SiameseXML: Siamese Networks meet Extreme Classifiers with 100M Labels

Supplementary Material

February 8, 2022

A Proof of Lemma 1

Postulate 2 states that $p_l(+1, \mathcal{E}_{\boldsymbol{\theta}^*}(\mathbf{z}_l)^\top \mathbf{w}_l^*) \geq p_l(+1, 1) - \epsilon_l$. Let $q : [0, 1] \to [-1, 1]$ denote the inverse of $p_l(+1, \cdot)$. By assumption, q is C_{lip}-Lipschitz over the set $\mathcal{R}_p \stackrel{\text{def}}{=} \{p_l(+1, v) : v \in [-1, +1]\}$. This gives us

$$\left|\mathcal{E}_{\boldsymbol{\theta}^*}(\mathbf{z}_l)^\top \mathbf{w}_l^* - 1\right| = \left|q(p_l(+1, \mathcal{E}_{\boldsymbol{\theta}^*}(\mathbf{z}_l)^\top \mathbf{w}_l^*)) - q(p_l(+1, 1))\right| \leq C_{\text{lip}} \cdot \left|p_l(+1, \mathcal{E}_{\boldsymbol{\theta}^*}(\mathbf{z}_l)^\top \mathbf{w}_l^*) - p_l(+1, 1)\right|.$$

Since $\mathcal{E}_{\boldsymbol{\theta}^*}(\mathbf{z}_l)$ and $\mathbf{w}_l^*$ are both unit vectors, $\left|\mathcal{E}_{\boldsymbol{\theta}^*}(\mathbf{z}_l)^\top \mathbf{w}_l^*\right| \leq 1$. As $p_l(+1, \cdot)$ is monotonically increasing by assumption,

$$\left|p_l(+1, \mathcal{E}_{\boldsymbol{\theta}^*}(\mathbf{z}_l)^\top \mathbf{w}_l^*) - p_l(+1, 1)\right| = p_l(+1, 1) - p_l(+1, \mathcal{E}_{\boldsymbol{\theta}^*}(\mathbf{z}_l)^\top \mathbf{w}_l^*) \leq \epsilon_l,$$

by using the postulate. This in turn gives us

$$\|\mathcal{E}_{\boldsymbol{\theta}^*}(\mathbf{z}_l) - \mathbf{w}_l^*\|_2^2 = 2(1 - \mathcal{E}_{\boldsymbol{\theta}^*}(\mathbf{z}_l)^\top \mathbf{w}_l^*) = 2\left|\mathcal{E}_{\boldsymbol{\theta}^*}(\mathbf{z}_l)^\top \mathbf{w}_l^* - 1\right| \leq 2C_{\text{lip}} \cdot \epsilon_l$$

A straightforward application of the Cauchy-Schwartz inequality then tells us that for any $\mathbf{x} \in \mathcal{X}$, we have

$$\left|\mathcal{E}_{\boldsymbol{\theta}^*}(\mathbf{x})^\top \mathbf{w}_l^* - \mathcal{E}_{\boldsymbol{\theta}^*}(\mathbf{x})^\top \mathcal{E}_{\boldsymbol{\theta}^*}(\mathbf{z}_l)\right| \leq \|\mathcal{E}_{\boldsymbol{\theta}^*}(\mathbf{x})\|_2 \cdot \|\mathbf{w}_l^* - \mathcal{E}_{\boldsymbol{\theta}^*}(\mathbf{z}_l)\|_2 \leq \sqrt{2C_{\text{lip}} \cdot \epsilon_l},$$

which establishes the first result in the lemma. For the second result, we notice that the above calculation shows that for any data point $\mathbf{x}$, the use of the label embedding $\mathcal{E}_{\boldsymbol{\theta}^*}(\mathbf{z}_l)$ instead of the gold 1-vs-all classifier $\mathbf{w}_l^*$ perturbs the score for label l by a quantity that has magnitude at most $\sqrt{2C_{\text{lip}} \cdot \epsilon_l}$. The (q, D_{lip})-Lipschitzness of the joint likelihood function then immediately establishes the second result.

For the special case of decomposable likelihoods, we notice that for any $i \in [N]$

$$\begin{aligned}
&\left|\frac{1}{L}\sum_{l=1}^{L} \ln p_l\left(y_{il}, \mathcal{E}_{\boldsymbol{\theta}}(\mathbf{x}_i)^\top \mathbf{w}_l^*\right) - \frac{1}{L}\sum_{l=1}^{L} \ln p_l\left(y_{il}, \mathcal{E}_{\boldsymbol{\theta}}(\mathbf{x}_i)^\top \mathcal{E}_{\boldsymbol{\theta}^*}(\mathbf{z}_l)\right)\right| \\
&\leq \frac{1}{L}\sum_{l=1}^{L} \left|\ln p_l\left(y_{il}, \mathcal{E}_{\boldsymbol{\theta}}(\mathbf{x}_i)^\top \mathbf{w}_l^*\right) - \ln p_l\left(y_{il}, \mathcal{E}_{\boldsymbol{\theta}}(\mathbf{x}_i)^\top \mathcal{E}_{\boldsymbol{\theta}^*}(\mathbf{z}_l)\right)\right| \leq \frac{D_{\text{lip}}\sqrt{2C_{\text{lip}}}}{L}\sum_{l=1}^{L}\sqrt{\epsilon_l} \leq D_{\text{lip}}\sqrt{2C_{\text{lip}} \cdot \bar{\epsilon}},
\end{aligned}$$

where the last inequality follows from Jensen's inequality. Taking an average over all $i \in [N]$ and applying triangle inequality yet again yields

$$\begin{aligned}
&\left|\frac{1}{NL}\sum_{i=1}^{N}\sum_{l=1}^{L} \ln p_l\left(y_{il}, \mathcal{E}_{\boldsymbol{\theta}}(\mathbf{x}_i)^\top \mathbf{w}_l^*\right) - \frac{1}{NL}\sum_{i=1}^{N}\sum_{l=1}^{L} \ln p_l\left(y_{il}, \mathcal{E}_{\boldsymbol{\theta}}(\mathbf{x}_i)^\top \mathcal{E}_{\boldsymbol{\theta}^*}(\mathbf{z}_l)\right)\right| \\
&\leq \frac{1}{NL}\sum_{i=1}^{N}\sum_{l=1}^{L} \left|\ln p_l\left(y_{il}, \mathcal{E}_{\boldsymbol{\theta}}(\mathbf{x}_i)^\top \mathbf{w}_l^*\right) - \ln p_l\left(y_{il}, \mathcal{E}_{\boldsymbol{\theta}}(\mathbf{x}_i)^\top \mathcal{E}_{\boldsymbol{\theta}^*}(\mathbf{z}_l)\right)\right| \leq D_{\text{lip}}\sqrt{2C_{\text{lip}} \cdot \bar{\epsilon}},
\end{aligned}$$

which establishes the second result in the lemma.

B Proof of Lemma 2

We note that the objectives $\mathcal{L}^1$ and $\mathcal{L}^2$ are identical but for the parameters over which they desire optimization, as both of them are incomplete NLL expressions derived form $\mathcal{L}$. Suppose $\hat{\theta}$ is the embedding model obtained at the end of Module I. Thus, initializing $\boldsymbol{\eta}_l = \mathbf{0}$ for all $l \in [L]$ and using monotonicity of training guarantee us that the refinement vectors $\hat{\boldsymbol{\eta}}_l$ learnt in Module III satisfy

$$\mathcal{L}^2(\{\hat{\boldsymbol{\eta}}_l\}) \leq \mathcal{L}^2(\{\mathbf{0}\})$$

However, note that we have

$$\mathcal{L}^2(\{\hat{\boldsymbol{\eta}}_l\}) = \mathcal{L}(\hat{\boldsymbol{\theta}}, \{\hat{\mathbf{w}}_l\})$$
$$\mathcal{L}^2(\{\mathbf{0}\}) = \mathcal{L}^1(\hat{\boldsymbol{\theta}}),$$

where the second equality follows since $\mathfrak{N}(\mathcal{E}_{\hat{\boldsymbol{\theta}}}(\mathbf{z}_l)+\mathbf{0}) = \mathfrak{N}(\mathcal{E}_{\hat{\boldsymbol{\theta}}}(\mathbf{z}_l)) = \mathcal{E}_{\hat{\boldsymbol{\theta}}}(\mathbf{z}_l)$ since $\mathcal{E}$ already outputs normalized embeddings and normalization is an idempotent operation. This tells us that

$$\mathcal{L}(\hat{\boldsymbol{\theta}}, \{\hat{\mathbf{w}}_l\}) \leq \mathcal{L}^1(\hat{\boldsymbol{\theta}})$$

Now, $\hat{\boldsymbol{\theta}}$ is a δ_{opt}-approximate solution to $\mathcal{L}^1$. Moreover, by construction, we have

$$\mathcal{L}^1(\boldsymbol{\theta}^*) = \mathcal{L}(\boldsymbol{\theta}^*, \{\mathcal{E}_{\boldsymbol{\theta}^*}(\mathbf{z}_l)\})$$

This tells us that

$$\mathcal{L}^1(\hat{\boldsymbol{\theta}}) \leq \mathcal{L}(\boldsymbol{\theta}^*, \{\mathcal{E}_{\boldsymbol{\theta}^*}(\mathbf{z}_l)\}) + \delta_{\text{opt}}$$

However, Lemma 1 tells us that

$$\mathcal{L}(\boldsymbol{\theta}^*, \{\mathcal{E}_{\boldsymbol{\theta}^*}(\mathbf{z}_l)\}) \leq \mathcal{L}(\boldsymbol{\theta}^*, \{\mathbf{w}_l^*\}) + D_{\text{lip}} \cdot \sqrt{2C_{\text{lip}} \cdot \epsilon},$$

where ϵ is the sub-optimality (either ϵ_{eff} or $\bar{\epsilon}$) in Lemma 1. Putting the above chain of results together establishes

$$\mathcal{L}(\hat{\boldsymbol{\theta}}, \{\hat{\mathbf{w}}_l\}) \leq \mathcal{L}(\boldsymbol{\theta}^*, \{\mathbf{w}_l^*\}) + D_{\text{lip}} \cdot \sqrt{2C_{\text{lip}} \cdot \epsilon} + \delta_{\text{opt}}.$$

We note that appealing to the monotonicity of training once more shows that the above results are unaffected even if the components of the embedding architecture, for example the residual layer and the token embeddings, are jointly fine-tuned with the refinement vectors in Module IV.

C Inverse and Lipschitzness Properties of p_l

Recall that SiameseXML uses

$$p_l(+1, v) = \frac{c \cdot \exp(d \cdot v)}{\exp(d)}$$
$$p_l(-1, v) = 1 - p_l(+1, v)$$

Clearly $p_l(+1, \cdot)$ is monotonically increasing with the function $q(p) = 1 + \frac{1}{d} \ln \frac{p}{c}$ as its inverse. We also have the derivative $q'(p) = \frac{1}{dp}$. Now it is easy to see that $R_p \subseteq \left[\frac{c}{\exp(2d)}, c\right]$ which tells us that $|q'(p)| \leq \frac{\exp(2d)}{cd}$ for all $p \in R_p$. Since the function q is continuously differentiable on R_p, it is $\frac{\exp(2d)}{cd}$-Lipschitz on R_p.

We move on to the log-likelihood functions

$$\ln p_l(+1, v) = \ln c + d(v - 1)$$
$$\ln p_l(-1, v) = \ln \left(1 - \frac{c \cdot \exp(d \cdot v)}{\exp(d)}\right)$$

Clearly $\ln p_l(+1, \cdot)$ is d-Lipschitz on the interval $[-1, 1]$. It is similarly simply to see that we have

$$\left|\frac{d \ln p_l(-1, v)}{dv}\right| = \frac{cd \exp(d(v-1))}{1 - c \exp(d(v-1))}$$

The above is an increasing function of v and achieves its highest value at $v = 1$ which is $\frac{dc}{1-c}$. This establishes that $\ln p_l(-1, \cdot)$ is $\frac{dc}{1-c}$-Lipschitz on the interval $[-1, 1]$.

D Time Complexity Calculations

For the purpose of the discussion below, we introduce some additional notation.

Notation: As Section 4 explains, SiameseXML learns D-dimensional embeddings for all V tokens ($\mathbf{e}_t, t \in [V]$) in the vocabulary. These are used to embed all L labels $\hat{\mathbf{z}}_l^1, l \in [L]$, as well as all N training documents $\hat{\mathbf{x}}_i, i \in [N]$. $\|\cdot\|_0$ denotes the sparsity "norm" that gives the number of non-zero elements in a vector. Let $\hat{V}_x$ be the average number of unique tokens present in a document i.e. $\hat{V}_x = \frac{1}{N}\sum_{i=1}^{N} \|\mathbf{x}_i\|_0$ and $\hat{V}_y = \frac{1}{L}\sum_{l=1}^{L} \|\mathbf{z}_l\|_0$ similarly be the average number of tokens in a label text. As Table 5 in Appendix F indicates, short text documents exhibit token sparsity with $\hat{V}_x, \hat{V}_y \leq 10$. We also let $\hat{L} = \frac{1}{N}\sum_{i=1}^{N} \|\mathbf{y}_i\|_0$ be the average number of labels per document and $\hat{N} = \frac{N\hat{L}}{L}$ be the average number of documents per label. Table 5 presents these statistics for all datasets.

Embedding Complexity: Given a piece of text in bag-of-words representation $\mathbf{x} \in \mathbb{R}^V$ (corresponding to either a document or a label) containing, say $\hat{V} = \|\mathbf{x}\|_0$ tokens, the embedding block $\mathcal{E}_{\boldsymbol{\theta}}$ requires $\hat{V}D$ operations to aggregate token embedding vectors and calculate $\mathbf{Ex}$ and D^2 operations to apply the residual block $\mathbf{R}$ and $\mathcal{O}(D)$ operations to perform operations such as normalization and applying ReLU, thus totalling $\mathcal{O}\left(\hat{V}D + D^2\right)$ operations. Thus, encoding a label takes $\mathcal{O}\left(\hat{V}_y D + D^2\right)$ operations on average (respectively $\mathcal{O}\left(\hat{V}_x D + D^2\right)$ for a document).

Prediction: Given a test data point $\mathbf{x} \in \mathbb{R}^V$ with say, $\hat{V}_x$ unique tokens, obtaining the intermediate embedding $f_{\mathbf{E}}(\mathbf{x})$ takes $\mathcal{O}\left(\hat{V}_x D\right)$ time since it only involves aggregating token embeddings followed by ReLU and normalization operations which take $\mathcal{O}(D)$ time each. Obtaining the final embedding $\mathcal{E}_{\boldsymbol{\theta}}(\mathbf{x})$ on the other hand, takes $\mathcal{O}\left(\hat{V}_x D + D^2\right)$ time as described above. Since search is over L items, ANNS structures such as HNSW [6] offer $\mathcal{O}(\log L)$ time. Thus, identifying the shortlist $\mathcal{S}$ of labels and applying the 1-vs-all and centroid models corresponding to each label $l \in \mathcal{S}$ in the shortlist takes $\mathcal{O}(D \log L)$ time since $|\mathcal{S}| = \mathcal{O}(\log L)$ by design. Since $\hat{V}_x \leq D$, this brings the overall prediction time complexity to $\mathcal{O}\left(D^2 + D \log L\right)$.

Training Module I: Creation of the offline negative mining ANNS structure, that is periodically refreshed after every few epochs, requires feature embeddings to be recomputed which takes $\mathcal{O}\left(N(\hat{V}_x D + D^2)\right)$ time whereas computing the ANNS search graph itself takes $\mathcal{O}(ND \log N)$ time [6]. Thus, $\mathcal{O}\left(N(\hat{V}_x D + D^2)\right)$ operations are required each time this graph needs to be recomputed since $\log N \leq D$. We note that the time taken to (re)create the document embeddings using the updated parameters is the dominating complexity.

We additionally discuss the amount of time it takes to process each mini-batch of B data points. It takes $\mathcal{O}\left(B((\hat{V}_x + \hat{V}_y)D + D^2)\right)$ time to obtain the embeddings of all B labels and their corresponding positive documents. Obtaining the hardest $\kappa - 1$ "online" negative documents for each label takes $\mathcal{O}(BD)$ time, taking a total of $\mathcal{O}\left(B^2 D\right)$ time for the entire batch. Obtaining the remaining hard negative coming from the offline ANNS structure takes $\mathcal{O}(D \log N)$ per label and $\mathcal{O}(BD \log N)$ time for the entire batch. Since $N \leq L^{\mathcal{O}(1)}$, this is $\mathcal{O}(BD \log L)$. Training with respect to the $B(\kappa + 1)$ pairs takes $\mathcal{O}(\kappa BD)$ time. Thus, it takes $\mathcal{O}\left(BD^2 + B^2 D\right)$ time to process each mini-batch (typically $\kappa, \hat{V}_y, \hat{V}_x \leq \min\{B, D\}$ and $\log L \leq D$).

Training Module II: Once Module I is completed, computing the intermediate embeddings of all labels $f_{\hat{\mathbf{E}}}(\mathbf{z}^l)$ takes $\mathcal{O}\left(L\hat{V}_y D\right)$ time which can be simplified to $\mathcal{O}(ND \log L)$ since $L \leq \mathcal{O}(N), \hat{V}_y \leq \mathcal{O}(\log L)$.

Computing the label centroids $\mathbf{v}_l$ takes $\mathcal{O}\left(LD\hat{N}\right)$ time. This is simplified to $\mathcal{O}\left(ND\hat{L}\right)$, where we recall that $\hat{N}$ and $\hat{L}$ denote the average number of documents per label and labels per document respectively and thus $L\hat{N} = N\hat{L}$. Since $\hat{L} \leq \mathcal{O}(\log L)$, this simplifies to $\mathcal{O}(ND \log L)$ as well. Constructing the two search graphs themselves takes $\mathcal{O}(LD \log L)$ time which simplifies to $\mathcal{O}(ND \log L)$ for $L = \mathcal{O}(N)$. Shortlisting the labels from the two graphs as well as the the random ones takes $\mathcal{O}(D \log L)$ time per document for a total of $\mathcal{O}(ND \log L)$ time. Thus, the entire module is executed in $\mathcal{O}(ND \log L)$ time.

Training Module III: We discuss the time complexity of processing a single mini-batch of B documents, each document i accompanied with $\hat{L}$ positive labels $\mathcal{P}_i$ on average, and $\mathcal{O}(\log L)$ negative labels $\hat{\mathcal{N}}_i$. Computing the document embeddings takes $\mathcal{O}\left(B(\hat{V}_x D + D^2)\right)$ time whereas doing so for the labels takes $\mathcal{O}\left(B \log L(\hat{V}_y D + D^2)\right)$, since $|\mathcal{P}_i|, |\hat{\mathcal{N}}_i| \leq \mathcal{O}(\log L)$. We note that the $\mathcal{O}\left(\hat{V}_x D\right), \mathcal{O}\left(\hat{V}_y D\right)$ time spent doing token embedding aggregations for documents and labels can be avoided in Module III by precomputing them since Module III freezes the token embeddings $\hat{\mathbf{E}}$. However, doing so incurs an additional $\mathcal{O}((N + L)D)$ memory overhead. Computing gradients with respect to these $B \log L$ document-label pairs takes $\mathcal{O}(BD \log L)$ time. Thus, processing the entire mini-batch takes $\mathcal{O}\left(BD^2 \log L\right)$ time since $\hat{V}_y, \hat{V}_x \leq D$.

E Proof of Theorem 3 and Additional Discussion

Appendix E.1 begins by revisiting relevant notation and definitions. Appendix E.2 presents a qualitative discussion on Theorem 3 in the context of related works. Appendix E.3 then outlines a proof for Theorem 3. The further subsections E.4 and E.5 provide the key arguments establishing Parts I and II respectively of Theorem 3. These arguments themselves require supporting results presented in Appendices G and H.

E.1 Notation and Definitions

Given a label vector $\mathbf{y} \in \{-1, +1\}^L$, let $P_\mathbf{y} := \{l : y_l = +1\}$ and $N_\mathbf{y} := \{l : y_l = -1\}$ denote the sets of positive and negative labels respectively. Given a score vector $\mathbf{s} = [s_1, \ldots, s_L] \in [-1, 1]^L$, let $\pi_\mathbf{s} \in \text{Sym}([L])$ be the permutation that ranks labels in decreasing order of their scores according to $\mathbf{s}$ i.e. $s_{\pi_\mathbf{s}(1)} \geq s_{\pi_\mathbf{s}(2)} \geq \ldots$. We will also let $\pi^+_\mathbf{s} \in \text{Sym}(P_\mathbf{y})$ denote the permutation that ranks the positive labels in decreasing order of their scores according to $\mathbf{s}$ i.e. $\pi^+_\mathbf{s}(t) \in P_\mathbf{y}$ for all $t \in |P_\mathbf{y}|$ and $\pi^+_\mathbf{s}(1) \geq \pi^+_\mathbf{s}(2) \geq \ldots$.

Given a matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$, $\|\mathbf{A}\|_\sigma$ denotes its spectral norm i.e.

$$\|\mathbf{A}\|_\sigma := \sup_{\mathbf{x} \in \mathbb{R}^n} \frac{\|\mathbf{A}\mathbf{x}\|_2}{\|\mathbf{x}\|_2}$$

Moreover, for any $p, q \in [1, \infty]$, we define the mixed norm $\|\mathbf{A}\|_{p,q}$ as follows:

$$\|\mathbf{A}\|_{p,q} := \left\| \left[\|\mathbf{A}_{1,:}\|_p, \|\mathbf{A}_{2,:}\|_p, \ldots, \|\mathbf{A}_{m,:}\|_p \right] \right\|_q$$

We clarify that the above first takes the p-th norm over all rows, then the q-th norm over the columns. Note that this is slightly different from popular convention that sometimes takes norms over columns first and then over rows. However, this slight change of convention will greatly simplify our notation and avoid clutter due to transpose symbols everywhere.

Definition 1 (prec@k Loss Function)**.** *We define the prec@k loss function as follows*

$$\wp_k(\mathbf{s}, \mathbf{y}) := 1 - \frac{1}{k} \sum_{t=1}^{k} \mathbb{I}\{\pi_\mathbf{s}(t) \in P_\mathbf{y}\}$$

Definition 2 (γ-ramp Function)**.** *Given a margin parameter $\gamma > 0$, define the γ-ramp function for any $v \in \mathbb{R}$ as follows*

$$r_\gamma(v) = \begin{cases} 0 & \text{if } v < 0 \\ \frac{v}{\gamma} & \text{if } v \in [0, \gamma] \\ 1 & \text{if } v > \gamma \end{cases}$$

Definition 3 (prec@k Surrogate Loss Function)**.** *For any $k \in \mathbb{N}$, we define the prec@k surrogate loss function as follows*

$$\ell^{prec}_{\gamma,k}(\mathbf{s}, \mathbf{y}) := 1 - \frac{1}{k} \sum_{t=1}^{\min\{k, |P_\mathbf{y}|\}} r_\gamma \left(s_{\pi^+_\mathbf{s}(t)} - \max_{l' \in N_\mathbf{y}} s_{l'} \right)$$

Definition 4 (Contrastive Loss Function)**.** *For any $d \geq 1, c < 1$, we define the following contrastive loss function*

$$\ell^{cont}_{c,d}(\mathbf{s}, \mathbf{y}) = \sum_{l \in P_\mathbf{y}} \ln \frac{1}{c} + d \cdot (1 - s_l) - \sum_{l' \in N_\mathbf{y}} \ln \left(1 - c \cdot \exp(d \cdot (s_{l'} - 1))\right)$$

We note that the contrastive loss function defined above is identical to the negative log-likelihood function with respect to the likelihood model p_l used by SiameseXML.

E.2 Discussion on the Bounds offered by Theorem 3

1. The bound is offers generalization guarantees for deep multi-label learning with label features. Previous works mostly address multi-label learning using networks that have one output node per label and do not take label features into account.

2. The bound for module I (Part I of Theorem 3) is entirely independent of the number of labels. Although this is made possible by the fact that the parameters learnt in Module I (i.e. $\boldsymbol{\theta}$) have size independent of L, such a result is missing from previous analyses which incur a dependence on the number of labels either explicitly e.g. [11] or implicitly e.g. [1]. To be sure, Part I of the bound avoids even an implicit dependence on the number of labels. Part II of the bound that considers the model with extreme 1-vs-all classifiers incurs a weak $\mathcal{O}(\log L)$ direct dependence on the number of labels.

3. [1] states that it is a "tantalizing" open question to explore bounds that are better adapted to the neural networks encountered in practice. We show that it is possible, e.g. by adapting to the datatype at various layers of the network (sparse, high dimensional at one layer but dense low-dimensional unit norm vectors in another layer). In particular, the bound exploits sparsity of the input vectors, something not done by previous generalization bounds for deep architectures such as [1, 9]. However this requires adaptations to the proof technique mentioned below.

   (a) Sidestepping the usual Talagrand-Ledoux contraction step and instead directly invoking a variant of Dudley's integral argument to bound the Rademacher complexity. The standard approach of using contraction explicitly is unwieldy for Siamese architectures such as those used by SiameseXML since in these architectures, the neural architecture itself does not output label-wise scores (for example as the output of a fully-connected final layer). Rather, label-wise scores are obtained via dot-products of label embeddings obtained using the neural architecture itself.

   (b) Distinct from previous bounds such as [1] which rely on pushing empirical covers across layers of the network, the bound we present instead relies on uniform covers. This turns out to be critical in avoiding an implicit dependence on the number of labels.

   (c) A novel *uniform* Maurey-type sparsification lemma (see Lemma 16) that uses powerful Bernstein-style arguments, as opposed to the standard *empirical* Maurey-type lemmata that use much less powerful Chevyshev-style arguments and may be of independent interest.

E.3 Proof for Theorem 3

We prove the two parts separately below and then show how the NLL expression $\mathcal{L}$ can be used in place of $\hat{\ell}_N$ in the generalization bounds.

Proof of Part I: Recall from above and the discussion in Section 5 in the main paper that $\ell(\boldsymbol{\theta}), \hat{\ell}_N(\boldsymbol{\theta})$ denote, respectively, the population and empirical loss of a model $\boldsymbol{\theta}$ with respect to the surrogate prec@k loss function defined using the ramp function. Note that Module I considers no 1-vs-all classifiers and thus, the number of parameters in $\boldsymbol{\theta}$ is independent of L. Also recall that $\wp_k(\boldsymbol{\theta})$ denotes the population prec@k risk for a model $\boldsymbol{\theta}$. Theorem 4 below establishes that for the model $\hat{\boldsymbol{\theta}}^0$ learnt by Module I, we have, with probability at least $1-\delta$,

$$\ell(\hat{\boldsymbol{\theta}}^0) \leq \hat{\ell}_N(\hat{\boldsymbol{\theta}}^0) + \frac{1}{\gamma} \cdot \frac{P \ln(N)}{\sqrt{N}} + \sqrt{\frac{\ln \frac{1}{\delta}}{N}},$$

However, the surrogacy result Claim 17 establishes that the prec@k surrogate unconditionally upper bounds the prec@k loss i.e. for any $\gamma > 0$, we have $\ell^{\text{prec}}_{\gamma,k}(\mathbf{s}, \mathbf{y}) \geq \wp_k(\mathbf{s}, \mathbf{y})$. This establishes that

$$\wp_k(\hat{\boldsymbol{\theta}}^0) \leq \ell(\hat{\boldsymbol{\theta}}^0),$$

establishing Part I of Theorem 3.

Proof of Part II: This follows in a manner similar to Part I but instead uses Theorem 5 below that establishes that for the model $\hat{\boldsymbol{\xi}} := \left\{\hat{\boldsymbol{\theta}}, \hat{\mathbf{H}}\right\}$ learnt by Module III, we have, with probability at least $1-\delta$,

$$\ell(\hat{\boldsymbol{\xi}}) \leq \hat{\ell}_N(\hat{\boldsymbol{\xi}}) + \frac{1}{\gamma} \cdot \frac{Q \ln(N)}{\sqrt{N}} + \sqrt{\frac{\ln \frac{1}{\delta}}{N}}$$

Using Claim 17 as before establishes Part II of Theorem 3.

Incorporating NLL Objective into the Bounds: Recall that $\ell^{\text{cont}}_{c,d}$ denotes the loss function equivalent to the negative-log likelihood w.r.t p_l used by SiameseXML as its optimization objective in Modules I and III. Claim 18 tells us that for any $c \in (0,1), d \geq 1$, we have $\frac{1}{\ln(4c)} \cdot \ell^{\text{cont}}_{c,d}(\mathbf{s}, \mathbf{y}) \geq \ell^{\text{prec}}_{\gamma,k}(\mathbf{s}, \mathbf{y})$. In particular, if $c \in (0.7, 1), d \geq 1$, we have $\ln(4c) > 1$ and thus $\ell^{\text{cont}}_{c,d}(\mathbf{s}, \mathbf{y}) \geq \ell^{\text{prec}}_{\gamma,k}(\mathbf{s}, \mathbf{y})$. This shows that the likelihood expression $\mathcal{L}$ can be substituted (along with an appropriate scaling constant $\frac{1}{\ln(4c)}$) for $\hat{\ell}_N$ in the bounds in Theorem 3.

E.4 A Generalization Bound for Module I

The network architecture is given below with parameters $\boldsymbol{\theta} = \{\mathbf{E}, \mathbf{R}\}$ where $\mathbf{E} \in \mathbb{R}^{D \times V}, \mathbf{R} \in \mathbb{R}^{(D+1) \times (D+1)}$

$$\begin{aligned}
\mathcal{E}_{\boldsymbol{\theta}}(\mathbf{x}) &= \mathfrak{N}_\nu(f(\mathbf{x}; \mathbf{E}) + g(f(\mathbf{x}; \mathbf{E}); \mathbf{R})) \in \mathbb{R}^{D+2} \\
f(\mathbf{x}; \mathbf{E}) &= \mathfrak{N}_\nu(\text{ReLU}(\mathbf{E}\mathbf{x})) \in \mathbb{R}^{D+1} \\
g(\mathbf{v}; \mathbf{R}) &= \text{ReLU}(\mathbf{R}\mathbf{v}) \in \mathbb{R}^{D+1} \\
\mathfrak{N}_\nu(\mathbf{v}) &= \frac{1}{\sqrt{\|\mathbf{v}\|_2^2 + \nu^2}} \cdot [\mathbf{v}, \nu]
\end{aligned}$$

The Normalization Operator $\mathfrak{N}_\nu$: Projection onto the surface of the unit sphere S^{D-1} is a non-Lipschitz operation, with an irremovable discontinuity at the origin. To address this, SiameseXML uses a Lipschitz-surrogate $\mathfrak{N}_\nu : \mathbb{R}^D \to S^D$ as defined above. Claim 11 establishes that over vectors with Euclidean norm at least r, the function is at least $\left(\frac{2}{\sqrt{r^2+\nu^2}}\right)$ Lipschitz. Thus, even in the worst case, it is at least $\frac{2}{\nu}$-Lipschitz. Note that the normalization operation increments the dimensionality of the input vector by unity i.e. if $\mathbf{v} \in \mathbb{R}^D$, then $\mathfrak{N}_\nu(\mathbf{v}) \in \mathbb{R}^{D+1}$ but is a true normalization operator i.e. $\|\mathfrak{N}_\nu(\mathbf{v})\|_2 = 1$ for all

$\mathbf{v} \in \mathbb{R}^D$. However, alternate formulations are also possible, for instance the formulation found in popular packages such as `PyTorch` which instead uses

$$\mathfrak{N}_\rho(\mathbf{v}) = \frac{1}{\max\{\|\mathbf{v}\|_2, \rho\}} \cdot \mathbf{v}.$$

Although the above notion is also Lipschitz, it is not a true normalization operator since it does not output unit norm vectors if $\|\mathbf{v}\|_2 < \rho$. Nevertheless, we stress that the proofs of Theorems 4 and 5, and consequently, the proof of Theorem 3, do not rely on the choice of $\mathfrak{N}_\nu$ as the Lipschitz variant of the normalization operator, and any other Lipschitz variant, for example $\mathfrak{N}_\rho$, could have been used just as well. All that the proofs require is that the variant being used be Lipschitz.

Below we present a generalization bound for the above architecture used by SiameseXML in Module I. For any $R_1^E, R_\infty^E, R_1^R, R_\infty^R, R_\sigma^R > 0$, we define the model class $\mathbf{\Theta}_{R_1^E, R_\infty^E, R_1^R, R_\infty^R, R_\sigma^R}$ as follows

$$\left\{(\mathbf{E}, \mathbf{R}) \in \mathbb{R}^{D \times V} \times \mathbb{R}^{(D+1)\times(D+1)} : \|\mathbf{E}\|_{1,1} \leq R_1^E, \|\mathbf{E}\|_{\infty,1} \leq R_\infty^E, \|\mathbf{R}\|_{1,1} \leq R_1^R, \|\mathbf{R}\|_{\infty,1} \leq R_\infty^R, \|\mathbf{R}\|_\sigma \leq R_\sigma^R\right\}$$

In the sequel, we will abbreviate $\mathbf{\Theta}_{R_1^E, R_\infty^E, R_1^R, R_\infty^R, R_\sigma^R}$ as simply $\mathbf{\Theta}$ to avoid notational clutter. Let $\mathcal{X}$ be the set of V-dimensional s-sparse unit-norm vectors i.e. $\mathcal{X} = \left\{\mathbf{x} \in \mathbb{R}^V, \|\mathbf{x}\|_2 = 1, \|\mathbf{x}\|_\infty \leq s\right\}$. Given training N data points of the form $(\mathbf{x}_i, \mathbf{y}_i) \in \mathcal{X} \times \{-1, +1\}^L, i = 1, \ldots, N$ sampled from some distribution $\mathcal{D}$, as well as label features $\mathbf{z}_l \in \mathcal{X}, l = 1, \ldots, L$, we first define the score function. For any $\mathbf{x} \in \mathbb{R}^d$ and any model $\boldsymbol{\theta} \in \mathbf{\Theta}$, we have

$$s^{\boldsymbol{\theta}}(\mathbf{x}) = [s_1^{\boldsymbol{\theta}}, s_2^{\boldsymbol{\theta}}, \ldots, s_L^{\boldsymbol{\theta}}]^\top \in \mathbb{R}^L,$$

where $s_l^{\boldsymbol{\theta}} = \mathcal{E}_{\boldsymbol{\theta}}(\mathbf{x})^\top \mathcal{E}_{\boldsymbol{\theta}}(\mathbf{z}_l)$ for all $l \in [L]$. We then define the empirical risk for any model $\boldsymbol{\theta}$ as follows

$$\hat{\ell}_N(\boldsymbol{\theta}) = \frac{1}{N} \sum_{i=1}^N \ell_{\gamma,k}^{\text{prec}}(s^{\boldsymbol{\theta}}(\mathbf{x}_i), \mathbf{y}_i).$$

Similarly, the population (test) risk is defined for any model $\boldsymbol{\theta}$ as follows

$$\ell(\boldsymbol{\theta}) = \mathop{\mathbb{E}}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}} \left[\ell_{\gamma,k}^{\text{prec}}(s^{\boldsymbol{\theta}}(\mathbf{x}), \mathbf{y})\right].$$

Theorem 4 (Module-I Generalization Bound)**.** *Suppose the learning procedure in module I learns a model $\hat{\boldsymbol{\theta}} \in \mathbf{\Theta}$. Then with probability at least $1 - \delta$, we have*

$$\ell(\boldsymbol{\theta}) \leq \hat{\ell}_N(\boldsymbol{\theta}) + \frac{1}{\gamma} \cdot \frac{P \ln(N)}{\sqrt{N}} + \sqrt{\frac{\ln \frac{1}{\delta}}{N}},$$

where $P := C \cdot \frac{1}{\nu} \left(\sqrt{D \ln(D)} \sqrt{R_\infty^R R_1^R} + \frac{1}{\nu} \cdot \sqrt{s} \ln(DV) R_\sigma^R \sqrt{R_\infty^E R_1^E}\right)$ for some universal constant C.

Notice that the generalization bound is entirely independent of the number of labels, depends directly on V only logarithmically, and instead depends directly on s, the sparsity of the input vectors. Also, to be sure, the proof below establishes this result for a fixed model class, i.e. $\mathbf{\Theta}_{R_1^E, R_\infty^E, R_1^R, R_\infty^R, R_\sigma^R}$ with fixed bounds on various norms such as R_1^R, R_σ^R etc. However, this can be easily extended to admit empirical bounds on these norms (i.e. the norms of the model parameters obtained after training) using a standard stratification step as used by [1]. Before proving this theorem, we need to introduce the concept of model covers.

Definition 5 (Uniform ϵ-covers for Model Classes)**.** *For any $\epsilon > 0$ and a model class $\mathbf{\Theta}$, we say that a model class $\mathbf{\Omega}_\epsilon$ is an ϵ-cover if for any $\boldsymbol{\theta} \in \mathbf{\Theta}$, there exists an $\boldsymbol{\omega} \in \mathbf{\Omega}_\epsilon$, such that $\sup_{\mathbf{x}\in\mathcal{X}} \|\mathcal{E}_{\boldsymbol{\theta}}(\mathbf{x}) - \mathcal{E}_{\boldsymbol{\omega}}(\mathbf{x})\|_2 \leq \epsilon$. Moreover, for any $\epsilon > 0$, we let $\mathcal{N}_\infty(\mathbf{\Theta}, \epsilon)$ denote the size of the smallest such ϵ-cover.*

Proof of Theorem 4. Since the prec@k surrogate loss function is takes values in the bounded interval $[0, 1]$, a standard application of the McDiarmid's inequality followed by a symmetrization argument (see for example [7, Theorem 3.1]) tells us that with probability at least $1 - \delta$, we have, for every $\boldsymbol{\theta} \in \boldsymbol{\Theta}$

$$\ell(\boldsymbol{\theta}) \leq \hat{\ell}_N(\boldsymbol{\theta}) + 2\mathfrak{R}_N(\boldsymbol{\Theta}) + \sqrt{\frac{\ln \frac{1}{\delta}}{N}},$$

where the Rademacher complexity term is defined as

$$\mathfrak{R}_N(\boldsymbol{\Theta}) := \mathop{\mathbb{E}}_{\substack{(\mathbf{x}_i, \mathbf{y}_i) \sim \mathcal{D} \\ \tau_i \sim \{-1, +1\}}} \left[\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \frac{1}{N} \sum_{i=1}^{N} \tau_i \cdot \ell^{\text{prec}}_{\gamma,k}(s^{\boldsymbol{\theta}}(\mathbf{x}_i), \mathbf{y}_i) \right]$$

For standard analyses (of even deep networks), the next step is a Talagrand-Ledoux-style contraction argument. This is because those analyses, for example [1, 9], consider architectures where the final layer outputs are the label-wise scores. However, in our case, the label-wise scores used to calculate the loss function in our case are not direct outputs of the neural architecture, rather they are obtained via dot-products of label embeddings $\mathcal{E}_{\boldsymbol{\theta}}(\mathbf{z}_l)$ that are obtained using the architecture itself.

This makes the standard approach unweildy. We overcome this difficulty by directly upper bounding the Rademacher complexity term using a chaining bound, implicitly folding in a Talagrand-Ledoux-style contraction argument in the process. The steps involved in the chaining argument are often routine but presented here for sake of completeness. The novelty in the proof lies largely in establishing the uniform covering number bounds (see Lemma 8 and Claim 15) that are used in the chaining argument. It is in establishing these covering numbers that the uniform Maurey sparsification lemma (see Lemma 16) is required.

We note that the fact that the cover defined above acts uniformly over all inputs in the set $\mathcal{X}$ is crucial in avoiding the drawbacks of the empirical covering bounds used in previous works that lead to a direct dependence on the number of labels in the bound. Lemma 8 below will establish model covers for $\boldsymbol{\Omega}_\epsilon$ of bounded size. However, for now we use proceed assuming the existence of appropriate model covers to bound the Rademacher complexity term.

We note that since our architecture outputs unit norm vectors i.e. $\|\mathcal{E}_{\boldsymbol{\theta}}(\mathbf{x})\|_2 = 1$ for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}, \mathbf{x} \in \mathcal{X}$, we can obtain trivial ϵ-covers for all $\epsilon \geq 2$. For example, we may simply take $\boldsymbol{\Omega}_\epsilon = \{(\mathbf{0}, \mathbf{0})\}$ for all $\epsilon \geq 2$. Note that due to the regularization step, our normalization operator $\mathfrak{N}_\nu(\cdot)$ does not suffer divide-by-zero errors or instability even in these degenerate settings.

We now pick scales $\epsilon = 2, 1, \frac{1}{2}, \ldots$, i.e. $\epsilon_j = 2^{-(j-1)}, j = 0, \ldots, K$ for some $K \in \mathbb{N}$ to be decided later, and let $\boldsymbol{\Omega}_{\epsilon_j}$ denote an ϵ_j-cover of size $\mathcal{N}_\infty(\boldsymbol{\Theta}, \epsilon_j)$. Then, for any $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, select a covering element from each one of these covers i.e. select $\boldsymbol{\omega}_j \in \boldsymbol{\Omega}_{\epsilon_j}$ such that $\boldsymbol{\omega}_j$ is an ϵ_j-covering element for $\boldsymbol{\theta}$ for $j = 0, \ldots, K$. Note that $\boldsymbol{\Omega}_0 = \{(\mathbf{0}, \mathbf{0})\}$ and thus $\boldsymbol{\omega}_0 = (\mathbf{0}, \mathbf{0})$ itself as $\epsilon_0 = 2$ allowing a trivial cover. Also, to avoid notational clutter, we abbreviate $\ell_i(\boldsymbol{\theta}) := \ell^{\text{prec}}_{\gamma,k}(s^{\boldsymbol{\theta}}(\mathbf{x}_i), \mathbf{y}_i)$. Then we have

$$\ell_i(\boldsymbol{\theta}) = \ell_i(\boldsymbol{\theta}) - \ell_i(\boldsymbol{\omega}_K) + \sum_{j=1}^{K} (\ell_i(\boldsymbol{\omega}_j) - \ell_i(\boldsymbol{\omega}_{j-1})) + \ell_i(\boldsymbol{\omega}_0)$$

Since the choice of $\boldsymbol{\omega}_0$ as a 2-covering element is independent of the model $\boldsymbol{\theta}$ to be covered, we have

$$\mathop{\mathbb{E}}_{\tau_i} \left[\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \frac{1}{N} \sum_{i=1}^{N} \tau_i \cdot \ell_i(\boldsymbol{\omega}_0) \right] = \mathop{\mathbb{E}}_{\tau_i} \left[\frac{1}{N} \sum_{i=1}^{N} \tau_i \cdot \ell_i(\boldsymbol{\omega}_0) \right] = 0$$

This gives us

$$\mathop{\mathbb{E}}_{\tau_i} \left[\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \frac{1}{N} \sum_{i=1}^{N} \tau_i \cdot \ell_i(\boldsymbol{\theta}) \right] \leq \underbrace{\mathop{\mathbb{E}}_{\tau_i} \left[\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \frac{1}{N} \sum_{i=1}^{N} \tau_i (\ell_i(\boldsymbol{\theta}) - \ell_i(\boldsymbol{\omega}_K)) \right]}_{(A)} + \sum_{j=1}^{K} \underbrace{\mathop{\mathbb{E}}_{\tau_i} \left[\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \frac{1}{N} \sum_{i=1}^{N} \tau_i (\ell_i(\boldsymbol{\omega}_j) - \ell_i(\boldsymbol{\omega}_{j-1})) \right]}_{(B_j)}$$

To bound the term (A) we notice that Lemma 6 tells us that $|\ell_i(\boldsymbol{\theta}) - \ell_i(\boldsymbol{\omega}_K)| \leq \frac{4\epsilon_K}{\gamma}$ for all choices of $(\mathbf{x}_i, \mathbf{y}_i)$. This allows a straightforward application of the Cauchy-Schwartz inequality to give us

$$(A) \leq \frac{1}{N} \cdot \|[\tau_1, \ldots, \tau_N]\|_2 \cdot \frac{4\epsilon_K}{\gamma}\sqrt{N} = \frac{4\epsilon_K}{\gamma}$$

Bounding the terms (B_j) requires us to notice that

$$\ell_i(\boldsymbol{\omega}_j) - \ell_i(\boldsymbol{\omega}_{j-1}) \leq \ell_i(\boldsymbol{\omega}_j) - \ell_i(\boldsymbol{\theta}) + \ell_i(\boldsymbol{\theta}) - \ell_i(\boldsymbol{\omega}_{j-1}) \leq \frac{4}{\gamma}(\epsilon_j + \epsilon_{j-1}) = \frac{4}{\gamma}(2^{-(j-1)} + 2^{-(j-2)}) = \frac{12\epsilon_j}{\gamma}$$

Also, we notice that the even if a union is taken over all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, the number of possible pairs $(\boldsymbol{\omega}_j, \boldsymbol{\omega}_{j-1})$ is at most $\mathcal{N}_\infty(\boldsymbol{\Theta}, \epsilon_j)\mathcal{N}_\infty(\boldsymbol{\Theta}, \epsilon_{j-1}) \leq (\mathcal{N}_\infty(\boldsymbol{\Theta}, \epsilon_j))^2$. An application of Massart's Finite Class Lemma now tells us that

$$\mathop{\mathbb{E}}_{\tau_i}\left[\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \frac{1}{N}\sum_{i=1}^{N} \tau_i(\ell_i(\boldsymbol{\omega}_j) - \ell_i(\boldsymbol{\omega}_{j-1}))\right] \leq \frac{24\epsilon_j}{\gamma}\sqrt{\frac{\ln(\mathcal{N}_\infty(\boldsymbol{\Theta}, \epsilon_j))}{N}} \leq \frac{48(\epsilon_j - \epsilon_{j-1})}{\gamma}\sqrt{\frac{\ln(\mathcal{N}_\infty(\boldsymbol{\Theta}, \epsilon_j))}{N}}$$

Taking a sum over $j = 1, \ldots, K$ and upper bounding the sum by an integral give us

$$\mathfrak{R}_N(\boldsymbol{\Theta}) \leq \frac{4}{\gamma}\left(\epsilon_K + \int_{\epsilon_{K+1}}^{2} \sqrt{\frac{\ln(\mathcal{N}_\infty(\boldsymbol{\Theta}, \epsilon))}{N}}\, d\epsilon\right)$$

Now, Lemma 8 offers us model covers of all scales with sizes at most

$$\ln(\mathcal{N}_\infty(\boldsymbol{\Theta}, \epsilon)) \leq \mathcal{O}\left(\frac{1}{\nu^2}\left(D\ln(D)R_\infty^R R_1^R + \frac{1}{\nu^2}\cdot s\ln^2(DV)R_\infty^E R_1^E (R_\sigma^R)^2\right)\cdot\frac{1}{\epsilon^2}\right) =: \frac{B}{\epsilon^2}$$

where we abbreviate $B := \mathcal{O}\left(\frac{1}{\nu^2}\left(D\ln(D)R_\infty^R R_1^R + \frac{1}{\nu^2}\cdot s\ln^2(DV)R_\infty^E R_1^E (R_\sigma^R)^2\right)\right)$. Thus,

$$\int_{\epsilon_{K+1}}^{2} \sqrt{\frac{\ln(\mathcal{N}_\infty(\boldsymbol{\Theta}, \epsilon))}{N}}\, d\epsilon \leq \sqrt{\frac{B}{N}}\ln\frac{2}{\epsilon_{K+1}}$$

Choosing $K > \frac{\ln N}{2}$ i.e. $\epsilon_{K+1} < \frac{1}{\sqrt{N}}$ grants us

$$\mathfrak{R}_N(\boldsymbol{\Theta}) \leq \frac{4}{\gamma}\left(\sqrt{\frac{1}{N}} + \sqrt{\frac{B\ln^2(2N)}{N}}\right) = \mathcal{O}\left(\sqrt{\frac{B\ln^2(2N)}{\gamma^2 N}}\right)$$

This finishes the claimed generalization bound. □

E.5 A Generalization Bound for Module III

The network architecture for documents remains identical in this module. However, the architecture for labels now includes a free vector parameter (the "extreme" 1-vs-all classifier $\boldsymbol{\eta}_l, l \in [L]$) and is given below. It is notable that the parameters $\boldsymbol{\theta} = \{\mathbf{E}, \mathbf{R}\}$ are shared with documents whereas the free parameters $\boldsymbol{\eta}_l$ are learnt one-per label. Let us abbreviate $\mathbf{H} = [\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_L]^\top \in \mathbb{R}^{L\times(D+2)}$ and let $\boldsymbol{\xi} := \{\boldsymbol{\theta}, \mathbf{H}\}$ denote the model augmented with these free parameters.

$$\mathcal{F}_{\boldsymbol{\xi}}(\mathbf{z}_l) = \mathfrak{N}_\nu(\mathcal{E}_{\boldsymbol{\theta}}(\mathbf{z}_l) + \boldsymbol{\eta}_l) \in \mathbb{R}^{D+3}$$

We define a corresponding model class as follows

$$\Xi_{R_1^E, R_\infty^E, R_1^R, R_\infty^R, R_\sigma^R, R_1^C, R_\infty^C} = \left\{(\boldsymbol{\theta}, \mathbf{H}) : \boldsymbol{\theta} \in \boldsymbol{\Theta}_{R_1^E, R_\infty^E, R_1^R, R_\infty^R, R_\sigma^R}, \mathbf{H} \in \mathbb{R}^{L\times(D+2)}, \|\mathbf{H}\|_{1,1} \leq R_1^C, \|\mathbf{H}\|_{1,\infty} \leq R_\infty^C\right\}.$$

In the sequel, we will abbreviate $\Xi_{R_1^E,R_\infty^E,R_1^R,R_\infty^R,R_\sigma^R,R_1^C,R_\infty^C}$ as simply Ξ, as well as continue to refer to abbreviate $\Theta_{R_1^E,R_\infty^E,R_1^R,R_\infty^R,R_\sigma^R}$ as simply Θ, to avoid notational clutter. Scores are now calculated as follows: for any $\mathbf{x} \in \mathbb{R}^d$ and any model $\boldsymbol{\xi} \in \Xi$, we have

$$s^{\boldsymbol{\xi}}(\mathbf{x}) = [s_1^{\boldsymbol{\xi}}, s_2^{\boldsymbol{\xi}}, \ldots, s_L^{\boldsymbol{\xi}}]^\top \in \mathbb{R}^L,$$

where $s_l^{\boldsymbol{\xi}} = \mathfrak{N}_0(\mathcal{E}_{\boldsymbol{\theta}}(\mathbf{x}))^\top \mathcal{F}_{\boldsymbol{\xi}}(\mathbf{z}_l)$ for all $l \in [L]$. Note that $\mathfrak{N}_0(\mathcal{E}_{\boldsymbol{\theta}}(\mathbf{x}))$ simply appends a zero to the vector $\mathcal{E}_{\boldsymbol{\theta}}(\mathbf{x})$ to make it a $D+3$ dimensional vector compatible for dot products with $\mathcal{F}_{\boldsymbol{\xi}}(\mathbf{z}_l) \in \mathbb{R}^{D+3}$. We do not require regularization for this normalization step since $\mathcal{E}_{\boldsymbol{\theta}}(\mathbf{x})$ is already a unit vector.

Theorem 5 (Module-III Generalization Bound)**.** *Suppose the learning procedure in module III learns a model $\hat{\boldsymbol{\xi}} \in \Xi$. Then with probability at least $1-\delta$, we have*

$$\ell(\boldsymbol{\xi}) \leq \hat{\ell}_N(\boldsymbol{\xi}) + \frac{1}{\gamma} \cdot \frac{Q \ln(N)}{\sqrt{N}} + \sqrt{\frac{\ln \frac{1}{\delta}}{N}},$$

where $Q := C \cdot \frac{1}{\nu}\left(\ln(DL)\sqrt{R_\infty^C R_1^C} + \frac{1}{\nu}\left(\sqrt{D\ln(D)}\sqrt{R_\infty^R R_1^R} + \frac{1}{\nu} \cdot \sqrt{s}\ln(DV) R_\sigma^R \sqrt{R_\infty^E R_1^E}\right)\right)$ for some universal constant C.

Before proceeding with the proof, we need to appropriately extend the notion of a model cover for these augmented models which we do below. As before, this result can be readily extended to admit empirical bounds on various norms using a standard stratification step [1].

Definition 6 (Uniform ϵ-covers for Augmented Model Classes)**.** *For any $\epsilon > 0$ and model class Ξ, we say that a model class Ψ_ϵ is an ϵ-cover if for any $\boldsymbol{\xi} = (\boldsymbol{\theta}, \mathbf{H}) \in \Xi$, there exists $\boldsymbol{\psi} = (\boldsymbol{\omega}, \mathbf{G}) \in \Psi_\epsilon$, such that $\sup_{\mathbf{x}\in\mathcal{X}} \|\mathcal{E}_{\boldsymbol{\theta}}(\mathbf{x}) - \mathcal{E}_{\boldsymbol{\omega}}(\mathbf{x})\|_2 \leq \epsilon$ as well as $\|\mathbf{H}-\mathbf{G}\|_{2,\infty} \leq \epsilon$. Moreover, for any $\epsilon > 0$, we let $\mathcal{N}_\infty(\Xi, \epsilon)$ denote the size of the smallest such ϵ-cover.*

Note that if we let $\mathbf{H} = [\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_L]^\top \in \mathbb{R}^{L\times(D+2)}$ and $\mathbf{G} = [\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_L]^\top \in \mathbb{R}^{L\times(D+2)}$, then the condition $\|\mathbf{H}-\mathbf{G}\|_{2,\infty} \leq \epsilon$ above translates to $\sup_{l\in[L]} \|\boldsymbol{\eta}_l - \boldsymbol{\zeta}_l\|_2 \leq \epsilon$.

Proof of Theorem 5. We start by noticing that Lemma 7 tells us that the loss function is Lipschitz with respect to the augmented models as well, but with a Lipschitz constant $\frac{10}{\gamma\nu}$ that now includes a factor involving ν due to the additional normalization step carried out in $\mathcal{F}_{\boldsymbol{\xi}}(\cdot)$.

Thus, to produce a generalization bound in the presence of the augmented free parameters, all we need to do is establish an ϵ-cover for the model class Ξ. To do so we notice that $\Xi = \Theta \times \Lambda$ where $\Lambda := \Lambda_{R_1^C, R_\infty^C} = \left\{\mathbf{H} \in \mathbb{R}^{L\times(D+2)} : \|\mathbf{H}\|_{1,1} \leq R_1^C, \|\mathbf{H}\|_{1,\infty} \leq R_\infty^C\right\}$. Thus, to construct a cover for Ξ, we take the following steps

1. Invoke Lemma 8 to obtain an ϵ-cover Ω_ϵ for Θ
2. Construct Π_ϵ such that for every $\mathbf{H} \in \Lambda$, there exists a $\mathbf{G} \in \Pi_\epsilon$ such that $\|\mathbf{H}-\mathbf{G}\|_{2,\infty} \leq \epsilon$
3. Construct the overall cover as $\Psi_\epsilon := \Omega_\epsilon \times \Pi_\epsilon$

The size of such a cover would be $\ln(|\Psi_\epsilon|) \leq \ln(|\Omega_\epsilon|) + \ln(|\Pi_\epsilon|)$ which also gives us

$$\ln(\mathcal{N}_\infty(\Xi, \epsilon)) \leq \ln(\mathcal{N}_\infty(\Theta, \epsilon)) + \ln(|\Pi_\epsilon|).$$

Now, Claim 15 offers us a cover Π_ϵ of size at most

$$\ln(|\Pi_\epsilon|) \leq 64 \ln(3eL) \ln(2(D+2)L) R_\infty^C R_1^C \cdot \frac{1}{\epsilon^2}$$

Then, following the same steps as in the proof of Theorem 4 and using Lemma 7 tells us that

$$\begin{aligned}\mathfrak{R}_N(\Xi) &\leq \frac{10}{\gamma\nu}\left(\epsilon_K + \int_{\epsilon_{K+1}}^{2}\sqrt{\frac{\ln(\mathcal{N}_\infty(\Xi,\epsilon))}{N}}\,d\epsilon\right)\\ &\leq \frac{10}{\gamma\nu}\left(\epsilon_K + \int_{\epsilon_{K+1}}^{2}\sqrt{\frac{\ln(\mathcal{N}_\infty(\boldsymbol{\Theta},\epsilon)) + \ln(|\boldsymbol{\Pi}_\epsilon|)}{N}}\right)\,d\epsilon\\ &\leq \frac{10}{\gamma\nu}\left(\epsilon_K + \sqrt{\frac{A}{N}}\ln\frac{2}{\epsilon_{K+1}}\right)\end{aligned}$$

where $A = \mathcal{O}\left(\frac{1}{\nu^2}\left(D\ln(D)R_\infty^R R_1^R + \frac{1}{\nu^2}\cdot s\ln^2(DV)R_\infty^E R_1^E (R_\sigma^R)^2\right) + \ln^2(DL)R_\infty^C R_1^C\right)$. As before, choosing $K > \frac{\ln N}{2}$ i.e. $\epsilon_{K+1} < \frac{1}{\sqrt{N}}$ finishes the proof. □

F Experiments

Table 5 presents the statistics for all datasets used in the experiments.

F.1 Evaluation metrics

Precision and Normalized Discounted Cumulative Gain (nDCG) are widely used evaluation metrics in extreme multi-label learning. Results for various methods are reported with respect to vanilla precision (P@k) and nDCG (N@k), as well as propensity scored precision (PSP@k) and nDCG (PSN@k) with $k = 1$, 3 and 5. The propensity scoring model and values taken from the Extreme Classification Repository [2] were used to evaluate the methods. For proprietary datasets, the method outlined in [5] was followed to obtain propensity scores for labels. For a predicted score vector $\hat{\mathbf{y}} \in R^L$ and ground truth label vector $\mathbf{y} \in \{0,1\}^L$, the metrics are defined below. In the following, p_l is propensity score of the label l as described in [5].

$$\begin{aligned}P@k &= \frac{1}{k}\sum_{l\in rank_k(\hat{\mathbf{y}})} y_l & PSP@k &= \frac{1}{k}\sum_{l\in rank_k(\hat{\mathbf{y}})}\frac{y_l}{p_l}\\ D@k &= \frac{1}{k}\sum_{l\in rank_k(\hat{\mathbf{y}})}\frac{y_l}{\log(l+1)} & PSD@k &= \frac{1}{k}\sum_{l\in rank_k(\hat{\mathbf{y}})}\frac{y_l}{p_l\log(l+1)}\\ N@k &= \frac{D@k}{\sum_{l=1}^{\min(k,||\mathbf{y}||_0)}\frac{1}{\log(l+1)}} & PSN@k &= \frac{PSD@k}{\sum_{l=1}^{k}\frac{1}{\log l+1}}\end{aligned},$$

F.2 Hyper-parameters

SiameseXML uses a few hyper-parameters described below.

1. Scaling parameters $c \in (0,1), d \geq 1$ in the probability model p_l (see Section 4, "Architecture Details").
2. The dimensionality D of the embeddings.

Module I used $c = 0.9, d = 1.5$ whereas Module III used $c = 0.75, d = 3.0$. $D = 300$ was used for the publicly available benchmarks datasets and $D = 128$ was used for the (larger) proprietary datasets. A shortlist of size $|\mathcal{S}| = 500$ was used and the weighting constant α (see Section 4 "Log-time Prediction") was picked from the set $\{0.75, 0.9, 0.95\}$ across the datasets. Finally, the hyper-parameters for the optimizer include learning rate and batch size which were set to default values across all datasets. In particular, Module-I used a batch size of 4096 and learning rate of 0.005 whereas Module II used a batch size of 256 and learning rate 0.0005.

Table 5: Dataset Statistics. A ‡ sign denotes information that was redacted for proprietary datasets.

Dataset	Num train docs N	Num labels L	Num tokens V	Num test docs N'	Avg labels per doc	Avg docs per label	Avg tokens per doc	Avg tokens per label
Benchmark public datasets								
LF-AmazonTitles-131K	294,805	131,073	40,000	134,835	2.29	5.15	7.46	7.15
LF-WikiSeeAlsoTitles-320K	693,082	312,330	40,000	177,515	2.11	4.68	3.97	3.92
Proprietary datasets								
Q2BP-4M	20,973,324	5,246,101	‡	4,000,000	‡	‡	‡	‡
Q2BP-40M	64,308,169	40,000,000	‡	16,075,850	‡	‡	‡	‡
Q2BP-100M	187,355,925	100,000,000	‡	80,289,870	‡	‡	‡	‡

Table 6: An extension of Table 1 from the main paper displaying nDCG and propensity scored nDCG, as well as model sizes for all methods. SiameseXML could be significantly more accurate and scalable than leading deep extreme classifiers including Astec, Decaf, and XTransformer on repository datasets. Results are only presented for datasets to which an algorithm could scale with the timeout described in Section 6.

Method	N@1	N@3	N@5	PSN@1	PSN@3	PSN@5	Model Size (GB)	Training time (hr)
LF-AmazonTitles-131K								
SiameseXML	39.43	40.21	42.16	33.83	36.83	39.03	1.16	0.66
SiameseXML-3	**40.26**	**41.14**	**43.21**	**34.51**	**37.62**	**39.93**	3.48	1.97
Astec	37.12	38.17	40.16	29.22	32.73	35.03	3.31	1.83
Decaf	38.40	39.43	41.46	30.85	34.69	37.13	0.83	2.16
MACH	33.49	34.36	36.16	24.97	28.41	30.54	2.41	3.30
SLICE+FastText	30.43	31.07	32.76	23.08	26.11	28.13	0.40	0.08
AttentionXML	32.25	32.83	34.42	23.97	26.88	28.75	2.67	20.73
Parabel	32.60	32.96	34.47	23.27	26.36	28.21	0.35	0.03
Bonsai	34.11	34.81	36.57	24.75	28.32	30.47	0.24	0.10
DiSMEC	35.14	36.17	38.06	25.86	30.09	32.47	0.11	3.10
LF-WikiSeeAlsoTitles-320K								
SiameseXML	27.63	27.03	27.87	22.68	24.20	25.51	2.69	1.05
SiameseXML-3	**28.51**	**28.01**	**28.91**	**23.29**	**24.92**	**26.28**	8.07	3.15
Astec	22.72	22.16	22.87	13.69	15.56	16.75	7.47	4.17
Decaf	25.14	24.99	25.95	16.73	19.18	20.75	1.80	11.16
MACH	18.06	17.57	18.17	9.68	11.19	12.14	2.57	8.23
SLICE+FastText	18.55	18.29	19.07	11.24	13.03	14.23	0.96	0.20
AttentionXML	17.56	16.58	17.07	9.45	10.45	11.24	6.17	56.12
Parabel	17.68	16.96	17.44	9.24	10.49	11.32	0.61	0.07
Bonsai	19.31	18.74	19.32	10.69	12.29	13.29	0.38	0.37
DiSMEC	19.12	18.93	19.71	10.56	12.70	14.02	0.20	15.56

Table 7: A subjective comparison of the top 5 predictions made by various algorithms on selected examples from the LF-WikiSeeAlsoTitles-320K dataset. SiameseXML's predictions are more accurate as compared to leading methods including DECAF and AttentionXML. Mispredictions are typeset in light gray.

Method	**Top 5 Predictions**
Document	**Sinhala script**
SiameseXML	Dutch loanwords in Sinhala, Portuguese loanwords in Sinhala, Tamil loanwords in Sinhala, English loanwords in Sinhala, History of Sinhala software
DECAF	Portuguese loanwords in Sinhala, Tamil loanwords in Sinhala, Dutch loanwords in Sinhala Gupta script, Pre-Islamic scripts in Afghanistan
Astec	Portuguese loanwords in Sinhala, *Mongolian script*, Filipino orthography, Mongolian writing systems, Thai honorifics
AttentionXML	Mongolian writing systems, Sindhi language, Romanization of Khmer, Greater India Indosphere
Document	**List of Go players**
SiameseXML	List of Go organizations, Go players, International Go Federation, Go professional, List of professional Go tournaments
DECAF	Go players, List of Go organizations, Music of the Republic of Macedonia, Players, List of all-female bands
Astec	List of NHL players, List of Israeli chess players, List of chess players, Hardball squash Gibson Guitar Corporation product list
AttentionXML	List of NHL players, List of professional Go & tournaments, List of foreign NBA players, List of chess grandmasters, List of Israeli chess players

G Supporting Results

Lemma 6. *Let $\mathbf{\Omega}_\epsilon$ be an ϵ-cover for the model class $\mathbf{\Theta}$ as defined in Definition 5. For any $\boldsymbol{\theta} \in \mathbf{\Theta}$, suppose $\boldsymbol{\omega} \in \mathbf{\Omega}_\epsilon$ is the ϵ-covering element, i.e. $\sup_{\mathbf{x}\in\mathcal{X}} \|\mathcal{E}_{\boldsymbol{\theta}}(\mathbf{x}) - \mathcal{E}_{\boldsymbol{\omega}}(\mathbf{x})\|_2 \leq \epsilon$. Then for any $(\mathbf{x}, \mathbf{y})$, we have*

$$\left|\ell^{prec}_{\gamma,k}(s^{\boldsymbol{\theta}}(\mathbf{x}), \mathbf{y}) - \ell^{prec}_{\gamma,k}(s^{\boldsymbol{\omega}}(\mathbf{x}), \mathbf{y})\right| \leq \frac{4\epsilon}{\gamma}.$$

Proof. Since the γ-ramp function is $\frac{1}{\gamma}$-Lipschitz, we have

$$\begin{aligned}
\left|\ell^{\text{prec}}_{\gamma,k}(s^{\boldsymbol{\theta}}(\mathbf{x}), \mathbf{y}) - \ell^{\text{prec}}_{\gamma,k}(s^{\boldsymbol{\omega}}(\mathbf{x}), \mathbf{y})\right| &\leq \frac{1}{\gamma}\left(\max_{l\in P_{\mathbf{y}}} \left|s^{\boldsymbol{\theta}}_l - s^{\boldsymbol{\omega}}_l\right| + \max_{l'\in N_{\mathbf{y}}} \left|s^{\boldsymbol{\theta}}_{l'} - s^{\boldsymbol{\omega}}_{l'}\right|\right) \leq \frac{2}{\gamma}\cdot\max_{l\in[L]} \left|s^{\boldsymbol{\theta}}_l - s^{\boldsymbol{\omega}}_l\right| \\
&= \frac{2}{\gamma}\cdot\max_{l\in[L]} \left|\mathcal{E}_{\boldsymbol{\theta}}(\mathbf{x})^\top \mathcal{E}_{\boldsymbol{\theta}}(\mathbf{z}_l) - \mathcal{E}_{\boldsymbol{\omega}}(\mathbf{x})^\top \mathcal{E}_{\boldsymbol{\omega}}(\mathbf{z}_l)\right| \\
&\leq \frac{2}{\gamma}\cdot\max_{l\in[L]} \left\{\left|\mathcal{E}_{\boldsymbol{\theta}}(\mathbf{x})^\top (\mathcal{E}_{\boldsymbol{\theta}}(\mathbf{z}_l) - \mathcal{E}_{\boldsymbol{\omega}}(\mathbf{z}_l))\right| + \left|(\mathcal{E}_{\boldsymbol{\theta}}(\mathbf{x}) - \mathcal{E}_{\boldsymbol{\omega}}(\mathbf{x}))^\top \mathcal{E}_{\boldsymbol{\omega}}(\mathbf{z}_l)\right|\right\} \\
&\leq \frac{4\epsilon}{\gamma},
\end{aligned}$$

where we used the fact that our architecture outputs normalized vectors i.e. $\|\mathcal{E}_{\boldsymbol{\theta}}(\mathbf{x})\|_2 = 1 = \|\mathcal{E}_{\boldsymbol{\theta}}(\mathbf{z}_l)\|_2$ for all $l \in [L]$, the fact that $\mathbf{z}_l \in \mathcal{X}$ for all $l \in [L]$, as well as the fact that $\boldsymbol{\omega}$ is an ϵ-covering element for $\boldsymbol{\theta}$. This concludes the proof. □

Lemma 7. *Let $\mathbf{\Psi}_\epsilon$ be an ϵ-cover for the model class $\mathbf{\Xi}$ as defined in Definition 6. For any $\boldsymbol{\xi} = (\boldsymbol{\theta}, \mathbf{H}) \in \mathbf{\Xi}$, suppose $\boldsymbol{\psi} = (\boldsymbol{\xi}, \mathbf{G}) \in \mathbf{\Psi}_\epsilon$ is the ϵ-covering element, i.e. $\sup_{\mathbf{x}\in\mathcal{X}} \|\mathcal{E}_{\boldsymbol{\theta}}(\mathbf{x}) - \mathcal{E}_{\boldsymbol{\omega}}(\mathbf{x})\|_2 \leq \epsilon$ and $\|\mathbf{H} - \mathbf{G}\|_{2,\infty} \leq \epsilon$. Then for any $(\mathbf{x}, \mathbf{y})$, we have*

$$\left|\ell^{prec}_{\gamma,k}(s^{\boldsymbol{\xi}}(\mathbf{x}), \mathbf{y}) - \ell^{prec}_{\gamma,k}(s^{\boldsymbol{\psi}}(\mathbf{x}), \mathbf{y})\right| \leq \frac{10\epsilon}{\gamma\nu}.$$

Proof. Similar to the proof of Lemma 6, if we let $\mathbf{H} = [\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_L]$ and $\mathbf{G} = [\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_L]$, we have

$$\begin{aligned}
\left|\ell^{\text{prec}}_{\gamma,k}(s^{\boldsymbol{\xi}}(\mathbf{x}), \mathbf{y}) - \ell^{\text{prec}}_{\gamma,k}(s^{\boldsymbol{\psi}}(\mathbf{x}), \mathbf{y})\right| &\leq \frac{2}{\gamma}\cdot\max_{l\in[L]} \left|s^{\boldsymbol{\xi}}_l - s^{\boldsymbol{\psi}}_l\right| = \frac{2}{\gamma}\cdot\max_{l\in[L]} \left|\mathfrak{N}_0(\mathcal{E}_{\boldsymbol{\theta}}(\mathbf{x}))^\top \mathcal{F}_{\boldsymbol{\xi}}(\mathbf{z}_l) - \mathfrak{N}_0(\mathcal{E}_{\boldsymbol{\omega}}(\mathbf{x}))^\top \mathcal{F}_{\boldsymbol{\psi}}(\mathbf{z}_l)\right| \\
&\leq \frac{2}{\gamma}\cdot\max_{l\in[L]} \left\{\left|\mathfrak{N}_0(\mathcal{E}_{\boldsymbol{\theta}}(\mathbf{x}))^\top (\mathcal{F}_{\boldsymbol{\xi}}(\mathbf{z}_l) - \mathcal{F}_{\boldsymbol{\psi}}(\mathbf{z}_l))\right| + \left|(\mathfrak{N}_0(\mathcal{E}_{\boldsymbol{\theta}}(\mathbf{x})) - \mathfrak{N}_0(\mathcal{E}_{\boldsymbol{\omega}}(\mathbf{x})))^\top \mathcal{F}_{\boldsymbol{\psi}}(\mathbf{z}_l)\right|\right\} \\
&\leq \frac{2}{\gamma}\left(\|\mathcal{F}_{\boldsymbol{\xi}}(\mathbf{z}_l) - \mathcal{F}_{\boldsymbol{\psi}}(\mathbf{z}_l)\|_2 + \|\mathfrak{N}_0(\mathcal{E}_{\boldsymbol{\theta}}(\mathbf{x})) - \mathfrak{N}_0(\mathcal{E}_{\boldsymbol{\omega}}(\mathbf{x}))\|_2\right) \\
&= \frac{2}{\gamma}\left(\|\mathfrak{N}_\nu(\mathcal{E}_{\boldsymbol{\theta}}(\mathbf{z}_l) + \boldsymbol{\eta}_l) - \mathfrak{N}_\nu(\mathcal{E}_{\boldsymbol{\omega}}(\mathbf{z}_l) + \boldsymbol{\zeta}_l)\|_2 + \|\mathcal{E}_{\boldsymbol{\theta}}(\mathbf{x}) - \mathcal{E}_{\boldsymbol{\omega}}(\mathbf{x})\|_2\right) \\
&\leq \frac{2}{\gamma}\left(\frac{2}{\nu}\left(\|\mathcal{E}_{\boldsymbol{\theta}}(\mathbf{z}_l) - \mathcal{E}_{\boldsymbol{\omega}}(\mathbf{z}_l)\|_2 + \|\boldsymbol{\eta}_l - \boldsymbol{\zeta}_l\|_2\right) + \|\mathcal{E}_{\boldsymbol{\theta}}(\mathbf{x}) - \mathcal{E}_{\boldsymbol{\omega}}(\mathbf{x})\|_2\right) \\
&\leq \frac{2}{\gamma}\left(\frac{4\epsilon}{\nu} + \epsilon\right) \leq \frac{10\epsilon}{\gamma\nu},
\end{aligned}$$

where in the third step, we used the fact that our architecture outputs normalized vectors i.e. $\|\mathfrak{N}_0(\mathcal{E}_{\boldsymbol{\theta}}(\mathbf{x}))\|_2 = 1 = \|\mathcal{F}_{\boldsymbol{\psi}}(\mathbf{z}_l)\|_2$ for all $l \in [L]$, in the next step we used the fact that $\mathfrak{N}_0(\mathcal{E}_{\boldsymbol{\theta}}(\mathbf{x})) = [\mathcal{E}_{\boldsymbol{\theta}}(\mathbf{x}), 0]$, in the next step we used Corollary 12 and in the final step, we used the fact that $\mathbf{z}_l \in \mathcal{X}$ for all $l \in [L]$, the fact that $\|\mathbf{H} - \mathbf{G}\|_{2,\infty} \leq \epsilon$, and then simplified the bound for $\nu \leq 1$ w.l.o.g. which will always be the case whenever this bound is applied. This concludes the proof. □

Lemma 8 (Model Covering)**.** *For any $R_1^E, R_\infty^E, R_1^R, R_\infty^R, R_\sigma^R, \epsilon > 0$, the model class $\mathbf{\Theta}_{R_1^E, R_\infty^E, R_1^R, R_\infty^R, R_\sigma^R}$ admits an ϵ-cover $\mathbf{\Omega}_\epsilon$ of size at most*

$$\ln(|\mathbf{\Omega}_\epsilon|) \leq c \cdot \left(\frac{1}{\nu^2} \left(D \ln(D) R_\infty^R R_1^R + \frac{1}{\nu^2} \cdot s \ln^2(DV) R_\infty^E R_1^E (R_\sigma^R)^2 \right) \cdot \frac{1}{\epsilon^2} \right),$$

where $c > 0$ is a universal constant.

Proof. Let $\epsilon_E, \epsilon_R > 0$ be constants that shall be fixed later. We first invoke Claim 13 to obtain an ϵ_E-cover $\mathcal{C}_E$ over the set of matrices $\left\{ \mathbf{E} \in \mathbb{R}^{D \times V} : \|\mathbf{E}\|_{1,1} \leq R_1^E, \|\mathbf{E}\|_{\infty,1} \leq R_\infty^E \right\}$. We then invoke Claim 14 to obtain an ϵ_R-cover $\mathcal{C}_R$ over the set of matrices $\left\{ \mathbf{R} \in \mathbb{R}^{(D+1)\times(D+1)} : \|\mathbf{R}\|_{1,1} \leq R_1^R, \|\mathbf{R}\|_{\infty,1} \leq R_\infty^R \right\} \supseteq \left\{ \mathbf{R} \in \mathbb{R}^{(D+1)\times(D+1)} : \|\mathbf{R}\|_{1,1} \leq R_1^R, \|\mathbf{R}\|_{\infty,1} \leq R_\infty^R, \|\mathbf{R}\|_\sigma \leq R_\sigma^R \right\}$. Note that we are assured that

$$\ln(|\mathcal{C}_E|) \leq 64 s \ln(3e^2 V) \ln(2DV) R_\infty^E R_1^E \cdot \frac{1}{\epsilon_E^2}$$
$$\ln(|\mathcal{C}_R|) \leq 128 (D+1) \ln(2(D+1)^2) R_\infty^R R_1^R \cdot \frac{1}{\epsilon_R^2}$$

Claim 9 shows that the model $\tilde{\boldsymbol{\theta}} := (\tilde{\mathbf{E}}, \tilde{\mathbf{R}})$ is a $\left(\frac{2}{\nu} \left(\frac{2(R_\sigma^R + 1)}{\nu} \cdot \epsilon_E + \epsilon_R \right) \right)$-cover for $\boldsymbol{\theta}$. Setting $\epsilon_E = \frac{\nu^2 \epsilon}{8(R_\sigma^R + 1)}$ and $\epsilon_R = \frac{\nu \epsilon}{4}$ gives us

$$\frac{2}{\nu} \left(\frac{2(R_\sigma^R + 1)}{\nu} \cdot \epsilon_E + \epsilon_R \right) = \epsilon.$$

Constructing the overall cover as

$$\mathbf{\Omega}_\epsilon := \mathcal{C}_E \times \mathcal{C}_R = \left\{ (\tilde{\mathbf{E}}, \tilde{\mathbf{R}}) : \tilde{\mathbf{E}} \in \mathcal{C}_E, \tilde{\mathbf{R}} \in \mathcal{C}_R \right\}$$

gives us $\ln(|\mathbf{\Omega}_\epsilon|) \leq \ln(|\mathcal{C}_E|) + \ln(|\mathcal{C}_R|)$ which finishes the proof after simplifications. □

Claim 9. *For an arbitrary model $\boldsymbol{\theta} = (\mathbf{E}, \mathbf{R}) \in \mathbf{\Theta}_{R_1^E, R_\infty^E, R_1^R, R_\infty^R, R_\sigma^R}$, let $\tilde{\mathbf{E}} \in \mathcal{C}_E$ (respectively $\tilde{\mathbf{R}} \in \mathcal{C}_R$) be an ϵ_E-covering element for $\mathbf{E}$ (respectively ϵ_R-covering element for $\mathbf{R}$). Then we have*

$$\sup_{\mathbf{x} \in \mathcal{X}} \left\| \mathcal{E}_{\boldsymbol{\theta}}(\mathbf{x}) - \mathcal{E}_{\tilde{\boldsymbol{\theta}}}(\mathbf{x}) \right\|_2 \leq \frac{2}{\nu} \left(\frac{2(R_\sigma^R + 1)}{\nu} \cdot \epsilon_E + \epsilon_R \right).$$

Proof. Choose an arbitrary data point $\mathbf{x} \in \mathcal{X}$ and note that

$$\begin{aligned}
\left\| f(\mathbf{x}; \mathbf{E}) - f(\mathbf{x}; \tilde{\mathbf{E}}) \right\|_2 &= \left\| \mathfrak{N}_\nu(\mathrm{ReLU}(\mathbf{E}\mathbf{x})) - \mathfrak{N}_\nu(\mathrm{ReLU}(\tilde{\mathbf{E}}\mathbf{x})) \right\|_2 \\
&\leq \frac{2}{\nu} \cdot \left\| \mathrm{ReLU}(\mathbf{E}\mathbf{x}) - \mathrm{ReLU}(\tilde{\mathbf{E}}\mathbf{x}) \right\|_2 \\
&\leq \frac{2}{\nu} \cdot \left\| \mathbf{E}\mathbf{x} - \tilde{\mathbf{E}}\mathbf{x} \right\|_2 \leq \frac{2}{\nu} \cdot \epsilon_E,
\end{aligned}$$

where the second step follows from Claim 11, the third step follows from Claim 10, and the last step follows from Claim 13 and using the fact that by construction, $\tilde{\mathbf{E}}$ is an ϵ_E-covering element for $\mathbf{E}$ and that $\mathbf{x} \in \mathcal{X}$. Similarly, we have

$$\begin{aligned}
\left\| g(f(\mathbf{x}; \mathbf{E}); \mathbf{R}) - g(f(\mathbf{x}; \tilde{\mathbf{E}}); \tilde{\mathbf{R}}) \right\|_2 &= \left\| \mathrm{ReLU}(\mathbf{R} \cdot f(\mathbf{x}; \mathbf{E})) - \mathrm{ReLU}(\tilde{\mathbf{R}} \cdot f(\mathbf{x}; \tilde{\mathbf{E}})) \right\|_2 \\
&\leq \left\| \mathbf{R} \cdot f(\mathbf{x}; \mathbf{E}) - \tilde{\mathbf{R}} \cdot f(\mathbf{x}; \tilde{\mathbf{E}}) \right\|_2 \\
&\leq \left\| \mathbf{R} \cdot f(\mathbf{x}; \mathbf{E}) - \mathbf{R} \cdot f(\mathbf{x}; \tilde{\mathbf{E}}) \right\|_2 + \left\| \mathbf{R} \cdot f(\mathbf{x}; \tilde{\mathbf{E}}) - \tilde{\mathbf{R}} \cdot f(\mathbf{x}; \tilde{\mathbf{E}}) \right\|_2 \\
&\leq \|\mathbf{R}\|_\sigma \cdot \left\| f(\mathbf{x}; \mathbf{E}) - f(\mathbf{x}; \tilde{\mathbf{E}}) \right\|_2 + \epsilon_R \\
&\leq \frac{2R_\sigma^R}{\nu} \cdot \epsilon_E + \epsilon_R,
\end{aligned}$$

where the second step follows from Claim 10, the fourth step follows from Claim 14 and using the fact that $\tilde{\mathbf{R}}$ is an ϵ_R-covering element for $\mathbf{R}$ and that $\left\|f(\mathbf{x};\tilde{\mathbf{E}})\right\|_2 = 1$ by the definition of the function f. The last step follows since we proved $\left\|f(\mathbf{x};\mathbf{E}) - f(\mathbf{x};\tilde{\mathbf{E}})\right\|_2 \leq \frac{2\epsilon_E}{\nu}$ earlier. Finally, we note that

$$\begin{aligned}
\left\|\mathcal{E}_{\boldsymbol{\theta}}(\mathbf{x}) - \mathcal{E}_{\tilde{\boldsymbol{\theta}}}(\mathbf{x})\right\|_2 &= \left\|\mathfrak{N}_\nu(f(\mathbf{x};\mathbf{E}) + g(f(\mathbf{x};\mathbf{E});\mathbf{R})) - \mathfrak{N}_\nu(f(\mathbf{x};\tilde{\mathbf{E}}) + g(f(\mathbf{x};\tilde{\mathbf{E}});\tilde{\mathbf{R}}))\right\|_2 \\
&\leq \frac{2}{\nu} \cdot \left\|f(\mathbf{x};\mathbf{E}) + g(f(\mathbf{x};\mathbf{E});\mathbf{R}) - f(\mathbf{x};\tilde{\mathbf{E}}) - g(f(\mathbf{x};\tilde{\mathbf{E}});\tilde{\mathbf{R}})\right\|_2 \\
&\leq \frac{2}{\nu}\left(\left\|f(\mathbf{x};\mathbf{E}) - f(\mathbf{x};\tilde{\mathbf{E}})\right\|_2 + \left\|g(f(\mathbf{x};\mathbf{E});\mathbf{R}) - g(f(\mathbf{x};\tilde{\mathbf{E}});\tilde{\mathbf{R}})\right\|_2\right) \\
&\leq \frac{2}{\nu}\left(\frac{2(R_\sigma^R+1)}{\nu}\cdot\epsilon_E + \epsilon_R\right)
\end{aligned}$$

where in the second step, we use Claim 11 and in the last step we use the above calculations. Since the point $\mathbf{x} \in \mathcal{X}$ we chose was completely arbitrary, the result holds uniformly over the entire set $\mathcal{X}$. This concludes the proof. □

Claim 10. *For any $\epsilon > 0, d \in \mathbb{N}$, and any $\mathbf{x}, \mathbf{c} \in \mathbb{R}^d$, we always have $\|ReLU(\mathbf{x}) - ReLU(\mathbf{c})\|_2 \leq \|\mathbf{x} - \mathbf{c}\|_2$.*

Proof. The proof follows from the fact that $\max\{x, 0\}$ is 1-Lipschitz function which ensures that for all $a, b \in \mathbb{R}$, we have $|\mathrm{ReLU}(a) - \mathrm{ReLU}(b)| \leq |a - b|$. This gives us

$$\|\mathrm{ReLU}(\mathbf{x}) - \mathrm{ReLU}(\mathbf{c})\|_2^2 = \sum_{j=1}^{d}(\mathrm{ReLU}(x_j) - \mathrm{ReLU}(c_j))^2 \leq \sum_{j=1}^{d}(x_j - c_j)^2 = \|\mathbf{x} - \mathbf{c}\|_2^2$$ □

Claim 11. *For any $\epsilon > 0, d \in \mathbb{N}$, and any $\mathbf{x}, \mathbf{c} \in \mathbb{R}^d$, we always have $\|\mathfrak{N}_\nu(\mathbf{x}) - \mathfrak{N}_\nu(\mathbf{c})\|_2 \leq \frac{2}{\nu} \cdot \|\mathbf{x} - \mathbf{c}\|_2$.*

Proof. We have

$$\begin{aligned}
\|\mathfrak{N}_\nu(\mathbf{x}) - \mathfrak{N}_\nu(\mathbf{c})\|_2 &= \left\|\frac{[\mathbf{x},\nu] - [\mathbf{c},\nu]}{\sqrt{\|\mathbf{x}\|_2^2 + \nu^2}} - [\mathbf{c},\nu]\left(\frac{1}{\sqrt{\|\mathbf{c}\|_2^2+\nu^2}} - \frac{1}{\sqrt{\|\mathbf{x}\|_2^2+\nu^2}}\right)\right\|_2 \\
&\leq \frac{\|[\mathbf{x},\nu] - [\mathbf{c},\nu]\|_2}{\sqrt{\|\mathbf{x}\|_2^2+\nu^2}} + \|[\mathbf{c},\nu]\|_2 \cdot \left|\frac{1}{\sqrt{\|\mathbf{c}\|_2^2+\nu^2}} - \frac{1}{\sqrt{\|\mathbf{x}\|_2^2+\nu^2}}\right| \\
&= \frac{\|\mathbf{x}-\mathbf{c}\|_2 + \left|\sqrt{\|\mathbf{x}\|_2^2+\nu^2} - \sqrt{\|\mathbf{c}\|_2^2+\nu^2}\right|}{\sqrt{\|\mathbf{x}\|_2^2+\nu^2}} \\
&\leq \frac{\|\mathbf{x}-\mathbf{c}\|_2 + \left|\sqrt{\|\mathbf{x}\|_2^2+\nu^2} - \sqrt{\|\mathbf{c}\|_2^2+\nu^2}\right|}{\nu},
\end{aligned}$$

where the last step uses the fact that $\|\mathbf{x}\|_2^2 \geq 0$. Now, the function $\sqrt{t^2+\nu^2}$ is 1-Lipschitz for all $\nu > 0$ which gives us

$$\left|\sqrt{\|\mathbf{x}\|_2^2+\nu^2} - \sqrt{\|\mathbf{c}\|_2^2+\nu^2}\right| \leq |\|\mathbf{x}\|_2 - \|\mathbf{c}\|_2|,$$

whereas the reverse triangle inequality gives us $|\|\mathbf{x}\|_2 - \|\mathbf{c}\|_2| \leq \|\mathbf{x} - \mathbf{c}\|_2$ which finishes the proof. □

Corollary 12. *For any $\epsilon > 0, d \in \mathbb{N}$, and any $\mathbf{x}, \mathbf{y}, \mathbf{c}, \mathbf{d} \in \mathbb{R}^d$, we always have $\|\mathfrak{N}_\nu(\mathbf{x}+\mathbf{y}) - \mathfrak{N}_\nu(\mathbf{c}+\mathbf{d})\|_2 \leq \frac{2}{\nu}\left(\|\mathbf{x}-\mathbf{c}\|_2 + \|\mathbf{y}-\mathbf{d}\|_2\right)$.*

Proof. We have, by applied the triangle inequality,

$$\begin{aligned}\|\mathfrak{N}_\nu(\mathbf{x}+\mathbf{y}) - \mathfrak{N}_\nu(\mathbf{c}+\mathbf{d})\|_2 &\leq \|\mathfrak{N}_\nu(\mathbf{x}+\mathbf{y}) - \mathfrak{N}_\nu(\mathbf{x}+\mathbf{d})\|_2 + \|\mathfrak{N}_\nu(\mathbf{x}+\mathbf{d}) - \mathfrak{N}_\nu(\mathbf{c}+\mathbf{d})\|_2 \\ &\leq \frac{2}{\nu}\left(\|\mathbf{y}-\mathbf{d}\|_2 + \|\mathbf{x}-\mathbf{c}\|_2\right),\end{aligned}$$

where in the the last step we applied Claim 11 twice. □

Claim 13. *For $s \in \mathbb{N}, r \geq 1$, let $\mathcal{X} = \left\{\mathbf{x} \in \mathbb{R}^V : \|\mathbf{x}\|_0 \leq s, \|\mathbf{x}\|_2 = 1\right\} \subset \mathbb{R}^V$ denote the set of s-sparse unit-norm vectors, and let $\mathcal{M} := \left\{\mathbf{M} \in \mathbb{R}^{D\times V} : \|\mathbf{M}\|_{1,1} \leq R_1, \|\mathbf{M}\|_{\infty,1} \leq R_\infty\right\}$ be a set of matrices. Then there exists an ϵ-cover $\mathcal{C}_M \subset \mathbb{R}^{D\times V}$ of $\mathcal{M}$ w.r.t $\mathcal{X}$ of size at most*

$$\ln(|\mathcal{C}_M|) \leq 64 s \ln(3e^2 V)\ln(2DV) R_\infty R_1 \cdot \frac{1}{\epsilon^2}$$

Proof. We will establish this result by applying Lemma 16. To do so we need to establish a cover over $\mathcal{X}$ (note that we have $r = \sup_{\mathbf{x}\in\mathcal{X}} \|\mathbf{x}\|_\infty \leq 1$ since $\|\cdot\|_\infty \leq \|\cdot\|_2$. Standard results, for example [4] show that for any fixed support $T \subset [V], |T| = s$, there exists a $\frac{1}{2}$-cover with at most 6^s elements. Taking a union over all possible supports, of which there are $\binom{V}{s} \leq \left(\frac{eV}{s}\right)^s$ in number, tells us that there exists an ϵ-cover $\mathcal{C}_X$ of $\mathcal{X}$ with at most $\left(\frac{eV}{s}\right)^s 6^s$ elements. We also notice that the set $\mathcal{X}$ is indeed "closed" in the sense required by Lemma 16 since for any $\mathbf{x} \in \mathcal{X}$, we can always find a covering element $\mathbf{c}$ with an identical support. This means that $\mathbf{x} - \mathbf{c}$ is s-sparse as well. Since $\mathcal{X}$ contains all s-sparse unit-norm vectors, we are assured that $\frac{\mathbf{x}-\mathbf{c}}{\|\mathbf{x}-\mathbf{c}\|_2} \in \mathcal{X}$. Moreover, we have

$$Q = \sum_{i=1}^{D}\sum_{j=1}^{V} |m_{ij}|\, x_j^2 \leq \sum_{i=1}^{D} \|\mathbf{M}_{i,:}\|_\infty \cdot \sum_{j=1}^{V} x_j^2 = \sum_{i=1}^{D} \|\mathbf{M}_{i,:}\|_\infty \cdot \|\mathbf{x}\|_2^2 = \|\mathbf{M}\|_{\infty,1} \cdot \|\mathbf{x}\|_2^2 \leq R_\infty$$

Applying Lemma 16 and simplifying the expression using $s \geq 1$ then finishes the proof. □

Claim 14. *Suppose we let $\mathcal{X} \subset \mathbb{R}^{D+1}$ denote the set of unit norm vectors i.e. $\mathcal{X} = \left\{\mathbf{x} \in \mathbb{R}^{D+1} : \|\mathbf{x}\|_2 = 1\right\}$ and let $\mathcal{M} := \left\{\mathbf{M} \in \mathbb{R}^{(D+1)\times(D+1)} : \|\mathbf{M}\|_{1,1} \leq R_1, \|\mathbf{M}\|_{\infty,1} \leq R_\infty\right\}$ be a set of matrices. Then there exists an ϵ-cover $\mathcal{C}_M \subset \mathbb{R}^{(D+1)\times(D+1)}$ of $\mathcal{M}$ w.r.t $\mathcal{X}$ of size at most*

$$\ln(|\mathcal{C}_M|) \leq 128(D+1)\ln(2(D+1)^2) R_\infty R_1 \cdot \frac{1}{\epsilon^2}$$

Proof. We will establish this result by applying Lemma 16. We first notice that we have $r = 1$ in this setting since $\|\cdot\|_\infty \leq \|\cdot\|_2$ and we have $\|\mathbf{x}\|_2 \leq 1$. Standard results on covers of the unit sphere using spherical caps, for instance [3, 10], show that a $\frac{1}{2}$ cover $\mathcal{C}_X$ exists for $\mathcal{X}$ with at most $(D+1)\cdot 4^{D+1}$ elements (for $D \geq 3$). The set $\mathcal{X}$ is also "closed" in the sense required by Lemma 16 since $\mathcal{X}$ is the set of unit vectors itself hence any L_2-normalized vector must lie in $\mathcal{X}$. We also have

$$Q = \sum_{i=1}^{D+1}\sum_{j=1}^{D+1} |m_{ij}|\, x_j^2 \leq \sum_{i=1}^{D+1} \|\mathbf{M}_{i,:}\|_\infty \cdot \|\mathbf{x}\|_2^2 = \|\mathbf{M}\|_{\infty,1},$$

where the last step follows since $\|\mathbf{x}\|_2 = 1$. Applying Lemma 16 and simplifying the expression for $D > 3$ then finishes the proof. □

Claim 15. *For some $L \in \mathbb{N}$, let $\mathcal{M} := \left\{\mathbf{M} \in \mathbb{R}^{L\times(D+2)} : \|\mathbf{M}\|_{1,1} \leq R_1, \|\mathbf{M}\|_{1,\infty} \leq R_\infty\right\}$ be a set of matrices. Then there exists a set of matrices $\mathcal{C}_M \subset \mathbb{R}^{L\times(D+2)}$ of size at most*

$$\ln(|\mathcal{C}_M|) \leq 64\ln(3eL)\ln(2(D+2)L) R_\infty R_1 \cdot \frac{1}{\epsilon^2}$$

such that for any $\mathbf{M} \in \mathcal{M}$, there exists a $\mathbf{C} \in \mathcal{C}_M$ such that $\|\mathbf{M} - \mathbf{C}\|_{2,\infty} \leq \epsilon$.

Proof. We will establish this result by applying Lemma 16. Let $\mathcal{X} = \{\mathbf{e}_1, \ldots, \mathbf{e}_L\} \subset \mathbb{R}^L$ denote the set of the canonical L-dimensional vectors i.e. $\mathbf{e}_l = (0, 0, \ldots, 0, 1, 0, \ldots, 0)$ with a 1 at the l^{th} position and 0 everywhere else. We first notice that we have $r = \sup_{\mathbf{x}\in\mathcal{X}} \|\mathbf{x}\|_\infty = 1$ since $\mathcal{X}$ contains only canonical vectors. Establishing a cover over $\mathcal{X}$ is straightforward since $\mathcal{X}$ is finite and admits itself as a trivial 0-cover, giving us a 0-cover of size $|\mathcal{X}| = L$. We now apply Lemma 16 with the set of transposed matrices i.e. to $\mathcal{M}^\top := \left\{\mathbf{M}^\top : \mathbf{M} \in \mathcal{M}\right\}$. It can be verified that the lemma continues to hold here even without the "closedness" condition since we are able to provide a 0-cover. For any $\mathbf{x} = \mathbf{e}_l \in \mathcal{X}$ and any $\mathbf{M}^\top \in \mathcal{M}^\top$, we have

$$Q = \sum_{i=1}^{D+2}\sum_{j=1}^{L} |m_{ji}|\, x_j^2 = \sum_{i=1}^{D+2} |m_{li}| = \|\mathbf{M}_{l,:}\|_1 \leq \|\mathbf{M}\|_{1,\infty}$$

Applying Lemma 16 assures us of a cover $\mathcal{C}_M^\top \subset \mathbb{R}^{(D+2)\times L}$ containing only

$$\ln(|\mathcal{C}_M|) \leq 64 \ln(3eL) \ln(2(D+2)L) R_\infty R_1 \cdot \frac{1}{\epsilon^2}$$

such that for every $\mathbf{M}^\top \in \mathcal{M}^\top$, there exists a $\mathbf{C}^\top \in \mathcal{C}_M^\top$ such that $\sup_{\mathbf{v}\in\mathcal{X}} \left\|(\mathbf{M}^\top - \mathbf{C}^\top)\mathbf{v}\right\|_2 \leq \epsilon$. However, since $\mathcal{X} = \{\mathbf{e}_1, \ldots, \mathbf{e}_L\}$ by construction, this guarantee translates to $\|\mathbf{M} - \mathbf{C}\|_{2,\infty} \leq \epsilon$ as required. □

Lemma 16 (Uniform Maurey-type Sparsification)**.** *For any constants $p, q \in \mathbb{N}$ and $R > 0$, let $\mathcal{M} \subseteq \left\{\mathbf{M} \in \mathbb{R}^{p\times q} : \|\mathbf{M}\|_{1,1} \leq R\right\}$ be a set of $L_{1,1}$-norm bounded matrices. Let $\mathcal{X} \subset \mathbb{R}^q$ be a set of vectors that admits a $\frac{1}{2}$-cover $\mathcal{C}_X \subset \mathcal{X}$ i.e. for every $\mathbf{x} \in \mathcal{X}$, there exists some $\mathbf{c} \in \mathcal{C}_X$ such that $\|\mathbf{x} - \mathbf{c}\|_2 \leq \frac{1}{2}$. To avoid an unnecessary log factor, we will also assume that our set $\mathcal{X}$ is "closed" in a way such the L_2-normalized vector $\frac{\mathbf{x}-\mathbf{c}}{\|\mathbf{x}-\mathbf{c}\|_2} \in \mathcal{X}$ for all $\mathbf{x} \in \mathcal{X}$ and their corresponding covering element $\mathbf{c} \in \mathcal{C}$. This assumption will hold in the application settings of this lemma. Also denote $r := \sup_{\mathbf{x}\in\mathcal{X}} \|\mathbf{x}\|_\infty$ and $Q := \sup_{\mathbf{x}\in\mathcal{X},\mathbf{M}\in\mathcal{M}} \sum_{i=1}^p \sum_{j=1}^q |m_{ij}|\, x_j^2$. Then $\mathcal{M}$ admits an ϵ-cover $\mathcal{C}_M$ w.r.t $\mathcal{X}$ of size at most*

$$\ln(|\mathcal{C}_M|) \leq 16 \ln(3e\,|\mathcal{C}_X|) \ln(2pq) Q R (r+1)^2 \cdot \frac{1}{\epsilon^2}$$

i.e., for any $\mathbf{M} \in \mathcal{M}$, there exists a $\mathbf{C} \in \mathcal{C}_M$ such that $\sup_{\mathbf{x}\in\mathcal{X}} \|\mathbf{M}\mathbf{x} - \mathbf{C}\mathbf{x}\|_2 \leq \epsilon$.

Proof. We will first prove the result for matrices with $\|\mathbf{M}\|_{1,1} = R$, then show that essentially the same result holds even for matrices with $\|\mathbf{M}\|_{1,1} \leq R$. Fix a matrix $\mathbf{M} \in \mathcal{M}$ with $\|\mathbf{M}\|_{1,1} = R$ and consider the following set of *basis* matrices

$$\mathcal{V} = \left\{h \cdot \mathbf{e}_i \mathbf{e}_j^\top, i \in [p], j \in [q], h \in \pm 1\right\}$$

Note that $|\mathcal{V}| = 2pq$ and $\|\mathbf{V}\|_{1,1} \leq 1$ for all $\mathbf{V} \in \mathcal{V}$. Let us set up a distribution over $\mathcal{V}$ as (with m_{ij} denoting the i,j-th entry of the matrix $\mathbf{M}$)

$$\mathbb{P}\left[V = \operatorname{sign}(m_{ij}) \cdot \mathbf{e}_i \mathbf{e}_j^\top\right] = \frac{|m_{ij}|}{R}$$
$$\mathbb{P}\left[V = -\operatorname{sign}(m_{ij}) \cdot \mathbf{e}_i \mathbf{e}_j^\top\right] = 0$$

We now draw T samples $\mathbf{V}_t = \operatorname{sign}(m_{ij}) \cdot \mathbf{e}_{i_t} \mathbf{e}_{j_t}^\top \in \mathbb{R}^{p\times q}, t = 1, \ldots, T$ from this distribution. Consider a fixed vector $\mathbf{x} \in \mathcal{X}$ and notice that for all $t \in [T]$, we have $\mathbb{E}\left[\mathbf{V}_t\mathbf{x}\right] = \sum_{i,j} \frac{|m_{ij}|}{R} \cdot \operatorname{sign}(m_{ij}) \cdot x_j \cdot \mathbf{e}_i = \frac{1}{R} \cdot \mathbf{M}\mathbf{x}$. Let us use the shorthand $\mathbf{u}_t := \mathbf{V}_t\mathbf{x} - \frac{1}{R} \cdot \mathbf{M}\mathbf{x} \in \mathbb{R}^p$ and immediately conclude that $\mathbb{E}\left[\mathbf{u}_t\right] = \mathbf{0}$. We also have

$$\mathbb{E}\left[\|\mathbf{u}_t\|_2^2\right] \leq \mathbb{E}\left[\|\mathbf{V}_t\mathbf{x}\|_2^2\right] = \mathbb{E}\left[x_{j_t}^2\right] = \sum_{i=1}^p \sum_{j=1}^q \frac{|m_{ij}|}{R} x_j^2 = \frac{Q}{R},$$

as well as that

$$\|\mathbf{u}_t\|_2 \leq \|\mathbf{V}_t\mathbf{x}\|_2 + \frac{1}{R} \cdot \|\mathbf{M}\mathbf{x}\|_2 \leq 2r$$

almost surely since $\|\mathbf{Mx}\|_2 \leq \|\mathbf{M}\|_{1,1} \cdot \|\mathbf{x}\|_\infty$ and $\|\mathbf{x}\|_\infty \leq r$. This gives us, for $B = \sqrt{\frac{QT}{R}}, H = 2r$, the following result for $m = 2, 3, \ldots$

$$\sum_{t=1}^{T} \mathbb{E}\left[\|\mathbf{u}_t\|_2^m\right] \leq \sum_{t=1}^{T} \mathbb{E}\left[\|\mathbf{u}_t\|_2^2\right] (2r)^{m-2} \leq B^2 H^{m-2} \leq \frac{m!}{2} \cdot B^2 H^{m-2}$$

Then Bernstein-style bounds on Hilbert spaces, such as [8, Corollary 1], tell us that

$$\mathbb{P}\left[\left\|\frac{1}{T}\sum_{t=1}^{T} \mathbf{u}_t\right\|_2 \geq \eta\right] \leq 2 \cdot \exp\left(-\frac{T^2\eta^2}{2B^2 + 2H\eta T}\right) = 2 \cdot \exp\left(-\frac{T\eta^2 R}{2Q + 4rR\eta}\right)$$

An alternate way of writing this result is to construct a matrix $\hat{\mathbf{M}}_T := \frac{R}{T}\sum_{t=1}^{T} \mathbf{V}_t$ and see that

$$\mathbb{P}\left[\left\|\hat{\mathbf{M}}_T\mathbf{x} - \mathbf{Mx}\right\|_2 \geq R\eta\right] \leq 2 \cdot \exp\left(-\frac{T\eta^2 R}{2Q + 4rR\eta}\right) \tag{1}$$

Denote $\mathbf{\Delta}_T := \hat{\mathbf{M}}_T - \mathbf{M} \in \mathbb{R}^{p\times q}$ to simplify the notation. We wish to bound $\sup_{\mathbf{x}\in\mathcal{X}} \|\mathbf{\Delta}_T\mathbf{x}\|_2$. Let $\mathbf{y}$ be a vector that achieves this limit i.e. $\|\mathbf{\Delta}_T\mathbf{y}\|_2 = \sup_{\mathbf{x}\in\mathcal{X}} \|\mathbf{\Delta}_T\mathbf{x}\|_2$ and let $\mathbf{c} \in \mathcal{C}_X$ be its covering element. Note that this assures us that $\|\mathbf{y} - \mathbf{c}\|_2 \leq \frac{1}{2}$ since $\mathcal{C}$ is a $\frac{1}{2}$-cover for $\mathcal{X}$, and also that $\frac{\mathbf{y}-\mathbf{c}}{\|\mathbf{y}-\mathbf{c}\|_2} \in \mathcal{X}$ by the closure assumption. Then we have, by applying the triangle inequality,

$$\begin{aligned}
\|\mathbf{\Delta}_T\mathbf{c}\|_2 &\geq \|\mathbf{\Delta}_T\mathbf{y}\|_2 - \|\mathbf{\Delta}_T(\mathbf{y} - \mathbf{c})\|_2 \\
&= \sup_{\mathbf{x}\in\mathcal{X}} \|\mathbf{\Delta}_T\mathbf{x}\|_2 - \|\mathbf{y} - \mathbf{c}\|_2 \cdot \left\|\mathbf{\Delta}_T \cdot \frac{\mathbf{y} - \mathbf{c}}{\|\mathbf{y} - \mathbf{c}\|_2}\right\| \\
&\geq \sup_{\mathbf{x}\in\mathcal{X}} \|\mathbf{\Delta}_T\mathbf{x}\|_2 - \frac{1}{2} \cdot \sup_{\mathbf{x}\in\mathcal{X}} \|\mathbf{\Delta}_T\mathbf{x}\|_2 .
\end{aligned}$$

This tells us that $\sup_{\mathbf{x}\in\mathcal{X}} \|\mathbf{\Delta}_T\mathbf{x}\|_2 \leq 2 \cdot \sup_{\mathbf{c}\in\mathcal{C}_X} \|\mathbf{\Delta}_T\mathbf{c}\|_2$. Applying the result from equation (1) with a union bound over all cover elements $\mathbf{c} \in \mathcal{C}_X$ grants us the following result

$$\mathbb{P}\left[\exists \mathbf{c} \in \mathcal{C}_X : \|\mathbf{\Delta}_T\mathbf{c}\|_2 \geq R\eta\right] \leq 2\left|\mathcal{C}_X\right| \cdot \exp\left(-\frac{T\eta^2 R}{2Q + 4rR\eta}\right)$$

which in turn tells us that

$$\mathbb{P}\left[\sup_{\mathbf{x}\in\mathcal{X}} \left\|\hat{\mathbf{M}}_T\mathbf{x} - \mathbf{Mx}\right\|_2 \geq 2R\eta\right] \leq 2\left|\mathcal{C}_X\right| \cdot \exp\left(-\frac{T\eta^2 R}{2Q + 4rR\eta}\right)$$

Now we apply the probabilistic method to argue that so long as the right hand side is less than unity, there must always exist a matrix $\hat{\mathbf{M}}_T = \frac{R}{T}\sum_{t=1}^{T} \mathbf{V}_t$ that satisfies $\sup_{\mathbf{x}\in\mathcal{X}} \left\|\mathbf{Mx} - \hat{\mathbf{M}}_T\mathbf{x}\right\|_2 \leq 2R\eta$. Note that $\left\|\hat{\mathbf{M}}_T\right\|_{1,1} \leq R$ which implies that $\hat{\mathbf{M}}_T \in \mathcal{M}$. Letting the right hand side equal $\frac{1}{e} < 1$ and setting $\eta = \frac{\epsilon}{2R(r+1)}$ shows us that having $T \geq \ln(2e\left|\mathcal{C}_X\right|)\left(\frac{8QR(r+1)^2}{\epsilon^2} + \frac{8Rr(r+1)}{\epsilon}\right)$ guarantees that

$$\mathbb{P}\left[\sup_{\mathbf{x}\in\mathcal{X}} \left\|\hat{\mathbf{M}}_T\mathbf{x} - \mathbf{Mx}\right\|_2 \geq \frac{\epsilon}{r+1}\right] \leq \frac{1}{e}$$

To avoid notational clutter, we will use a stricter bound $T \geq \ln(2e\left|\mathcal{C}_X\right|)\left(\frac{16(r+1)^2QR}{\epsilon^2}\right)$. We now note that the preceding argument shows that if T is large enough as established above, then for any $\mathbf{M} \in \mathcal{M}$ such that

$\|\mathbf{M}\|_{1,1} = R$, a covering element $\hat{\mathbf{M}}_T$ can always be constructed of the form $\frac{R}{T}\sum_{t=1}^{T}\mathbf{V}_t$ where each $\mathbf{V}_t \in \mathcal{V}$. This leads us to define the following cover

$$\mathcal{C}_M^R := \left\{ \frac{R}{T}\sum_{t=1}^{T}\mathbf{V}_t : \mathbf{V}_t \in \mathcal{V} \right\}$$

and note that $\mathcal{C}_M^R$ is indeed an $\frac{\epsilon}{r+1}$-cover for all matrices whose norm satisfies $\|\mathbf{M}\|_{1,1} = R$. We also note that by construction, we have $\left|\mathcal{C}_M^R\right| \leq |\mathcal{V}|^T = (2pq)^{\frac{cR}{\epsilon^2}}$ where $c = 16\ln(2e\,|\mathcal{C}_X|)Q(r+1)^2$.

To cover matrices of smaller norms, we now establish such covers for matrices of various norms. In particular consider the following set set of positive real numbers $\mathcal{R} = \left\{\frac{\epsilon}{r+1}, \frac{2\epsilon}{r+1}, \frac{3\epsilon}{r+1}, \ldots, R\right\}$. For each value $g \in \mathcal{R}$, using the same argument as above, we can construct an $\frac{\epsilon}{r+1}$-cover $\mathcal{C}_M^g$ for all matrices with norm $\|\mathbf{M}\|_{1,1} = g$. The size of this cover will be at most $|\mathcal{C}_M^g| \leq (2pq)^{\frac{cg}{\epsilon^2}}$, again by the above argument.

Our final ϵ-cover for $\mathcal{M}$ is defined as the union of all these $|\mathcal{R}| = \frac{R(r+1)}{\epsilon}$ covers i.e.

$$\mathcal{C}_M := \bigcup_{t=1}^{\frac{R(r+1)}{\epsilon}} \mathcal{C}_M^{\frac{t\epsilon}{r+1}}.$$

We first show that $\mathcal{C}_M$ is indeed an ϵ-cover for $\mathcal{M}$. Given a matrix $\mathbf{M} \in \mathcal{M}$, we first find the number $g \in \mathcal{R}$ closest to $\|\mathbf{M}\|_{1,1}$. Clearly, by the construction of $\mathcal{R}$, we are assured that $\left|g - \|\mathbf{M}\|_{1,1}\right| \leq \frac{\epsilon}{r+1}$. Given this, we first rescale the matrix to get $\tilde{\mathbf{M}} := \frac{g}{\|\mathbf{M}\|_{1,1}} \cdot \mathbf{M}$ and obtain the covering element, say $\mathbf{C} \in \mathcal{C}_M^g$, for this rescaled matrix. Note that by construction of $\mathcal{C}_M^g$, we are assured that $\sup_{\mathbf{x}\in\mathcal{X}} \left\|\tilde{\mathbf{M}}\mathbf{x} - \mathbf{C}\mathbf{x}\right\|_2 \leq \frac{\epsilon}{r+1}$. This allows us to show, for any $\mathbf{x} \in \mathcal{X}$,

$$\begin{aligned}
\|M\mathbf{x} - \mathbf{C}\mathbf{x}\|_2 &\leq \left\|\mathbf{M}\mathbf{x} - \tilde{\mathbf{M}}\mathbf{x}\right\|_2 + \left\|\tilde{\mathbf{M}}\mathbf{x} - \mathbf{C}\mathbf{x}\right\|_2 \leq \left|1 - \frac{g}{\|\mathbf{M}\|_{1,1}}\right| \cdot \|\mathbf{M}\mathbf{x}\|_2 + \frac{\epsilon}{r+1} \\
&\leq \left|1 - \frac{g}{\|\mathbf{M}\|_{1,1}}\right| \cdot \|\mathbf{M}\|_{1,1} \cdot \|\mathbf{x}\|_\infty + \frac{\epsilon}{r+1} \\
&\leq \frac{r\epsilon}{r+1} + \epsilon = \epsilon,
\end{aligned}$$

where in the last step, we used the fact that $\left|g - \|\mathbf{M}\|_{1,1}\right| \leq \frac{\epsilon}{r+1}$. This establishes that $\mathcal{C}_M$ is indeed an ϵ-cover for all matrices in $\mathcal{M}$. We now bound the number of elements in $\mathcal{C}_M$. We note that

$$|\mathcal{C}_M| \leq \sum_{t=1}^{\frac{R(r+1)}{\epsilon}} (2pq)^{\frac{c}{\epsilon^2}\left(\frac{t\epsilon}{r+1}\right)} \leq 2 \cdot (2pq)^{\frac{cR}{\epsilon^2}},$$

where the last step simplifies the bound by using $2pq \geq 2$ as $p, q \in \mathbb{N}$. Thus, $\ln(|\mathcal{C}_M|) \leq \ln 2 + \frac{cR}{\epsilon^2}\ln(2pq) \leq 16\ln(3e\,|\mathcal{C}_X|)\ln(2pq)QR(r+1)^2 \cdot \frac{1}{\epsilon^2}$ which completes the proof. As a concluding note, we observe that such an argument covering matrices of various norms is required in the proof of [1, Lemma 3.2] as well. Fortunately, although that paper omits it, the above argument seems to be applicable there as well. □

H Surrogacy Results

Claim 17. *For any $\gamma > 0$, we have $\ell_{\gamma,k}^{prec}(\mathbf{s}, \mathbf{y}) \geq \wp_k(\mathbf{s}, \mathbf{y})$.*

Proof. Note that $r_\gamma(v) \leq \mathbb{I}\{v > 0\}$ for all values of $\gamma > 0$. Thus, we have

$$\frac{1}{k}\sum_{t=1}^{\min\{k,|P_\mathbf{y}|\}} r_\gamma\left(s_{\pi_\mathbf{s}^+(t)} - \max_{l' \in N_\mathbf{y}} \; s_{l'}\right) \leq \frac{1}{k}\sum_{t=1}^{\min\{k,|P_\mathbf{y}|\}} \mathbb{I}\left\{s_{\pi_\mathbf{s}^+(t)} > \max_{l' \in N_\mathbf{y}} \; s_{l'}\right\}.$$

Now, by definition, we have $\mathbb{I}\left\{s_{\pi_\mathbf{s}^+(t)} > \max_{l' \in N_\mathbf{y}} \; s_{l'}\right\} \leq \mathbb{I}\{\pi_\mathbf{s}(t) \in P_\mathbf{y}\}$ for all $t \leq |P_\mathbf{y}|$ which gives us

$$\frac{1}{k}\sum_{t=1}^{\min\{k,|P_\mathbf{y}|\}} r_\gamma\left(s_{\pi_\mathbf{s}^+(t)} - \max_{l' \in N_\mathbf{y}} \; s_{l'}\right) \leq \frac{1}{k}\sum_{t=1}^{\min\{k,|P_\mathbf{y}|\}} \mathbb{I}\{\pi_\mathbf{s}(t) \in P_\mathbf{y}\} \leq \frac{1}{k}\sum_{t=1}^{k} \mathbb{I}\{\pi_\mathbf{s}(t) \in P_\mathbf{y}\},$$

where the last step follows since the indicator function takes only non-negative values. Subtracting both sides from unity gives us the claimed result. □

Claim 18. *For any $c \in (0,1), d \geq 1$, we have $\frac{1}{\ln(4c)} \cdot \ell^{cont}_{c,d}(\mathbf{s},\mathbf{y}) \geq \ell^{prec}_{\gamma,k}(\mathbf{s},\mathbf{y})$. In particular, if $c \in (0.7,1), d \geq 1$, we have $\ln(4c) > 1$ and thus $\ell^{cont}_{c,d}(\mathbf{s},\mathbf{y}) \geq \ell^{prec}_{\gamma,k}(\mathbf{s},\mathbf{y})$.*

Proof. For this proof, we need to introduce an intermediate surrogate loss function that we define below for any $\lambda_+, \lambda_- \in \mathbb{R}$ (not necessarily positive),

$$\ell^{\text{hinge}}_{\lambda_+,\lambda_-}(\mathbf{s},\mathbf{y}) = \sum_{l \in P_\mathbf{y}} \max\{\lambda_+ - s_l, 0\} + \sum_{l' \in N_\mathbf{y}} \max\{s_{l'} - \lambda_-, 0\}$$

Using the statement of Claim 21 with $\gamma = \frac{\ln(4c)}{d}$ and dividing throughout by $d\gamma$ tells us that we have

$$\frac{1}{d\gamma} \cdot \ell^{\text{cont}}_{c,d}(\mathbf{s},\mathbf{y}) \geq \frac{1}{\gamma} \cdot \ell^{\text{hinge}}_{1,1-\gamma}(\mathbf{s},\mathbf{y})$$

Using Claim 20 then finishes the proof. □

Corollary 19. *Since Claim 17 holds for all values of $\gamma > 0$, we have $\frac{1}{\ln(4c)} \cdot \ell^{cont}_{c,d}(\mathbf{s},\mathbf{y}) \geq \wp_k(\mathbf{s},\mathbf{y})$ for all values of $c \in (0,1), d \geq 1$ and, in particular, $\ell^{cont}_{c,d}(\mathbf{s},\mathbf{y}) \geq \wp_k(\mathbf{s},\mathbf{y})$ for values of $c \in (0.7,1), d \geq 1$.*

For the rest of the discussion, we will assume that $k \leq |P_\mathbf{y}|$. This to avoid a situation where there are less positive labels than positions at which precision is being measured which put an artificial lower bound on the prec@k loss function value.

Claim 20. *Whenever $\lambda_+ > \lambda_-$ and , we have $\frac{1}{\lambda_+ - \lambda_-} \cdot \ell^{hinge}_{\lambda_+,\lambda_-}(\mathbf{s},\mathbf{y}) \geq \ell^{prec}_{(\lambda_+ - \lambda_-),k}(\mathbf{s},\mathbf{y})$.*

Proof. Let h denote a negative label that achieves the maximum hinge loss among all negative labels i.e.

$$\max_{l' \in N_\mathbf{y}} \; \{\max\{s_{l'} - \lambda_-, 0\}\} =: \max\{s_h - \lambda_-, 0\}.$$

Now consider a single positive label $l \in P_\mathbf{y}$ and the following cases

1. Case 1: $s_l - s_h > \lambda_+ - \lambda_-$. In this case, $r_{(\lambda_+ - \lambda_-)}(s_l - s_h) = 1$ and we use the fact that the hinge loss is a non-negative function to get

$$\max\{\lambda_+ - s_l, 0\} + \max\{s_h - \lambda_-, 0\} \geq 0 = (\lambda_+ - \lambda_-)\left(1 - r_{(\lambda_+ - \lambda_-)}(s_l - s_h)\right)$$

2. Case 2: $s_l - s_h \in (0, \lambda_+ - \lambda_-)$. In this case we use the fact that $\max\{x, 0\} \geq x$ to get

$$\begin{aligned}\max\{\lambda_+ - s_l, 0\} + \max\{s_h - \lambda_-, 0\} &\geq (\lambda_+ - \lambda_-) - (s_l - s_h) \\ &= (\lambda_+ - \lambda_-)\left(1 - \frac{s_l - s_h}{\lambda_+ - \lambda_-}\right) \\ &= (\lambda_+ - \lambda_-)\left(1 - r_{(\lambda_+ - \lambda_-)}(s_l - s_h)\right)\end{aligned}$$

3. Case 3: $s_l - s_h \leq 0$. In this case $r_{(\lambda_+ - \lambda_-)}(s_l - s_h) = 0$ and we similarly use $\max\{x, 0\} \geq x$ to get

$$\begin{aligned}\max\{\lambda_+ - s_l, 0\} + \max\{s_h - \lambda_-, 0\} &\geq (\lambda_+ - \lambda_-) - (s_l - s_h) \\ &\geq (\lambda_+ - \lambda_-) \\ &= (\lambda_+ - \lambda_-)\left(1 - r_{(\lambda_+ - \lambda_-)}(s_l - s_h)\right)\end{aligned}$$

Notice we make two observations

1. Since the hinge loss takes only non-negatively values, we have

$$\sum_{l' \in N_{\mathbf{y}}} \max\{s_{l'} - \lambda_-, 0\} \geq \max_{l' \in N_{\mathbf{y}}} \{\max\{s_{l'} - \lambda_-, 0\}\} = \max\{s_h - \lambda_-, 0\},$$

2. Since $\max\{v - \lambda_-, 0\}$ is an increasing function, we have $s_h = \max_{l' \in N_{\mathbf{y}}} s_{l'}$.

Taking the cases analyzed above along with the above two observations gives us

$$\frac{1}{\lambda_+ - \lambda_-}\left(\max\{\lambda_+ - s_l, 0\} + \sum_{l' \in N_{\mathbf{y}}} \max\{s_{l'} - \lambda_-, 0\}\right) \geq 1 - r_{(\lambda_+ - \lambda_-)}\left(s_l - \max_{l' \in N_{\mathbf{y}}} s_{l'}\right).$$

Rearranging, summing over the top positives, and dividing by k gives us

$$\begin{aligned}&\frac{1}{k}\sum_{t=1}^{\min\{k, |P_{\mathbf{y}}|\}} r_{(\lambda_+ - \lambda_-)}\left(s_{\pi^+_{\mathbf{s}}(t)} - \max_{l' \in N_{\mathbf{y}}} s_{l'}\right) \\ &\geq \frac{\min\{k, |P_{\mathbf{y}}|\}}{k} - \frac{1}{\lambda_+ - \lambda_-}\left(\frac{1}{k}\sum_{t=1}^{\min\{k, |P_{\mathbf{y}}|\}} \max\left\{\lambda_+ - s_{\pi^+_{\mathbf{s}}(t)}, 0\right\} + \frac{\min\{k, |P_{\mathbf{y}}|\}}{k} \cdot \sum_{l' \in N_{\mathbf{y}}} \max\{s_{l'} - \lambda_-, 0\}\right)\end{aligned}$$

Subtracting both sides from unity gives us, for $k \leq |P_{\mathbf{y}}|$,

$$\begin{aligned}&\ell^{\text{prec}}_{(\lambda_+ - \lambda_-), k}(\mathbf{s}, \mathbf{y}) \\ &\leq 1 - \frac{\min\{k, |P_{\mathbf{y}}|\}}{k} + \frac{1}{\lambda_+ - \lambda_-}\left(\frac{1}{k}\sum_{t=1}^{\min\{k, |P_{\mathbf{y}}|\}} \max\left\{\lambda_+ - s_{\pi^+_{\mathbf{s}}(t)}, 0\right\} + \frac{\min\{k, |P_{\mathbf{y}}|\}}{k} \cdot \sum_{l' \in N_{\mathbf{y}}} \max\{s_{l'} - \lambda_-, 0\}\right) \\ &\leq \frac{1}{\lambda_+ - \lambda_-}\left(\frac{1}{k}\sum_{t=1}^{\min\{k, |P_{\mathbf{y}}|\}} \max\left\{\lambda_+ - s_{\pi^+_{\mathbf{s}}(t)}, 0\right\} + \sum_{l' \in N_{\mathbf{y}}} \max\{s_{l'} - \lambda_-, 0\}\right) \\ &\leq \frac{1}{\lambda_+ - \lambda_-}\left(\sum_{l \in P_{\mathbf{y}}} \max\{\lambda_+ - s_l, 0\} + \sum_{l' \in N_{\mathbf{y}}} \max\{s_{l'} - \lambda_-, 0\}\right) = \frac{1}{\lambda_+ - \lambda_-} \cdot \ell^{\text{hinge}}_{\lambda_+, \lambda_-}(\mathbf{s}, \mathbf{y}),\end{aligned}$$

which finishes the proof. □

Claim 21. *For all $c \in (0, 1), d \geq 1$, we have $\ell^{cont}_{c,d}(\mathbf{s}, \mathbf{y}) \geq d \cdot \ell^{hinge}_{1, 1-\gamma}(\mathbf{s}, \mathbf{y})$, where $\gamma = \frac{\ln(4c)}{d}$.*

Proof. We can safely ignore the $\ln \frac{1}{c}$ terms in expression for $\ell^{\text{cont}}_{c,d}(\mathbf{s}, \mathbf{y})$ since they are always non-negative as $c \in (0, 1)$. Since since $d \geq 1$ and scores are contained in the interval $[-1, 1]$, using $\lambda_+ = 1$ gives us,

$$\sum_{l \in P_{\mathbf{y}}} d \cdot (1 - s_l) \geq \sum_{l \in P_{\mathbf{y}}} d \cdot \max\{\lambda_+ - s_l, 0\}$$

Now consider the negative portion of the contrastive loss i.e. $f(x) = -\ln(1 - c \cdot \exp(d \cdot (x-1)))$. This function is defined in the interval $\mathcal{V} := (-\infty, 1 + \frac{1}{d}\ln\frac{1}{c})$ and is a strictly convex function on that interval. Note that $1 + \frac{1}{d}\ln\frac{1}{c} > 1$ for any $c \in (0,1), d \geq 1$ and thus $\mathcal{V} \supset [-1,1]$ i.e. $f(\cdot)$ is well-defined for all scores in the interval $[-1,1]$. However, the convexity of $f(\cdot)$ tells us that for any $x_0 \in \mathcal{V}$, we must have for all $x \in \mathcal{V}$

$$f(x) \geq f(x_0) + f'(x_0) \cdot (x - x_0)$$

However, for any $c \in (0,1), d \geq 1$, since $f(\cdot)$ takes only non-negative values i.e. $f(x) > 0$ for all $x \in \mathcal{V}$, we additionally have, for all $x \in \mathcal{V}$,

$$f(x) \geq \max\{f(x_0) + f'(x_0) \cdot (x - x_0), 0\}$$

We choose x_0 to be the point where $f'(x_0) = d$ (this point turns out to be unique since $f(\cdot)$ is strictly convex). Thus we choose $x_0 = 1 - \frac{\ln(2c)}{d}$ where we have $f(x_0) = \ln(2)$. Note that for any $c \in (0,1), d \geq 1$, we always have $x_0 \in \mathcal{V}$ i.e. $f(\cdot)$ is well-defined on x_0. This gives us

$$f(x) \geq \max\left\{\ln(2) + d\left(x - \left(1 - \frac{\ln(2c)}{d}\right)\right), 0\right\} = d \cdot \max\left\{x - \left(1 - \frac{\ln(4c)}{d}\right), 0\right\}$$

Using a value of $\lambda_- = 1 - \gamma$ for the hinge loss where $\gamma = \frac{\ln(4c)}{d}$ finishes the proof. □

References

[1] Peter L Bartlett, Dylan J Foster, and Matus Telgarsky. Spectrally-normalized margin bounds for neural networks. In *NIPS*, 2017.

[2] K. Bhatia, K. Dahiya, H. Jain, A. Mittal, Y. Prabhu, and M. Varma. The extreme classification repository: Multi-label datasets & code, 2016.

[3] Károly Böröczky and Gergely Wintsche. Covering the Sphere by Equal Spherical Balls. In *Algorithms and Combinatorics*, volume 25, pages 235–251. Springer Berlin Heidelberg, 2003.

[4] Felipe Cucker and Steve Smale. On the mathematical foundations of learning. *Bulletin of the American Mathematical Society*, 39(01):1–50, October 2001.

[5] H. Jain, Y. Prabhu, and M. Varma. Extreme multi-label loss functions for recommendation, tagging, ranking and other missing label applications. In *KDD*, August 2016.

[6] A. Y. Malkov and D. A. Yashunin. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *CoRR*, 2016.

[7] Mehryar Mohri and Afshin Rostamizadeh Ameet Talwalkar. *Foundations of Machine Learning*. MIT Press Ltd, 2018.

[8] I. F. Pinelis and A. I. Sakhanenko. Remarks on Inequalities for Large Deviation Probabilities. *Theory of Probability & Its Applications*, 30(1):143–148, March 1986.

[9] Colin Wei and Tengyu Ma. Data-dependent Sample Complexity of Deep Neural Networks via Lipschitz Augmentation. In *NeurIPS*, 2019.

[10] A. D. Wyner. Random Packings and Coverings of the Unit n-Sphere. *Bell System Technical Journal*, 46(9):2111–2118, November 1967.

[11] Tong Zhang. Statistical Analysis of Some Multi-Category Large Margin Classification Methods. *Journal of Machine Learning Research*, 5:1225–1251, October 2004.