RISAN: Robust Instance Specific Deep Abstention Network (Supplementary Material)

Bhavya Kalra[1] **Kulin Shah**[2] **Naresh Manwani**[1]

[1]Machine Learning Lab, International Institute of Technology, Hyderabad, India
[2]Microsoft Research, Bangalore, India,

A APPENDIX

A.1 PROOF OF THEOREM 1

Theorem 1. *For a fixed cost of rejection d, the risk under double sigmoid loss is minimized by the generalized Bayes classifier $f_d^*(.)$*

Proof. The generalized bayes discriminant for reject option classifier (0-d-1 loss) is defined as

$$f_d^*(x) = 1 \cdot \mathbb{I}_{\eta(x)>1-d} + 0 \cdot \mathbb{I}_{d\leq\eta(x)\leq 1-d} - 1 \cdot \mathbb{I}_{\eta(x)<d} \quad (1)$$

and the risk for double sigmoid loss is defined as,

$$R_{ds}(f,\rho) = \mathbb{E}\left[L_{ds}(yf(\mathbf{x}),\rho)\right]$$

If $r_\eta(z) = \mathbb{E}_{y|\mathbf{x}}\left[L_{dr}(yf(\mathbf{x}),\rho)\right]$ and $z = f(\mathbf{x})$. Then

$$r_\eta(z) = \eta L_{ds}(z,\rho) + (1-\eta)L_{ds}(-z,\rho)$$

or

$$r_\eta(z) = 2(1-\eta)+(2\eta-2d)\sigma(z+\rho)+2(\eta+d-1)\sigma(z-\rho)$$

This can also be written as,

$$r_\eta(z) = 2(1-\eta) + (\eta-d)\left(1 - tanh\left(\frac{z+\rho}{2}\right)\right) + (\eta+d-1)\left(1 - tanh\left(\frac{z-\rho}{2}\right)\right) \quad (2)$$

where ρ is the rejection parameter, d is the cost of rejection and $\eta = P(Y=1|X)$. We observe that the function $r_\eta(z)$ can take different values corresponding to the value of parameter η. The parameter η can be majorly broken into 3 intervals for the reject option classifier, $\eta \in [0,d]$, $\eta \in [d,1-d]$ and $\eta \in [1-d,1]$. Thus we study, $r_\eta(z)$ in these 3 intervals. To find the minima of equation 2, we take it's derivative w.r.t. z. First we expand 2 using,

$$tanh(A \pm B) = \frac{tanh(A) \pm tanh(B)}{1 \pm tanh(A)tanh(B)}$$

with $A = \frac{z}{2}$ and $B = \frac{\rho}{2}$. Further on differentiating w.r.t z we get,

$$(K^2-1)(1-\zeta^2)\frac{(2\eta-1)K^2\zeta^2 + (4d-2)K\zeta + 2\eta - 1}{\left(1-K^2\zeta^2\right)^2} \quad (3)$$

where $K = tanh(\frac{z}{2})$, $\zeta = tanh(\frac{\rho}{2})$. We equate the derivative in equation 3 to 0 and find solutions for K as ± 1 from $K^2 - 1 = 0$ and $\frac{(1-2d)\pm\sqrt{(1-2d)^2-(2\eta-1)^2}}{(2\eta-1)\zeta}$ from

$$(2\eta-1)K^2\zeta^2 + (4d-2)K\zeta + 2\eta - 1$$

The numerator of eqn 3 contains two quadratic equations. We check if minima exists at $K = \pm 1$ by taking the second derivative of eqn 3 and evaluating the sign at $K = \pm 1$. We observe that the second derivative is positive hence minima exists for both the values.

However, for the roots $K = \frac{(1-2d)\pm\sqrt{(1-2d)^2-(2\eta-1)^2}}{(2\eta-1)\zeta}$ we look at the curve of the quadratic which yields the two solutions. We observe that the curve is opening upwards curve when $2\eta - 1 < 0$ and opening downwards when $2\eta - 1 > 0$ because $K^2 - 1 \leq 0$.

These curves suggest $K_1 = \frac{(1-2d)+\sqrt{(1-2d)^2-(2\eta-1)^2}}{(2\eta-1)\zeta}$ is the minima when $2\eta - 1 < 0$ since the slope for $r_\eta(z)$ changes from negative to positive at K_1. Similarly, $K_2 = \frac{(1-2d)-\sqrt{(1-2d)^2-(2\eta-1)^2}}{(2\eta-1)\zeta}$ is the minima when $2\eta - 1 > 0$.

Thus $r_\eta^*(z)$ would be,

$$r_\eta^*(z) = \min \begin{cases} 2\eta \\ 1-(\eta-d)\left(\frac{T_1+\zeta^2(2\eta-1)}{(2\eta-1)\zeta+T_1\zeta}\right) \\ \quad -(\eta+d-1)\left(\frac{T_1-\zeta^2(2\eta-1)}{(2\eta-1)\zeta-T_1\zeta}\right) \\ 1-(\eta-d)\left(\frac{T_2+\zeta^2(2\eta-1)}{(2\eta-1)\zeta+T_2\zeta}\right) \\ \quad -(\eta+d-1)\left(\frac{T_2-\zeta^2(2\eta-1)}{(2\eta-1)\zeta-T_2\zeta}\right) \\ 2(1-\eta) \end{cases}$$

Supplement for the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence (UAI 2021).

where $T_1 = (1-2d) + \sqrt{(1-2d)^2 - (2\eta-1)^2}$ and $T_2 = (1-2d) - \sqrt{(1-2d)^2 - (2\eta-1)^2}$.

Moreover the complex roots K_1 and K_2 are real only when

$$(1-2d)^2 - (2\eta-1)^2 \geq 0$$

Therefore, the solutions K_1 and K_2 are real only when $d \leq \eta \leq 1-d$.

Thus, when $\eta < d$, we have two candidates for minimum value. And we realise 2η is the minimum value since $2\eta \leq 2(1-\eta)$ and $\eta < d \leq 0.5$. Similarly when $\eta > 1-d$, we have two candidates for minimum value. And we observe, $2(1-\eta)$ would be the minimum value since $2(1-\eta) \leq 2\eta$.

However, when $\eta \in [d, 0.5]$ we have 3 candidates for minima 2η, $2(1-\eta)$ and
$1-(\eta-d)\left(\frac{T_1+\zeta^2(2\eta-1)}{(2\eta-1)\zeta+T_1\zeta}\right) - (\eta+d-1)\left(\frac{T_1-\zeta^2(2\eta-1)}{(2\eta-1)\zeta-T_1\zeta}\right)$.

Note: Even though $1-(\eta-d)\left(\frac{T_2+\zeta^2(2\eta-1)}{(2\eta-1)\zeta+T_2\zeta}\right) - (\eta+d-1)\left(\frac{T_2-\zeta^2(2\eta-1)}{(2\eta-1)\zeta-T_2\zeta}\right)$ is a minima, it's only a minima when $2\eta - 1 > 0$.

So we first show that,

$$\begin{aligned} 2\eta \geq 1 &- (\eta-d)\left(\frac{T_1+\zeta^2(2\eta-1)}{(2\eta-1)\zeta+T_1\zeta}\right) \\ &- (\eta+d-1)\left(\frac{T_1-\zeta^2(2\eta-1)}{(2\eta-1)\zeta-T_1\zeta}\right) \end{aligned}$$

Which can be rewritten and compared as,

$$\begin{aligned} 2(1-\eta) &+ (\eta-d)(2) + (\eta+d-1)(2) \geq \\ &2(1-\eta) + (\eta-d)\left(1 - \frac{T_1+\zeta^2(2\eta-1)}{(2\eta-1)\zeta+T_1\zeta}\right) \\ &+ (\eta+d-1)\left(1 - \frac{T_1-\zeta^2(2\eta-1)}{(2\eta-1)\zeta-T_1\zeta}\right) \end{aligned}$$

$$\begin{aligned} 0 &\geq (\eta-d)\left(-1 - \frac{T_1+\zeta^2(2\eta-1)}{(2\eta-1)\zeta+T_1\zeta}\right) \\ &\quad + (\eta+d-1)\left(-1 - \frac{T_1-\zeta^2(2\eta-1)}{(2\eta-1)\zeta-T_1\zeta}\right) \\ &= (\eta-d)\left(\frac{-T_1(1+\zeta) - (2\eta-1)\zeta(\zeta+1)}{(2\eta-1)\zeta+T_1\zeta}\right) + \\ &\quad (\eta+d-1)\left(\frac{T_1(\zeta-1) + (2\eta-1)\zeta(\zeta-1)}{(2\eta-1)\zeta-T_1\zeta}\right) \\ &= \left(\frac{T_1\left((1-2d)T_1 - (2\eta-1)^2\right)}{(2\eta-1)^2\zeta - T_1^2\zeta}\right) \\ &\quad + \left(\frac{(2\eta-1)^2(\zeta)^2(T_1+(2d-1))}{(2\eta-1)^2\zeta - T_1^2\zeta}\right) + \\ &\quad \left(\frac{(2\eta-1)\zeta((2\eta-1)^2 - T_1^2)}{(2\eta-1)^2\zeta - T_1^2\zeta}\right) \end{aligned} \tag{4}$$

We also find further relationship between $2\eta-1$, T_1 and T_2 based on the value of η. We observe that $(2\eta-1)^2 \leq T_1^2$ when $d < \eta \leq 0.5$ and $(2\eta-1)^2 \geq T_2^2$ when $0.5 < \eta \leq 1-d$. Using these facts we can see that equation 4 holds true, even at maximum value of $\zeta = 1$, for $\eta < 0.5$.

Since $\eta < 0.5$, $2\eta < 2(1-\eta)$ and hence $1-(\eta-d)\left(\frac{T_2+\zeta^2(2\eta-1)}{(2\eta-1)\zeta+T_2\zeta}\right) - (\eta+d-1)\left(\frac{T_2-\zeta^2(2\eta-1)}{(2\eta-1)\zeta-T_2\zeta}\right)$ is the minimum value of $r_\eta(z)$ when $d \leq \eta < 0.5$.

Due to the symmetry of $r_\eta(z)$ we can similarly show that $1-(\eta-d)\left(\frac{T_2+\zeta^2(2\eta-1)}{(2\eta-1)\zeta+T_2\zeta}\right) - (\eta+d-1)\left(\frac{T_2-\zeta^2(2\eta-1)}{(2\eta-1)\zeta-T_2\zeta}\right)$ is the minimum when $0.5 < \eta \leq 1-d$ by comparing it with $2(1-\eta)$. Since $2\eta > 2(1-\eta)$ for $0.5 < \eta \leq 1-d$, $1-(\eta-d)\left(\frac{T_2+\zeta^2(2\eta-1)}{(2\eta-1)\zeta+T_2\zeta}\right) - (\eta+d-1)\left(\frac{T_2-\zeta^2(2\eta-1)}{(2\eta-1)\zeta-T_2\zeta}\right)$ is the minimum value of $r_\eta(z)$ when $0.5 < \eta \leq 1-d$. Following from above established minimum values for each region of η, the $r^*_\eta(z)$ becomes,

$$r^*_\eta(z) = \begin{cases} 2\eta & \eta < d \\ 1-(\eta-d)\left(\frac{T_1+\zeta^2(2\eta-1)}{(2\eta-1)\zeta+T_1\zeta}\right) - & \\ \quad (\eta+d-1)\left(\frac{T_1-\zeta^2(2\eta-1)}{(2\eta-1)\zeta-T_1\zeta}\right) & d \leq \eta < 0.5 \\ 1-(\eta-d)\left(\frac{T_2+\zeta^2(2\eta-1)}{(2\eta-1)\zeta+T_2\zeta}\right) - & \\ \quad (\eta+d-1)\left(\frac{T_2-\zeta^2(2\eta-1)}{(2\eta-1)\zeta-T_2\zeta}\right) & 0.5 < \eta \leq 1-d \\ 2(1-\eta) & \eta > 1-d \end{cases}$$

The reject option in reject option classifiers is exercised when $z \in (-\rho, \rho)$ thus we need to show that z^* for double sigmoid loss for region $d < \eta < 1-d$ lies in $-\rho, \rho$. Let $\theta = 2\eta - 1$ and we show that $\frac{-\rho}{2} \leq K_1 \leq 0$ for $2d-1 \leq \theta < 0$ and $0 \leq K_2 \leq \rho$ for $0 < \theta \leq 1-2d$.

$$\begin{aligned} -\rho &\leq 2\tanh^{-1}(K_1) \\ -\zeta &\leq \frac{1-2d+\sqrt{(1-2d)^2-\theta^2}}{\theta\zeta} \\ -\theta\zeta^2 &\leq 1-2d+\sqrt{(1-2d)^2-\theta^2} \\ 0 &\geq \theta^2 + \theta^2\zeta^4 + 2(1-2d)\zeta^2 \end{aligned}$$

which says that for $K_1 \geq -\rho$, $\theta \in \left[\frac{-2(1-2d)\zeta^2}{1+\zeta^4}, 0\right]$. Similarly, for $K_2 \leq \rho$ we get $\theta \in \left[\frac{2(1-2d)\zeta^2}{1+\zeta^4}, 0\right]$. We also verify that our current solutions of θ lie between $[2d-1, 0)$ and $(0, 1-2d]$.

$$\begin{aligned} \frac{-2(1-2d)\zeta^2}{1+\zeta^4)} &\geq 2d-1 \\ \zeta^2(2-\zeta^2) &\leq 1 \end{aligned}$$

which is true for all ζ. Similarly we can show that θ with respect to K_2 also lies in $(0, 1-2d]$. Thus our z^* would

become,

$$z^* = \begin{cases} -\infty & \eta < d \\ [-\rho, 0) & d \leq \eta < 0.5 \\ (0, \rho] & 0.5 < \eta \leq 1 - d \\ \infty & \eta > 1 - d \end{cases}$$

Thus, our f^*_{ds} or the discriminant function for double sigmoid loss would be

$$f^*_{ds} = \begin{cases} -1 & \eta < d \\ 0 & d \leq \eta \leq 1 - d \\ 1 & \eta > 1 - d \end{cases}$$

which is similar to the bayes discriminant function for 0-d-1 loss. Therefore, bayes discriminant function minimizes the double sigmoid risk. □

A.2 PROOF OF THEOREM 2

Theorem 2. *Let $0 \leq d \leq 1/2$ and a measurable function z. Then we have the excess risk relation as*

$$\psi\left(R_d(f, \rho) - R_d(f^*_d)\right) \leq \left(R_{ds}(f, \rho) - R_{ds}(f^*_d)\right)$$

where

$$\psi(\theta) = \begin{cases} 0 & \theta = 0 \\ (2d-1)\zeta + \left(\frac{\theta+1-2d}{2}\right)\left(\frac{T+\zeta^2\theta}{\zeta\theta+T\zeta}\right) & \\ \qquad + \left(\frac{\theta+2d-1}{2}\right)\left(\frac{T-\zeta^2\theta}{\zeta\theta-T\zeta}\right) & \theta \in (0, 1-2d] \\ \theta + (2d-1)\zeta & \theta \in [1-2d, 1] \end{cases}$$

and $\theta = R_d(f, \rho) - R_d(f^*_d)$. Also, $\zeta = tanh(\frac{\rho}{2})$ and $T = (1-2d) - \sqrt{(1-2d)^2 - \theta^2}$.

Proof. We follow the approach described in Bartlett et al. [2006] and define the $\psi : [0, 1] \to [0, \infty)$ transform of a loss function as $\psi(\theta) = \text{co}\tilde{\psi}(\theta)$, where

$$\tilde{\psi}(\theta) = H^-\left(\frac{1+\theta}{2}\right) - H\left(\frac{1+\theta}{2}\right)$$

and *co* represents convex hull of the function. This implies that $\psi = \tilde{\psi}$ if and only if $\tilde{\psi}$ is convex. Also, $\theta \in [0, 1]$ with

$$H^-(\eta) = \inf_{z(2\eta-1)\leq 0} r_\eta(z) \quad \text{and} \quad H(\eta) = \inf_{z \in R} r_\eta(z)$$

From the definition H^- is the optimal conditional risk such that sign of z disagrees with sign of $2\eta - 1$.

$$\begin{aligned} H^-\left(\frac{1+\theta}{2}\right) &= \inf_{z\theta \leq 0} r_\eta(z) \\ &= \inf_{z \in (-\infty, 0)} 2(1-\eta) + (2\eta - 2d)\sigma(z+\rho) \\ &\qquad + 2(\eta + d - 1)\sigma(z-\rho) \\ &= \inf_{z \in (-\infty, 0)} 1 - \theta + (1 + \theta - 2d)\sigma(z+\rho) \\ &\qquad + (\theta + 2d - 1)\sigma(z-\rho) \\ &= \inf_{z \in (-\infty, 0)} 1 - \theta + \left(\frac{1+\theta-2d}{2}\right)\left(1 - tanh\left(\frac{z+\rho}{2}\right)\right) \\ &\qquad + \left(\frac{\theta+2d-1}{2}\right)\left(1 - tanh\left(\frac{z-\rho}{2}\right)\right) \\ &= 1 - \theta + \left(\frac{1+\theta-2d}{2}\right)\left(1 - tanh\left(\frac{\rho}{2}\right)\right) \\ &\qquad + \left(\frac{\theta+2d-1}{2}\right)\left(1 - tanh\left(\frac{-\rho}{2}\right)\right) \end{aligned}$$

Let $\tanh(\frac{\rho}{2}) = \zeta$ and thus $tanh(\frac{-\rho}{2}) = -\zeta$

$$\begin{aligned} H^-\left(\frac{1+\theta}{2}\right) &= 1 - \theta + \left(\frac{1+\theta-2d}{2}\right)(1-\zeta) \\ &\qquad + \left(\frac{\theta+2d-1}{2}\right)(1+\zeta) \\ &= 1 - \zeta + 2d\zeta \end{aligned}$$

Similarly, from the definition H is the optimal conditional risk,

$$\begin{aligned} H\left(\frac{1+\theta}{2}\right) &= \inf_{z \in R} r_\eta(z) \\ &= \inf_{z \in R} 2(1-\eta) + (2\eta - 2d)\sigma(z+\rho) + 2(\eta + d - 1)\sigma(z-\rho) \\ &= \inf_{z \in R} 1 - \theta + \left(\frac{1+\theta-2d}{2}\right)\left(1 - tanh\left(\frac{z+\rho}{2}\right)\right) \\ &\qquad + \left(\frac{\theta+2d-1}{2}\right)\left(1 - tanh\left(\frac{z-\rho}{2}\right)\right) \end{aligned}$$

$$\begin{aligned} H\left(\frac{1+\theta}{2}\right) &= \inf_{z \in R} 1 - \theta + \\ &\left(\frac{1+\theta-2d}{2}\right)\left(\frac{1 + tanh(\frac{z}{2})tanh(\frac{\rho}{2}) - tanh(\frac{z}{2}) - tanh(\frac{\rho}{2})}{1 + tanh(\frac{z}{2})tanh(\frac{\rho}{2})}\right) \\ &+ \left(\frac{\theta+2d-1}{2}\right)\left(\frac{1 - tanh(\frac{z}{2})tanh(\frac{\rho}{2}) - tanh(\frac{z}{2}) + tanh(\frac{\rho}{2})}{1 - tanh(\frac{z}{2})tanh(\frac{\rho}{2})}\right) \end{aligned}$$

Since $r^*_\eta(z) = H(\eta)$, we follow the definition of $r^*_\eta(z)$, hence

$$H(\eta) = r^*_\eta(z) = \begin{cases} 2\eta & \eta < d \\ 1 - (\eta - d)\left(\frac{T+\zeta^2(2\eta-1)}{(2\eta-1)\zeta+T\zeta}\right) & \\ -(\eta + d - 1)\left(\frac{T-\zeta^2(2\eta-1)}{(2\eta-1)\zeta-T\zeta}\right) & d \leq \eta \leq 1-d \\ 2(1-\eta) & \eta > 1 - d \end{cases}$$

Also, since $\theta \geq 0$ and $\eta = \frac{1+\theta}{2}$, we use definition of $r^*_\eta(z)$ for $\eta \geq 0.5$. Thus, $H\left(\frac{1+\theta}{2}\right)$ is defined over different intervals as,

$$H\left(\frac{1+\theta}{2}\right) = \begin{cases} 1+(2d-1)\zeta & \theta = 0 \\ 1-\left(\frac{\theta+1-2d}{2}\right)\left(\frac{T+\zeta^2\theta}{\zeta\theta+T\zeta}\right) & \\ \quad -\left(\frac{\theta+2d-1}{2}\right)\left(\frac{T-\zeta^2\theta}{\zeta\theta-T\zeta}\right) & \theta \in (0, 1-2d] \\ 1-\theta & \theta \in [1-2d, 1] \end{cases}$$

where $T = (1-2d) - \sqrt{(1-2d)^2 - \theta^2}$ and $\zeta = tanh(\frac{\rho}{2})$.

Thus, our $\tilde{\psi}(\theta)$ would be

$$\tilde{\psi}(\theta) = \begin{cases} 0 & \theta = 0 \\ (2d-1)\zeta + \left(\frac{\theta+1-2d}{2}\right)\left(\frac{T+\zeta^2\theta}{\zeta\theta+T\zeta}\right) & \\ \quad +\left(\frac{\theta+2d-1}{2}\right)\left(\frac{T-\zeta^2\theta}{\zeta\theta-T\zeta}\right) & \theta \in (0, 1-2d] \\ \theta + (2d-1)\zeta & \theta \in [1-2d, 1] \end{cases}$$

The function $\tilde{\psi}(\theta)$ is continuous in θ. Moreover, the corresponding term of $\tilde{\psi}(\theta)$ to $\theta \in (0, 1-2d]$, achieves indeterminate values at $\theta = 0$ and $\theta = 1-2d$, which is resolved by finding the limit value, which shows the continuity of $\tilde{\psi}(\theta)$.

$$\begin{aligned} &\lim_{\theta\to 0}\left(\frac{\theta+1-2d}{2}\right)\left(\frac{T+\zeta^2\theta}{\zeta\theta+T\zeta}\right) \\ &\qquad + \left(\frac{\theta+2d-1}{2}\right)\left(\frac{T-\zeta^2\theta}{\zeta\theta-T\zeta}\right) \\ &= \frac{1}{2}\left(\frac{\theta T'+2\zeta^2\theta}{\zeta+\zeta T'}\right) + \frac{1-2d}{2}\left(\frac{T'+\zeta^2}{\zeta+\zeta T'}\right) \\ &\qquad + \frac{1}{2}\left(\frac{\theta T'-2\zeta^2\theta}{\zeta-\zeta T'}\right) + \frac{1-2d}{2}\left(\frac{T'-\zeta^2}{\zeta-\zeta T'}\right) \\ &= \frac{1}{2}\left(\frac{\theta^2+2\zeta^2\theta\sqrt{(1-2d)^2-\theta^2}}{\zeta\sqrt{(1-2d)^2-\theta^2}+\zeta\theta}\right) \\ &\qquad + \frac{1-2d}{2}\left(\frac{\theta+\zeta^2\sqrt{(1-2d)^2-\theta^2}}{\zeta\sqrt{(1-2d)^2-\theta^2}+\zeta\theta}\right) \\ &+ \frac{1}{2}\left(\frac{\theta^2-2\zeta^2\theta\sqrt{(1-2d)^2-\theta^2}}{\zeta\sqrt{(1-2d)^2-\theta^2}-\zeta\theta}\right) \\ &\qquad + \frac{2d-1}{2}\left(\frac{\theta-\zeta^2\sqrt{(1-2d)^2-\theta^2}}{\zeta\sqrt{(1-2d)^2-\theta^2}-\zeta\theta}\right) \\ &= \frac{1-2d}{2}(\zeta) + \frac{2d-1}{2}(-\zeta) = (1-2d)\zeta \end{aligned}$$

where $T' = \frac{dT}{d\theta} = \frac{\theta}{\sqrt{(1-2d)^2-\theta^2}}$. However, at $\theta = 1-2d$ we can first need to look at value of $H(\frac{1+\theta}{2})$ at $\theta = 1-2d$, The value of K_1 at $\theta = 1-2d$,

$$K_1 = \frac{(1-2d) - \sqrt{(1-2d)^2 - (1-2d)^2}}{(1-2d)\zeta} = \frac{1}{\zeta}$$

which means K_1 will be valid only when $\zeta = 1$ since $K_1 \leq 1$. Thus,

$$\lim_{\theta\to 1-2d}\zeta = 1$$

Using this information when finding the value at the limit

$$\begin{aligned} &\lim_{\theta\to 1-2d}\left(\frac{\theta+1-2d}{2}\right)\left(\frac{T+\zeta^2\theta}{\zeta\theta+T\zeta}\right) \\ &\qquad + \left(\frac{\theta+2d-1}{2}\right)\left(\frac{T-\zeta^2\theta}{\zeta\theta-T\zeta}\right) \\ &= \frac{\theta+1-2d}{2}(1) + \frac{\theta+2d-1}{2}(-1) = 1-2d = \theta \end{aligned}$$

We can easily see that for the intervals $\theta = 0$ and $\theta \in [1-2d, 1]$, $\tilde{\psi}(\theta)$ is convex. However, we show the convexity for the interval $\theta \in (0, 1-2d)$ by taking the second derivative of the corresponding $\tilde{\psi}(\theta)$.

We first show the convexity of $C_1 = \left(\frac{\theta+1-2d}{2}\right)\left(\frac{T+\zeta^2\theta}{\zeta\theta+T\zeta}\right)$. Since our functions both the numerator $f(\theta) = (\theta+1-2d)(T+\zeta^2\theta)$ and denominator $g(\theta) = (2)(\zeta\theta+T\zeta)$ are convex and $g(\theta) > 0$. We can say that C_1 is convex if

$$\left(\frac{f(\theta)}{g(\theta)}\right)'' = \frac{f''g^2 - 2f'gg' - fgg'' + 2f(g')^2}{g^3} \geq 0 \quad (5)$$

i.e. the numerator of eqn 5 is greater than 0. Let $M = \theta+1-2d$, then

$$\begin{aligned} &(2T' + MT'' + 2\zeta^2)(2\zeta)^2(\theta+T)^2 \\ &- 2(T + MT' + \zeta^2 M + \zeta^2\theta)(2\zeta)^2(\theta+T)(1+T') \\ &- (MT + M\zeta^2\theta)(2\zeta)^2(\theta+T)T'' \\ &+ 2(MT + \zeta^2 M\theta)(2\zeta)^2(1+T')^2 \geq 0 \end{aligned}$$

On further solving we get,

$$\begin{aligned} &4\zeta^2(\theta+T)(MT''\theta)(1-\zeta^2) + 2(4\zeta^2)(\theta+T)(1-\zeta^2)(T'\theta - T) \\ &\qquad + 2(4\zeta^2)(1+T')(1-\zeta^2)M(T - T'\theta) \geq 0 \\ &4\zeta^2(\theta+T)(MT''\theta)(1-\zeta^2) \\ &\qquad + 2(4\zeta^2)(1-\zeta^2)(M)(T - T'\theta)(1+T'-\theta-T) \geq 0 \\ &(4\zeta^2)(1-\zeta^2)M[(\theta+T)(T''\theta) + 2(T - T'\theta)(1+T') \\ &\qquad + 2(\theta+T)(T'\theta - T)] \geq 0 \end{aligned}$$

This can be rearranged to get,

$$\begin{aligned} &(4\zeta^2)(1-\zeta^2)M[(\theta+T)(2T(1-T) \\ &\qquad + 2T(T'-\theta) + 2\theta(T'\theta - T') \\ &\qquad\qquad + (T''T\theta + T''\theta^2 + 2TT'\theta - 2(T')^2\theta))] \geq 0 \end{aligned}$$

which is true for $\theta \in (0, 1-2d]$. Where

$$T' = \frac{\theta\left((1-2d) - T\right)}{(1-2d)^2 - \theta^2} \quad \text{and} \quad T'' = \frac{(1-2d)^2\left((1-2d) - T\right)}{((1-2d)^2 - \theta^2)^2}$$

with $T'' \geq T' \geq T$ and $T' \geq \theta \geq T$. These definitions can be used to verify

$$T''\theta^2 + 2T - 2T^2 - 2T'\theta \geq 0 \tag{6}$$

$$T''T\theta + 2T'\theta^2 - 2\theta(T')^2 \geq 0 \tag{7}$$

$$2TT' + 2TT'\theta - 2T\theta \geq 0 \tag{8}$$

The inequality in eq. (8) is straightforward using the conditions on T, T', T'' and θ. The quadratic inequality in eq. (6) is a upward opening curve and the solutions are at $\theta = 0$. Hence, eq. (6) holds true. The same goes for eq. (7), which is a quadratic in θ, an upward opening curve with solutions at $\theta = 0$.

Similarly, we can show the convexity of $C_2 = \left(\frac{\theta-1+2d}{2}\right)\left(\frac{T-\zeta^2\theta}{\zeta\theta-T\zeta}\right)$. Also, since sum of convex functions is a convex function, we establish that $\tilde{\psi}(\theta)$ is convex when $\theta \in (0, 1-2d)$.

Now, $\tilde{\psi}(\theta)$ is individually convex in all the 3 intervals of θ, $\psi(\theta) = \tilde{\psi}(\theta)$ and continuous in θ. The convexity of $\tilde{\psi}(\theta)$ also depends on the slope of $\tilde{\psi}(\theta)$ for these intervals. While $\tilde{\psi}(\theta)$ has a slope of 0 when $\theta = 0$ and 1 when $\theta \in [1-2d, 1]$. The slope of $\tilde{\psi}(\theta)$ for $\theta \in (0, 1-2d]$ should be between $(0, 1]$ since it's an increasing convex function which will achieve it's maximum at $\theta = 1-2d$. So,

$$\lim_{\theta\to 1-2d}(2d-1)\zeta + \left(\frac{\theta+1-2d}{2}\right)\left(\frac{T+\zeta^2\theta}{\zeta\theta+T\zeta}\right) + \left(\frac{\theta+2d-1}{2}\right)\left(\frac{T-\zeta^2\theta}{\zeta\theta-T\zeta}\right) = (1-2d) + (2d-1)\zeta$$

which is equal to $\theta + (2d-1)\zeta$ when $\theta = 1-2d$, slope at $\theta = 1-2d$ is 1 for $\tilde{\psi}(\theta)$ corresponding to $\theta = 1-2d$. We can say that $\tilde{\psi}(\theta)$ is convex in it's domain $\theta \in [0, 1]$. Thus, $\tilde{\psi}(\theta) = \psi(\theta)$ and this suggests our excess risk relationship is

$$\psi\left(R_d(f,\rho) - R_d(f_d^*)\right) \leq \left(R_{ds}(f,\rho) - R_{ds}(f_d^*)\right)$$

where

$$\psi(\theta) = \begin{cases} 0 & \theta = 0 \\ (2d-1)\zeta + \left(\frac{\theta+1-2d}{2}\right)\left(\frac{T+\zeta^2\theta}{\zeta\theta+T\zeta}\right) & \\ \qquad + \left(\frac{\theta+2d-1}{2}\right)\left(\frac{T-\zeta^2\theta}{\zeta\theta-T\zeta}\right) & \theta \in (0, 1-2d] \\ \theta + (2d-1)\zeta & \theta \in [1-2d, 1] \end{cases}$$

□

A.3 PROOF OF THEOREM 3

Theorem 3. *Let $\mathcal{D}$ be any distribution on $\mathcal{X} \times \{-1, +1\}$. Let $0 < \delta \leq 1$. Then for any n, $q \geq 1$, $1 \leq p < \infty$ and any set $S = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$; with probability at least $1-\delta$ (over $S \sim \mathcal{D}^m$), all functions $f \in \mathcal{F}$ satisfy*

$$R_{ds}(f,\rho) \leq \hat{R}_{ds}(f,\rho) + \frac{\bar{\rho}}{\sqrt{m}} + \sqrt{\frac{8\ln\left(\frac{4}{\delta}\right)}{m}} + \sqrt{\frac{2\ln\left(\frac{2}{\delta}\right)}{m}} + \left(\frac{2\beta}{\sqrt{m}}\max_i \|\mathbf{x}_i\|_{p'}\right)\left(2H^{\left[\frac{1}{p'}-\frac{1}{q}\right]_+}\right)^{n-1}$$

where n is the number of layers in the network, H is the number of neurons in the hidden layers, rejection region parameter is bounded as $\rho \leq \bar{\rho}$. Also $\frac{1}{p'} + \frac{1}{p} = 1$ and $[a]_+ = \max(0, a)$. $\hat{R}_{ds}(f,\rho)$ is the empirical error and $\beta_{p,q}(W) = \prod_{k=1}^n \|W_k\|_{p,q} \leq \beta$.

Proof. We follow lemma 4,

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

where $R(\mathcal{F})$ is the rademachar complexity and f_s is a function belonging to function class $\mathcal{F}$. Since the bounds are described for loss $\ell : \mathcal{Y} \times [a, b] \to [0, B]$. For double sigmoid loss, we get $B = 2$,$a = -1$ and $b = 1$. So, we bound the generalization error with probability atleast $1-\delta$ by

$$R_{ds}(f,\rho) \leq \hat{R}_{ds}(f,\rho) + 2L\hat{R}_m(\mathcal{F}) + \sqrt{\frac{8\ln\left(\frac{4}{\delta}\right)}{m}} + \sqrt{\frac{2\ln\left(\frac{2}{\delta}\right)}{m}}$$

We now find an upper bound for the rademachar complexity $\hat{R}_m(\mathcal{F})$, following theorem 1 in Neyshabur et al. [2015].

Hence, let $\hat{R}_m(\mathcal{F}) = R(\mathcal{F}^{n,H}_{\beta_{p,q}\leq\beta})$ where $\mathcal{F}(\mathbf{x}) = \left|\mathbf{w}^T(\phi\left(W_{n-1}\phi\left(W_{n-2}\left(\ldots\phi\left(W_1\mathbf{x}\right)\right)\right)\right)\right| - \rho$, $\beta_{p,q}(W) = \prod_{k=1}^n \|W_k\|_{p,q} \leq \beta$ and ϕ is ReLU activation function. Also, $\mathbf{w}$ is an H dimensional vector. We prove the bound by

induction

$$R(\mathcal{F}^{n,H}_{\beta_{p,q}\leq\beta}) = \mathbb{E}_\lambda\left[\frac{1}{m}\sup_{f\in\mathcal{F}^{n,H}_{\beta_{p,q}\leq\beta}}\sup_{\rho}\left|\sum_{i=1}^{m}\lambda_i(|f(\mathbf{x}_i)|-\rho)\right|\right]$$

$$R(\mathcal{F}^{n,H}_{\beta_{p,q}\leq\beta}) \leq \mathbb{E}_\lambda\left[\frac{1}{m}\sup_{f\in\mathcal{F}^{n,H}_{\beta_{p,q}\leq\beta}}\sup_{\rho}\left|\sum_{i=1}^{m}\lambda_i|f(\mathbf{x}_i)|\right|\right] + \mathbb{E}_\lambda\left[\frac{1}{m}\sup_{f\in\mathcal{F}^{n,H}_{\beta_{p,q}\leq\beta}}\sup_{\rho}\left|\sum_{i=1}^{m}\lambda_i\rho\right|\right]$$

$$\begin{aligned}R(\mathcal{F}^{n,H}_{\beta_{p,q}\leq\beta}) &= \mathbb{E}_\lambda\left[\frac{1}{m}\sup_{f\in\mathcal{F}^{n,H}_{\beta_{p,q}\leq\beta}}\sup_{\rho}\left|\sum_{i=1}^{m}\lambda_i|f(\mathbf{x}_i)|\right|\right] + \frac{1}{m}\bar{\rho}\mathbb{E}_\lambda\left[\left|\sum_{i=1}^{m}\lambda_i\right|\right]\\ &\leq \mathbb{E}_\lambda\left[\frac{1}{m}\sup_{f\in\mathcal{F}^{n,H}_{\beta_{p,q}\leq\beta}}\sup_{\rho}\left|\sum_{i=1}^{m}\lambda_i|f(\mathbf{x}_i)|\right|\right] + \frac{1}{m}\bar{\rho}\sqrt{m}(1)\\ &= \mathbb{E}_\lambda\left[\frac{1}{m}\sup_{f\in\mathcal{F}^{n,H}_{\beta_{p,q}\leq\beta}}\left|\sum_{i=1}^{m}\lambda_i|f(\mathbf{x}_i)|\right|\right] + \frac{\bar{\rho}}{\sqrt{m}}\end{aligned}$$

Let, $\mathcal{R}_{rec}$ be defined as,

$$\mathcal{R}_{rec} = \mathbb{E}_\lambda\left[\frac{1}{m}\sup_{f\in\mathcal{N}^{n,H}}\frac{\beta}{\beta_{p,q}(f)}\left|\sum_{i=1}^{m}\lambda_i|f(\mathbf{x}_i)|\right|\right]$$

Also, $\mathcal{R}_{rec} = R(\mathcal{N}^{n,H}_{\beta_{p,q}\leq\beta})$ where $\mathcal{N}(\mathbf{x}) = \left|\mathbf{w}^T(\phi\left(W_{n-1}\phi\left(W_{n-2}\left(\ldots\phi\left(W_1\mathbf{x}\right)\right)\right)\right)\right|$ and ϕ is ReLU activation function.

$$\begin{aligned}\mathcal{R}_{rec} &= \mathbb{E}_\lambda\left[\frac{1}{m}\sup_{f\in\mathcal{N}^{n,H}}\frac{\beta}{\beta_{p,q}(f)}\left|\sum_{i=1}^{m}\lambda_i|f(\mathbf{x}_i)|\right|\right]\\ &= \mathbb{E}_\lambda\left[\frac{1}{m}\sup_{g\in\mathcal{N}^{n-1,H,H}}\sup_{\mathbf{w}}\frac{\beta}{\beta_{p,q}(g)\|\mathbf{w}\|_p}\left|\sum_{i=1}^{m}\lambda_i|\mathbf{w}^T[g(\mathbf{x}_i)]_+|\right|\right]\\ &= \mathbb{E}_\lambda\left[\frac{1}{m}\sup_{g\in\mathcal{N}^{n-1,H,H}}\sup_{\mathbf{w}}\frac{\beta}{\beta_{p,q}(g)\|\mathbf{w}\|_p}\left|\sum_{i=1}^{m}\lambda_i\|\mathbf{w}\|_p\|[g(\mathbf{x}_i)]_+\|_{p'}\right|\right]\\ &= \mathbb{E}_\lambda\left[\frac{1}{m}\sup_{g\in\mathcal{N}^{n-1,H,H}}\frac{\beta}{\beta_{p,q}(g)}\left|\sum_{i=1}^{m}\lambda_i\|\left[g(\mathbf{x}_i)\right]_+\|_{p'}\right|\right]\\ &= \mathbb{E}_\lambda\left[\frac{1}{m}\sup_{h\in\mathcal{N}^{n-2,H,H}}\frac{\beta}{\beta_{p,q}(h)}\sup_{W}\frac{1}{\|W\|_{p,q}}\left|\sum_{i=1}^{m}\lambda_i\left\|\left[W\left[h(\mathbf{x}_i)\right]_+\right]_+\right\|_{p'}\right|\right]\\ &= \mathbb{E}_\lambda\left[\frac{1}{m}\sup_{h\in\mathcal{N}^{n-2,H,H}}\frac{\beta}{\beta_{p,q}(h)}\sup_{W}\frac{1}{\|W\|_{p,q}}\left\|\sum_{i=1}^{m}\lambda_i\left\|\left[W\left[h(\mathbf{x}_i)\right]_+\right]_+\right\|_{p'}\right\|_{p'}\right]\end{aligned}$$

We use Lemma 6 to obtain the following result,

$$\begin{aligned}R(\mathcal{N}^{n,H}_{\beta_{p,q}\leq\beta}) &= H^{\left[\frac{1}{p'}-\frac{1}{q}\right]_+}\mathbb{E}_\lambda\left[\frac{1}{m}\sup_{h\in\mathcal{N}^{n-2,H,H}}\frac{\beta}{\beta_{p,q}(h)}\sup_{\mathbf{w}}\frac{1}{\|\mathbf{w}\|_p}\left|\sum_{i=1}^{m}\lambda_i\left\|\left[\mathbf{w}^\top\left[h\left(\mathbf{x}_i\right)\right]_+\right]_+\right\|_{p'}\right|\right]\\ &= H^{\left[\frac{1}{p'}-\frac{1}{q}\right]_+}\mathbb{E}_\lambda\left[\frac{1}{m}\sup_{h\in\mathcal{N}^{n-1,H,H}}\frac{\beta}{\beta_{p,q}(g)}\left|\sum_{i=1}^{m}\lambda_i\left\|\left[g(\mathbf{x}_i)\right]_+\right\|_{p'}\right|\right]\\ &\leq H^{\left[\frac{1}{p'}-\frac{1}{q}\right]_+}\mathbb{E}_\lambda\left[\frac{1}{m}\sup_{h\in\mathcal{N}^{n-1,H,H}}\frac{\beta}{\beta_{p,q}(g)}\left|\sum_{i=1}^{m}\lambda_i\left|\left[g(\mathbf{x}_i)\right]_+\right|\right|\right]\\ &\leq H^{\left[\frac{1}{p'}-\frac{1}{q}\right]_+}\mathbb{E}_\lambda\left[\frac{1}{m}\sup_{h\in\mathcal{N}^{n-1,H,H}}\frac{\beta}{\beta_{p,q}(g)}\left|\sum_{i=1}^{m}\lambda_i\left[|g(\mathbf{x}_i)|\right]_+\right|\right]\end{aligned}$$

We use Lemma 16 (Contraction Lemma) Neyshabur et al. [2015] result directly to obtain the following result,

$$\begin{aligned}R(\mathcal{N}^{n,H}_{\beta_{p,q}\leq\beta}) &\leq 2(1)H^{\left[\frac{1}{p'}-\frac{1}{q}\right]_+}\mathbb{E}_\lambda\left[\frac{1}{m}\sup_{h\in\mathcal{N}^{n-1,H,H}}\frac{\beta}{\beta_{p,q}(g)}\left|\sum_{i=1}^{m}\lambda_i\left|g(\mathbf{x}_i)\right|\right|\right]\\ R(\mathcal{N}^{n,H}_{\beta_{p,q}\leq\beta}) &\leq 2H^{\left[\frac{1}{p'}-\frac{1}{q}\right]_+}R\left(\mathcal{N}^{n-1,H}_{\beta_{p,q}\leq\beta}\right)\end{aligned}$$

We now use the recurrence relationship and Rademacher complexity obtained from Theorem 5 to get,

$$R(\mathcal{F}^{n,H}_{\beta_{p,q}\leq\beta}) \leq \left(\frac{\beta}{\sqrt{m}}\max_i\|\mathbf{x}_i\|_{p'}\right)\left(2H^{\left[\frac{1}{p'}-\frac{1}{q}\right]_+}\right)^{n-1} + \frac{\bar{\rho}}{\sqrt{m}}$$

The Lipschitz constant for double sigmoid loss with cost of rejection d_r can be computed as

$$\begin{aligned}L = \sup|&2d_r\sigma(z-\rho)\left(1-\sigma(z-\rho)\right)\\ &+2(1-d_r)\sigma(z+\rho)\left(1-\sigma(z+\rho)\right)|\end{aligned} \quad (9)$$

The maximum value of the product $\sigma(z-\rho)\left(1-\sigma(z-\rho)\right)$ and $\sigma(z+\rho)\left(1-\sigma(z+\rho)\right)$ is at $z=\rho$ and $z=-\rho$

respectively. However, the maximum value of the equation 9 is achieved at $z = 0$ and $\rho = 0$. Thus,

$$L = 2d_r\sigma(-\rho)(1-\sigma(-\rho)) + 2(1-d_r)\sigma(\rho)(1-\sigma(\rho))$$
$$L = 2d_r\sigma(-\rho)\sigma(\rho) + 2(1-d_r)\sigma(\rho)\sigma(-\rho)$$
$$L = 2\sigma(\rho)\sigma(-\rho)$$
$$L = 2\sigma(\rho)\sigma(-\rho)$$

when $\rho = 0$, we get $L = 0.5$. We now use the above result to get the generalization bound where the lipschitz constant L for double sigmoid loss would be $L = 0.5$.

$$R_{ds}(f,\rho) \leq \hat{R}_{ds}(f,\rho) + \left(\frac{2\beta}{\sqrt{m}}\max_i \|\mathbf{x}_i\|_{p'}\right)\left(2H^{\left[\frac{1}{p'}-\frac{1}{q}\right]_+}\right)^{n-1} + \frac{\bar{\rho}}{\sqrt{m}} + \sqrt{\frac{8\ln\left(\frac{4}{\delta}\right)}{m}} + \sqrt{\frac{2\ln\left(\frac{2}{\delta}\right)}{m}}$$

□

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

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

Proof. First we define

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

where λ 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.

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

We directly use the result from the Bartlett and Mendelson [2002] and using the results directly with probability atleast $1-\delta$, we bound the generalization error as,

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

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

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

$$\begin{aligned}
&\left|\phi\left(\mathbf{x}_1,\ldots,\mathbf{x}_j,\ldots,\mathbf{x}_m\right) - \phi\left(\mathbf{x}_1,\ldots,\mathbf{x}'_j,\ldots,\mathbf{x}_m\right)\right| \\
&= \hat{R}_m(\mathcal{F}) - R_{\left(\mathbf{x}_1,\ldots,\mathbf{x}'_j,\ldots,\mathbf{x}_m\right)}(\mathcal{F}) \\
&= \mathbf{E}_{\lambda\in\{\pm1\}^m}\left[\sup_{f\in\mathcal{F}}\frac{1}{m}\sum_{i=1}^{m}\lambda_i f\left(\mathbf{x}_i\right)\right. \\
&\qquad \left. - \sup_{f\in\mathcal{F}}\left(\frac{1}{m}\sum_{i\neq j}\lambda_i f\left(\mathbf{x}_i\right) + \frac{1}{m}\lambda_j f\left(\mathbf{x}'_j\right)\right)\right] \\
&\leq \frac{b-a}{m}
\end{aligned}$$

Thus by McDarmid's inequality we get,

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

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

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

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 $1-\delta$,

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

□

Theorem 5. *The rademachar complexity for RISAN with a single layer $(n = 1)$, is bounded as*

$$R(\mathcal{F}^1_{\|\mathbf{w}\|_p\leq\beta}) \leq \frac{\beta}{\sqrt{m}}\max_i \|\mathbf{x}_i\|_{p'}$$

Proof. For a network with a single layer, it is important to

notice that $\beta_{p,q}(\mathbf{w}) = \|\mathbf{w}\|_p$.

$$R(\mathcal{F}^1_{\|\mathbf{w}\|_p \le \beta}) \le \mathbb{E}_\lambda \left[\frac{1}{m} \sup_{\|\mathbf{w}\|_p \le \beta} \sup_{\rho} \left| \sum_{i=1}^{m} \lambda_i |\mathbf{w}^T \mathbf{x}_i| \right| \right]$$
$$R(\mathcal{F}^1_{\|\mathbf{w}\|_p \le \beta}) \le \mathbb{E}_\lambda \left[\frac{1}{m} \sup_{\|\mathbf{w}\|_p \le \beta} \sup_{\rho} \left| \sum_{i=1}^{m} \lambda_i |\mathbf{w}^T \mathbf{x}_i| \right| \right]$$
$$R(\mathcal{F}^1_{\|\mathbf{w}\|_p \le \beta}) \le \mathbb{E}_\lambda \left[\frac{1}{m} \sup_{\|\mathbf{w}\|_p \le \beta} \sup_{\rho} \left| \sum_{i=1}^{m} \lambda_i |\mathbf{w}^T \mathbf{x}_i| \right| \right]$$
$$R(\mathcal{F}^1_{\|\mathbf{w}\|_p \le \beta}) \le \mathbb{E}_\lambda \left[\frac{1}{m} \sup_{\|\mathbf{w}\|_p \le \beta} \sup_{\rho} \left| \sum_{i=1}^{m} \lambda_i \|\mathbf{w}\|_p \|\mathbf{x}_i\|_{p'} \right| \right]$$
$$R(\mathcal{F}^1_{\|\mathbf{w}\|_p \le \beta}) \le \beta \mathbb{E}_\lambda \left[\frac{1}{m} \left| \sum_{i=1}^{m} \lambda_i \|\mathbf{x}_i\|_{p'} \right| \right]$$
$$R(\mathcal{F}^1_{\|\mathbf{w}\|_p \le \beta}) \le \frac{\beta}{m} \left(\sum_{i=1}^{m} \|\mathbf{x}_i\|^2_{p'} \right)^{\frac{1}{2}}$$
$$R(\mathcal{F}^1_{\|\mathbf{w}\|_p \le \beta}) \le \frac{\beta}{m} \left(m \max_i \|\mathbf{x}_i\|^2_{p'} \right)^{\frac{1}{2}} = \frac{\beta}{\sqrt{m}} \max_i \|\mathbf{x}_i\|_{p'}$$

□

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

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

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

Proof.

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

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

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

Thus,

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

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

$$\frac{g(\mathbf{w}^*)}{\|\mathbf{w}^*\|_p} \ge \frac{g(V_i)}{\|V_i\|_p}$$
$$\left(\frac{g(\mathbf{w}^*)}{\|\mathbf{w}^*\|_p} \right)^{p'} \ge \left(\frac{g(V_i)}{\|V_i\|_p} \right)^{p'}$$
$$\left(\frac{g(\mathbf{w}^*)\|V_i\|_p}{\|\mathbf{w}^*\|_p} \right)^{p'} \ge \left(g(V_i) \right)^{p'}$$
$$\sum_{i=1}^{H} \left(\frac{g(\mathbf{w}^*)\|V_i\|_p}{\|\mathbf{w}^*\|_p} \right)^{p'} \ge \sum_{i=1}^{H} \left(g(V_i) \right)^{p'}$$
$$\left(\frac{g(\mathbf{w}^*)}{\|\mathbf{w}^*\|_p} \right)^{p'} \sum_{i=1}^{H} \|V_i\|_p^{p'} \ge \sum_{i=1}^{H} \left(g(V_i) \right)^{p'}$$
$$\frac{g(\mathbf{w}^*)}{\|\mathbf{w}^*\|_p} \left(\sum_{i=1}^{H} \|V_i\|_p^{p'} \right)^{\frac{1}{p'}} \ge \left(\sum_{i=1}^{H} \left(g(V_i) \right)^{p'} \right)^{\frac{1}{p'}}$$
$$\frac{g(\mathbf{w}^*)}{\|\mathbf{w}^*\|_p} \ge \frac{\|g(V)\|_{p'}}{\|V\|_{p,p'}} \tag{13}$$

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

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

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

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

And from eqn.(13) we get,

$$\frac{g(\mathbf{w}^*)}{\|\mathbf{w}^*\|_p} \ge \frac{\|g(V)\|_{p'}}{\|V\|_{p,p'}} \ge \frac{\|g(V)\|_{p'}}{H^{[\frac{1}{p'} - \frac{1}{q}]} \|V\|_{p,q}}$$
$$H^{[\frac{1}{p'} - \frac{1}{q}]} \frac{g(\mathbf{w}^*)}{\|\mathbf{w}^*\|_p} \ge \frac{\|g(V)\|_{p'}}{\|V\|_{p,q}}$$

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. □

A.4 ARCHITECTURE DETAILS AND HYPERPARAMETER SELECTION

Small Datasets Experiments: The experiments with regular small dimensional data are conducted with the network architecture shown in Figure 1. We use 3 fully connected layers in the main body block of the network architecture with batch normalization [Ioffe and Szegedy, 2015] and dropout [Srivastava et al., 2014] at each layer. Each layer uses ReLU (Rectified Linear Units) as the activation function with 64 neurons in each layer. We further use Adagrad (adaptive gradient) optimizer with a learning rate of $1\text{e}{-3}$

	RISAN Double Sigmoid + Cross Entropy	**RISAN-NA** Double Sigmoid	**SNN** Selective Loss + Cross Entropy	**SNN-NA** Selective Loss	**DAC** DAC Loss
Architecture					
FC Layers Prediction head Rejection head/ Selective head	512,256,128 512,256,64	512,256,128 512,256,64	512,256,128 512,256,64	512,256,128 512,256,64	512,256,128
Weight Decay CNN FC	1e-4 1e-7	1e-4 1e-7	1e-4 1e-7	1e-4 1e-7	1e-4 1e-7
Datasets					
Cats vs Dogs	Figure 2b α = 0.9 $\gamma = 1e-3$ 250 epochs	Figure 2a $\gamma = 1e-3$ 250 epochs	Figure 2b α=0.5 250 epochs	Figure 2a 250 epochs	Figure 2a (no rejection head) 250 epochs
CIFAR	Figure 2b α = 0.7 $\gamma = 1e-3$ 250 epochs	Figure 2a $\gamma = 1e-3$ 250 epochs	Figure 2b α=0.5 250 epochs	Figure 2a 250 epochs	Figure 2a (no rejection head) 250 epochs
MNIST	Figure 2b α = 0.7 $\gamma = 1e-3$ 150 epochs	Figure 2a $\gamma = 1e-3$ 150 epochs	Figure 2b α=0.5 150 epochs	Figure 2a 150 epochs	Figure 2a (no rejection head) 150 epochs
CBIS-DDSM	Figure 2b α = 0.7 $\gamma = 1.0$ 250 epochs	Figure 2a $\gamma = 1e-3$ 250 epochs	Figure 2b α=0.5 250 epochs	Figure 2a 250 epochs	Figure 2a (no rejection head) 250 epochs

Table 1: Architecture and Hyperparameters for Large datasets

and run it for 100 epochs. We fix the batch size as 32. We use a γ value of 2 in the double sigmoid loss function for Ionosphere dataset whereas a value of 1 for ILPD dataset. For both SDR-SVM and DH-SVM we use a Gaussian kernel. We select the best values of regularization parameter λ and kernel parameter γ using 10-fold cross validation. We also use $\mu = 1$ for SDR-SVM.

Large Datasets Experiments: For phishing dataset, we use RISAN with input dependent rejection. However, the rejection head is dependant on the input, hence it gets input not from a constant valued neuron but the fully connected (FC) layers. We used 4 FC layers with 64 neurons each, and followed each layer with dropout and batch normalization layers. The same architecture is used for experiments with SNN-NA and DAC with phishing dataset. The phishing dataset being a small dimensional dataset, the auxiliary loss becomes redundant and hence we remove the auxiliary loss for this experiment for SNN.

In the CNN based experiments, we used and followed the architecture and hyperparameters similar to the ones used in Geifman and El-Yaniv [2019]. The VGG-16 architecture from Simonyan and Zisserman [2014] was optimized for the small datasets and image sizes as suggested in Liu and Deng [2015] with following alterations: (i) used only one fully connected layer with 512 neurons (the original VGG-16 has two fully connected layers of 4096 neurons). (ii) added batch normalization Ioffe and Szegedy [2015] (iii) added dropout Srivastava et al. [2014]. Also, the standard data augmentation consisting of horizontal flips, vertical and horizontal shifts, and rotations were included. The network was optimized using stochastic gradient descent (SGD) with a momentum of 0.9, an initial learning rate of 0.1, and a weight decay of 5e-4. The learning rate was reduced by 0.5 every 25 epochs.

We made further amendments to it by incorporating a separate stack of fully connected layers for each head. While the main body block of across all algorithms is the VGG-16 architecture. We added individual hidden layers to both prediction head and rejection head. We used three fully con-

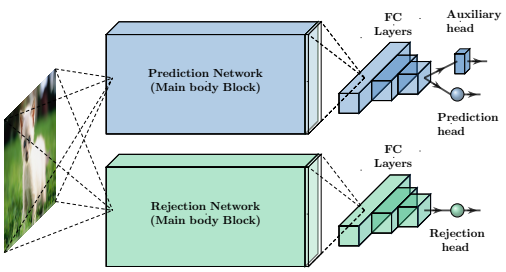

Figure 9: GradCAM Implementation

nected layers of sizes 512, 256 and 128, followed by a single neuron (prediction head). We also used additional three fully connected layers of size 512, 256 and 64 neurons followed by a single neuron for the rejection head. Moreover, for the rejection head, ReLU activation was used to ensure a positive rejection region parameter. A separate weight decay of $1e{-}4$ for CNN layers and $1e{-}7$ for FC layers. The data for individual CNN based datasets and algorithm is provided in Table 1. The learning rate scheduler was used which reduces learning rate by 0.5 once the validation loss stagnates. We utilize the same learning rate scheduler across all algorithms and datasets.

A.5 GRADCAM IMPLEMENTATION

One exciting prospect of deep learning models is their ability to help us understand why a classification decision was made or the important generic features learned during the training process. We followed the GradCAM technique of Selvaraju et al. [2017] that produces a localization map highlighting important regions in the image corresponding to particular predictions(class) to evaluate our model. We used the architecture described in Fig. 9, with VGG-16 in both the networks, and executed the GradCAM technique on the sigmoid outputs of the auxiliary head associated with the prediction network. We explored the possibility of having two separate networks, prediction network and rejection network, for each head separately. The use of separate networks was adopted to minimize the sharing of features, and subsequently, important features learned by each network could be examined independently. This helps in visualizing features learned by the prediction network to produce highlighted regions corresponding to the image's different classes.

References

Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

Peter L Bartlett, Michael I Jordan, and Jon D McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

Yonatan Geifman and Ran El-Yaniv. Selectivenet: A deep neural network with an integrated reject option. *arXiv preprint arXiv:1901.09192*, 2019.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

Shuying Liu and Weihong Deng. Very deep convolutional neural network based image classification using small training sample size. In *2015 3rd IAPR Asian conference on pattern recognition (ACPR)*, pages 730–734. IEEE, 2015.

Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. Norm-based capacity control in neural networks. In *Conference on Learning Theory*, pages 1376–1401, 2015.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.