Contents of Appendix

A. Symmetrization

We use the following lemmas from (Cortes et al., 2019) in our proofs.

Lemma 8 ((Cortes et al., 2019))**.** *Fix $\eta > 0$ and α with $1 < \alpha \le 2$. Let $f\colon (0, +\infty) \times (0, +\infty) \to \mathbb{R}$ be the function defined by $f\colon (x, y) \mapsto \frac{x-y}{\sqrt[\alpha]{x+y+\eta}}$. Then, f is a strictly increasing function of x and a strictly decreasing function of y.*

Lemma 9 ((Greenberg and Mohri, 2013))**.** *Let X be a random variable distributed according to the binomial distribution $B(m, p)$ with m a positive integer (the number of trials) and $p > \frac{1}{m}$ (the probability of success of each trial). Then, the following inequality holds:*

$$\mathbb{P}[X \ge \mathbb{E}[X]] > \frac{1}{4}, \tag{6}$$

and, if instead of requiring $p > \frac{1}{m}$ we require $p < 1 - \frac{1}{m}$, then

$$\mathbb{P}[X \le \mathbb{E}[X]] > \frac{1}{4}, \tag{7}$$

where in both cases $\mathbb{E}[X] = mp$.

The following symmetrization lemma in terms of empirical margin loss is proven using the previous lemmas.

Lemma 1. *Fix $\rho \ge 0$ and $1 < \alpha \le 2$ and assume that $m\epsilon^{\frac{\alpha}{\alpha-1}} > 1$. Then, for any $\epsilon, \tau > 0$, the following inequality holds:*

$$\mathbb{P}_{S \sim \mathcal{D}^m}\left[\sup_{h \in \mathcal{H}} \frac{R(h) - \widehat{R}^{\rho}_S(h)}{\sqrt[\alpha]{R(h) + \tau}} > \epsilon\right] \le 4 \mathbb{P}_{S, S' \sim \mathcal{D}^m}\left[\sup_{h \in \mathcal{H}} \frac{\widehat{R}_{S'}(h) - \widehat{R}^{\rho}_S(h)}{\sqrt[\alpha]{\frac{1}{2}[\widehat{R}_{S'}(h) + \widehat{R}^{\rho}_S(h) + \frac{1}{m}]}} > \epsilon\right].$$

Proof. We will use the function F defined over $(0, +\infty) \times (0, +\infty)$ by $F\colon (x, y) \mapsto \frac{x-y}{\sqrt[\alpha]{\frac{1}{2}[x+y+\frac{1}{m}]}}$.

Fix $S, S' \in \mathcal{Z}^m$. We first show that the following implication holds for any $h \in \mathcal{H}$:

$$\left(\frac{R(h) - \widehat{R}^{\rho}_S(h)}{\sqrt[\alpha]{R(h) + \tau}} > \epsilon\right) \wedge \left(\widehat{R}_{S'}(h) > R(h)\right) \Rightarrow F(\widehat{R}_{S'}(h), \widehat{R}^{\rho}_S(h)) > \epsilon. \tag{8}$$

The first condition can be equivalently rewritten as $\widehat{R}^{\rho}_S(h) < R(h) - \epsilon \sqrt[\alpha]{(R(h) + \tau)}$, which implies

$$\widehat{R}^{\rho}_S(h) < R(h) - \epsilon \sqrt[\alpha]{R(h)} \qquad \wedge \qquad \epsilon^{\frac{\alpha}{\alpha-1}} < R(h), \tag{9}$$

since $\widehat{R}^{\rho}_S(h) \ge 0$. Assume that the antecedent of the implication (8) holds for $h \in \mathcal{H}$. Then, in view of the monotonicity properties of function F (Lemma 8), we can write:

$$\begin{aligned}
F(\widehat{R}_{S'}(h), \widehat{R}^{\rho}_S(h)) &\ge F(R(h), R(h) - \epsilon \sqrt[\alpha]{R(h)}) && (\widehat{R}_{S'}(h) > R(h) \text{ and 1st ineq. of (9)})\\
&= \frac{R(h) - (R(h) - \epsilon R(h)^{\frac{1}{\alpha}}}{\sqrt[\alpha]{\frac{1}{2}[2R(h) - \epsilon R(h)^{\frac{1}{\alpha}} + \frac{1}{m}]}}\\
&\ge \frac{\epsilon R(h)^{\frac{1}{\alpha}}}{\sqrt{\frac{1}{2}[2R(h) - \epsilon^{\frac{\alpha}{\alpha-1}} + \frac{1}{m}]}} && (\text{second ineq. of (9)})\\
&> \frac{\epsilon R(h)^{\frac{1}{\alpha}}}{\sqrt[\alpha]{\frac{1}{2}[2R(h)]}} = \epsilon, && (m\epsilon^{\frac{\alpha}{\alpha-1}} > 1)
\end{aligned}$$

which proves (8).

Now, by definition of the supremum, for any $\eta > 0$, there exists $h_S \in \mathcal{H}$ such that

$$\sup_{h \in \mathcal{H}} \frac{R(h) - \widehat{R}^{\rho}_S(h)}{\sqrt[\alpha]{R(h) + \tau}} - \frac{R(h_S) - \widehat{R}^{\rho}_S(h_S)}{\sqrt[\alpha]{R(h_S) + \tau}} \le \eta. \tag{10}$$

Using the definition of h_S and the implication (8), we can write

$$\begin{aligned}
&\mathop{\mathbb{P}}_{S,S'\sim\mathcal{D}^m}\left[\sup_{h\in\mathcal{H}} \frac{\widehat{R}_{S'}(h)-\widehat{R}^\rho_S(h)}{\sqrt[\alpha]{\frac{1}{2}[\widehat{R}^\rho_S(h)+\widehat{R}_{S'}(h)+\frac{1}{m}]}} > \epsilon\right]\\
&\geq \mathop{\mathbb{P}}_{S,S'\sim\mathcal{D}^m}\left[\frac{\widehat{R}_{S'}(h_S)-\widehat{R}^\rho_S(h_S)}{\sqrt[\alpha]{\frac{1}{2}[\widehat{R}^\rho_S(h_S)+\widehat{R}_{S'}(h_S)+\frac{1}{m}]}} > \epsilon\right] && \text{(def. of sup)}\\
&\geq \mathop{\mathbb{P}}_{S,S'\sim\mathcal{D}^m}\left[\left(\frac{R(h_S)-\widehat{R}^\rho_S(h_S)}{\sqrt[\alpha]{R(h_S)+\tau}} > \epsilon\right)\wedge\left(\widehat{R}_{S'}(h_S) > R(h_S)\right)\right] && \text{(implication (8))}\\
&= \mathop{\mathbb{E}}_{S,S'\sim\mathcal{D}^m}\left[1_{\frac{R(h_S)-\widehat{R}^\rho_S(h_S)}{\sqrt[\alpha]{R(h_S)+\tau}}>\epsilon}1_{\widehat{R}_{S'}(h_S)>R(h_S)}\right] && \text{(def. of expectation)}\\
&= \mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}\left[1_{\frac{R(h_S)-\widehat{R}^\rho_S(h_S)}{\sqrt[\alpha]{R(h_S)+\tau}}>\epsilon}\mathop{\mathbb{P}}_{S'\sim\mathcal{D}^m}\left[\widehat{R}_{S'}(h_S) > R(h_S)\right]\right]. && \text{(linearity of expectation)}
\end{aligned}$$

Now, observe that, if $R(h_S) \leq \epsilon^{\frac{\alpha}{\alpha-1}}$, then the following inequalities hold:

$$\frac{R(h_S)-\widehat{R}^\rho_S(h_S)}{\sqrt[\alpha]{R(h_S)+\tau}} \leq \frac{R(h_S)}{\sqrt[\alpha]{R(h_S)}} = R(h_S)^{\frac{\alpha-1}{\alpha}} \leq \epsilon. \tag{11}$$

In light of that, we can write

$$\begin{aligned}
&\mathop{\mathbb{P}}_{S,S'\sim\mathcal{D}^m}\left[\sup_{h\in\mathcal{H}} \frac{\widehat{R}_{S'}(h)-\widehat{R}^\rho_S(h)}{\sqrt[\alpha]{\frac{1}{2}[\widehat{R}^\rho_S(h)+\widehat{R}_{S'}(h)+\frac{1}{m}]}} > \epsilon\right]\\
&\geq \mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}\left[1_{\frac{R(h_S)-\widehat{R}^\rho_S(h_S)}{\sqrt[\alpha]{R(h_S)+\tau}}>\epsilon}1_{R(h_S)>\epsilon^{\frac{\alpha}{\alpha-1}}}\mathop{\mathbb{P}}_{S'\sim\mathcal{D}^m}\left[\widehat{R}_{S'}(h_S) > R(h_S)\right]\right]\\
&\geq \frac{1}{4}\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}\left[1_{\frac{R(h_S)-\widehat{R}^\rho_S(h_S)}{\sqrt[\alpha]{R(h_S)+\tau}}>\epsilon}\right] && (\epsilon^{\frac{\alpha}{\alpha-1}} > \tfrac{1}{m} \text{ and Lemma 9})\\
&\geq \frac{1}{4}\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}\left[1_{\sup_{h\in\mathcal{H}}\frac{R(h)-\widehat{R}^\rho_S(h)}{\sqrt[\alpha]{R(h)+\tau}}>\epsilon+\eta}\right] && (\text{def. of } h_S)\\
&= \frac{1}{4}\mathop{\mathbb{P}}_{S\sim\mathcal{D}^m}\left[\sup_{h\in\mathcal{H}}\frac{R(h)-\widehat{R}^\rho_S(h)}{\sqrt[\alpha]{R(h)+\tau}} > \epsilon+\eta\right]. && \text{(def. of expectation)}
\end{aligned}$$

Now, since this inequality holds for all $\eta > 0$, we can take the limit $\eta \to 0$ and use the right-continuity of the cumulative distribution to obtain

$$\mathop{\mathbb{P}}_{S,S'\sim\mathcal{D}^m}\left[\sup_{h\in\mathcal{H}} \frac{\widehat{R}_{S'}(h)-\widehat{R}^\rho_S(h)}{\sqrt[\alpha]{\frac{1}{2}[\widehat{R}^\rho_S(h)+\widehat{R}_{S'}(h)+\frac{1}{m}]}} > \epsilon\right] \geq \frac{1}{4}\mathop{\mathbb{P}}_{S\sim\mathcal{D}^m}\left[\sup_{h\in\mathcal{H}}\frac{R(h)-\widehat{R}^\rho_S(h)}{\sqrt[\alpha]{R(h)+\tau}} > \epsilon\right],$$

which completes the proof. □

Lemma 2. *Fix $\rho \geq 0$ and $1 < \alpha \leq 2$. Then, the following inequality holds:*

$$\mathop{\mathbb{P}}_{S,S'\sim\mathcal{D}^m}\left[\sup_{h\in\mathcal{H}} \frac{\widehat{R}_{S'}(h)-\widehat{R}^\rho_S(h)}{\sqrt[\alpha]{\frac{1}{2}[\widehat{R}_{S'}(h)+\widehat{R}^\rho_S(h)+\frac{1}{m}]}} > \epsilon\right] \leq \mathop{\mathbb{P}}_{S,S'\sim\mathcal{D}^m}\left[\sup_{g\in\mathcal{G}} \frac{\widehat{R}_{S'}(g)-\widehat{R}_S(g)}{\sqrt[\alpha]{\frac{1}{2}[\widehat{R}_{S'}(g)+\widehat{R}_S(g)+\frac{1}{m}]}} > \epsilon\right].$$

Further for $g(z) = 1_{yh(x)<\rho/2}$, using the shorthand $\mathcal{K} = \mathcal{C}(\mathcal{H}_\rho, \frac{\rho}{2}, S \cup S')$, the following holds:

$$\underset{S,S'\sim\mathcal{D}^m}{\mathbb{P}}\left[\sup_{h\in\mathcal{H}} \frac{\widehat{R}_{S'}(h) - \widehat{R}^\rho_S(h)}{\sqrt[\alpha]{\frac{1}{2}[\widehat{R}_{S'}(h) + \widehat{R}^\rho_S(h) + \frac{1}{m}]}} > \epsilon\right] \le \underset{S,S'\sim\mathcal{D}^m}{\mathbb{P}}\left[\sup_{h\in\mathcal{K}} \frac{\widehat{R}^{\frac{\rho}{2}}_{S'}(h) - \widehat{R}^{\frac{\rho}{2}}_S(h)}{\sqrt[\alpha]{\frac{1}{2}[\widehat{R}^{\frac{\rho}{2}}_{S'}(h) + \widehat{R}^{\frac{\rho}{2}}_S(h) + \frac{1}{m}]}} > \epsilon\right].$$

Proof. For the first part of the lemma, note that for any given h and the corresponding g, and sample $z \in S \cup S'$, using inequalities

$$1_{yh(x)<0} \le g(z) \le 1_{yh(x)<\rho}.$$

and taking expectations yields for any sample S:

$$\widehat{R}_S(h) \le R_S(g) \le \widehat{R}^\rho_S(h).$$

The result then follows by Lemma 8.

For the second part of the lemma, observe that restricting the output of $h \in \mathcal{H}$ to be in $[-\rho, \rho]$ does not change its binary or margin-loss: $1_{yh(x)<\rho} = 1_{yh_\rho(x)<\rho}$ and $1_{yh(x)\le 0} = 1_{yh_\rho(x)\le 0}$. Thus, we can write

$$\underset{S,S'\sim\mathcal{D}^m}{\mathbb{P}}\left[\sup_{h\in\mathcal{H}} \frac{\widehat{R}_{S'}(h) - \widehat{R}^\rho_S(h)}{\sqrt[\alpha]{\frac{1}{2}[\widehat{R}_{S'}(h) + \widehat{R}^\rho_S(h) + \frac{1}{m}]}} > \epsilon\right] = \underset{S,S'\sim\mathcal{D}^m}{\mathbb{P}}\left[\sup_{h\in\mathcal{H}_\rho} \frac{\widehat{R}_{S'}(h) - \widehat{R}^\rho_S(h)}{\sqrt[\alpha]{\frac{1}{2}[\widehat{R}_{S'}(h) + \widehat{R}^\rho_S(h) + \frac{1}{m}]}} > \epsilon\right].$$

Now, by definition of $\mathcal{C}(\mathcal{H}_\rho, \frac{\rho}{2}, S \cup S')$, for any $h \in \mathcal{H}_\rho$ there exists $g \in \mathcal{C}(\mathcal{H}_\rho, \frac{\rho}{2}, S \cup S')$ such that for any $x \in S \cup S'$,

$$|g(x) - h(x)| \le \frac{\rho}{2}.$$

Thus, for any $y \in \{-1, +1\}$ and $x \in S \cup S'$, we have $|yg(x) - yh(x)| \le \frac{\rho}{2}$, which implies:

$$1_{yh(x)\le 0} \le 1_{yg(x)\le\frac{\rho}{2}} \le 1_{yh(x)\le\rho}.$$

Hence, we have $\widehat{R}_{S'}(h) \le \widehat{R}^{\frac{\rho}{2}}_{S'}(g)$ and $\widehat{R}^\rho_S(h) \ge \widehat{R}^{\frac{\rho}{2}}_S(g)$ and, by the monotonicity properties of Lemma 8:

$$\frac{\widehat{R}_{S'}(h) - \widehat{R}^\rho_S(h)}{\sqrt[\alpha]{\frac{1}{2}[\widehat{R}_{S'}(h) + \widehat{R}^\rho_S(h) + \frac{1}{m}]}} \le \frac{\widehat{R}^{\frac{\rho}{2}}_{S'}(g) - \widehat{R}^{\frac{\rho}{2}}_S(g)}{\sqrt[\alpha]{\frac{1}{2}[\widehat{R}^{\frac{\rho}{2}}_{S'}(g) + \widehat{R}^{\frac{\rho}{2}}_S(g) + \frac{1}{m}]}}.$$

Taking the supremum over both sides yields the result. □

B. Relative Deviation Margin Bounds – Covering Numbers

Theorem 1 (General relative deviation margin bound)**.** *Fix $\rho \geq 0$ and $1 < \alpha \leq 2$. Then, for any hypothesis set $\mathcal{H}$ of functions mapping from $\mathcal{X}$ to $\mathbb{R}$ and any $\tau > 0$, the following inequality holds:*

$$\mathop{\mathbb{P}}_{S\sim\mathcal{D}^m}\left[\sup_{h\in\mathcal{H}}\frac{R(h)-\widehat{R}^{\rho}_S(h)}{\sqrt[\alpha]{R(h)+\tau}}>\epsilon\right]\leq 4\mathop{\mathbb{E}}_{x_1^{2m}\sim\mathcal{D}^{2m}}[\mathcal{N}_\infty(\mathcal{H}_\rho,\tfrac{\rho}{2},x_1^{2m})]\ \exp\left[\frac{-m^{\frac{2(\alpha-1)}{\alpha}}\epsilon^2}{2^{\frac{\alpha+2}{\alpha}}}\right].$$

Proof. Consider first the case where $m\epsilon^{\frac{\alpha}{\alpha-1}}\leq 1$. The bound then holds trivially since we have:

$$4\exp\left(\frac{-m^{\frac{2(\alpha-1)}{\alpha}}\epsilon^2}{2^{\frac{\alpha+2}{\alpha}}}\right)\geq 4\exp\left(\frac{-1}{2^{\frac{\alpha+2}{\alpha}}}\right)>1.$$

On the other hand, when $m\epsilon^{\frac{\alpha}{\alpha-1}}>1$, by Lemmas 1 and 2 we can write:

$$\mathop{\mathbb{P}}_{S\sim\mathcal{D}^m}\left[\sup_{h\in\mathcal{H}}\frac{R(h)-\widehat{R}^{\rho}_S(h)}{\sqrt[\alpha]{R(h)+\tau}}>\epsilon\right]\leq 4\mathop{\mathbb{P}}_{S,S'\sim\mathcal{D}^m}\left[\sup_{h\in\mathcal{C}(\mathcal{H}_\rho,\frac{\rho}{2},S\cup S')}\frac{\widehat{R}^{\frac{\rho}{2}}_{S'}(h)-\widehat{R}^{\frac{\rho}{2}}_{S}(h)}{\sqrt[\alpha]{\frac{1}{2}[\widehat{R}^{\frac{\rho}{2}}_{S'}(h)+\widehat{R}^{\frac{\rho}{2}}_{S}(h)+\frac{1}{m}]}}>\epsilon\right].$$

To upper bound the probability that the symmetrized expression is larger than ϵ, we begin by introducing a vector of Rademacher random variables $\sigma=(\sigma_1,\sigma_2,\ldots,\sigma_m)$, where σ_is are independent identically distributed random variables each equally likely to take the value $+1$ or -1. Let $x_1,x_2,\ldots x_m$ be samples in S and $x_{m+1},x_{m+2},\ldots x_{2m}$ be samples in S'. Using the shorthands $z=(x,y)$, $g(z)=1_{yh(x)\leq\frac{\rho}{2}}$, and $\mathcal{G}(x_1^{2m})=\mathcal{C}(\mathcal{H}_\rho,\frac{\rho}{2},S\cup S')$, we can then write the above quantity as

$$\begin{aligned}&\mathop{\mathbb{P}}_{S,S'\sim\mathcal{D}^m}\left[\sup_{h\in\mathcal{C}(\mathcal{H}_\rho,\frac{\rho}{2},S\cup S')}\frac{\widehat{R}^{\frac{\rho}{2}}_{S'}(h)-\widehat{R}^{\frac{\rho}{2}}_{S}(h)}{\sqrt[\alpha]{\frac{1}{2}[\widehat{R}^{\frac{\rho}{2}}_{S'}(h)+\widehat{R}^{\frac{\rho}{2}}_{S}(h)+\frac{1}{m}]}}>\epsilon\right]\\
&=\mathop{\mathbb{P}}_{z_1^{2m}\sim\mathcal{D}^{2m}}\left[\sup_{g\in\mathcal{G}(x^{2m})}\frac{\frac{1}{m}\sum_{i=1}^m(g(z_{m+i})-g(z_i))}{\sqrt[\alpha]{\frac{1}{2m}[\sum_{i=1}^m(g(z_{m+i})+g(z_i))+1]}}>\epsilon\right]\\
&=\mathop{\mathbb{P}}_{z_1^{2m}\sim\mathcal{D}^{2m},\boldsymbol{\sigma}}\left[\sup_{g\in\mathcal{G}(x^{2m})}\frac{\frac{1}{m}\sum_{i=1}^m\sigma_i(g(z_{m+i})-g(z_i))}{\sqrt[\alpha]{\frac{1}{2m}[\sum_{i=1}^m(g(z_{m+i})+g(z_i))+1]}}>\epsilon\right]\\
&=\mathop{\mathbb{E}}_{z_1^{2m}\sim\mathcal{D}^{2m}}\left[\mathop{\mathbb{P}}_{\boldsymbol{\sigma}}\left[\sup_{g\in\mathcal{G}(x^{2m})}\frac{\frac{1}{m}\sum_{i=1}^m\sigma_i(g(z_{m+i})-g(z_i))}{\sqrt[\alpha]{\frac{1}{2m}[\sum_{i=1}^m(g(z_{m+i})+g(z_i))+1]}}>\epsilon\,\middle|\,z_1^{2m}\right]\right].\end{aligned}$$

Now, for a fixed z_1^{2m}, we have $\mathbb{E}_{\boldsymbol{\sigma}}\left[\frac{\frac{1}{m}\sum_{i=1}^m\sigma_i(g(z_{m+i})-g(z_i))}{\sqrt[\alpha]{\frac{1}{2m}[\sum_{i=1}^m(g(z_{m+i})+g(z_i))+1]}}\right]=0$, thus, by Hoeffding's inequality, we can write

$$\begin{aligned}\mathop{\mathbb{P}}_{\boldsymbol{\sigma}}\left[\frac{\frac{1}{m}\sum_{i=1}^m\sigma_i(g(z_{m+i})-g(z_i))}{\sqrt[\alpha]{\frac{1}{2m}[\sum_{i=1}^m(g(z_{m+i})+g(z_i))+1]}}>\epsilon\,\middle|\,z_1^{2m}\right]&\leq\exp\left(\frac{-[\sum_{i=1}^m(g(z_{m+i})+g(z_i))+1]^{\frac{2}{\alpha}}m^{\frac{2(\alpha-1)}{\alpha}}\epsilon^2}{2^{\frac{\alpha+2}{\alpha}}\sum_{i=1}^m(g(z_{m+i})-g(z_i))^2}\right)\\
&\leq\exp\left(\frac{-[\sum_{i=1}^m(g(z_{m+i})+g(z_i))]^{\frac{2}{\alpha}}m^{\frac{2(\alpha-1)}{\alpha}}\epsilon^2}{2^{\frac{\alpha+2}{\alpha}}\sum_{i=1}^m(g(z_{m+i})-g(z_i))^2}\right).\end{aligned}$$

Since the variables $g(z_i)$, $i\in[1,2m]$, take values in $\{0,1\}$, we can write

$$\begin{aligned}\sum_{i=1}^m(g(z_{m+i})-g(z_i))^2&=\sum_{i=1}^m g(z_{m+i})+g(z_i)-2g(z_{m+i})g(z_i)\\
&\leq\sum_{i=1}^m g(z_{m+i})+g(z_i)\\
&\leq\sum_{i=1}^m[g(z_{m+i})+g(z_i)]^{\frac{2}{\alpha}}\ ,\end{aligned}$$

where the last inequality holds since $\alpha \leq 2$ and since the sum is either zero or greater than or equal to one. In view of this identity, we can write

$$\mathop{\mathbb{P}}_{\boldsymbol{\sigma}}\left[\frac{\frac{1}{m}\sum_{i=1}^{m}\sigma_i\big(g(z_{m+i})-g(z_i)\big)}{\sqrt[\alpha]{\frac{1}{2m}\big[\sum_{i=1}^{m}\big(g(z_{m+i})+g(z_i)\big)\big]}}>\epsilon\,\middle|\,z_1^{2m}\right]\leq\exp\left(\frac{-m^{\frac{2(\alpha-1)}{\alpha}}\epsilon^2}{2^{\frac{\alpha+2}{\alpha}}}\right).$$

The number of such hypotheses is $\mathcal{N}_\infty(\mathcal{H}_\rho,\frac{\rho}{2},x_1^{2m})$, thus, by the union bound, the following holds:

$$\mathop{\mathbb{P}}_{\boldsymbol{\sigma}}\left[\sup_{g\in\mathcal{G}(x^{2m})}\frac{\sum_{i=1}^{m}\sigma_i\big(g(z_{m+i})-g(z_i)\big)}{\sqrt[\alpha]{\frac{1}{2}\big[\sum_{i=1}^{m}\big(g(z_{m+i})+g(z_i)\big)\big]}}>\epsilon\,\middle|\,z_1^{2m}\right]\leq\mathcal{N}_\infty(\mathcal{H}_\rho,\tfrac{\rho}{2},x_1^{2m})\exp\left(\frac{-m^{\frac{2(\alpha-1)}{\alpha}}\epsilon^2}{2^{\frac{\alpha+2}{\alpha}}}\right).$$

The result follows by taking expectations with respect to z_1^{2m} and applying the previous lemmas. □

C. Relative Deviation Margin Bounds – Rademacher Complexity

The following lemma relates the symmetrized expression of Lemma 2 to a Rademacher average quantity.

Lemma 3. *Fix $1 < \alpha \leq 2$. Then, the following inequality holds:*

$$\mathop{\mathbb{P}}_{S,S'\sim\mathcal{D}^m}\left[\sup_{g\in\mathcal{G}}\frac{\widehat{R}_{S'}(g)-\widehat{R}_S(g)}{\sqrt[\alpha]{\frac{1}{2}[\widehat{R}_{S'}(g)+\widehat{R}_S(g)+\frac{1}{m}]}}>\epsilon\right]\leq 2\mathop{\mathbb{P}}_{z_1^m\sim\mathcal{D}^m,\boldsymbol{\sigma}}\left[\sup_{g\in\mathcal{G}}\frac{\frac{1}{m}\sum_{i=1}^m\sigma_i g(z_i)}{\sqrt[\alpha]{\frac{1}{m}[\sum_{i=1}^m g(z_i)+1]}}>\frac{\epsilon}{2\sqrt{2}}\right].$$

Proof. To upper bound the probability that the symmetrized expression is larger than ϵ, we begin by introducing a vector of Rademacher random variables $\sigma = (\sigma_1, \sigma_2, \ldots, \sigma_m)$, where σ_is are independent identically distributed random variables each equally likely to take the value $+1$ or -1. Let $z_1, z_2, \ldots z_m$ be samples in S and $z_{m+1}, z_{m+2}, \ldots z_{2m}$ be samples in S'. We can then write the above quantity as

$$\begin{aligned}&\mathop{\mathbb{P}}_{S,S'\sim\mathcal{D}^m}\left[\sup_{g\in\mathcal{G}}\frac{\widehat{R}_{S'}(g)-\widehat{R}_S(g)}{\sqrt[\alpha]{\frac{1}{2}[\widehat{R}_{S'}(g)+\widehat{R}_S(g)+\frac{1}{m}]}}>\epsilon\right]\\&=\mathop{\mathbb{P}}_{z_1^{2m}\sim\mathcal{D}^{2m}}\left[\sup_{g\in\mathcal{G}}\frac{\frac{1}{m}\sum_{i=1}^m(g(z_{m+i})-g(z_i))}{\sqrt[\alpha]{\frac{1}{2m}[\sum_{i=1}^m(g(z_{m+i})+g(z_i))+1]}}>\epsilon\right]\\&=\mathop{\mathbb{P}}_{z_1^{2m}\sim\mathcal{D}^{2m},\boldsymbol{\sigma}}\left[\sup_{g\in\mathcal{G}}\frac{\frac{1}{m}\sum_{i=1}^m\sigma_i(g(z_{m+i})-g(z_i))}{\sqrt[\alpha]{\frac{1}{2m}[\sum_{i=1}^m(g(z_{m+i})+g(z_i))+1]}}>\epsilon\right].\end{aligned}$$

If $a + b \geq \epsilon$, then either $a \geq \epsilon/2$ or $b \geq \epsilon/2$, hence

$$\begin{aligned}&\mathop{\mathbb{P}}_{z_1^{2m}\sim\mathcal{D}^{2m},\boldsymbol{\sigma}}\left[\sup_{g\in\mathcal{G}}\frac{\frac{1}{m}\sum_{i=1}^m\sigma_i(g(z_{m+i})-g(z_i))}{\sqrt[\alpha]{\frac{1}{2m}[\sum_{i=1}^m(g(z_{m+i})+g(z_i))+1]}}>\epsilon\right]\\&\leq\mathop{\mathbb{P}}_{z_1^{2m}\sim\mathcal{D}^{2m},\boldsymbol{\sigma}}\left[\sup_{g\in\mathcal{G}}\frac{\frac{1}{m}\sum_{i=1}^m\sigma_i(g(z_{m+i}))}{\sqrt[\alpha]{\frac{1}{2m}[\sum_{i=1}^m(g(z_{m+i})+g(z_i))+1]}}>\frac{\epsilon}{2}\right]\\&+\mathop{\mathbb{P}}_{z_1^{2m}\sim\mathcal{D}^{2m},\boldsymbol{\sigma}}\left[\sup_{g\in\mathcal{G}}\frac{\frac{1}{m}\sum_{i=1}^m\sigma_i(-g(z_i))}{\sqrt[\alpha]{\frac{1}{2m}[\sum_{i=1}^m(g(z_{m+i})+g(z_i))+1]}}>\frac{\epsilon}{2}\right]\\&=2\mathop{\mathbb{P}}_{z_1^{2m}\sim\mathcal{D}^{2m},\boldsymbol{\sigma}}\left[\sup_{g\in\mathcal{G}}\frac{\frac{1}{m}\sum_{i=1}^m\sigma_i g(z_i)}{\sqrt[\alpha]{\frac{1}{2m}[\sum_{i=1}^m(g(z_{m+i})+g(z_i))+1]}}>\frac{\epsilon}{2}\right]\\&\leq 2\mathop{\mathbb{P}}_{z_1^{2m}\sim\mathcal{D}^{2m},\boldsymbol{\sigma}}\left[\sup_{g\in\mathcal{G}}\frac{\frac{1}{m}\sum_{i=1}^m\sigma_i g(z_i)}{\sqrt[\alpha]{\frac{1}{2m}[\sum_{i=1}^m(g(z_i))+1]}}>\frac{\epsilon}{2}\right]\\&\leq 2\mathop{\mathbb{P}}_{z_1^{2m}\sim\mathcal{D}^{2m},\boldsymbol{\sigma}}\left[\sup_{g\in\mathcal{G}}\frac{\frac{1}{m}\sum_{i=1}^m\sigma_i g(z_i)}{\sqrt[\alpha]{\frac{1}{m}[\sum_{i=1}^m(g(z_i))+1]}}>\frac{\epsilon}{2\sqrt{2}}\right]\\&=2\mathop{\mathbb{P}}_{z_1^{m}\sim\mathcal{D}^{m},\boldsymbol{\sigma}}\left[\sup_{g\in\mathcal{G}}\frac{\frac{1}{m}\sum_{i=1}^m\sigma_i g(z_i)}{\sqrt[\alpha]{\frac{1}{m}[\sum_{i=1}^m(g(z_i))+1]}}>\frac{\epsilon}{2\sqrt{2}}\right],\end{aligned}$$

where the penultimate inequality follow by observing that if $a/c \geq \epsilon$, then $a/c' \geq \epsilon$, for all $c' \leq c$ and the last inequality follows by observing $\alpha \geq 1$. □

We will use the following bounded difference inequality (van Handel, 2016, Theorem 3.18), which provide us with a finer tool that McDiarmid's inequality.

Lemma 10 ((van Handel, 2016)). *Let $f(x_1, x_2, \ldots, x_n)$ be a function of n independent samples $x_1, x_2, \ldots x_n$. Let*

$$c_i = \max_{x_i'} f(x_1, x_2, \ldots, x_n) - f(x_1, x_2, \ldots, x_{i-1}, x_i', x_{i+1}, \ldots, x_n).$$

Then,

$$\mathbb{P}\left(f(x_1, x_2, \ldots, x_n) \geq \mathbb{E}[f(x_1, x_2, \ldots, x_n)] + \epsilon\right) \leq \exp\left(-\frac{\epsilon^2}{4\sum_i c_i^2}\right).$$

Using the above inequality and a peeling argument, we show the following upper bound expressed in terms of Rademacher complexities.

Lemma 4. *Fix $1 < \alpha \leq 2$ and $z_1^m \in \mathcal{Z}^m$. Then, the following inequality holds:*

$$\mathbb{P}_{\boldsymbol{\sigma}}\left[\sup_{g \in \mathcal{G}} \frac{\frac{1}{m}\sum_{i=1}^m \sigma_i g(z_i)}{\sqrt[\alpha]{\frac{1}{m}[\sum_{i=1}^m (g(z_i)) + 1]}} > \epsilon \,\middle|\, z^m\right] \leq 2\sum_{k=0}^{\lfloor \log_2 m \rfloor} \exp\left[\frac{m^2 \widehat{\mathfrak{R}}_m^2(\mathcal{G}_k(z_1^m))}{2^{k+5}} - \frac{\epsilon^2}{64\frac{2^{k(1-2/\alpha)}}{m^{2-2/\alpha}}}\right] 1_{\epsilon \leq 2[\frac{2^k}{m}]^{1-\frac{1}{\alpha}}}.$$

Proof. By definition of $\mathcal{G}_k$, the following inequality holds:

$$\sup_{g \in \mathcal{G}_k(z_1^m)} \frac{\frac{1}{m}\sum_{i=1}^m \sigma_i g(z_i)}{\sqrt[\alpha]{\frac{1}{m}[\sum_{i=1}^m (g(z_i)) + 1]}} \leq \frac{\frac{2^{k+1}}{m}}{\sqrt[\alpha]{\frac{1}{m}[\sum_{i=1}^m (g(z_i)) + 1]}} \leq \frac{\frac{2^{k+1}}{m}}{\left(\frac{2^k}{m}\right)^{1/\alpha}}.$$

Thus, for $\epsilon > 2\left(\frac{2^k}{m}\right)^{1-1/\alpha}$, the left-hand side probability is zero. This leads to the indicator function factor in the right-hand side of the expression. We now prove the non-indicator part.

By the union bound,

$$\begin{aligned}
\mathbb{P}_{\boldsymbol{\sigma}}\left[\sup_{g \in \mathcal{G}} \frac{\frac{1}{m}\sum_{i=1}^m \sigma_i g(z_i)}{\sqrt[\alpha]{\frac{1}{m}[\sum_{i=1}^m (g(z_i)) + 1]}} > \epsilon \,\middle|\, z^m\right] &= \mathbb{P}_{\boldsymbol{\sigma}}\left[\sup_k \sup_{g \in \mathcal{G}_k(z_1^m)} \frac{\frac{1}{m}\sum_{i=1}^m \sigma_i g(z_i)}{\sqrt[\alpha]{\frac{1}{m}[\sum_{i=1}^m (g(z_i)) + 1]}} > \epsilon \,\middle|\, z^m\right] \\
&\leq \sum_k \mathbb{P}_{\boldsymbol{\sigma}}\left[\sup_{g \in \mathcal{G}_k(z_1^m)} \frac{\frac{1}{m}\sum_{i=1}^m \sigma_i g(z_i)}{\sqrt[\alpha]{\frac{1}{m}[\sum_{i=1}^m (g(z_i)) + 1]}} > \epsilon \,\middle|\, z^m\right] \\
&\leq \sum_k \mathbb{P}_{\boldsymbol{\sigma}}\left[\sup_{g \in \mathcal{G}_k(z_1^m)} \frac{\frac{1}{m}|\sum_{i=1}^m \sigma_i g(z_i)|}{\sqrt[\alpha]{\frac{1}{m}[\sum_{i=1}^m (g(z_i)) + 1]}} > \epsilon \,\middle|\, z^m\right] \\
&\overset{(a)}{\leq} \sum_k \mathbb{P}_{\boldsymbol{\sigma}}\left[\sup_{g \in \mathcal{G}_k(z_1^m)} \frac{1}{m}|\sum_{i=1}^m \sigma_i g(z_i)| > \epsilon\sqrt[\alpha]{\frac{2^k}{m}} \,\middle|\, z^m\right] \\
&\overset{(b)}{\leq} \sum_k 2\,\mathbb{P}_{\boldsymbol{\sigma}}\left[\sup_{g \in \mathcal{G}_k(z_1^m)} \frac{1}{m}\sum_{i=1}^m \sigma_i g(z_i) > \epsilon\sqrt[\alpha]{\frac{2^k}{m}} \,\middle|\, z^m\right],
\end{aligned}$$

where the (a) follows by observing that for all $g \in \mathcal{G}_k$, $[\sum_{i=1}^m (g(z_i)) + 1] \geq 2^k/m$ and (b) follows by observing that for a particular $\boldsymbol{\sigma}$, $\frac{1}{m}\sum_{i=1}^m \sigma_i g(z_i) < \epsilon\sqrt[\alpha]{\frac{2^k}{m}}$, then for $\boldsymbol{\sigma}' = -\boldsymbol{\sigma}$, the value would be $\frac{1}{m}\sum_{i=1}^m \sigma_i' g(z_i) > \epsilon\sqrt[\alpha]{\frac{2^k}{m}}$. Hence it suffices to bound

$$\mathbb{P}_{\boldsymbol{\sigma}}\left[\sup_{g \in \mathcal{G}_k(z_1^m)} \frac{1}{m}\sum_{i=1}^m \sigma_i g(z_i) > \epsilon\sqrt[\alpha]{\frac{2^k}{m}} \,\middle|\, z^m\right],$$

for a given k. We will apply the bounded difference inequality ((van Handel, 2016, Theorem 3.18)), which is a finer concentration bound than McDiarmid's inequality in this context, to the random variable $\sup_{g \in \mathcal{G}_k(z_1^m)} \frac{1}{m}\sum_{i=1}^m \sigma_i g(z_i)$. For any $\boldsymbol{\sigma}$, let $g_{\boldsymbol{\sigma}}$ denote the function in $\mathcal{G}_k(z_1^m)$ that achieves the supremum. For simplicity, we assume that the supremum can be achieved. The proof can be extended to the case when its not achieved. Then, for any two vectors of Rademacher variables $\boldsymbol{\sigma}$ and $\boldsymbol{\sigma}'$ that differ only in the j^{th} coordinate, the difference of suprema can be bounded as follows:

$$\begin{aligned}\frac{1}{m}\sum_{i=1}^{m}\sigma_i g_{\boldsymbol{\sigma}}(z_i) - \frac{1}{m}\sum_{i=1}^{m}\sigma'_i g_{\boldsymbol{\sigma}'}(z_i) &\leq \frac{1}{m}\sum_{i=1}^{m}\sigma_i g_{\boldsymbol{\sigma}}(z_i) - \frac{1}{m}\sum_{i=1}^{m}\sigma'_i g_{\boldsymbol{\sigma}}(z_i)\\ &= \frac{1}{m}(\sigma_j - \sigma'_j)g_{\boldsymbol{\sigma}}(z_j)\\ &\leq \frac{2g_{\boldsymbol{\sigma}}(z_j)}{m}.\end{aligned}$$

The sum of the squares of the changes is therefore bounded by

$$\frac{4}{m^2}\sum_{i=1}^{m} g_{\boldsymbol{\sigma}}^2(z_i) \leq \frac{4}{m^2}\sup_{g\in\mathcal{G}_k(z_1^m)}\sum_{i=1}^{m} g^2(z_i) \leq \frac{4}{m^2}\sup_{g\in\mathcal{G}_k(z_1^m)}\sum_{i=1}^{m} g(z_i) \leq \frac{4}{m^2}m2^{k+1} = \frac{2^{k+3}}{m}.$$

Since $\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{g\in\mathcal{G}_k(z_1^m)}\frac{1}{m}\sum_{i=1}^{m}\sigma_i g(z_i)\right] = \widehat{\mathfrak{R}}_{z_1^m}(\mathcal{G}_k(z_1^m))$, by the Lemma 10, for $\epsilon \geq \frac{\widehat{\mathfrak{R}}_{z_1^m}(\mathcal{G}_k(z_1^m))}{\sqrt[\alpha]{2^k/m}}$, the following holds:

$$\begin{aligned}&\mathbb{P}_{\boldsymbol{\sigma}}\left[\sup_{g\in\mathcal{G}_k(z_1^m)}\frac{1}{m}\sum_{i=1}^{m}\sigma_i g(z_i) > \epsilon\sqrt[\alpha]{\frac{2^k}{m}}\,\middle|\, z^m\right]\\ &= \mathbb{P}_{\boldsymbol{\sigma}}\left[\sup_{g\in\mathcal{G}_k(z_1^m)}\frac{1}{m}\sum_{i=1}^{m}\sigma_i g(z_i) - \widehat{\mathfrak{R}}_m(\mathcal{G}_k(z_1^m)) > \epsilon\sqrt[\alpha]{\frac{2^k}{m}} - \widehat{\mathfrak{R}}_m(\mathcal{G}_k(z_1^m))\,\middle|\, z^m\right]\\ &\leq \exp\left(-\frac{m\left[\epsilon\sqrt[\alpha]{\frac{2^k}{m}} - \widehat{\mathfrak{R}}_{z_1^m}(\mathcal{G}_k(z_1^m))\right]^2}{2^{k+5}}\right) = \exp\left(-\frac{\left(\epsilon - \frac{\widehat{\mathfrak{R}}_{z_1^m}(\mathcal{G}_k(z_1^m))}{\sqrt[\alpha]{\frac{2^k}{m}}}\right)^2}{32\frac{2^{k(1-2/\alpha)}}{m^{2-2/\alpha}}}\right).\end{aligned}$$

Since, $-(\epsilon - a)^2 \leq a^2 - \epsilon^2/2$, for $\epsilon \geq \frac{\widehat{\mathfrak{R}}_m(\mathcal{G}_k(z_1^m))}{\sqrt[\alpha]{2^k/m}}$, we can write:

$$\begin{aligned}\mathbb{P}_{\boldsymbol{\sigma}}\left[\sup_{g\in\mathcal{G}_k(z_1^m)}\frac{1}{m}\sum_{i=1}^{m}\sigma_i g(z_i) > \epsilon\sqrt[\alpha]{\frac{2^k}{m}}\,\middle|\, z^m\right] &\leq \exp\left(\frac{\left(\frac{\widehat{\mathfrak{R}}_m(\mathcal{G}_k(z_1^m))}{\sqrt[\alpha]{2^k/m}}\right)^2}{32\frac{2^{k(1-2/\alpha)}}{m^{2-2/\alpha}}}\right)\cdot\exp\left(-\frac{\epsilon^2}{64\frac{2^{k(1-2/\alpha)}}{m^{2-2/\alpha}}}\right)\\ &= \exp\left(\frac{m^2\widehat{\mathfrak{R}}_m^2(\mathcal{G}_k(z_1^m))}{2^{k+5}}\right)\cdot\exp\left(-\frac{\epsilon^2}{64\frac{2^{k(1-2/\alpha)}}{m^{2-2/\alpha}}}\right).\end{aligned}$$

For $\epsilon < \frac{\widehat{\mathfrak{R}}_m(\mathcal{G}_k(z_1^m))}{\sqrt[\alpha]{2^k/m}}$, the bound holds trivially since the right-hand side is at most one. □

The following is a margin-based relative deviation bound expressed in terms of Rademacher complexities.

Theorem 2. *Fix $1 < \alpha \leq 2$. Then, with probability at least $1 - \delta$, for all hypothesis $h \in \mathcal{H}$, the following inequality holds:*

$$R(h) - \widehat{R}_S^{\rho}(h) \leq 16\sqrt{2}\sqrt[\alpha]{R(h)}\left[\frac{\mathfrak{r}_m(\mathcal{G}) + \log\log m + \log\frac{16}{\delta}}{m}\right]^{1-\frac{1}{\alpha}}.$$

Proof. Let $\mathfrak{r}_m^k(\mathcal{G})$ be the k-peeling-based Rademacher complexity of $\mathcal{G}$ defined as follows:

$$\mathfrak{r}_m^k(\mathcal{G}) = \log\mathop{\mathbb{E}}_{z_1^m}\left[\exp\left(\frac{m^2\widehat{\mathfrak{R}}_m^2(\mathcal{G}_k(z_1^m))}{2^{k+5}}\right)\right].$$

Combining Lemmas 1, 2, 3, and 4 yields:

$$
\begin{aligned}
&\mathop{\mathbb{P}}_{S\sim\mathcal{D}^m}\left[\sup_{h\in\mathcal{H}} \frac{R(h)-\widehat{R}^{\rho}_S(h)}{\sqrt[\alpha]{R(h)+\tau}} > \epsilon\right]\\
&\leq 8 \mathop{\mathbb{P}}_{z_1^m\sim\mathcal{D}^m,\boldsymbol{\sigma}}\left[\sup_{g\in\mathcal{G}} \frac{\frac{1}{m}\sum_{i=1}^m \sigma_i g(z_i)}{\sqrt[\alpha]{\frac{1}{m}[\sum_{i=1}^m (g(z_i))+1]}} > \frac{\epsilon}{2\sqrt{2}}\right]\\
&= 8 \mathop{\mathbb{E}}_{z^m\sim\mathcal{D}^m}\left[\mathop{\mathbb{P}}_{\boldsymbol{\sigma}}\left[\sup_{g\in\mathcal{G}} \frac{\frac{1}{m}\sum_{i=1}^m \sigma_i g(z_i)}{\sqrt[\alpha]{\frac{1}{m}[\sum_{i=1}^m (g(z_i))+1]}} > \frac{\epsilon}{2\sqrt{2}} \,\middle|\, z^m\right]\right]\\
&\leq 16 \mathop{\mathbb{E}}_{z^m\sim\mathcal{D}^m}\left[\sum_k \exp\left(\frac{m^2\widehat{\mathfrak{R}}^2_m(\mathcal{G}_k(z_1^m))}{2^{k+5}}\right)\cdot\exp\left(-\frac{\epsilon^2}{512\frac{2^{k(1-2/\alpha)}}{m^{2-2/\alpha}}}\right)1_{\epsilon\leq 4\sqrt{2}\left(\frac{2^k}{m}\right)^{1-1/\alpha}}\right]\\
&= 16 \sum_k \mathop{\mathbb{E}}_{z^m\sim\mathcal{D}^m}\left[\exp\left(\frac{m^2\widehat{\mathfrak{R}}^2_m(\mathcal{G}_k(z_1^m))}{2^{k+5}}\right)\right]\cdot\exp\left(-\frac{\epsilon^2}{512\frac{2^{k(1-2/\alpha)}}{m^{2-2/\alpha}}}\right)1_{\epsilon\leq 4\sqrt{2}\left(\frac{2^k}{m}\right)^{1-1/\alpha}}\\
&\leq 16(\log_2 m) \mathop{\mathbb{E}}_{z^m\sim\mathcal{D}^m}\left[\exp\left(\frac{m^2\widehat{\mathfrak{R}}^2_m(\mathcal{G}_k(z_1^m))}{2^{k+5}}\right)\right]\cdot\exp\left(-\frac{\epsilon^2}{512\frac{2^{k(1-2/\alpha)}}{m^{2-2/\alpha}}}\right)1_{\epsilon\leq 4\sqrt{2}\left(\frac{2^k}{m}\right)^{1-1/\alpha}}\\
&\leq 16(\log_2 m)\sup_k e^{\mathfrak{r}^k_m(\mathcal{G})}\cdot\exp\left(-\frac{\epsilon^2}{512\frac{2^{k(1-2/\alpha)}}{m^{2-2/\alpha}}}\right)1_{\epsilon\leq 4\sqrt{2}\left(\frac{2^k}{m}\right)^{1-1/\alpha}}
\end{aligned}
$$

Hence, with probability at least $1-\delta$,

$$
\sup_{h\in\mathcal{H}} \frac{R(h)-\widehat{R}^{\rho}_S(h)}{\sqrt[\alpha]{R(h)+\tau}} \leq \sup_k \min\left(16\sqrt{2}\frac{2^{k(1/2-1/\alpha)}}{m^{1-1/\alpha}}\sqrt{\mathfrak{r}^k_m(\mathcal{G})+\log\log m+\log\frac{16}{\delta}}, 4\sqrt{2}\left(\frac{2^k}{m}\right)^{1-1/\alpha}\right).
$$

For $\alpha\leq 2$, the first term in the minimum decreases with k and the second term increases with k. Let k_0 be such that

$$
2^{k_0} = 16\left(\sup_k \mathfrak{r}^k_m(\mathcal{G})+\log\log m+\log\frac{16}{\delta}\right) = 16\left(\mathfrak{r}_m(\mathcal{G})+\log\log m+\log\frac{16}{\delta}\right).
$$

Then for any k,

$$
\begin{aligned}
&\sup_k \min\left(16\sqrt{2}\frac{2^{k(1/2-1/\alpha)}}{m^{1-1/\alpha}}\sqrt{\mathfrak{r}^k_m(\mathcal{G})+\log\log m+\log\frac{16}{\delta}}, 4\sqrt{2}\left(\frac{2^k}{m}\right)^{1-1/\alpha}\right)\\
&\leq \sup_k \max\left(16\sqrt{2}\frac{2^{k_0(1/2-1/\alpha)}}{m^{1-1/\alpha}}\sqrt{\mathfrak{r}^k_m(\mathcal{G})+\log\log m+\log\frac{16}{\delta}}, 4\sqrt{2}\left(\frac{2^{k_0}}{m}\right)^{1-1/\alpha}\right)\\
&\leq \max\left(16\sqrt{2}\frac{2^{k_0(1/2-1/\alpha)}}{m^{1-1/\alpha}}\sqrt{\mathfrak{r}_m(\mathcal{G})+\log\log m+\log\frac{16}{\delta}}, 4\sqrt{2}\left(\frac{2^{k_0}}{m}\right)^{1-1/\alpha}\right)\\
&\leq 4\sqrt{2}\left(\frac{2^{k_0}}{m}\right)^{1-1/\alpha}\\
&\leq 16\sqrt{2}\left(\frac{\mathfrak{r}_m(\mathcal{G})+\log\log m+\log\frac{16}{\delta}}{m}\right)^{1-1/\alpha}.
\end{aligned}
$$

Rearranging and taking the limit as $\tau\to 0$ yields the result. □

Lemma 11. *For any $x, y, z\geq 0$, if $(x - y\sqrt[\alpha]{x}\leq z)$, then the following inequality holds:*

$$
x \leq z + 2y\sqrt[\alpha]{z} + (2y)^{\frac{\alpha}{\alpha-1}}.
$$

Proof. In view of the assumption, we can write:

$$x \leq z + y\sqrt[\alpha]{x} \leq 2\max(z, y\sqrt[\alpha]{x}),$$

If $z \geq y\sqrt[\alpha]{x}$, then $x \leq 2z$. if $z \leq y\sqrt[\alpha]{x}$, then $x \leq (2y)^{\alpha/(\alpha-1)}$. This shows that we have $x \leq 2\max(z, (2y)^{1-1/\alpha})$. Plugging in the right-hand side in the previous inequality and using the sub-additivity of $x \mapsto \sqrt[\alpha]{x}$ gives:

$$x \leq z + y\sqrt[\alpha]{x} \leq z + y\sqrt[\alpha]{2\max(z, (2y)^{\alpha/(\alpha-1)})} \leq z + y\sqrt[\alpha]{2z} + y^{\frac{\alpha}{\alpha-1}} 2^{\frac{1}{\alpha}+\frac{1}{\alpha-1}}.$$

The lemma follows by observing that $2^{\frac{1}{\alpha}} \leq 2$ for $\alpha \geq 1$. □

Corollary 6. *Let $\mathcal{G}$ be defined as above. Then, with probability at least $1-\delta$, for all hypothesis $h \in \mathcal{H}$ and $\alpha \in (1, 2]$,*

$$R(h) - \widehat{R}_S^\rho(h) \leq 32\sqrt{2}\sqrt[\alpha]{R(h)}\left[\frac{\mathfrak{r}_m(\mathcal{G}) + \log\frac{16\log m}{\delta}}{m}\right]^{1-\frac{1}{\alpha}}.$$

Proof. By Theorem 2,

$$R(h) - \widehat{R}_S^\rho(h) \leq 16\sqrt[\alpha]{R(h)}\left(\frac{\mathfrak{r}_m(\mathcal{G}) + \log\log m + \log\frac{16}{\delta}}{m}\right)^{1-1/\alpha}.$$

Let $B = \mathfrak{r}_m(\mathcal{G}) + \log\log m + \log\frac{16}{\delta}$. Let $\alpha_k = 1 + e^{-\epsilon k}$. Let $\delta_k = \delta/k^2$. Then, by the union bound, for all α_k, with probability at least $1-\delta$,

$$R(h) - \widehat{R}_S^\rho(h) \leq 16\sqrt{2}\sqrt[\alpha_k]{R(h)}\left(\frac{B + 2\log k}{m}\right)^{1-1/\alpha_k}.$$

Let $\alpha_k \geq \alpha \geq \alpha_{k+1}$. Then $(k+1) \leq \frac{1}{\epsilon}\log\frac{1}{\alpha-1}$. Then,

$$\begin{aligned}
&\sqrt[\alpha]{R(h)}\left(\frac{B + \log\frac{1}{\alpha-1}}{m}\right)^{1-1/\alpha} \\
&\sqrt[\alpha]{R(h)}\left(\frac{B + 2\log(k+1)}{m}\right)^{1-1/\alpha} \\
&\geq \min\left(\sqrt[\alpha_k]{R(h)}\left(\frac{B + 2\log(k+1)}{m}\right)^{1-1/\alpha_k}, \sqrt[\alpha_{k+1}]{R(h)}\left(\frac{B + 2\log(k+1)}{m}\right)^{1-1/\alpha_{k+1}}\right).
\end{aligned}$$

Hence, with probability at least $1-\delta$, for all $\alpha \in (1, 2]$,

$$R(h) - \widehat{R}_S^\rho(h) \leq 16\sqrt{2}\sqrt[\alpha]{R(h)}\left(\frac{B + 2\log\frac{1}{\alpha-1}}{m}\right)^{1-1/\alpha}.$$

The lemma follows by observing that

$$\left(\frac{B + 2\log\frac{1}{\alpha-1}}{m}\right)^{1-1/\alpha} \leq \left(\frac{B}{m}\right)^{1-1/\alpha} + \left(2\frac{\log\frac{1}{\alpha-1}}{m}\right)^{1-1/\alpha} \leq \left(\frac{B}{m}\right)^{1-1/\alpha} + \left(\frac{1}{m}\right)^{1-1/\alpha} \leq 2\left(\frac{B}{m}\right)^{1-1/\alpha}.$$

□

D. Upper Bounds on Peeling-Based Rademacher Complexity

Lemma 5. *If the functions in $\mathcal{G}$ take values in $\{0,1\}$, then the following upper bounds hold for the peeling-based Rademacher complexity of $\mathcal{G}$:*

$$\mathfrak{r}_m(\mathcal{G}) \leq \frac{1}{8} \log \mathop{\mathbb{E}}_{z_1^m}[\mathbb{S}_{\mathcal{G}}(z_1^m)].$$

Proof. By definition,

$$\mathfrak{r}_m(\mathcal{G}) = \sup_k \log \mathop{\mathbb{E}}_{z_1^m}\left[\exp\left(\frac{m^2 \widehat{\mathfrak{R}}_m^2(\mathcal{G}_k(z_1^m))}{2^{k+5}}\right)\right].$$

For any $g \in \mathcal{G}_k(z_1^m)$, since g takes values in $[0,1]$, we have:

$$\sum_{i=1}^m g^2(z_i) \leq \sum_{i=1}^m g(z_i) \leq \frac{2^{k+1}}{m}.$$

Thus, by Massart's lemma and Jensen's inequality, the following inequality holds:

$$\begin{aligned}\widehat{\mathfrak{R}}_m(\mathcal{G}_k(z_1^m)) &\leq \sqrt{2 \log \mathop{\mathbb{E}}_{z_1^m}[|\mathcal{G}_k(z_1^m)|]}\sqrt{\frac{2^{k+1}}{m}} \\ &\leq \sqrt{2 \log \mathop{\mathbb{E}}_{z_1^m}[\mathbb{S}_{\mathcal{G}}(z_1^m)]}\sqrt{\frac{2^{k+1}}{m^2}}.\end{aligned}$$

Hence,

$$\mathfrak{r}_m(\mathcal{G}) \leq \sup_k \frac{1}{2^3} \log \mathop{\mathbb{E}}_{z_1^m}[\mathbb{S}_{\mathcal{G}}(z_1^m)] = \frac{1}{8} \log \mathop{\mathbb{E}}_{z_1^m}[\mathbb{S}_{\mathcal{G}}(z_1^m)].$$

□

Lemma 6. *For a set of hypotheses $\mathcal{G}$,*

$$\mathfrak{r}_m(\mathcal{G}) \leq \sup_{0 \leq k \leq \log_2(m)} \log\left[\mathop{\mathbb{E}}_{z_1^m \sim \mathcal{D}^m}\left[\exp\{f_k(z_1^m, \mathcal{G})\}\right]\right].$$

where

$$f_k(z_1^m, \mathcal{G}) = \frac{1}{16}\left[1 + \int_{\frac{1}{\sqrt{m}}}^1 \log \mathcal{N}_2\left(\mathcal{G}_k(z_1^m), \sqrt{\tfrac{2^k}{m}}\epsilon, z_1^m\right) d\epsilon\right].$$

Proof. By Dudley's integral,

$$\widehat{\mathfrak{R}}_m(\mathcal{G}_k(z_1^m)) = \min_\tau \tau + \int_{\epsilon=\tau}^{2^k/m} \sqrt{\frac{\log N_2(\mathcal{G}_k(z_1^m), \epsilon)}{m}} d\epsilon.$$

Choosing $\tau = \frac{2^{k/2}}{m}$ and changing variables from ϵ to $\epsilon \frac{2^{k/2}}{\sqrt{m}}$ yields,

$$\widehat{\mathfrak{R}}_m(\mathcal{G}_k(z_1^m)) = \frac{2^{k/2}}{m} + \frac{2^{k/2}}{m} \int_{\epsilon=1/\sqrt{m}}^1 \sqrt{\log N_2(\mathcal{G}_k(z_1^m), \epsilon\sqrt{2^k/m})} d\epsilon.$$

Using $(a+b)^2 \leq 2a^2 + 2b^2$ and the Cauchy-Schwarz inequality yields,

$$\begin{aligned}\frac{m^2 \widehat{\mathfrak{R}}_m^2(\mathcal{G}_k(z_1^m))}{2^{k+5}} &\leq \frac{1}{16}\left(1 + \left(\int_{\epsilon=1/\sqrt{m}}^1 \sqrt{\log N_2(\mathcal{G}_k(z_1^m), \epsilon\sqrt{2^k/m})} d\epsilon\right)^2\right) \\ &\leq \frac{1}{16}\left(1 + \int_{\epsilon=1/\sqrt{m}}^1 \log N_2(\mathcal{G}_k(z_1^m), \epsilon\sqrt{2^k/m}) d\epsilon\right).\end{aligned}$$

□

Recall that the worst case Rademacher complexity is defined as follows.

$$\widehat{\mathfrak{R}}_m^{\max}(\mathcal{H}) = \sup_{z_1^m} \widehat{\mathfrak{R}}_m(\mathcal{H})$$

E. Unbounded Margin Losses

Theorem 3. *Fix $\rho \geq 0$. Let $1 < \alpha \leq 2$, $0 < \epsilon \leq 1$, and $0 < \tau^{\frac{\alpha-1}{\alpha}} < \epsilon^{\frac{\alpha}{\alpha-1}}$. For any loss function L (not necessarily bounded) and hypothesis set $\mathcal{H}$ such that $\mathcal{L}_\alpha(h) < +\infty$ for all $h \in \mathcal{H}$,*

$$\mathbb{P}\left[\sup_{h\in H} \mathcal{L}(h) - \widehat{\mathcal{L}}_S(h) > \Gamma_\tau(\alpha,\epsilon)\, \epsilon \sqrt[\alpha]{\mathcal{L}_\alpha(h) + \tau} + \rho\right] \quad \leq \quad \mathbb{P}\left[\sup_{h\in\mathcal{H}, t\in\mathbb{R}} \frac{\mathbb{P}[L(h,z) > t] - \widehat{\mathbb{P}}[L(h,z) > t-\rho]}{\sqrt[\alpha]{\mathbb{P}[L(h,z) > t] + \tau}} > \epsilon\right],$$

where $\Gamma_\tau(\alpha,\epsilon) = \frac{\alpha-1}{\alpha}(1+\tau)^{\frac{1}{\alpha}} + \frac{1}{\alpha}\left(\frac{\alpha}{\alpha-1}\right)^{\alpha-1}\left(1 + \left(\frac{\alpha-1}{\alpha}\right)^{\alpha}\tau^{\frac{1}{\alpha}}\right)^{\frac{1}{\alpha}}\left[1 + \frac{\log(1/\epsilon)}{\left(\frac{\alpha}{\alpha-1}\right)^{\alpha-1}}\right]^{\frac{\alpha-1}{\alpha}}$.

Proof. Fix $1 < \alpha \leq 2$ and $\epsilon > 0$ and $\mathcal{S}$ assume that for any $h \in H$ and $t \geq 0$, the following holds:

$$\frac{\mathbb{P}[L(h,z) > t] - \widehat{\mathbb{P}}[L(h,z) > t-\rho]}{\sqrt[\alpha]{\mathbb{P}[L(h,z) > t] + \tau}} \leq \epsilon. \tag{12}$$

Let $t_1 = \frac{\alpha-1}{\alpha}\sqrt[\alpha]{\mathcal{L}_\alpha(h) + \tau}\left[\frac{1}{\epsilon}\right]^{\frac{1}{\alpha-1}}$. We show that this implies that for any $h \in H$, $\mathcal{L}(h) - \widehat{\mathcal{L}}_S(h) \leq \Gamma_\tau(\alpha,\epsilon)\epsilon\sqrt[\alpha]{\mathcal{L}_\alpha(h) + \tau} + \min(\rho, t_1)$. By the properties of the Lebesgue integral, we can write

$$\mathcal{L}(h) = \mathrm{E}_{z\sim D}[L(h,z)] = \int_0^{+\infty} \mathbb{P}[L(h,z) > t]\, dt.$$

Similarly, we can write

$$\begin{aligned}
\widehat{\mathcal{L}}(h) = \mathrm{E}_{z\sim\widehat{D}}[L(h,z)] &= \int_0^{+\infty} \widehat{\mathbb{P}}[L(h,z) > u]\, du \\
&= \int_\rho^{+\infty} \widehat{\mathbb{P}}[L(h,z) > t-\rho]\, dt \\
&= \int_0^{+\infty} \widehat{\mathbb{P}}[L(h,z) > t-\rho]\, dt - \int_0^{\rho} \widehat{\mathbb{P}}[L(h,z) > t-\rho]\, dt \\
\text{and} \quad \mathcal{L}_\alpha(h) &= \int_0^{+\infty} \mathbb{P}[L^\alpha(h,z) > t]\, dt = \int_0^{+\infty} \alpha t^{\alpha-1}\, \mathbb{P}[L(h,z) > t]\, dt.
\end{aligned}$$

To bound $\mathcal{L}(h) - \widehat{\mathcal{L}}(h)$, we simply bound $\mathbb{P}[L(h,z) > t] - \widehat{\mathbb{P}}[L(h,z) > t-\rho]$ by $\mathbb{P}[L(h,z) > t]$ for large values of t, that is $t > t_1$, and use inequality (12) for smaller values of t:

$$\begin{aligned}
&= \mathcal{L}(h) - \widehat{\mathcal{L}}(h) \\
&= \int_0^{+\infty} \mathbb{P}[L(h,z) > t] - \widehat{\mathbb{P}}[L(h,z) > t-\rho]\, dt + \int_0^{\rho} \widehat{\mathbb{P}}[L(h,z) > t-\rho]\, dt \\
&\leq \int_0^{+\infty} \mathbb{P}[L(h,z) > t] - \widehat{\mathbb{P}}[L(h,z) > t-\rho]\, dt + \rho \\
&\leq \int_0^{t_1} \epsilon\sqrt[\alpha]{\mathbb{P}[L(h,z) > t] + \tau}\, dt + \int_{t_1}^{+\infty} \mathbb{P}[L(h,z) > t]\, dt + \min(t_1, \rho),
\end{aligned}$$

where the last two inequalities use the fact that L is non-negative. The rest of the proof is similar to (Cortes et al., 2019, Theorem 3). □

Corollary 9. *Let $\epsilon < 1$, $1 < \alpha \leq 2$. and hypothesis set $\mathcal{H}$ such that $\mathcal{L}_\alpha(h) < +\infty$ for all $h \in \mathcal{H}$,*

$$\mathcal{L}(h) - \widehat{\mathcal{L}}_S(h) \leq \min_{\rho\leq r} \gamma\sqrt[\alpha]{\mathcal{L}_\alpha(h)}\sqrt{\frac{\log\mathbb{E}[\mathcal{N}_\infty(\mathcal{L}(\mathcal{H}), \frac{\rho}{2}, x_1^{2m})] + \log\frac{1}{\delta} + \log\log\frac{2r}{\rho}}{m^{\frac{2(\alpha-1)}{\alpha}}}} + \rho,$$

where $\gamma = \Gamma_0\left(\alpha, \sqrt{\frac{\log\mathbb{E}[\mathcal{N}_\infty(\mathcal{L}(\mathcal{H}), \frac{\rho}{2}, x_1^{2m})] + \log\frac{1}{\delta} + \log\log\frac{2r}{\rho}}{m^{\frac{2(\alpha-1)}{\alpha}}}}\right) = \mathcal{O}(\log m)$.

The proof of Corollary 9 is similar to that of Corollary 3 and is omitted.

F. Applications

F.1. Algorithms

As discussed in Section 5.2, our results can help derive tighter guarantees for margin-based algorithms such as Support Vector Machines (SVM) (Cortes and Vapnik, 1995) and other algorithms such as those based on neural networks that can be analyzed in terms of their margin. But, another potential application of our learning bounds is to design new algorithms, either by seeking to directly minimize the resulting upper bound, or by using the bound as an inspiration for devising a new algorithm.

In this sub-section, we briefly initiate this study in the case of linear hypotheses. We describe an algorithm seeking to minimize the upper bound of Corollary 4 (or Corollary 7) in the case of linear hypotheses. Let R be the radius of the sphere containing the data. Then, the bound of the corollary holds with high probability for any function $h\colon \mathbf{x} \mapsto \mathbf{w} \cdot \mathbf{x}$ with $\mathbf{w} \in \mathbb{R}^d$, $\|\mathbf{w}\|_2 \leq 1$, and for any $\rho > 0$ for $d = (R/\rho)^2$. Ignoring lower order terms and logarithmic factors, the guarantee suggests seeking to choose $\mathbf{w}$ with $\|\mathbf{w}\| \leq 1$ and $\rho > 0$ to minimize the following:

$$\widehat{R}_S^\rho(\mathbf{w}) + \frac{\lambda}{\rho}\sqrt{\widehat{R}_S^\rho(\mathbf{w})},$$

where we denote by $\widehat{R}_S^\rho(\mathbf{w})$ the empirical margin loss of $h\colon \mathbf{x} \mapsto \mathbf{w} \cdot \mathbf{x}$. Thus, using the so-called ramp loss $\Phi_\rho\colon u \mapsto \min(1, \max(0, 1 - \frac{u}{\rho}))$, this suggests choosing $\mathbf{w}$ with $\|\mathbf{w}\| \leq 1$ and $\rho > 0$ to minimize the following:

$$\frac{1}{m}\sum_{i=1}^m \Phi_\rho(y_i \mathbf{w} \cdot \mathbf{x}_i) + \frac{\lambda}{\rho}\sqrt{\frac{1}{m}\sum_{i=1}^m \Phi_\rho(y_i \mathbf{w} \cdot \mathbf{x}_i)}.$$

This optimization problem is closely related to that of SVM but it is distinct. The problem is non-convex, even if Φ_ρ is upper bounded by the hinge loss. The solution may also not coincide with that of SVM in general. As an example, when the training sample is linearly separable, any pair $(\mathbf{w}^*, \rho^*)$ with a weight vector $\mathbf{w}^*$ defining a separating hyperplane and ρ^* sufficiently large is solution, since we have $\sum_{i=1}^m \Phi_{\rho^*}(y_i \mathbf{w}^* \cdot \mathbf{x}_i) = 0$. In contrast, for (non-separable) SVM, in general the solution may not be a hyperplane with zero error on the training sample, even when the training sample is linearly separable. Furthermore, the SVM solution is unique (Cortes and Vapnik, 1995).

In the above, we used the ramp loss since it is closest to the hinge loss used in SVM and it has been shown recently that a slightly modified version of the ramp loss can also benefit from favorable adversarial loss guarantees in the context of linear hypotheses (Bao et al., 2020). Furthermore, it can of course be upper-bounded by the hinge loss. We note that our margin-based results hold for several loss functions highlighted in Figure 1.

F.2. Margin-Based Bounds for Known Hypothesis Sets

Ensembles of predictors in base hypothesis set $\mathcal{H}$: let d be the VC-dimension of $\mathcal{H}$ and consider the family of ensembles $\mathcal{F} = \{x \mapsto \sum_{k=1}^p w_k h_k(x)\colon h_k \in \mathcal{H}, w_k \geq 0, \sum_{k=1}^p w_k = 1\}$. The most well known existing margin bound for ensembles such as AdaBoost in terms of the VC-dimension of the base hypothesis given by Schapire et al. (1997) is:

$$R(h) \leq \widehat{R}_S^\rho(h) + c'\sqrt{\beta'_m}, \tag{13}$$

where c' is some universal constant and where $\beta'_m = \widetilde{O}\left(\frac{(d/\rho)^2}{m}\right)$. Gao and Zhou (2013) showed that

$$R(h) \leq \widehat{R}_S^\rho(h) + 2\sqrt{\widehat{R}_S^\rho(h)\,\beta''_m} + \beta''_m, \tag{14}$$

where $\beta''_m = \widetilde{O}\left(\frac{(d/\rho)^2}{m}\right)$. However, their proof technique depends crucially on the fact that the underlying hypothesis set is an ensemble of predictors. We can directly apply our relative deviation margin bounds to recover their result, up to logarithmic factors. The following upper bound on the fat-shattering dimension holds (Bartlett and Shawe-Taylor, 1998): $\mathrm{fat}_\rho(\mathcal{F}) \leq c(d/\rho)^2 \log(1/\rho)$, for some universal constant c. Plugging in this upper bound in the bound of Corollary 4 yields the following:

$$R(h) \leq \widehat{R}_S^\rho(h) + 2\sqrt{\widehat{R}_S^\rho(h)\,\beta_m} + \beta_m, \tag{15}$$

with $\beta_m = \widetilde{O}\left(\frac{(d/\rho)^2}{m}\right)$. The margin bound in (15) is thus more favorable than (13) and comparable to (14).

Feed-forward neural networks of depth d: let $\mathcal{H}_0 = \{\mathbf{x} \mapsto \mathbf{x}_i : i \in \{0, 1, \dots n\}, \mathbf{x} \in [-1, 1]^n\} \cup \{0, 1\}$ and

$$\mathcal{H}_i = \left\{ \sigma\left(\sum_{h \in \cup_{j<i} \mathcal{H}_j} \mathbf{w} \cdot h \right) : \|\mathbf{w}\|_1 \leq R \right\}$$

for $i \in [d]$, where σ is a μ-Lipschitz activation function. Then, the following upper bound holds for the fat-shattering dimension of $\mathcal{H}$ (Bartlett and Shawe-Taylor, 1998): $\operatorname{fat}_\rho(\mathcal{H}_d) \leq \frac{c^{d^2} (R\mu)^{d(d+1)}}{\rho^{2d}} \log n$. Plugging in this upper bound in the bound of Corollary 4 gives the following:

$$R(h) \leq \widehat{R}_S^\rho(h) + 2\sqrt{\widehat{R}_S^\rho(h)\, \beta_m} + \beta_m, \tag{16}$$

with $\beta_m = \widetilde{O}\left(\frac{c^{d^2} (R\mu)^{d(d+1)} / \rho^{2d}}{m}\right)$. In comparison, the best existing margin bound for neural networks by (Bartlett and Shawe-Taylor, 1998, Theorem 1.5 , Theorem 1.11) is

$$R(h) \leq \widehat{R}_S^\rho(h) + c' \sqrt{\beta'_m}, \tag{17}$$

where c' is some universal constant and where $\beta'_m = \widetilde{O}\left(\frac{c^{d^2} (R\mu)^{d(d+1)} / \rho^{2d}}{m}\right)$. The margin bound in (16) is thus more favorable than (17).