-a)\sqrt{\frac{ln\left(\frac{2}{\delta}\right)}{m}}$$

Proof. First we define

$$\hat{\mathfrak{R}}_m(\mathcal{F}) = \mathbb{E}_{\{\sigma\in\pm1\}}\left[\sup_{f\in\mathcal{F}}\frac{1}{m}\sum_{i=1}^{m}\sigma_i f(\mathbf{x}_i)\right]$$

where σ is the Rademachar variable and $R_m(\mathcal{F})$ is defined as expectation over data samples of size m obtained in an i.i.d fashion from probability distribution μ i.e.

$$\mathfrak{R}_m(\mathcal{F}) = \mathbb{E}_{\mathbf{x}^m\sim\mu^m}\left[\hat{R}_m(\mathcal{F})\right]$$

We directly use the result from the [1] and using the results directly with probability atleast $1-\delta$, we bound the generalization error as,

$$R_{BMA}(f,\rho) \leq \hat{R}_{BMA}(f,\rho) + 2L\mathfrak{R}_m(\mathcal{F}) + 2(b-a)\sqrt{\frac{\ln\left(\frac{2}{\delta}\right)}{2m}} \tag{10}$$

Now for any set $\mathcal{S} = \{\mathbf{x}_1, \mathbf{x}_2...., \mathbf{x}_m\}$, and a function $\Phi : \mathcal{X}^m \rightarrow \mathbb{R}$ such that $\Phi(\mathbf{x}_1, \mathbf{x}_2, ...\mathbf{x}_m) = \hat{\mathfrak{R}}_m(\mathcal{F})$. Hence, $\mathfrak{R}_m(\mathcal{F}) = \mathbf{E}_{\mathbf{x}^m \sim \mu^m}\left[\Phi\left(\mathbf{x}_1, \ldots, \mathbf{x}_m\right)\right]$

Then, for any $j \in [m]$, and any $\mathbf{x}_1, \ldots, \mathbf{x}_m, \mathbf{x}_j' \in \mathcal{X}$

$$\begin{aligned}
&\left|\Phi\left(\mathbf{x}_1, \ldots, \mathbf{x}_j, \ldots, \mathbf{x}_m\right) - \Phi\left(\mathbf{x}_1, \ldots, \mathbf{x}_j', \ldots, \mathbf{x}_m\right)\right| \\
&= \hat{\mathfrak{R}}_m(\mathcal{F}) - \mathfrak{R}_{\left(\mathbf{x}_1, \ldots, \mathbf{x}_j', \ldots, \mathbf{x}_m\right)}(\mathcal{F}) \\
&= \mathbf{E}_{\sigma \in \{\pm 1\}^m}\left[\sup_{f \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^m \sigma_i f\left(\mathbf{x}_i\right)\right. \\
&\qquad\qquad \left. - \sup_{f \in \mathcal{F}}\left(\frac{1}{m} \sum_{i \neq j} \sigma_i f\left(\mathbf{x}_i\right) + \frac{1}{m} \sigma_j f\left(\mathbf{x}_j'\right)\right)\right] \\
&\leq \frac{b-a}{m}
\end{aligned}$$

Thus by McDarmid's inequality we get,

$$\mathbf{P}\left[\hat{\mathfrak{R}}_m(\mathcal{F}) - \mathfrak{R}_m(\mathcal{F}) \geq \epsilon\right] \leq e^{-2m\epsilon^2/(b-a)^2} \tag{11}$$

Now with probability atleast $1 - \frac{\delta}{2}$ we get,

$$\hat{\mathfrak{R}}_m(\mathcal{F}) - \mathfrak{R}_m(\mathcal{F}) \leq 2(b-a)\sqrt{\frac{\ln \frac{2}{\delta}}{2m}} \tag{12}$$

We also know that, a relationship exits between Using the combination of eqn. (12) and eqn. (10) each holding with a probability of atleast $1-\frac{\delta}{2}$, we get with a probability atleast $1-\delta$,

$$\begin{aligned}
R_{BMA}(f,\rho) \leq \hat{R}_{BMA}(f,\rho) &+ 2L\hat{\mathfrak{R}}_m(\mathcal{F}) \\
&+ 2(b-a)\sqrt{\frac{\ln\left(\frac{4}{\delta}\right)}{2m}} + (b-a)\sqrt{\frac{\ln\left(\frac{2}{\delta}\right)}{m}}
\end{aligned}$$

□

For completeness we provide the upper bounds on rademachar complexity for linear as well as d-layered neural networks. The upper bounds presented here are provided in [2]. First we present the paper rademachar complexity of linear separators with bounded ℓ_p norm.

Lemma 5. *For any d,q ≥ 1, for any $1 \geq p \geq 2$, the rademachar complexity for a linear predictor(single layer neural network) is given as,*

$$\mathfrak{R}_m\left(\mathcal{N}^1_{\beta_p,q\leq\beta}\right) \leq \sqrt{\frac{\beta^2 \min\{p^*, 4\log(2D)\} \max_i \|\mathbf{x}_i\|^2_{p^*}}{m}}$$

for any $2 < p < \infty$,

$$\mathfrak{R}_m\left(\mathcal{N}^1_{\beta_{p,q}\leq\beta}\right) \leq \frac{\sqrt{2}\beta\|X\|_{2,p^*}}{m} \leq \frac{\sqrt{2}\beta \max_i \|\mathbf{x}_i\|_{p^*}}{m^{\frac{1}{p}}} \tag{13}$$

where $\frac{1}{p} + \frac{1}{p*} = 1$.

Proof. The N^1 represents all linear class of functions and hence for any function

$$\begin{aligned}
\mathcal{R}_m\left(\mathcal{N}^1_{\beta_{p,q}\leq\beta}\right) &= \mathbb{E}_{\sigma\in\{\pm1\}^m}\left[\frac{1}{m}\sup_{\|w\|_p\leq\beta}\left|\sum_{i=1}^m \sigma_i w^\top \mathbf{x}_i\right|\right] \\
&= \mathbb{E}_{\sigma\in\{\pm1\}^m}\left[\frac{1}{m}\sup_{\|w\|_p\leq\beta}\left|w^\top\sum_{i=1}^m \sigma_i \mathbf{x}_i\right|\right] \\
&= \beta\mathbb{E}_{\sigma\in\{\pm1\}^m}\left[\frac{1}{m}\left\|\sum_{i=1}^m \sigma_i \mathbf{x}_i\right\|_{p^*}\right] \\
&= \beta\mathbb{E}_{\sigma\in\{\pm1\}^m}\left[\frac{1}{m}\left\|\sum_{i=1}^m \sigma_i \mathbf{x}_i\right\|_{p^*}\right]
\end{aligned}$$

For $1 \leq p \leq \min\{2, \frac{2log(2D)}{2log(2D)-1}\}$ and thus $2log(2D) \leq p^*$

$$\begin{aligned}
\mathcal{R}_m\left(\mathcal{N}^1_{\beta_{p,q}\leq\beta}\right) &\leq D^{\frac{1}{p^*}}\beta\mathbb{E}_{\sigma\in\{\pm1\}^m}\left[\frac{1}{m}\left\|\sum_{i=1}^m \sigma_i \mathbf{x}_i\right\|_\infty\right] \\
&\leq D^{\frac{1}{2\log(2D)}}\beta\mathbb{E}_{\sigma\in\{\pm1\}^m}\left[\frac{1}{m}\left\|\sum_{i=1}^m \sigma_i \mathbf{x}_i\right\|_\infty\right] \\
&\leq \sqrt{2}\beta\mathbb{E}_{\sigma\in\{\pm1\}^m}\left[\frac{1}{m}\left\|\sum_{i=1}^m \sigma_i \mathbf{x}_i\right\|_\infty\right]
\end{aligned}$$

The next step is to use lemma 8 viewing each feature as a member of finite hypothesis class.

$$\begin{aligned}
\mathcal{R}_m\left(\mathcal{N}^1_{\beta_{p,q}\leq\beta}\right) &\leq \sqrt{2}\beta\mathbb{E}_{\sigma\in\{\pm1\}^m}\left[\frac{1}{m}\left\|\sum_{i=1}^{m}\sigma_i\mathbf{x}_i\right\|_\infty\right] \\
&\leq 2\beta\frac{\sqrt{\log(2D)}}{m}\max_{j=1\dots,D}\|(\mathbf{x}_i[j])_{i=1}^m\|_2 \\
&\leq 2\beta\sqrt{\frac{\log(2D)}{m}}\max_{i=1,\dots,m}\|\mathbf{x}_i\|_\infty \\
&\leq 2\beta\sqrt{\frac{\log(2D)}{m}}\max_{i=1,\dots,m}\|\mathbf{x}_i\|_{p^*}
\end{aligned}$$

Thus,

$$\mathfrak{R}_m\left(\mathcal{N}^1_{\beta_{p,q}\leq\beta}\right) \leq \sqrt{\frac{\beta^2\left(4\log(2D)\right)\max_i\|\mathbf{x}_i\|_{p^*}^2}{m}}$$

If $\min\left\{2,\frac{2\log(2D)}{2\log(2D)-1}\right\} < p < \infty$, by Khintchine-Kahane inequality we have

$$\begin{aligned}
\mathcal{R}_m\left(\mathcal{N}^1_{\beta_p,q\leq\beta}\right) &= \beta\mathbb{E}_{\sigma\in\{\pm1\}^m}\left[\frac{1}{m}\left\|\sum_{i=1}^{m}\sigma_i\mathbf{x}_i\right\|_{p^*}\right] \\
&\leq \beta\frac{1}{m}\left(\sum_{j=1}^{D}\mathbb{E}_{\sigma\in\{\pm1\}^m}\left[|\sum_{i=1}^{m}\sigma_i\mathbf{x}_i[j]^{p^*}\right]\right)^{1/p^*} \\
&\leq \beta\frac{\sqrt{p^*}}{m}\left(\sum_{j=1}^{D}\|(\mathbf{x}_i[j])_{i=1}^m\|_2^{p^*}\right)^{1/p^*} \\
&= \beta\frac{\sqrt{p^*}}{m}\|X\|_{2,p^*}
\end{aligned}$$

If $p^* \geq 2$, by Minskowski inequality we have that $\|X\|_{2,p^*} \leq m^{1/2}\max_i\|\mathbf{x}_i\|_{p^*}$. Otherwise, by subadditivity of the function $f(z) = z^{\frac{p^*}{2}}$, we get $\|X\|_{2,p^*} \leq m^{1/p^*}\max_i\|\mathbf{x}_i\|_{p^*}$. Thus, at $p^* = 2$,

$$\mathfrak{R}_m\left(\mathcal{N}^1_{\beta_{p,q}\leq\beta}\right) \leq \frac{\sqrt{2}\beta\|X\|_{2,p^*}}{m}$$

and for $p^* > 2$, we get

$$\mathfrak{R}_m\left(\mathcal{N}^1_{\beta_{p,q}\leq\beta}\right) \leq \frac{\sqrt{2}\beta\max_i\|\mathbf{x}_i\|_{p^*}}{m^{\frac{1}{p}}}$$

This also establishes that,

$$\mathfrak{R}_m\left(\mathcal{N}^1_{\beta_{p,q}\leq\beta}\right) \leq \frac{\sqrt{2}\beta\|X\|_{2,p^*}}{m} \leq \frac{\sqrt{2}\beta\max_i\|\mathbf{x}_i\|_{p^*}}{m^{\frac{1}{p}}}$$

□

Lemma 6. *For any p, $q \geq 1$, $n \geq 2$, $\sigma \in \{\pm 1\}^m$ and $f \in \mathcal{N}^{n,H,H}$*

$$\begin{aligned}\sup_{W} \frac{1}{\|W\|_{p,q}} \left\| \sum_{i=1}^{m} \sigma_i \left\| \left[W \left[f\left(\mathbf{x}_i\right)\right]_+ \right]_+ \right\|_{p'} \right\|_{p'} = \\ U^{\left[\frac{1}{p'}-\frac{1}{q}\right]_+} \sup_{\mathbf{w}} \frac{1}{\|\mathbf{w}\|_p} \left| \sum_{i=1}^{m} \sigma_i \left\| \left[\mathbf{w}^\top \left[f\left(\mathbf{x}_i\right)\right]_+ \right]_+ \right\|_{p'} \right|\end{aligned}$$

where n is the depth of the network, U is the height of the layer and U is the no. of outputs.

Proof.

$$g(\mathbf{w}) = \left| \sum_{i=1}^{m} \sigma_i \|\mathbf{w}^\top \left[f(\mathbf{x}_i)\right]_+\|_{p'} \right|$$

We define $\mathbf{w}^*$ as

$$\mathbf{w}^* = \arg\max_{\mathbf{w}} \frac{g(\mathbf{w})}{\|\mathbf{w}\|_p}$$

Thus,

$$g(V_i) = \left| \sum_{i=1}^{m} \sigma_i \|V_i^\top \left[f(\mathbf{x}_i)\right]_+\|_{p'} \right|$$

where V_i is row of any matrix V. Now we know that,

$$\begin{aligned}\frac{g(\mathbf{w}^*)}{\|\mathbf{w}^*\|_p} &\geq \frac{g(V_i)}{\|V_i\|_p} \\ \left(\frac{g(\mathbf{w}^*)}{\|\mathbf{w}^*\|_p}\right)^{p'} &\geq \left(\frac{g(V_i)}{\|V_i\|_p}\right)^{p'} \\ \left(\frac{g(\mathbf{w}^*)\|V_i\|_p}{\|\mathbf{w}^*\|_p}\right)^{p'} &\geq \big(g(V_i)\big)^{p'} \\ \sum_{i=1}^{H} \left(\frac{g(\mathbf{w}^*)\|V_i\|_p}{\|\mathbf{w}^*\|_p}\right)^{p'} &\geq \sum_{i=1}^{H} \big(g(V_i)\big)^{p'} \\ \left(\frac{g(\mathbf{w}^*)}{\|\mathbf{w}^*\|_p}\right)^{p'} \sum_{i=1}^{H} \|V_i\|_p^{p'} &\geq \sum_{i=1}^{H} \big(g(V_i)\big)^{p'} \\ \frac{g(\mathbf{w}^*)}{\|\mathbf{w}^*\|_p} \left(\sum_{i=1}^{H} \|V_i\|_p^{p'}\right)^{\frac{1}{p'}} &\geq \left(\sum_{i=1}^{H} \big(g(V_i)\big)^{p'}\right)^{\frac{1}{p'}} \\ \frac{g(\mathbf{w}^*)}{\|\mathbf{w}^*\|_p} &\geq \frac{\|g(V)\|_{p'}}{\|V\|_{p,p'}} \qquad (14)\end{aligned}$$

We have 2 cases now, $q > p'$ and $q < p'$. If $q < p'$ $\|V\|_{p,p'} \leq \|V\|_{p,q}$ and $H^{[\frac{1}{p'}-\frac{1}{q}]_+} = 1$. Thus,

$$\frac{g(\mathbf{w}^*)}{\|\mathbf{w}^*\|_p} \geq \frac{\|g(V)\|_{p'}}{\|V\|_{p,q}}$$

We also know that $\|V\|_{p,p'} \leq U^{[\frac{1}{p'}-\frac{1}{q}]}\|V\|_{p,q}$ and thus,

$$\frac{\|g(V)\|_{p'}}{\|V\|_{p,q}} \leq U^{[\frac{1}{p'}-\frac{1}{q}]}\frac{\|g(V)\|_{p'}}{\|V\|_{p,p'}}$$

And from eqn.(14) we get,

$$\begin{aligned}\frac{g(\mathbf{w}^*)}{\|\mathbf{w}^*\|_p} \geq \frac{\|g(V)\|_{p'}}{\|V\|_{p,p'}} &\geq \frac{\|g(V)\|_{p'}}{U^{[\frac{1}{p'}-\frac{1}{q}]}\|V\|_{p,q}} \\ U^{[\frac{1}{p'}-\frac{1}{q}]}\frac{g(\mathbf{w}^*)}{\|\mathbf{w}^*\|_p} &\geq \frac{\|g(V)\|_{p'}}{\|V\|_{p,q}}\end{aligned}$$

The LHS of the lemma is greater than RHS is true for any given vector $\mathbf{w}$, not $\mathbf{w}^*$ in the RHS. Also, the equality exists when W matrix contains $\mathbf{w}^*$ as all of its rows. □

Lemma 7. *The rademachar complexity for d layered network*

Proof. Following the definition of Rademachar Complexity,

$$\begin{aligned}\mathcal{R}_m\left(\mathcal{N}^{d,U}_{\beta_{p,q}\leq\beta}\right) &= \mathbb{E}_\sigma\left[\frac{1}{m}\sup_{f\in\mathcal{N}^{d,U}_{\beta_{p,q}\leq\beta}}\left|\sum_{i=1}^{m}\sigma_i f\left(x_i\right)\right|\right] \\ &= \mathbb{E}_\sigma\left[\frac{1}{m}\sup_{f\in\mathcal{N}^{d,U}}\frac{\beta}{\beta_{p,q}(f)}\left|\sum_{i=1}^{m}\sigma_i f\left(x_i\right)\right|\right]\end{aligned}$$

On expanding our $f(x_i)$ we get,

$$\begin{aligned}&= \mathbb{E}_\sigma\left[\frac{1}{m}\sup_{g\in\mathcal{N}^{d-1},U,U}\sup_{w}\frac{\beta}{\beta_{p,q}(g)\|w\|_p}\left|\sum_{i=1}^{m}\sigma_i w^\top\left[g\left(x_i\right)\right]_+\right|\right] \\ &= \mathbb{E}_\sigma\left[\frac{1}{m}\sup_{g\in\mathcal{N}^{d-1,U,U}}\frac{\beta}{\beta_{p,q}(g)}\left\|\sum_{i=1}^{m}\sigma_i\left[g\left(x_i\right)\right]_+\right\|_{p^*}\right]\end{aligned}$$

We use Lemma 6,to get

$$\begin{aligned}
&= \mathbb{E}_\sigma \Bigg[\frac{1}{m} \sup_{h \in \mathcal{N}^{d-2,U,U}} \frac{\beta}{\beta_{p,q}(h)} \\
&\qquad \sup_W \frac{1}{\|W\|_{p,q}} \left\| \sum_{i=1}^m \sigma_i \left[W \left[h\left(x_i\right) \right]_+ \right]_+ \right\|_{p^*} \Bigg] \\
&= U^{\left[\frac{1}{p^*}-\frac{1}{q}\right]_+} \mathbb{E}_\sigma \Bigg[\frac{1}{m} \sup_{h \in \mathcal{N}^{d-2,U,U}} \frac{\beta}{\beta_{p,q}(h)} \\
&\qquad \sup_w \frac{1}{\|w\|_p} \left| \sum_{i=1}^m \sigma_i \left[w^\top \left[h\left(x_i\right) \right]_+ \right]_+ \right| \Bigg] \\
&= U^{\left[\frac{1}{p^*}-\frac{1}{q}\right]_+} \mathbb{E}_\sigma \left[\frac{1}{m} \sup_{g \in \mathcal{N}^{d-1,U}_{\beta_p,q \leq \beta}} \left| \sum_{i=1}^m \sigma_i \left[g\left(x_i\right) \right]_+ \right| \right] \\
&\leq 2U^{\left[\frac{1}{p^*}-\frac{1}{q}\right]_+} \mathbb{E}_\sigma \left[\frac{1}{m} \sup_{g \in \mathcal{N}^{d-1,U}_{\beta_{p,q} \leq \beta}} \left| \sum_{i=1}^m \sigma_i g\left(x_i\right) \right| \right] \\
&= 2U^{\left[\frac{1}{p^*}-\frac{1}{q}\right]_+} \mathfrak{R}_m \left(\mathcal{N}^{d-1,U}_{\beta_{p,q} \leq \beta} \right)
\end{aligned}$$

We now use the recurrence relationship and Rademachar complexity obtained from Lemma 5 to get,

$$\mathfrak{R}_m \left(\mathcal{N}^1_{\beta_p,q \leq \beta} \right) \leq \sqrt{\frac{\beta^2 \min\left\{p^*, 4\log(2D)\right\} \max_i \|x_i\|^2_{p^*}}{m}} \tag{15}$$

□

Lemma 8. *The Massart's Lemma: Let A be a finite set of m dimensional vectors. Then*

$$\mathbb{E}_\sigma \left[\max_{a \in A} \frac{1}{m} \sum_{i=1}^m \sigma_i a_i \right] \leq \max_{a \in A} \|a\|_2 \frac{\sqrt{2 \log |A|}}{m}$$

where $|A|$ is the cardinality of A.

References

1. Bartlett, P.L., Mendelson, S.: Rademacher and gaussian complexities: Risk bounds and structural results. Journal of Machine Learning Research **3**(Nov), 463–482 (2002)
2. Neyshabur, B., Tomioka, R., Srebro, N.: Norm-based capacity control in neural networks. In: Conference on Learning Theory. pp. 1376–1401 (2015)