A. Deferred Proofs

A.1. Class Collision Lemma

We prove a general Lemma, from which Lemma 4.4 can be derived directly.

Lemma A.1. *Let $c \in \mathcal{C}$ and $\ell : \mathbb{R}^t \to \mathbb{R}$ be either the t-way hinge loss or t-way logistic loss, as defined in Section 2. Let $x, x^+, x_1^-, ..., x_t^-$ be iid draws from $\mathcal{D}_c$. For all $f \in \mathcal{F}$, let*

$$L_{un,c}^{=}(f) = \mathop{\mathbb{E}}_{x,x^+,x_i^-}\left[\ell\left(\left\{f(x)^T\left(f(x^+) - f(x_i^-)\right)\right\}_{i=1}^{t}\right)\right]$$

Then

$$L_{un,c}^{=}(f) - \ell(\vec{0}) \leq c' t \sqrt{\|\Sigma(f,c)\|_2} \mathop{\mathbb{E}}_{x\sim\mathcal{D}_c}[\|f(x)\|] \tag{13}$$

where c' is a positive constant.

Lemma 4.4 is a direct consequence of the above Lemma, by setting $t = 1$ (which makes $\ell(0) = 1$), taking an expectation over $c \sim \nu$ in Equation (13) and noting that $\mathbb{E}_{c\sim\nu}[L_{un,c}^{=}(f)] = L_{un}^{=}(f)$.

Proof of Lemma A.1. Fix an $f \in \mathcal{F}$ and let $z_i = f(x)^T\left(f(x_i^-) - f(x^+)\right)$ and $z = \max_{i\in[t]} z_i$. First, we show that $L_{un,c}^{=}(f) - \ell(\vec{0}) \leq c'\mathbb{E}[|z|]$, for some constant c'. Note that $\mathbb{E}[|z|] = \mathbb{P}[z \geq 0]\mathbb{E}[z|z \geq 0] + \mathbb{P}[z \leq 0]\mathbb{E}[-z|z \leq 0] \geq \mathbb{P}[z \geq 0]\mathbb{E}[z|z \geq 0]$.

t**-way hinge loss**: By definition $\ell(\boldsymbol{v}) = \max\{0, 1+\max_{i\in[t]}\{-\boldsymbol{v}_i\}\}$. Here, $L_{un,c}^{=}(f) = \mathbb{E}[(1+z)_+] \leq \mathbb{E}[\max\{1+z, 1\}] = 1 + \mathbb{P}[z \geq 0]\mathbb{E}[z|z \geq 0] \leq 1 + \mathbb{E}[|z|]$.

t**-way logistic loss**: By definition $\ell(\boldsymbol{v}) = \log_2(1 + \sum_{i=1}^{t} e^{-\boldsymbol{v}_i})$, we have $L_{un,c}^{=}(f) = \mathbb{E}[\log_2(1 + \sum_i e^{z_i})] \leq \mathbb{E}[\log_2(1 + te^z)] \leq \max\{\frac{z}{\log 2} + \log_2(1+t), \log_2(1+t)\} = \frac{\mathbb{P}[z\geq 0]\mathbb{E}[z|z\geq 0]}{\log 2} + \log_2(1+t) \leq \frac{\mathbb{E}[|z|]}{\log 2} + \log_2(1+t)$.

Finally, $\mathbb{E}[|z|] \leq \mathbb{E}[\max_{i\in[t]} |z_i|] \leq t\mathbb{E}[|z_1|]$. But,

$$\begin{aligned}\mathbb{E}[|z_1|] &= \mathbb{E}_{x,x^+,x_1^-}\left[\,\left|f(x)^T\left(f(x_1^-) - f(x^+)\right)\right|\,\right] \\ &\leq \mathbb{E}_x\left[\|f(x)\|\sqrt{\mathbb{E}_{x^+,x_1^-}\left[\left(\frac{f(x)^T}{\|f(x)\|}\left(f(x_1^-) - f(x^+)\right)\right)^2\right]}\right] \leq \sqrt{2}\,\sqrt{\|\Sigma(f,c)\|_2}\mathop{\mathbb{E}}_{x\sim\mathcal{D}_c}[\|f(x)\|]\end{aligned}$$

□

A.2. Proof of Lemma 5.1

Fix an $f \in \mathcal{F}$ and suppose that within each class c, f is σ^2-subgaussian in every direction. [7] Let $\mu_c = \mathop{\mathbb{E}}_{x\sim\mathcal{D}_c}[f(x)]$. This means that for all $c \in \mathcal{C}$ and unit vectors v, for $x \sim D_c$, we have that $v^T(f(x) - \mu_c)$ is σ^2-subgaussian. Let $\epsilon > 0$ and $\gamma = 1 + 2R\sigma\sqrt{2\log R + \log 3/\epsilon}$. [8] Consider fixed c^+, c^-, x and let $f(x)^T(f(x^-) - f(x^+)) = \mu + z$, where

$$\mu = f(x)^T(\mu_{c^-} - \mu_{c^+}) \qquad \text{and} \qquad z = f(x)^T\left(f(x^-) - \mu_{c^-}\right) - f(x)^T\left(f(x^+) - \mu_{c^+}\right)$$

For $x^+ \sim \mathcal{D}_c^+$, $x^- \sim \mathcal{D}_c^-$ independently, z is the sum of two independent $R^2\sigma^2$-subgaussians (x is fixed), so z is $2R^2\sigma^2$-subgaussian and thus $p = \Pr[z \geq \gamma - 1] \leq e^{-\frac{4R^2\sigma^2(2\log R + \log 3/\epsilon)}{4R^2\sigma^2}} = \frac{\epsilon}{3R^2}$. So, $\mathbb{E}_z[(1 + \mu + z)_+] \leq (1-p)(\gamma + \mu)_+ + p(2R^2 + 1) \leq \gamma(1 + \frac{\mu}{\gamma})_+ + \epsilon$ (where we used that $\mu + z \leq 2R^2$). By taking expectation over $c^+, c^- \sim \rho^2$, $x \sim \mathcal{D}_{c^+}$ we

[7] A random variable X is called σ^2-subgaussian if $\mathbb{E}[e^{\lambda(X-\mathbb{E}[X])}] \leq e^{\lambda^2\sigma^2/2}$, $\forall\lambda \in \mathbb{R}$. A random vector $V \in \mathbb{R}^d$ is σ^2-subgaussian in every direction, if $\forall u \in \mathbb{R}^d$, $||u|| = 1$, the random variable $\langle u, V\rangle$ is σ^2-subgaussian.

[8] We implicitly assume here that $R \geq 1$, but for $R < 1$, we just set $\gamma = 1 + 2R\sigma\sqrt{\log 3/\epsilon}$ and the same argument holds.

have

$$\begin{aligned}
L^{\neq}_{un}(f) &\leq \mathop{\mathbb{E}}_{\substack{c^+,c^-\sim\rho^2\\ x\sim\mathcal{D}_{c^+}}}\left[\gamma\left(1+\frac{f(x)^T(\mu_{c^-}-\mu_{c^+})}{\gamma}\right)_+ \middle| c^+\neq c^-\right]+\epsilon\\
&= \gamma \mathop{\mathbb{E}}_{c^+,c^-\sim\rho^2}\left[\frac{1}{2}\mathop{\mathbb{E}}_{x\sim\mathcal{D}_{c^+}}\left[\left(1+\frac{f(x)^T(\mu_{c^-}-\mu_{c^+})}{\gamma}\right)_+\right]+\frac{1}{2}\mathop{\mathbb{E}}_{x\sim\mathcal{D}_{c^-}}\left[\left(1+\frac{f(x)^T(\mu_{c^+}-\mu_{c^-})}{\gamma}\right)_+\right]\middle| c^+\neq c^-\right]+\epsilon\\
&= \gamma \mathop{\mathbb{E}}_{c^+,c^-\sim\rho^2}\left[L^{\mu}_{\gamma,sup}(\{c^+,c^-\},f)\middle| c^+\neq c^-\right]+\epsilon
\end{aligned} \tag{14}$$

where $L^{\mu}_{\gamma,sup}(\{c^+,c^-\},f)$ is $L^{\mu}_{sup}(\{c^+,c^-\},f)$ when $\ell_\gamma(x)=(1-x/\gamma)_+$ is the loss function. Observe that in 14 we used that $\mathcal{D}_\mathcal{T}$ are uniform for binary $\mathcal{T}$, which is an assumption we work with in section 4, but we remove it in section 5. The proof finishes by observing that the last line in 14 is equal to $\gamma L^{\mu}_{\gamma,sup}(f)+\epsilon$.

□

A.3. Generalization Bound

We first state the following general Lemma in order to bound the generalization error of the function class $\mathcal{F}$ on the unsupervised loss function $L_{un}(\cdot)$. Lemma 4.2 can be directly derived from it.

Lemma A.2. *Let $\ell:\mathbb{R}^k\to\mathbb{R}$ be η-Lipschitz and bounded by B. Then with probability at least $1-\delta$ over the training set $\mathcal{S}=\{(x_j,x_j^+,x_{j1}^-,\ldots,x_{jk}^-)\}_{j=1}^M$, for all $f\in\mathcal{F}$*

$$L_{un}(\hat{f})\leq L_{un}(f)+O\left(\frac{\eta R\sqrt{k}\mathcal{R}_\mathcal{S}(\mathcal{F})}{M}+B\sqrt{\frac{\log\frac{1}{\delta}}{M}}\right) \tag{15}$$

where

$$\mathcal{R}_\mathcal{S}(\mathcal{F})=\mathop{\mathbb{E}}_{\sigma\sim\{\pm1\}^{(k+2)dM}}\left[\sup_{f\in\mathcal{F}}\langle\sigma,f_{|\mathcal{S}}\rangle\right] \tag{16}$$

and $f_{|\mathcal{S}}=\left(f_t(x_j),f_t(x_j^+),f_t(x_{j1}^-),\ldots,,f_t(x_{jk}^-)\right)_{\substack{j\in[M]\\ t\in[d]}}$

Note that for $k+1$-way classification, for hinge loss we have $\eta=1$ and $B=O(R^2)$, while for logistic loss $\eta=1$ and $B=O(R^2+\log k)$. Setting $k=1$, we get Lemma 4.2. We now prove Lemma A.2.

Proof of Lemma A.2. First, we use the classical bound for the generalization error in terms of the Rademacher complexity of the function class (see (Mohri et al., 2018) Theorem 3.1). For a real function class G whose functions map from a set Z to $[0,1]$ and for any $\delta>0$, if $\mathcal{S}$ is a training set composed by M iid samples $\{z_j\}_{j=1}^M$, then with probability at least $1-\frac{\delta}{2}$, for all $g\in G$

$$\mathbb{E}[g(z)]\leq\frac{1}{M}\sum_{j=1}^M g(z_i)+\frac{2\mathcal{R}_\mathcal{S}(G)}{M}+3\sqrt{\frac{\log\frac{4}{\delta}}{2M}} \tag{17}$$

where $\mathcal{R}_\mathcal{S}(G)$ is the usual Rademacher complexity. We apply this bound to our case by setting $Z=\mathcal{X}^{k+2}$, $\mathcal{S}$ is our training set and the function class is

$$G=\left\{g_f(x,x^+,x_1^-,...,x_k^-)=\frac{1}{B}\ell\left(\{f(x)^T\left(f(x^+)-f(x_i^-)\right)\}_{i=1}^k\right)\middle| f\in\mathcal{F}\right\} \tag{18}$$

We will show that for some universal constant c, $\mathcal{R}_\mathcal{S}(G)\leq c\frac{\eta R\sqrt{k}}{B}\mathcal{R}_\mathcal{S}(\mathcal{F})$ or equivalently

$$\mathop{\mathbb{E}}_{\sigma\sim\{\pm1\}^M}\left[\sup_{f\in\mathcal{F}}\langle\sigma,(g_f)_{|\mathcal{S}}\rangle\right]\leq c\frac{\eta R\sqrt{k}}{B}\mathop{\mathbb{E}}_{\sigma\sim\{\pm1\}^{d(k+2)M}}\left[\sup_{f\in\mathcal{F}}\langle\sigma,f_{|\mathcal{S}}\rangle\right] \tag{19}$$

where $(g_f)_{|\mathcal{S}} = \{g_f(x_j, x_j^+, x_{j1}^-, ..., x_{jk}^-)\}_{j=1}^M$. To do that we will use the following vector-contraction inequality.

Theorem A.3. *[Corollary 4 in (Maurer, 2016)] Let Z be any set, and $\mathcal{S} = \{z_j\}_{j=1}^M \in Z^M$. Let $\widetilde{\mathcal{F}}$ be a class of functions $\tilde{f} : Z \to \mathbb{R}^n$ and $h : \mathbb{R}^n \to \mathbb{R}$ be L-Lipschitz. For all $\tilde{f} \in \widetilde{\mathcal{F}}$, let $g_{\tilde{f}} = h \circ \tilde{f}$. Then*

$$\mathop{\mathbb{E}}_{\sigma\sim\{\pm1\}^M}\left[\sup_{\tilde{f}\in\widetilde{\mathcal{F}}}\left\langle \sigma, (g_{\tilde{f}})_{|\mathcal{S}}\right\rangle\right] \leq \sqrt{2}L \mathop{\mathbb{E}}_{\sigma\sim\{\pm1\}^{nM}}\left[\sup_{\tilde{f}\in\widetilde{\mathcal{F}}}\left\langle \sigma, \tilde{f}_{|\mathcal{S}}\right\rangle\right]$$

where $\tilde{f}_{|\mathcal{S}} = \left(\tilde{f}_t(z_j)\right)_{t\in[n], j\in[M]}$.

We apply Theorem A.3 to our case by setting $Z = \mathcal{X}^{k+2}$, $n = d(k+2)$ and

$$\widetilde{\mathcal{F}} = \left\{\tilde{f}(x, x^+, x_{j1}^-, ..., x_{jk}^-) = (f(x), f(x^+), f(x_{j1}^-), ..., f(x_{jk}^-)) | f \in \mathcal{F}\right\}$$

We also use $g_{\tilde{f}} = g_f$ where $\tilde{f}$ is derived from f as in the definition of $\widetilde{F}$. Observe that now A.3 is exactly in the form of 19 and we need to show that $L \leq \frac{c}{\sqrt{2}}\frac{\eta R\sqrt{k}}{B}$ for some constant c. But, for $z = (x, x^+, x_1^-, ..., x_k^-)$, we have $g_{\tilde{f}}(z) = \frac{1}{B}\ell(\phi(\tilde{f}(z)))$ where $\phi : \mathbb{R}^{(k+2)d} \to \mathbb{R}^k$ and $\phi\left((v_t, v_t^+, v_{t1}^-, ..., v_{tk}^-)_{t\in[d]}\right) = \left(\sum_t v_t(v_t^+ - v_{ti}^-)\right)_{i\in[k]}$. Thus, we may use $h = \frac{1}{B}\ell \circ \phi$ to apply Theorem A.3.

Now, we see that ϕ is $\sqrt{6k}R$-Lipschitz when $\sum_t v_t^2, \sum_t (v_t^+)^2, \sum_t (v_{tj}^-)^2 \leq R^2$ by computing its Jacobian. Indeed, for all $i, j \in [k]$ and $t \in [d]$, we have $\frac{\partial\phi_i}{\partial v_t} = v_t^+ - v_{ti}^-$, $\frac{\partial\phi_i}{\partial v_t^+} = v_t$ and $\frac{\partial\phi_i}{\partial v_{tj}^-} = -v_t 1\{i = j\}$. From triangle inequaltiy, the Frobenius norm of the Jacobian J of ϕ is

$$||J||_F = \sqrt{\sum_{i,t}(v_t^+ - v_{ti}^-)^2 + 2k\sum_t v_t^2} \leq \sqrt{4kR^2 + 2kR^2} = \sqrt{6k}R$$

Now, taking into account that $||J||_2 \leq ||J||_F$, we have that ϕ is $\sqrt{6k}R$-Lipschitz on its domain and since ℓ is η-Lipschitz, we have $L \leq \sqrt{6}\frac{\eta R\sqrt{k}}{B}$.

Now, we have that with probability at least $1 - \frac{\delta}{2}$

$$L_{un}(\hat{f}) \leq \widehat{L}_{un}(\hat{f}) + O\left(\frac{\eta R\sqrt{k}\mathcal{R}_{\mathcal{S}}(\mathcal{F})}{M} + B\sqrt{\frac{\log\frac{1}{\delta}}{M}}\right) \tag{20}$$

Let $f^* \in \arg\min_{f\in\mathcal{F}} L_{un}(f)$. With probability at least $1-\frac{\delta}{2}$, we have that $\widehat{L}_{un}(f^*) \leq L_{un}(f^*) + 3B\sqrt{\frac{\log\frac{2}{\delta}}{2M}}$ (Hoeffding's inequality). Combining this with Equation (20), the fact that $\widehat{L}_{un}(\hat{f}) \leq \widehat{L}_{un}(f^*)$ and applying a union bound, finishes the proof. □

A.4. Proof of Proposition 6.2

By convexity of ℓ,

$$\ell\left(f(x)^T\left(\frac{\sum_i f(x_i^+)}{b} - \frac{\sum_i f(x_i^-)}{b}\right)\right) = \ell\left(\frac{1}{b}\sum_i f(x)^T\left(f(x_i^+) - f(x_i^-)\right)\right) \leq \frac{1}{b}\sum_i \ell\left(f(x)^T\left(f(x_i^+) - f(x_i^-)\right)\right)$$

Thus,

$$L_{un}^{block}(f) = \mathop{\mathbb{E}}_{\substack{x,x_i^+\\x_i^-}}\left[\ell\left(f(x)^T\left(\frac{\sum_i f(x_i^+)}{b} - \frac{\sum_i f(x_i^-)}{b}\right)\right)\right] \leq \mathop{\mathbb{E}}_{\substack{x,x_i^+\\x_i^-}}\left[\frac{1}{b}\sum_i \ell\left(f(x)^T\left(f(x_i^+) - f(x_i^-)\right)\right)\right] = L_{un}(f)$$

The proof of the lower bound is analogous to that of Lemma 4.3.

□

B. Results for k Negative Samples

B.1. Formal theorem statement and proof

We now present Theorem B.1 as the formal statement of Theorem 6.1 and prove it. First we define some necessary quantities.

Let $(c^+, c_1^-, \ldots, c_k^-)$ be $k+1$ not necessarily distinct classes. We define $Q(c^+, c_1^-, \ldots, c_k^-)$ to be the set of distinct classes in this tuple. We also define $I^+(c_1^-, ..., c_k^-) = \{i \in [k] \mid c_i^- = c^+\}$ to be the set of indices where c^+ reappears in the negative samples. We will abuse notation and just write Q, I^+ when the tuple is clear from the context.

To define $L^{\neq}_{un}(f)$ consider the following tweak in the way the latent classes are sampled: sample $c^+, c_1^-, \ldots, c_k^- \sim \rho^{k+1}$ conditioning on $|I^+| < k$ and then remove all c_i^-, $i \in I^+$. The datapoints are then sampled as usual: $x, x^+ \sim \mathcal{D}^2_{c^+}$ and $x_i^- \sim \mathcal{D}_{c_i^-}$, $i \in [k]$, independently.

$$L^{\neq}_{un}(f) := \mathop{\mathbb{E}}_{\substack{c^+,c_i^-\\x,x^+,x_i^-}} \left[\ell\left(\left\{f(x)^T\left(f(x^+)-f(x_i^-)\right)\right\}_{i\notin I^+}\right)\middle| |I^+| < k\right]$$

which always contrasts points from different classes, since it only considers the negative samples that are not from c^+.

The generalization error is [9]

$$Gen_M = O\left(R\sqrt{k}\frac{\mathcal{R}_{\mathcal{S}}(\mathcal{F})}{M} + (R^2 + \log k)\sqrt{\frac{\log\frac{1}{\delta}}{M}}\right)$$

were $\mathcal{R}_{\mathcal{S}}(\mathcal{F})$ is the extension of the definition in Section 4: $\mathcal{R}_{\mathcal{S}}(\mathcal{F}) = \mathop{\mathbb{E}}_{\sigma\sim\{\pm1\}^{(k+2)dM}}\left[\sup_{f\in\mathcal{F}}\langle\sigma, f_{|\mathcal{S}}\rangle\right]$, where $f_{|\mathcal{S}} = \left(f_t(x_j), f_t(x_j^+), f_t(x_{j1}^-), \ldots, , f_t(x_{jk}^-)\right)_{j\in[M],t\in[d]}$.

For $c^+, c_1^-, ..., c_k^- \sim \rho^{k+1}$, let $\tau_k = \mathbb{P}[I^+ \neq \emptyset]$ and $\tau' = \mathbb{P}[c^+ = c_i^-, \forall i]$. Observe that τ_1, as defined in Section 4, is $\mathbb{P}[c^+ = c_1^-]$. Let $p_{max}(\mathcal{T}) = \max_c \mathcal{D}_{\mathcal{T}}(c)$ and

$$\rho^+_{min}(\mathcal{T}) = \min_{c\in\mathcal{T}} \mathbb{P}_{c^+,c_i^-\sim\rho^{k+1}}\left(c^+ = c | Q = \mathcal{T}, I^+ = \emptyset\right)$$

In Theorem B.1 we will upper bound the following quantity: $\mathop{\mathbb{E}}_{\mathcal{T}\sim\mathcal{D}}\left[\frac{\rho^+_{min}(\mathcal{T})}{p_{max}(\mathcal{T})}\, L^{\mu}_{sup}(\mathcal{T}, \hat{f})\right]$ ($\mathcal{D}$ was defined in Section 6.1).

Theorem B.1. *Let $\hat{f} \in \arg\min_{f\in\mathcal{F}} \widehat{L}_{un}(f)$. With probability at least $1-\delta$, for all $f \in \mathcal{F}$*

$$\mathop{\mathbb{E}}_{\mathcal{T}\sim\mathcal{D}}\left[\frac{\rho^+_{min}(\mathcal{T})}{p_{max}(\mathcal{T})}\, L^{\mu}_{sup}(\mathcal{T}, \hat{f})\right] \leq \frac{1-\tau'}{1-\tau_k} L^{\neq}_{un}(f) + c'k\frac{\tau_1}{1-\tau_k}s(f) + \frac{1}{1-\tau_k}Gen_M$$

where c' is a constant.

Note that the definition of $s(f)$ used here is defined in Section 4

Proof. First, we note that both hinge and logistic loss satisfy the following property: $\forall I_1, I_2$ such that $I_1 \cup I_2 = [t]$ we have that

$$\ell(\{\boldsymbol{v}_i\}_{i\in I_1}) \leq \ell(\{\boldsymbol{v}_i\}_{i\in[t]}) \leq \ell(\{\boldsymbol{v}_i\}_{i\in I_1}) + \ell(\{\boldsymbol{v}_i\}_{i\in I_2}) \tag{21}$$

We now prove the Theorem in 3 steps. First, we leverage the convexity of ℓ to upper bound a supervised-type loss with the unsupervised loss $L_{un}(f)$ of any $f \in \mathcal{F}$. We call it supervised-type loss because it also includes degenerate tasks: $|\mathcal{T}| = 1$.

[9] The $\log k$ term can be made $O(1)$ for the hinge loss.

Then, we decompose the supervised-type loss into an average loss over a distribution of supervised tasks, as defined in the Theorem, plus a degenerate/constant term. Finally, we upper bound the unsupervised loss $L_{un}(f)$ with two terms: $L^{\neq}_{un}(f)$ that measures how well f contrasts points from different classes and an intraclass deviation penalty, corresponding to $s(f)$.

Step 1 (convexity): When the class c is clear from context, we write $\hat{\mu}_c = \underset{x\sim c}{\mathbb{E}}[\hat{f}(x)]$. Recall that the sampling procedure for unsupervised data is as follows: sample $c^+, c^-_1, ..., c^-_k \sim \rho^{k+1}$ and then $x, x^+ \sim \mathcal{D}^2_{c^+}$ and $x^-_i \sim \mathcal{D}_{c^-_i}$, $i \in [k]$. So, we have

$$
\begin{aligned}
L_{un}(\hat{f}) &= \underset{\substack{c^+,c^-_i\sim\rho^{k+1}\\ x,x^+\sim\mathcal{D}^2_{c^+}\\ x^-_i\sim\mathcal{D}_{c^-_i}}}{\mathbb{E}} \left[\ell\left(\left\{\hat{f}(x)^T\left(\hat{f}(x^+)-\hat{f}(x^-_i)\right)\right\}_{i=1}^{k}\right)\right] \\
&= \underset{\substack{c^+,c^-_i\sim\rho^{k+1}\\ x\sim\mathcal{D}_{c^+}}}{\mathbb{E}} \; \underset{\substack{x^+\sim\mathcal{C}_{c^+}\\ x^-_i\sim\mathcal{D}_{c^-_i}}}{\mathbb{E}} \left[\ell\left(\left\{\hat{f}(x)^T\left(\hat{f}(x^+)-\hat{f}(x^-_i)\right)\right\}_{i=1}^{k}\right)\right] \geq \underset{\substack{c^+,c^-_i\sim\rho^{k+1}\\ x\sim\mathcal{D}_{c^+}}}{\mathbb{E}} \left[\ell\left(\left\{\hat{f}(x)^T\left(\hat{\mu}_{c^+}-\hat{\mu}_{c^-_i}\right)\right\}_{i=1}^{k}\right)\right]
\end{aligned}
\tag{22}
$$

where the last inequality follows by applying the usual Jensen's inequality and the convexity of ℓ. Note that in the upper bounded quantity, the $c^+, c^-_1, ..., c^-_k$ don't have to be distinct and so the tuple does not necessarily form a task.

Step 2 (decomposing into supervised tasks) We now decompose the above quantity to handle repeated classes.

$$
\begin{aligned}
&\underset{\substack{c^+,c^-_i\sim\rho^{k+1}\\ x\sim\mathcal{D}_{c^+}}}{\mathbb{E}} \left[\ell\left(\left\{\hat{f}(x)^T\left(\hat{\mu}_{c^+}-\hat{\mu}_{c^-_i}\right)\right\}_{i=1}^{k}\right)\right] \\
&\qquad \geq (1-\tau_k) \underset{\substack{c^+,c^-_i\sim\rho^{k+1}\\ x\sim\mathcal{D}_{c^+}}}{\mathbb{E}} \left[\ell\left(\left\{\hat{f}(x)^T\left(\hat{\mu}_{c^+}-\hat{\mu}_{c^-_i}\right)\right\}_{i=1}^{k}\right)\middle| I^+=\emptyset\right] + \tau_k \underset{c^+,c^-_i\sim\rho^{k+1}}{\mathbb{E}} [\ell(\underbrace{0,...,0}_{|I^+|\text{ times}})|I^+\neq\emptyset] \\
&\qquad \geq (1-\tau_k) \underset{\substack{c^+,c^-_i\sim\rho^{k+1}\\ x\sim\mathcal{D}_{c^+}}}{\mathbb{E}} \left[\ell\left(\left\{\hat{f}(x)^T\left(\hat{\mu}_{c^+}-\hat{\mu}_{c}\right)\right\}_{\substack{c\in Q\\ c\neq c^+}}\right)\middle| I^+=\emptyset\right] + \tau_k \underset{c^+,c^-_i\sim\rho^{k+1}}{\mathbb{E}} \left[\ell_{|I^+|}(\vec{0}) \;\middle|\; I^+\neq\emptyset\right]
\end{aligned}
\tag{23}
$$

where $\ell_t(\vec{0}) = \ell(0, \ldots, 0)$ (t times). Both inequalities follow from the LHS of Equation (21). Now we are closer to our goal of lower bounding an average supervised loss, since the first expectation in the RHS has a loss which is over a set of distinct classes. However, notice that this loss is for separating c^+ from $Q(c^+, c^-_1, ..., c^-_k) \setminus \{c^+\}$. We now proceed to a symmetrization of this term to alleviate this issue.

Recall that in the main paper, sampling $\mathcal{T}$ from $\mathcal{D}$ is defined as sampling the (k+1)-tuple from ρ^{k+1} conditioned on $I^+ = \emptyset$ and setting $\mathcal{T} = Q$. Based on this definition, by the tower property of expectation, we have

$$
\begin{aligned}
&\underset{\substack{c^+,c^-_i\sim\rho^{k+1}\\ x\sim\mathcal{D}_{c^+}}}{\mathbb{E}} \left[\ell\left(\left\{\hat{f}(x)^T\left(\hat{\mu}_{c^+}-\hat{\mu}_{c}\right)\right\}_{\substack{c\in Q\\ c\neq c^+}}\right)\middle| I^+=\emptyset\right] \\
&= \underset{\mathcal{T}\sim\mathcal{D}}{\mathbb{E}} \; \underset{\substack{c^+,c^-_i\sim\rho^{k+1}\\ x\sim\mathcal{D}_{c^+}}}{\mathbb{E}} \left[\ell\Big(\Big\{\hat{f}(x)^T\big(\hat{\mu}_{c^+}-\hat{\mu}_{c}\big)\Big\}_{\substack{c\in Q\\ c\neq c^+}}\Big)\Big| Q=\mathcal{T}, I^+=\emptyset\right] \\
&= \underset{\mathcal{T}\sim\mathcal{D}}{\mathbb{E}} \; \underset{\substack{c^+\sim\rho^+(\mathcal{T})\\ x\sim\mathcal{D}_{c^+}}}{\mathbb{E}} \left[\ell\Big(\Big\{\hat{f}(x)^T\big(\hat{\mu}_{c^+}-\hat{\mu}_{c}\big)\Big\}_{\substack{c\in\mathcal{T}\\ c\neq c^+}}\Big)\right]
\end{aligned}
\tag{24}
$$

where $\rho^+(\mathcal{T})$ is the distribution of c^+ when $(c^+, c^-_1, ..., c^-_k)$ are sampled from ρ^{k+1} conditioned on $Q = \mathcal{T}$ and $I^+ = \emptyset$. Recall that $\rho^+_{min}(\mathcal{T})$ from the theorem's statement is exactly the minimum out of these $|\mathcal{T}|$ probabilities. Now, to lower bound the last quantity with the LHS in the theorem statement, we just need to observe that for all tasks $\mathcal{T}$

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\substack{c^+\sim\rho^+(\mathcal{T})\\ x\sim\mathcal{D}_{c^+}}}\left[\ell\Big(\Big\{\hat f(x)^T\big(\hat\mu_{c^+}-\hat\mu_c\big)\Big\}_{\substack{c\in\mathcal{T}\\ c\neq c^+}}\Big)\right]\\
&\geq \frac{\rho^+_{min}(\mathcal{T})}{p_{max}(\mathcal{T})}\mathop{\mathbb{E}}_{\substack{c^+\sim\mathcal{D}_{\mathcal{T}}\\ x\sim\mathcal{D}_{c^+}}}\left[\ell\Big(\Big\{\hat f(x)^T\big(\hat\mu_{c^+}-\hat\mu_c\big)\Big\}_{\substack{c\in\mathcal{T}\\ c\neq c^+}}\Big)\right] \qquad (25)\\
&= \frac{\rho^+_{min}(\mathcal{T})}{p_{max}(\mathcal{T})}L_{sup}(\mathcal{T},\hat f)
\end{aligned}$$

By combining this with Equations (22), (23), (25) we get

$$(1-\tau_k)\mathop{\mathbb{E}}_{\mathcal{T}\sim\mathcal{D}}\left[\frac{\rho^+_{min}(T)}{p_{max}(T)}L_{sup}(\mathcal{T},\hat f)\right]\leq L_{un}(\hat f)-\tau_k\mathop{\mathbb{E}}_{c^+,c_i^-\sim\rho^{k+1}}\left[\ell_{|I^+|}(\vec{0})\;\middle|\;I^+\neq\emptyset\right] \tag{26}$$

Now, by applying Lemma A.2, we bound the generalization error: with probability at least $1-\delta$, $\forall f\in\mathcal{F}$

$$L_{un}(\hat f)\leq L_{un}(f)+Gen_M \tag{27}$$

However, $L_{un}(f)$ cannot be made arbitrarily small. One can see that for all $f\in\mathcal{F}$, $L_{un}(f)$ is lower bounded by the second term in Equation (22), which cannot be made arbitrarily small as $\tau_k>0$.

$$L_{un}(f)\geq\mathop{\mathbb{E}}_{\substack{c^+,c_i^-\sim\rho^{k+1}\\ x,x^+\sim\mathcal{D}_{c^+}\\ x_i^-\sim\mathcal{D}_{c_i^-}}}\left[\ell\left(\left\{f(x)^T\left(f(x^+)-f(x_i^-)\right)\right\}_{i\in I^+}\right)\right]\geq\tau\mathop{\mathbb{E}}_{c^+,c_i^-\sim\rho^{k+1}}\left[\ell_{|I^+|}(\vec{0})\;\middle|\;I^+\neq\emptyset\right] \tag{28}$$

where we applied Jensen's inequality. Since τ_k is not 0, the above quantity can never be arbitrarily close to 0 (no matter how rich $\mathcal{F}$ is).

Step 3 (L_{un} decomposition) Now, we decompose $L_{un}(f)$ by applying the RHS of Equation (21)

$$\mathcal{L}_{un}(f)\leq\mathop{\mathbb{E}}_{\substack{c^+,c_i^-\sim\rho^{k+1}\\ x,x^+\sim\mathcal{D}^2_{c^+}\\ x_i^-\sim\mathcal{D}_{c_i^-}}}\left[\ell\Big(\Big\{f(x)^T\big(f(x^+)-f(x_i^-)\big)\Big\}_{i\notin I^+}\Big)+\ell\Big(\Big\{f(x)^T\big(f(x^+)-f(x_i^-)\big)\Big\}_{i\in I^+}\Big)\right] \tag{29}$$

$$=\mathop{\mathbb{E}}_{\substack{c^+,c_i^-\sim\rho^{k+1}\\ x,x^+\sim\mathcal{D}^2_{c^+}\\ x_i^-\sim\mathcal{D}_{c_i^-},\,i\notin I^+}}\left[\ell\Big(\Big\{f(x)^T\big(f(x^+)-f(x_i^-)\big)\Big\}_{i\notin I^+}\Big)\right]+\mathop{\mathbb{E}}_{\substack{c^+,c_i^-\sim\rho^{k+1}\\ x,x^+\sim\mathcal{D}^2_{c^+}\\ x_i^-\sim\mathcal{D}_{c_i^-},\,i\in I^+}}\left[\ell\left(\left\{f(x)^T\left(f(x^+)-f(x_i^-)\right)\right\}_{i\in I^+}\right)\right] \tag{30}$$

$$\begin{aligned}
&=(1-\tau')\mathop{\mathbb{E}}_{\substack{c^+,c_i^-\sim\rho^{k+1}\\ x,x^+\sim\mathcal{D}^2_{c^+}\\ x_i^-\sim\mathcal{D}_{c_i^-},i\notin I^+}}\left[\ell\Big(\Big\{f(x)^T\big(f(x^+)-f(x_i^-)\big)\Big\}_{i\notin I^+}\Big)\Big||I^+|<k\right]\\
&\quad+\tau_k\mathop{\mathbb{E}}_{\substack{c^+,c_i^-\sim\rho^{k+1}\\ x,x^+\sim\mathcal{D}^2_{c^+}\\ x_i^-\sim\mathcal{D}_{c_i^-},\,i\in I^+}}\left[\ell\left(\left\{f(x)^T\left(f(x^+)-f(x_i^-)\right)\right\}_{i\in I^+}\right)\middle|\,I^+\neq\emptyset\right]
\end{aligned} \tag{31}$$

Observe that the first term is exactly $(1-\tau')L^{\neq}_{un}(f)$. Thus, combining (26), (27) and (31) we get

$$
\begin{aligned}
(1-\tau_k)\underset{\mathcal{T}\sim\mathcal{D}}{\mathbb{E}}\left[\frac{\rho^{+}_{min}(T)}{p_{max}(T)}L_{sup}(\mathcal{T},\hat{f})\right] &\leq (1-\tau')L^{\neq}_{un}(f)+Gen_M \\
&+\tau_k \underbrace{\underset{c^+,c_i^-\sim\rho^{k+1}}{\mathbb{E}}\left[\underset{\substack{x,x^+\sim\mathcal{D}^2_{c^+}\\ x_i^-\sim\mathcal{D}_{c_i^-},\ i\in I^+}}{\mathbb{E}}\left[\ell\Big(\Big\{f(x)^T\big(f(x^+)-f(x_i^-)\big)\Big\}_{i\in I^+}\Big)\right]-\ell_{|I^+|}(\vec{0})\,\middle|\, I^+\neq\emptyset\right]}_{\Delta(f)}
\end{aligned}
\tag{32}
$$

From the definition of I^+, $c_i^- = c^+, \forall i\in I^+$. Thus, from Lemma A.1, we get that

$$
\Delta(f)\leq c'\underset{c^+,c_i^-\sim\rho^{k+1}}{\mathbb{E}}\left[|I^+|\sqrt{\|\Sigma(f,c)\|_2}\underset{x\sim\mathcal{D}_c}{\mathbb{E}}[\|f(x)\|]\,\middle|\, I^+\neq\emptyset\right] \tag{33}
$$

for some constant c'.

Let u be a distribution over classes with $u(c)=\mathbb{P}_{c^+,c_i^-\sim\rho^{k+1}}[c^+=c|I^+\neq\emptyset]$ and it is easy to see that $u(c)\propto\rho(c)\big(1-(1-\rho(c))^k\big)$ By applying the tower property to Equation (33) we have

$$
\Delta(f)\leq c'\underset{c\sim u}{\mathbb{E}}\left[\underset{c^+,c_i^-\sim\rho^{k+1}}{\mathbb{E}}\big[|I^+|\big|c^+=c, I^+\neq\emptyset\big]\ \sqrt{\|\Sigma(f,c)\|_2}\underset{x\sim\mathcal{D}_c}{\mathbb{E}}[\|f(x)\|]\right] \tag{34}
$$

But,

$$
\begin{aligned}
\underset{c^+,c_i^-\sim\rho^{k+1}}{\mathbb{E}}\big[|I^+|\big|c^+=c, I^+\neq\emptyset\big] &= \sum_{i=1}^{k}\mathbb{P}_{c^+,c_i^-\sim\rho^{k+1}}\big(c_i^-=c^+\big|c^+=c, I^+\neq\emptyset\big) \\
&= k\mathbb{P}_{c^+,c_i^-\sim\rho^{k+1}}\big(c_1^-=c^+\big|c^+=c, I^+\neq\emptyset\big) \\
&= k\frac{\mathbb{P}_{c^+,c_i^-\sim\rho^{k+1}}\big(c_1^-=c^+=c\big)}{\mathbb{P}_{c^+,c_i^-\sim\rho^{k+1}}\big(c^+=c, I^+\neq\emptyset\big)} \\
&= k\frac{\rho^2(c)}{\rho(c)\big(1-(1-\rho(c))^k\big)} = k\frac{\rho(c)}{1-(1-\rho(c))^k}
\end{aligned}
\tag{35}
$$

Now, using the fact that $\tau_k = 1-\sum_{c'}\rho(c')(1-\rho(c'))^k = \sum_{c'}\rho(c')\left(1-(1-\rho(c'))^k\right)$ and $\tau_1=\sum_c\rho^2(c)$,

$$
\begin{aligned}
\frac{\tau_k}{1-\tau_k}\Delta(f) &\leq \frac{\tau_k}{1-\tau_k}c'\underset{c\sim u}{\mathbb{E}}\left[k\frac{\rho(c)}{1-(1-\rho(c))^k}\sqrt{\|\Sigma(f,c)\|_2}\underset{x\sim\mathcal{D}_c}{\mathbb{E}}[\|f(x)\|]\right] \\
&= c'k\frac{\tau_k}{1-\tau_k}\sum_c\frac{\rho^2(c)}{\sum_{c'}\rho(c')\left(1-(1-\rho(c'))^k\right)}\sqrt{\|\Sigma(f,c)\|_2}\underset{x\sim\mathcal{D}_c}{\mathbb{E}}[\|f(x)\|] \\
&= c'k\frac{\tau_1}{1-\tau_k}\underset{c\sim\nu}{\mathbb{E}}\left[\sqrt{\|\Sigma(f,c)\|_2}\underset{x\sim\mathcal{D}_c}{\mathbb{E}}[\|f(x)\|]\right] = c'k\frac{\tau_1}{1-\tau_k}s(f)
\end{aligned}
\tag{36}
$$

and we are done. □

B.2. Competitive Bound

As in Section 5.2, we prove a competitive type of bound, under similar assumptions. Let $\ell_\gamma(\boldsymbol{v}) = \max\{0, 1+\max_i\{-\boldsymbol{v}_i\}/\gamma\}$, $\boldsymbol{v}\in\mathbb{R}^k$, be the multiclass hinge loss with margin γ and for any $\mathcal{T}$ let $L^{\mu}_{\gamma,sup}(\mathcal{T},f)$ be $L^{\mu}_{sup}(\mathcal{T},f)$ when ℓ_γ is used as loss function. For all tasks $\mathcal{T}$, let ${\rho'}^{+}(\mathcal{T})$ is the distribution of c^+ when $(c^+, c_1^-, ..., c_k^-)$ are sampled from ρ^{k+1} conditioned on $Q=\mathcal{T}$ and $|I^+|<k$. Also, let ${\rho'}^{+}_{max}(\mathcal{T})$ be the maximum of these $|\mathcal{T}|$ probabilities and $p_{min}(\mathcal{T})=\min_{c\in\mathcal{T}}\mathcal{D}_{\mathcal{T}}(c)$.

We will show a competitive bound against the following quantity, for all $f \in \mathcal{F}$: $\underset{\mathcal{T}\sim\mathcal{D}'}{\mathbb{E}}\left[\frac{\rho'^{+}_{max}(\mathcal{T})}{p_{min}(\mathcal{T})}\mathcal{L}^{\mu}_{\gamma,sup}(\mathcal{T},f)\right]$, where $\mathcal{D}'$ is defined as follows: sample $c^+, c_1^-, ..., c_k^- \sim \rho^{k+1}$, conditioned on $|I^+| < k$. Then, set $\mathcal{T} = Q$. Observe that when $I^+ = \emptyset$ with high probability, we have $\mathcal{D}' \approx \mathcal{D}$.

Lemma B.2. *For all $f \in \mathcal{F}$ suppose the random variable $f(X)$, where $X \sim D_c$, is $\sigma^2(f)$-subgaussian in every direction for every class c and has maximum norm $R(f) = max_{x\in\mathcal{X}}\|f(x)\|$. Let $\widehat{f} \in \arg\min_{f\in\mathcal{F}} \widehat{L}_{un}(f)$. Then for all $\epsilon > 0$, with probability at least $1-\delta$, for all $f \in \mathcal{F}$*

$$\underset{\mathcal{T}\sim\mathcal{D}}{\mathbb{E}}\left[\frac{\rho^{+}_{min}(\mathcal{T})}{p_{max}(\mathcal{T})}\,L^{\mu}_{sup}(\mathcal{T},\hat{f})\right] \leq \alpha\gamma(f)\underset{\mathcal{T}\sim\mathcal{D}'}{\mathbb{E}}\left[\frac{\rho'^{+}_{max}(\mathcal{T})}{p_{min}(\mathcal{T})}\mathcal{L}^{\mu}_{\gamma,sup}(\mathcal{T},f)\right] + \beta s(f) + \eta Gen_M + \epsilon$$

where $\gamma(f) = 1 + c'R(f)\sigma(f)(\sqrt{\log k} + \sqrt{\log\frac{R(f)}{\epsilon}})$, c' is some constant, $\alpha = \frac{1-\tau'}{1-\tau_k}$, $\beta = k\frac{\tau_1}{1-\tau_k}$ and $\eta = \frac{1}{1-\tau_k}$.

Proof. We will show that $\forall f \in \mathcal{F}$

$$L^{\neq}_{un}(f) \leq \gamma(f)\underset{\mathcal{T}\sim\mathcal{D}'}{\mathbb{E}}\left[\frac{\rho'^{+}_{max}(\mathcal{T})}{p_{min}(\mathcal{T})}\mathcal{L}^{\mu}_{\gamma,sup}(\mathcal{T},f)\right] \tag{37}$$

and the Lemma follows from Theorem 6.1. Now, we fix an $\epsilon > 0$, an $f \in \mathcal{F}$ and we drop most of the arguments f in the rest of the proof. Also, fix $c^+, c_1^- \dots c_k^-, x$ and let $t = k - |I^+|$. We assume without loss of generality, that $c^+ \neq c_i^-, \forall i \in [t]$. Now,

$$\max_{i\in[t]} f(x)^T(f(x_i^-) - f(x^+)) \leq \mu + \max_i z_i^- - z^+ \tag{38}$$

where $\mu = \max_{i\in[t]} f(x)^T(\mu_{c_i^-} - \mu_{c^+})$, $z_i^- = f(x)^T(f(x_i^-) - \mu_{c_i^-})$ and $z^+ = f(x)^T(f(x^+) - \mu_{c^+})$. z_i are centered σ^2R^2-subgaussian, so from standard properties of subgaussian random variables $\mathbb{P}[\max_i z_i^- \geq \sqrt{2}\sigma R\sqrt{\log t} + \sqrt{2c_1}\sigma R\sqrt{\log R/\epsilon}] \leq (\epsilon/R)^{c_1}$ (again we consider here the case where $R \geq 1$ and for $R < 1$, the same arguments hold but with removing R from the log). z^+ is also centered σ^2R^2-subgaussian, so $\mathbb{P}[z^+ \geq \sqrt{2c_1}\sigma R\sqrt{\log R/\epsilon}] \leq (\epsilon/R)^{c_1}$. Let $\gamma = 1 + c'\sigma R(\sqrt{\log t} + \sqrt{\log R/\epsilon})$ for appropriate constant c'. By union bound, we have $p = \mathbb{P}[\max_i z_i^- - z^+ \geq \gamma - 1] \leq 2(\epsilon/R)^{c_1}$. Thus, $\mathbb{E}_{z^+,z_i^-}[(1 + \mu + \max_i z_i^- - z^+)_+] \leq (1-p)(\mu+\gamma)_+ + p(2R^2+1) \leq \gamma(1+\mu/\gamma)_+ + \epsilon$ (for appropriate constant c_1). By taking expectation over $c^+, c_i^- \sim \rho^{k+1}$, conditioned on $|I^+| < k$, and over $x \sim \mathcal{D}_{c^+}$ we get

$$\begin{aligned} L^{\neq}_{un}(f) &\leq \gamma \underset{\substack{c^+,c_i^-\sim\rho^{k+1}\\ x\sim\mathcal{D}_{c^+}}}{\mathbb{E}}\left[\left(1 + \frac{\max_{c\in Q, c\neq c^+} f(x)^T(\mu_c - \mu_{c^+})}{\gamma}\right)_+ \middle| |I^+| < k\right] \\ &= \gamma \underset{\mathcal{T}\sim\mathcal{D}'}{\mathbb{E}} \underset{\substack{c^+,c_i^-\sim\rho^{k+1}\\ x\sim\mathcal{D}_{c^+}}}{\mathbb{E}}\left[\left(1 + \frac{\max_{c\in Q, c\neq c^+} f(x)^T(\mu_c - \mu_{c^+})}{\gamma}\right)_+ \middle| Q = \mathcal{T}, |I^+| < k\right] \\ &= \gamma \underset{\mathcal{T}\sim\mathcal{D}'}{\mathbb{E}} \underset{\substack{c^+\sim\rho'^{+}(\mathcal{T})\\ x\sim\mathcal{D}_{c^+}}}{\mathbb{E}}\left[\left(1 + \frac{\max_{c\in T, c\neq c^+} f(x)^T(\mu_c - \mu_{c^+})}{\gamma}\right)_+\right] \leq \gamma \underset{\mathcal{T}\sim\mathcal{D}'}{\mathbb{E}}\left[\frac{\rho'^{+}_{max}(\mathcal{T})}{p_{min}(\mathcal{T})}\mathcal{L}^{\mu}_{\gamma,sup}(\mathcal{T},f)\right] \end{aligned} \tag{39}$$

□

C. Examples for Section 6.2

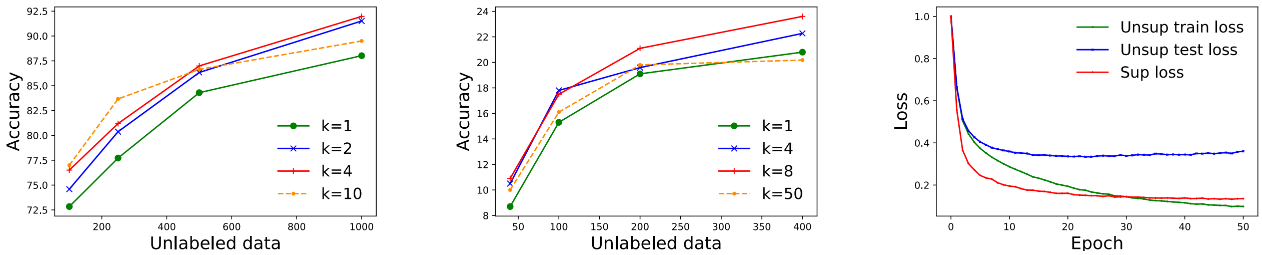
Here, we illustrate via examples two ways in which the increase of k can lead to suboptimal $\hat{f}$. We will consider the hinge loss as the loss function, while the examples carry over trivially for logistic loss.

1. The first example is the case where even though there exist representations in $\mathcal{F}$ that can separate every class, the suboptimal representation is picked by the algorithm when $k = \Omega(|\mathcal{C}|)$. Let $\mathcal{C} = \{c_i\}_{i\in[n]}$ where for each class, D_{c_i} is

uniform over two points $\{x_i^1, x_i^2\}$. Let e_i be the indicator vectors in $\mathbb{R}^n$ and let the class $\mathcal{F}$ consists of $\{f_0, f_1\}$ with $f_0, f_1 : \mathcal{X} \mapsto \mathbb{R}^n$ where $f_1(x_i^1) = 3/2re_i$ and $f_1(x_i^2) = 1/2re_i$ for all i, for some $r > 0$, and $f_0 = \vec{0}$. Finally, ρ is uniform over $\mathcal{C}$. Now, when the number of negative samples is $\Omega(n)$, the probability that $\exists j \in [k]$ such that $c^+ = c_j^-$ is constant, and therefore $L_{un}(f) = \Omega(r^2) > 1 = L_{un}(f_0)$ when r is large. This means that despite $L_{sup}(\mathcal{C}, f_1) = 0$, the algorithm will pick f_0 which is a suboptimal representation.

2. We can extend the first example to the case where, even when $k = o(|\mathcal{C}|)$, the algorithm picks suboptimal representations. To do so, we simply 'replicate' the first example to create clusters of classes. Formally, let $\mathcal{C} = \{c_{ij}\}_{i,j\in[n]}$ where for each class, $D_{c_{ij}}$ is uniform over two points $\{x_{ij}^1, x_{ij}^2\}$. Finally, same as above, let $\mathcal{F}$ consist of two functions $\{f_0, f_1\}$. The function f_1 maps $f_1(x_{ij}^1) = 3/2re_i$ and $f_1(x_{ij}^2) = 1/2re_i$ for all i, j and $f_0 = \vec{0}$. ρ is uniform over $\mathcal{C}$. Now, note that f_1 'clutsters' the n^2 classes and their points into n clusters, each along an e_i. Thus, it is only useful for contrasting classes from different clusters. However, note that the probability of intra-cluster collision with k negative samples is $1 - (1 - 1/n)^k$. When $k = o(n)$, we have that $L_{un}(f_1) = o(1) < 1 = L_{un}(f_0)$ so the algorithm will pick f_1. However, when $k = \Omega(n)$, $L_{un}(f) = \Omega(r^2) > 1 = L_{un}(f_0)$ and the algorithm will pick the suboptimal representation f_0. Thus, despite $|\mathcal{C}| = n^2$, having more than n negative samples can hurt performance, since even tough f_1 cannot solve all the tasks, the average supervised loss over t-way tasks, $t = o(n)$, is $L_{sup}(f) \leq O(1 - (1 - 1/n)^{t-1}) = o(1)$.

D. Controlled Experiments



(a) CIFAR-100 (b) Wiki-3029 (c) Wiki-3029

Figure D.1. Effect of amount of unlabeled data and # of negative samples on unsupervised representations, measured on binary classification for CIFAR100 in (a) and on top-1 performance on Wiki-3029 in Fig (b) (top-1 performance is used because avg binary was same for all k). Fig. (c) shows the dynamics of train/test loss; supervised loss roughly tracks unsupervised test loss, as suggested by Theorem 4.1

To simulate the data generation process described in Section 2, we generate similar pairs (blocks) of data points by sampling from the same class. Dissimilar pairs (negative samples) are selected randomly. Contrastive learning was done using our objectives (5), and compared to performance of standard supervised training, with both using the *same architecture* for representation f. For the Wiki-3029 experiment, we use a Gated Recurrent Network (GRU) (Chung et al., 2015) with output dimension 300 trained using dropout 0.3 and fix the word embedding layer with pre-trained GloVe embeddings (Pennington et al., 2014). For CIFAR-100 we use VGG-16 (Simonyan & Zisserman, 2014) with an additional 512x100 linear layer added at the end to make the final representations 100 dimensional, The unsupervised model for CIFAR-100 is trained with 500 blocks of size 2 per class with 4 negative samples, and for Wiki-3029 we use 20 blocks of size 10 per class with 8 negative samples. We test (1) learned representations on average tasks by using the mean classifier and compare to representations trained using labeled data; (2) the effect of various parameters like amount of unlabeled data (N)[10], number of negative samples (k) and block size (b) on representation quality; (3) whether the supervised loss tracks the unsupervised loss as suggested by Theorem 4.1; (4) performance of the mean classifier of the supervised model.

Results: These appear in Table D.1. For Wiki-3029 the unsupervised performance is very close to the supervised performance in all respects, while for CIFAR-100 the *avg-k* performance is respectable, rising to good for binary classification. One surprise is that the mean classifier, central to our analysis of unsupervised learning, performs well also with representations learned by supervised training on CIFAR-100. Even the mean computed by just 5 labeled samples performs well, getting within 2% accuracy of the 500 sample mean classifier on CIFAR-100. This suggests that representations learnt by

[10] If we used M similar blocks of size b and k negative blocks for each similar block, $N = Mb(k + 1)$. In practice, however, we reuse the blocks for negative sampling and lose the factor of $k + 1$.

Table D.1. Performance of supervised and unsupervised representations on average k-wise classification tasks (AVG-k) and for comparison, on full multiclass (TOP-R) which is not covered by our theory. Classifier can have a trained output layer (TR), or the mean classifier (μ) of Definition 2.1, with μ-5 indicating the mean was computed using only 5 labeled examples.

		SUPERVISED			UNSUPERVISED		
		TR	μ	μ-5	TR	μ	μ-5
WIKI-3029	AVG-2	97.8	97.7	97.0	97.3	97.7	96.9
	AVG-10	89.1	87.2	83.1	88.4	87.4	83.5
	TOP-10	67.4	59.0	48.2	64.7	59.0	45.8
	TOP-1	43.2	33.2	21.7	38.7	30.4	17.0
CIFAR-100	AVG-2	97.2	95.9	95.8	93.2	92.0	90.6
	AVG-5	92.7	89.8	89.4	80.9	79.4	75.7
	TOP-5	88.9	83.5	82.5	70.4	65.6	59.0
	TOP-1	72.1	69.9	67.3	36.9	31.8	25.0

standard supervised deep learning are actually quite concentrated. We also notice that the supervised representations have fairly low unsupervised training loss (as low as 0.4), even though the optimization is minimizing a different objective.

To measure the sample complexity benefit provided by contrastive learning, we train the supervised model on just 10% fraction of the dataset and compare it with an unsupervised model trained on unlabeled data whose mean classifiers are computed using the same amount of labeled data. We find that the unsupervised model beats the supervised model by almost 4% on the 100-way task and by 5% on the average binary task when only 50 labeled samples are used.

Figure D.1 highlights the positive effect of increasing number of negative samples as well as amount of data used by unsupervised algorithm. In both cases, using a lot of negative examples stops helping after a point, confirming our suspicions in Section 6.2. We also demonstrate how the supervised loss tracks unsupervised test loss in Figure D.1c.