A ANALYSIS OF HARD SAMPLING

A.1 HARD SAMPLING INTERPOLATES BETWEEN MARGINAL AND WORST-CASE NEGATIVES

We begin by proving Proposition 3. Recall that the proposition stated the following.

Proposition 6. *Let $\mathcal{L}^*(f) = \sup_{q \in \Pi} \mathcal{L}(f, q)$. Then for any $t > 0$ and measurable $f : \mathcal{X} \to \mathbb{S}^{d-1}/t$ we observe the convergence $\mathcal{L}(f, q_\beta^-) \longrightarrow \mathcal{L}^*(f)$ as $\beta \to \infty$.*

Proof. Consider the following essential supremum,

$$M(x) = \operatorname*{ess\,sup}_{x^- \in \mathcal{X} : x^- \nsim x} f(x)^T f(x^-) = \sup\{m > 0 : m \geq f(x)^T f(x^-) \text{ a.s. for } x^- \sim p^-\}.$$

The second inequality holds since $\text{supp}(p) = \mathcal{X}$. We may rewrite

$$\mathcal{L}^*(f) = \mathbb{E}_{\substack{x \sim p \\ x^+ \sim p_x^+}} \left[-\log \frac{e^{f(x)^T f(x^+)}}{e^{f(x)^T f(x^+)} + Q e^{M(x)}} \right],$$

$$\mathcal{L}(f, q_\beta^-) = \mathbb{E}_{\substack{x \sim p \\ x^+ \sim p_x^+}} \left[-\log \frac{e^{f(x)^T f(x^+)}}{e^{f(x)^T f(x^+)} + Q \mathbb{E}_{x^- \sim q_\beta^-}[e^{f(x)^T f(x^-)}]} \right].$$

The difference between these two terms can be bounded as follows,

$$\begin{aligned}
\left| \mathcal{L}^*(f) - \mathcal{L}(f, q_\beta^-) \right| &\leq \mathbb{E}_{\substack{x \sim p \\ x^+ \sim p_x^+}} \left| -\log \frac{e^{f(x)^T f(x^+)}}{e^{f(x)^T f(x^+)} + Q e^{M(x)}} + \log \frac{e^{f(x)^T f(x^+)}}{e^{f(x)^T f(x^+)} + Q \mathbb{E}_{x^- \sim q_\beta^-}[e^{f(x)^T f(x^-)}]} \right| \\
&= \mathbb{E}_{\substack{x \sim p \\ x^+ \sim p_x^+}} \left| \log \left(e^{f(x)^T f(x^+)} + Q \mathbb{E}_{x^- \sim q_\beta^-}[e^{f(x)^T f(x^-)}] \right) - \log \left(e^{f(x)^T f(x^+)} + Q e^{M(x)} \right) \right| \\
&\leq \frac{e^{1/t}}{Q+1} \cdot \mathbb{E}_{\substack{x \sim p \\ x^+ \sim p_x^+}} \left| e^{f(x)^T f(x^+)} + Q \mathbb{E}_{x^- \sim q_\beta^-}[e^{f(x)^T f(x^-)}] - e^{f(x)^T f(x^+)} - Q e^{M(x)} \right| \\
&= \frac{e^{1/t} Q}{Q+1} \cdot \mathbb{E}_{x \sim p} \left| \mathbb{E}_{x^- \sim q_\beta^-}[e^{f(x)^T f(x^-)}] - e^{M(x)} \right| \\
&\leq e^{1/t} \cdot \mathbb{E}_{x \sim p} \mathbb{E}_{x^- \sim q_\beta^-} \left| e^{M(x)} - e^{f(x)^T f(x^-)} \right|
\end{aligned}$$

where for the second inequality we have used the fact that f lies on the hypersphere of radius $1/t$ to restrict the domain of the logarithm to values greater than $(Q+1)e^{-1/t}$. Because of this the logarithm is Lipschitz with parameter $e^{1/t}/(Q+1)$. Using again the fact that f lies on the hypersphere we know that $\left| f(x)^T f(x^-) \right| \leq 1/t^2$ and hence have the following inequality,

$$\mathbb{E}_{x \sim p} \mathbb{E}_{q_\beta^-} \left| e^{M(x)} - e^{f(x)^T f(x^-)} \right| \leq e^{1/t^2} \mathbb{E}_{x \sim p} \mathbb{E}_{q_\beta^-} \left| M(x) - f(x)^T f(x^-) \right|$$

Let us consider the inner expectation $E_\beta(x) = \mathbb{E}_{q_\beta^-} \left| M(x) - f(x)^T f(x^-) \right|$. Note that since f is bounded, $E_\beta(x)$ is uniformly bounded in x. Therefore, in order to show the convergence $\mathcal{L}(f, q_\beta^-) \to \mathcal{L}^*(f)$ as $\beta \to \infty$, it suffices by the dominated convergence theorem to show that $E_\beta(x) \to 0$ pointwise as $\beta \to \infty$ for arbitrary fixed $x \in \mathcal{X}$.

From now on we denote $M = M(x)$ for brevity, and consider a fixed $x \in \mathcal{X}$. From the definition of q_β^- it is clear that $q_\beta^- \ll p^-$. That is, since $q_\beta^- = c \cdot p^-$ for some (non-constant) c, it is absolutely continuous with respect to p^-. So $M(x) \geq f(x)^T f(x^-)$ almost surely for $x^- \sim q_\beta^-$, and we

may therefore drop the absolute value signs from our expectation. Define the following event $\mathcal{G}_\varepsilon = \{x^- : f(x)^\top f(x^-) \geq M - \varepsilon\}$ where $\mathcal{G}$ is refers to a "good" event. Define its complement $\mathcal{B}_\varepsilon = \mathcal{G}_\varepsilon^c$ where $\mathcal{B}$ is for "bad". For a fixed $x \in \mathcal{X}$ and $\varepsilon > 0$ consider,

$$\begin{aligned}
E_\beta(x) &= \mathbb{E}_{x^- \sim q_\beta^-} \left| M(x) - f(x)^T f(x^-) \right| \\
&= \mathbb{P}_{x^- \sim q_\beta^-}(\mathcal{G}_\varepsilon) \cdot \mathbb{E}_{x^- \sim q_\beta^-} \left[\left| M(x) - f(x)^T f(x^-) \right| | \mathcal{G}_\varepsilon \right] + \mathbb{P}_{x^- \sim q_\beta^-}(\mathcal{B}_\varepsilon) \cdot \mathbb{E}_{x^- \sim q_\beta^-} \left[\left| M(x) - f(x)^T f(x^-) \right| | \mathcal{B}_\varepsilon \right] \\
&\leq \mathbb{P}_{x^- \sim q_\beta^-}(\mathcal{G}_\varepsilon) \cdot \varepsilon + 2\mathbb{P}_{x^- \sim q_\beta^-}(\mathcal{B}_\varepsilon) \\
&\leq \varepsilon + 2\mathbb{P}_{x^- \sim q_\beta^-}(\mathcal{B}_\varepsilon).
\end{aligned}$$

We need to control $\mathbb{P}_{x^- \sim q_\beta^-}(\mathcal{B}_\varepsilon)$. Expanding,

$$\mathbb{P}_{x^- \sim q_\beta^-}(\mathcal{B}_\varepsilon) = \int_{\mathcal{X}} \mathbf{1} \left\{ f(x)^T f(x^-) < M(x) - \varepsilon \right\} \frac{e^{\beta f(x)^T f(x^-)} \cdot p^-(x^-)}{Z_\beta} \mathrm{d}x^-$$

where $Z_\beta = \int_{\mathcal{X}} e^{\beta f(x)^T f(x^-)} p^-(x^-) \mathrm{d}x^-$ is the partition function of q_β^-. We may bound this expression by,

$$\begin{aligned}
\int_{\mathcal{X}} \mathbf{1} \left\{ f(x)^T f(x^-) < M - \varepsilon \right\} \frac{e^{\beta(M-\varepsilon)} \cdot p^-(x^-)}{Z_\beta} \mathrm{d}x^- &\leq \frac{e^{\beta(M-\varepsilon)}}{Z_\beta} \int_{\mathcal{X}} \mathbf{1} \left\{ f(x)^T f(x^-) < M - \varepsilon \right\} p^-(x^-) \mathrm{d}x^- \\
&= \frac{e^{\beta(M-\varepsilon)}}{Z_\beta} \mathbb{P}_{x^- \sim p^-}(\mathcal{B}_\varepsilon) \\
&\leq \frac{e^{\beta(M-\varepsilon)}}{Z_\beta}
\end{aligned}$$

Note that

$$Z_\beta = \int_{\mathcal{X}} e^{\beta f(x)^T f(x^-)} p^-(x^-) \mathrm{d}x^- \geq e^{\beta(M-\varepsilon/2)} \mathbb{P}_{x^- \sim p^-}(f(x)^T f(x^-) \geq M - \varepsilon/2).$$

By the definition of $M = M(x)$ the probability $\rho_\varepsilon = \mathbb{P}_{x^- \sim p^-}(f(x)^T f(x^-) \geq M - \varepsilon/2) > 0$, and we may therefore bound,

$$\begin{aligned}
\mathbb{P}_{x^- \sim q_\beta^-}(\mathcal{B}_\varepsilon) &= \frac{e^{\beta(M-\varepsilon)}}{e^{\beta(M-\varepsilon/2)} \rho_\varepsilon} \\
&= e^{-\beta\varepsilon/2} / \rho_\varepsilon \\
&\longrightarrow 0 \ \text{ as } \ \beta \to \infty.
\end{aligned}$$

We may therefore take β to be sufficiently big so as to make $\mathbb{P}_{x^- \sim q_\beta^-}(\mathcal{B}_\varepsilon) \leq \varepsilon$ and therefore $E_\beta(x) \leq 3\varepsilon$. In other words, $E_\beta(x) \longrightarrow 0$ as $\beta \to \infty$. □

A.2 Optimal Embeddings on the Hypersphere for Worst-Case Negative Samples

In order to study properties of global optima of the contrastive objective using the adversarial worst case hard sampling distribution recall that we have the following limiting objective,

$$\mathcal{L}_\infty(f, q) = \mathbb{E}_{\substack{x \sim p \\ x^+ \sim p_x^+}} \left[-\log \frac{e^{f(x)^T f(x^+)}}{\mathbb{E}_{x^- \sim q_\beta}[e^{f(x)^T f(x^-)}]} \right]. \tag{8}$$

We may separate the logarithm of a quotient into the sum of two terms plus a constant,

$$\mathcal{L}_\infty(f, q) = \mathcal{L}_{\text{align}}(f) + \mathcal{L}_{\text{unif}}(f, q) - 1/t^2$$

where $\mathcal{L}_{\text{align}}(f) = \mathbb{E}_{x,x^+}\|f(x) - f(x^+)\|^2/2$ and $\mathcal{L}_{\text{unif}}(f,q) = \mathbb{E}_{x\sim p}\log \mathbb{E}_{x^-\sim q} e^{f(x)^\top f(x^-)}$. Here we have used the fact that f lies on the boundary of the hypersphere of radius $1/t$, which gives us the following equivalence between inner products and squared Euclidean norm,

$$2/t^2 - 2f(x)^\top f(x^+) = \|f(x)\|^2 + \|f(x^+)\|^2 - 2f(x)^\top f(x^+) = \|f(x) - f(x^+)\|^2. \tag{9}$$

Taking supremum to obtain $\mathcal{L}^*_\infty(f) = \sup_{q\in\Pi} \mathcal{L}_\infty(f,q)$ we find that the second expression simplifies to,

$$\mathcal{L}^*_{\text{unif}}(f) = \sup_{q\in\Pi} \mathcal{L}_{\text{unif}}(f,q) = \mathbb{E}_{x\sim p}\log \sup_{x^-\nsim x} e^{f(x)^\top f(x^-)} = \mathbb{E}_{x\sim p} \sup_{x^-\nsim x} f(x)^\top f(x^-).$$

Using Eqn. (9), this can be re-expressed as,

$$\mathbb{E}_{x\sim p} \sup_{x^-\nsim x} f(x)^\top f(x^-) = -\mathbb{E}_{x\sim p} \inf_{x^-\nsim x} \|f(x) - f(x^-)\|^2/2 + 1/t^2. \tag{10}$$

The forthcoming theorem exactly characterizes the global optima of $\min_f \mathcal{L}^*_\infty(f)$

Theorem 7. *Suppose the downstream task is classification (i.e. $\mathcal{C}$ is finite), and let $\mathcal{L}^*_\infty(f) = \sup_{q\in\Pi} \mathcal{L}_\infty(f,q)$. The infimum $\inf_{f:\text{ measurable}} \mathcal{L}^*_\infty(f)$ is attained, and any f^* achieving the global minimum is such that $f^*(x) = f^*(x^+)$ almost surely. Furthermore, letting $\mathbf{v}_c = f^*(x)$ for any x such that $h(x) = c$ (so $\mathbf{v}_c$ is well defined up to a set of x of measure zero), f^* is characterized as being any solution to the following ball-packing problem,*

$$\max_{\{\mathbf{v}_c\in\mathbb{S}^{d-1}/t\}_{c\in\mathcal{C}}} \sum_{c\in\mathcal{C}} \rho(c)\cdot \min_{c'\neq c} \|\mathbf{v}_c - \mathbf{v}_{c'}\|^2. \tag{11}$$

Proof. Any minimizer of $\mathcal{L}_{\text{align}}(f)$ has the property that $f(x) = f(x^+)$ almost surely. So, in order to prove the first claim, it suffices to show that there exist functions $f \in \arg\inf_f \mathcal{L}^*_{\text{unif}}(f)$ for which $f(x) = f(x^+)$ almost surely. This is because, at that point, we have shown that $\arg\min_f \mathcal{L}_{\text{align}}(f)$ and $\arg\min_f \mathcal{L}^*_{\text{unif}}(f)$ intersect, and therefore any solution of $\mathcal{L}^*_\infty(f) = \mathcal{L}_{\text{align}}(f) + \mathcal{L}^*_{\text{unif}}(f)$ must lie in this intersection.

To this end, suppose that $f \in \arg\min_f \mathcal{L}^*_{\text{unif}}(f)$ but that $f(x) \neq f(x^+)$ with non-zero probability. We shall show that we can construct a new embedding $\hat{f}$ such that $f(x) = f(x^+)$ almost surely, and $\mathcal{L}^*_{\text{unif}}(\hat{f}) \leq \mathcal{L}^*_{\text{unif}}(f)$. Due to Eqn. (10) this last condition is equivalent to showing,

$$\mathbb{E}_{x\sim p} \inf_{x^-\nsim x} \|\hat{f}(x) - \hat{f}(x^-)\|^2 \geq \mathbb{E}_{x\sim p} \inf_{x^-\nsim x} \|f(x) - f(x^-)\|^2. \tag{12}$$

Fix a $c \in \mathcal{C}$, and let $x_c \in \arg\max_{x:h(x)=c} \inf_{x^-\nsim x} \|f(x) - f(x^-)\|^2$. The maximum is guaranteed to be attained, as we explain now. Indeed we know the maximum is attained at some point in the closure $\partial\{x : h(x) = c\} \cup \{x : h(x) = c\}$. Since $\mathcal{X}$ is compact and connected, any point $\bar{x} \in \partial\{x : h(x) = c\} \setminus \{x : h(x) = c\}$ is such that $\inf_{x^-\nsim \bar{x}} \|f(\bar{x}) - f(x^-)\|^2 = 0$ since $\bar{x}$ must belong to $\{x : h(x) = c'\}$ for some other c'. Such an $\bar{x}$ cannot be a solution unless all points in $\{x : h(x) = c\}$ also achieve 0, in which case we can simply take x_c to be a point in the interior of $\{x : h(x) = c\}$.

Now, define $\hat{f}(x) = f(x_c)$ for any x such that $h(x) = c$ and $\hat{f}(x) = f(x)$ otherwise. Let us first aim to show that Eqn. (12) holds for this $\hat{f}$. Let us begin to expand the left hand side of Eqn. (12),

$$
\begin{aligned}
&\mathbb{E}_{x\sim p} \inf_{x^- \nsim x} \|\hat{f}(x) - \hat{f}(x^-)\|^2 \\
&\qquad = \mathbb{E}_{\hat{c}\sim\rho}\mathbb{E}_{x\sim p(\cdot|\hat{c})} \inf_{x^- \nsim x} \|\hat{f}(x) - \hat{f}(x^-)\|^2 \\
&\qquad = \rho(c)\mathbb{E}_{x\sim p(\cdot|c)} \inf_{x^- \nsim x} \|\hat{f}(x) - \hat{f}(x^-)\|^2 \\
&\qquad\qquad + (1-\rho(c))\mathbb{E}_{\hat{c}\sim\rho(\cdot|\hat{c}\neq c)}\mathbb{E}_{x\sim p(\cdot|\hat{c})} \inf_{x^- \nsim x} \|\hat{f}(x) - \hat{f}(x^-)\|^2 \\
&\qquad = \rho(c)\mathbb{E}_{x\sim p(\cdot|c)} \inf_{x^- \nsim x} \|f(x_c) - f(x^-)\|^2 \\
&\qquad\qquad + (1-\rho(c))\mathbb{E}_{\hat{c}\sim\rho(\cdot|\hat{c}\neq c)}\mathbb{E}_{x\sim p(\cdot|\hat{c})} \inf_{x^- \nsim x} \|\hat{f}(x) - \hat{f}(x^-)\|^2 \\
&\qquad = \rho(c) \inf_{x^- \nsim x_c} \|f(x_c) - f(x^-)\|^2 \\
&\qquad\qquad + (1-\rho(c))\mathbb{E}_{\hat{c}\sim\rho(\cdot|\hat{c}\neq c)}\mathbb{E}_{x\sim p(\cdot|\hat{c})} \inf_{h(x^-)\neq\hat{c}} \|\hat{f}(x) - \hat{f}(x^-)\|^2 \qquad (13)
\end{aligned}
$$

By construction, the first term can be lower bounded by $\inf_{x^- \nsim x_c} \|f(x_c) - f(x^-)\|^2 \geq \mathbb{E}_{x\sim p(\cdot|c)} \inf_{h(x^-)\neq c} \|f(x) - f(x^-)\|^2$ for any x such that $h(x) = c$. To lower bound the second term, consider any fixed $\hat{c} \neq c$ and $x \sim p(\cdot|\hat{c})$ (so $h(x) = \hat{c}$). Define the following two subsets of the input space $\mathcal{X}$

$$
\mathcal{A} = \{f(x^-) : f(x^-) \neq \hat{c} \text{ for } x^- \in \mathcal{X}\} \qquad \widehat{\mathcal{A}} = \{f(x^-) \in \mathcal{X} : \hat{f}(x^-) \neq \hat{c} \text{ for } x^- \in \mathcal{X}\}.
$$

Since by construction the range of $\hat{f}$ is a subset of the range of f, we know that $\widehat{\mathcal{A}} \subseteq \mathcal{A}$. Combining this with the fact that $\hat{f}(x) = f(x)$ whenever $h(x) = \hat{c} \neq c$ we see,

$$
\begin{aligned}
\inf_{h(x^-)\neq\hat{c}} \|\hat{f}(x) - \hat{f}(x^-)\|^2 &= \inf_{h(x^-)\neq\hat{c}} \|f(x) - \hat{f}(x^-)\|^2 \\
&= \inf_{u\in\widehat{\mathcal{A}}} \|f(x) - u\|^2 \\
&\geq \inf_{u\in\mathcal{A}} \|f(x) - u\|^2 \\
&= \inf_{h(x^-)\neq\hat{c}} \|f(x) - f(x^-)\|^2
\end{aligned}
$$

Using these two lower bounds we may conclude that Eqn. (13) can be lower bounded by,

$$
\rho(c)\mathbb{E}_{x\sim p(\cdot|c)} \inf_{h(x^-)\neq c} \|f(x) - f(x^-)\|^2 + (1-\rho(c))\mathbb{E}_{\hat{c}\sim\rho(\cdot|\hat{c}\neq c)}\mathbb{E}_{x\sim p(\cdot|\hat{c})} \inf_{h(x^-)\neq\hat{c}} \|f(x) - f(x^-)\|^2
$$

which equals $\mathbb{E}_{x\sim p} \inf_{x^- \nsim x} \|f(x) - f(x^-)\|^2$. We have therefore proved Eqn. (12). To summarize the current progress; given an embedding f we have constructed a new embedding $\hat{f}$ that attains lower $\mathcal{L}_{\text{unif}}$ loss and which is constant on x such that $\hat{f}$ is constant on $\{x : h(x) = c\}$. Enumerating $\mathcal{C} = \{c_1, c_2 \ldots, c_{|\mathcal{C}|}\}$, we may repeatedly apply the same argument to construct a sequence of embeddings $f_1, f_2, \ldots, f_{|\mathcal{C}|}$ such that f_i is constant on each of the following sets $\{x : h(x) = c_j\}$ for $j \leq i$. The final embedding in the sequence $f^* = f_{|\mathcal{C}|}$ is such that $\mathcal{L}^*_{\text{unif}}(f^*) \leq \mathcal{L}^*_{\text{unif}}(f)$ and therefore f^* is a minimizer. This embedding is constant on each of $\{x : h(x) = c_j\}$ for $j = 1, 2, \ldots, |\mathcal{C}|$. In other words, $f^*(x) = f^*(x^+)$ almost surely. We have proved the first claim.

Obtaining the second claim is a matter of manipulating $\mathcal{L}^*_\infty(f^*)$. Indeed, we know that $\mathcal{L}^*_\infty(f^*) = \mathcal{L}^*_{\text{unif}}(f^*) - 1/t^2$ and defining $\mathbf{v}_c = f^*(x) = f(x_c)$ for each $c \in \mathcal{C}$, this expression is minimized if and only if f^* attains,

$$\begin{aligned}\max_f \mathbb{E}_{x\sim p} \inf_{x^-\nsim x} \|f(x) - f(x^-)\|^2 &= \max_f \mathbb{E}_{c\sim\rho}\mathbb{E}_{x\sim p(\cdot|c)} \inf_{h(x^-)\neq c} \|f(x) - f(x^-)\|^2 \\ &= \max_f \sum_{c\in\mathcal{C}} \rho(c) \cdot \inf_{h(x^-)\neq c} \|f(x) - f(x^-)\|^2 \\ &= \max_{\{\mathbf{v}_c\in\mathbb{S}^{d-1}/t\}_{c\in\mathcal{C}}} \sum_{c\in\mathcal{C}} \rho(c) \cdot \min_{c'\neq c} \|\mathbf{v}_c - \mathbf{v}_{c'}\|^2\end{aligned}$$

where the final equality inserts f^* as an optimal f and reparameterizes the maximum to be over the set of vectors $\{\mathbf{v}_c \in \mathbb{S}^{d-1}/t\}_{c\in\mathcal{C}}$. □

A.3 Downstream Generalization

Theorem 8. *Suppose ρ is uniform on $\mathcal{C}$ and f is such that $\mathcal{L}^*_\infty(f) - \inf_{\bar{f}\text{ measurable}} \mathcal{L}^*_\infty(\bar{f}) \leq \varepsilon$ with $\varepsilon \leq 1$. Let $\{\mathbf{v}_c^* \in \mathbb{S}^{d-1}/t\}_{c\in\mathcal{C}}$ be a solution to Problem 7, and define $\xi = \min_{c,c^-:c\neq c^-} \left\|\mathbf{v}_c^* - \mathbf{v}_{c^-}^*\right\| > 0$. Then there exists a set of vectors $\{\mathbf{v}_c \in \mathbb{S}^{d-1}/t\}_{c\in\mathcal{C}}$ such that the following 1-nearest neighbor classifier,*

$$\hat{h}(x) = \hat{c}, \quad \textit{where} \quad \hat{c} = \arg\min_{\bar{c}\in\mathcal{C}} \left\|f(x) - \mathbf{v}_{\bar{c}}\right\| \quad \textit{(ties broken arbitrarily)}$$

achieves misclassification risk,

$$\mathbb{P}(\hat{h}(x) \neq c) \leq \frac{8\varepsilon}{(\xi^2 - 2|\mathcal{C}|\,(1+1/t)\varepsilon^{1/2})^2}$$

Proof. To begin, using the definition of $\hat{h}$ we know that for any $0 < \delta < \xi$,

$$\begin{aligned}\mathbb{P}_{x,c}(\hat{h}(x) = c) &= \mathbb{P}_{x,c}\left(\left\|f(x) - \mathbf{v}_c\right\| \leq \min_{c^-:c^-\neq c} \left\|f(x) - \mathbf{v}_{c^-}\right\|\right) \\ &\geq \mathbb{P}_{x,c}\left(\left\|f(x) - \mathbf{v}_c\right\| \leq \delta, \quad \text{and} \quad \delta \leq \min_{c^-:c^-\neq c} \left\|f(x) - \mathbf{v}_{c^-}\right\|\right) \\ &\geq 1 - \mathbb{P}_{x,c}\left(\left\|f(x) - \mathbf{v}_c\right\| > \delta\right) - \mathbb{P}_{x,c}\Big(\min_{c^-:c^-\neq c} \left\|f(x) - \mathbf{v}_{c^-}\right\| < \delta\Big)\end{aligned}$$

So to prove the result, our goal is now to bound these two probabilities. To do so, we use the bound on the excess risk. Indeed, combining the fact $\mathcal{L}^*_\infty(f) - \inf_{\bar{f}\text{ measurable}} \mathcal{L}^*_\infty(\bar{f}) \leq \varepsilon$ with the notational rearrangements before Theorem 7 we observe that $\mathbb{E}_{x,x^+}\left\|f(x) - f(x^+)\right\|^2 \leq 2\varepsilon$.

We have,

$$2\varepsilon \geq \mathbb{E}_{x,x^+}\left\|f(x) - f(x^+)\right\|^2 = \mathbb{E}_{c\sim\rho}\mathbb{E}_{x^+\sim p(\cdot|c)}\mathbb{E}_{x\sim p(\cdot|c)}\left\|f(x) - f(x^+)\right\|^2.$$

For fixed c, x^+, let $x_c \in \arg\min_{\{x^+:h(x^+)=c\}} \mathbb{E}_{x\sim p(\cdot|c)}\left\|f(x) - f(x^+)\right\|^2$ where we extend the minimum to be over the closure, a compact set, to guarantee it is attained. Then we have

$$2\varepsilon \geq \mathbb{E}_{c\sim\rho}\mathbb{E}_{x^+\sim p(\cdot|c)}\mathbb{E}_{x\sim p(\cdot|c)}\left\|f(x) - f(x^+)\right\|^2 \geq \mathbb{E}_{c\sim\rho}\mathbb{E}_{x\sim p(\cdot|c)}\left\|f(x) - \mathbf{v}_c\right\|^2$$

where we have now defined $\mathbf{v}_c = f(x_c)$ for each $c \in \mathcal{C}$. Note in particular that $\mathbf{v}_c$ lies on the surface of the hypersphere $\mathbb{S}^{d-1}/t$. This enables us to obtain the follow bound using Markov's inequality,

$$\begin{aligned}\mathbb{P}_{x,c}\left(\left\|f(x)-\mathbf{v}_c\right\|>\delta\right) &= \mathbb{P}_{x,c}\left(\left\|f(x)-\mathbf{v}_c\right\|^2>\delta^2\right)\\ &\leq \frac{\mathbb{E}_{x,c}\left\|f(x)-\mathbf{v}_c\right\|^2}{\delta^2}\\ &\leq \frac{2\varepsilon}{\delta^2}.\end{aligned}$$

so it remains still to bound $\mathbb{P}_{x,c}\big(\min_{c^-:c^-\neq c}\left\|f(x)-\mathbf{v}_{c^-}\right\| < \delta\big)$. Defining $\xi' = \min_{c,c^-:c\neq c^-}\|\mathbf{v}_c-\mathbf{v}_{c-}\|$, we have the following fact (proven later).

Fact (see lemma 9): $\xi' \geq \sqrt{\xi^2-2|\mathcal{C}|\,(1+1/t)\sqrt{\varepsilon}}$.

Using this fact we are able to get control over the tail probability as follows,

$$\begin{aligned}\mathbb{P}_{x,c}\left(\min_{c^-:c^-\neq c}\left\|f(x)-\mathbf{v}_{c^-}\right\|<\delta\right) &\leq \mathbb{P}_{x,c}\left(\left\|f(x)-\mathbf{v}_c\right\|>\xi'-\delta\right)\\ &\leq \mathbb{P}_{x,c}\left(\left\|f(x)-\mathbf{v}_c\right\|>\xi-\sqrt{\xi^2-2|\mathcal{C}|\,(1+1/t)\varepsilon^{1/2}}-\delta\right)\\ &= \mathbb{P}_{x,c}\left(\left\|f(x)-\mathbf{v}_c\right\|^2>(\sqrt{\xi^2-2|\mathcal{C}|\,(1+1/t)\varepsilon^{1/2}}-\delta)^2\right)\\ &\leq \frac{2\varepsilon}{(\sqrt{\xi^2-2|\mathcal{C}|\,(1+1/t)\varepsilon^{1/2}}-\delta)^2}.\end{aligned}$$

where this inequality holds for for any $0\leq\delta\leq\sqrt{\xi^2-2|\mathcal{C}|\,(1+1/t)\varepsilon^{1/2}}$.

Gathering together our tail probability bounds we find that $\mathbb{P}_{x,c}(\hat{h}(x)=c)\geq 1-\frac{2\varepsilon}{\delta^2}-\frac{2\varepsilon}{(\sqrt{\xi^2-2|\mathcal{C}|(1+1/t)\varepsilon^{1/2}}-\delta)^2}$ for any $0\leq\delta\leq\sqrt{\xi^2-2|\mathcal{C}|\,(1+1/t)\varepsilon^{1/2}}$. That is,

$$\mathbb{P}_{x,c}(\hat{h}(x)\neq c)\leq\frac{2\varepsilon}{\delta^2}+\frac{2\varepsilon}{(\sqrt{\xi^2-2|\mathcal{C}|\,(1+1/t)\varepsilon^{1/2}}-\delta)^2}$$

Since this holds for any $0\leq\delta\leq\sqrt{\xi^2-2|\mathcal{C}|\,(1+1/t)\varepsilon^{1/2}}$,

$$\mathbb{P}_{x,c}(\hat{h}(x)\neq c)\leq\min_{0\leq\delta\leq\sqrt{\xi^2-2|\mathcal{C}|\varepsilon}}\left\{\frac{2\varepsilon}{\delta^2}+\frac{2\varepsilon}{(\sqrt{\xi^2-2|\mathcal{C}|\,(1+1/t)\varepsilon^{1/2}}-\delta)^2}\right\}.$$

Elementary calculus shows that the minimum is attained at $\delta=\frac{\sqrt{\xi^2-2|\mathcal{C}|(1+1/t)\varepsilon^{1/2}}}{2}$. Plugging this in yields the final bound,

$$\mathbb{P}(\hat{h}(x)\neq c)\leq\frac{8\varepsilon}{(\xi^2-2|\mathcal{C}|\,(1+1/t)\varepsilon^{1/2})^2}.$$

□

Lemma 9. *Consider the same setting as introduced in Theorem 5. In particular define*

$$\xi'=\min_{c,c^-:c\neq c^-}\|\mathbf{v}_c-\mathbf{v}_{c-}\|,\qquad \xi=\min_{c,c^-:c\neq c^-}\left\|\mathbf{v}_c^*-\mathbf{v}_{c-}^*\right\|.$$

where $\{\mathbf{v}_c^*\in\mathbb{S}^{d-1}/t\}_{c\in\mathcal{C}}$ is a solution to Problem 7, and $\{\mathbf{v}_c\in\mathbb{S}^{d-1}/t\}_{c\in\mathcal{C}}$ is defined via $\mathbf{v}_c=f(x_c)$ with $x_c\in\arg\min_{\{x^+:h(x^+)=c\}}\mathbb{E}_{x\sim p(\cdot|c)}\left\|f(x)-f(x^+)\right\|^2$ for each $c\in\mathcal{C}$. Then we have,

$$\xi'\geq\sqrt{\xi^2-2|\mathcal{C}|\,(1+1/t)\varepsilon^{1/2}}.$$

Proof. Define,

$$X = \min_{c^-:c^-\neq c} \|\mathbf{v}_c - \mathbf{v}_{c-}\|^2, \qquad X^* = \min_{c^-:c^-\neq c} \left\|\mathbf{v}_c^* - \mathbf{v}_{c-}^*\right\|^2.$$

X and X^* are random due to the randomness of $c \sim \rho$. We can split up the following expectation by conditioning on the event $\{X \leq X^*\}$ and its complement,

$$\mathbb{E}|X - X^*| = \mathbb{P}(X \geq X^*)\mathbb{E}[X - X^*] + \mathbb{P}(X \leq X^*)\mathbb{E}[X^* - X]. \tag{14}$$

Using $\mathcal{L}_\infty^*(f) - \inf_{\bar{f} \text{ measurable}} \mathcal{L}_\infty^*(\bar{f}) \leq \varepsilon$ and the notational re-writing of the objective $\mathcal{L}_\infty^*$ introduced before Theorem 7, we observe the following fact, whose proof we give in a separate lemma after the conclusion of this proof.

Fact (see lemma 10): $\mathbb{E}X^* - 2(1 + 1/t)\sqrt{\varepsilon} \leq \mathbb{E}X \leq \mathbb{E}X^*$.

This fact implies in particular $\mathbb{E}[X - X^*] \leq 0$ and $\mathbb{E}[X^* - X] \leq 2(1 + 1/t)\sqrt{\varepsilon}$. Inserting both inequalities into Eqn. 14 we find that $\mathbb{E}|X - X^*| \leq 2(1 + 1/t)\sqrt{\varepsilon}$. In other words, since ρ is uniform,

$$\frac{1}{|\mathcal{C}|}\sum_{c\in\mathcal{C}} \left| \min_{c^-:c^-\neq c} \|\mathbf{v}_c - \mathbf{v}_{c-}\|^2 - \min_{c^-:c^-\neq c} \left\|\mathbf{v}_c^* - \mathbf{v}_{c-}^*\right\|^2 \right| \leq 2(1 + 1/t)\sqrt{\varepsilon}.$$

From which we can say that for any $c \in \mathcal{C}$,

$$\left| \min_{c^-:c^-\neq c} \|\mathbf{v}_c - \mathbf{v}_{c-}\|^2 - \min_{c^-:c^-\neq c} \left\|\mathbf{v}_c^* - \mathbf{v}_{c-}^*\right\|^2 \right| \leq 2|\mathcal{C}|\,(1 + 1/t)\sqrt{\varepsilon}.$$

So $\min_{c^-:c^-\neq c}\|\mathbf{v}_c - \mathbf{v}_{c-}\| \geq \sqrt{\min_{c^-:c^-\neq c}\left\|\mathbf{v}_c^* - \mathbf{v}_{c-}^*\right\|^2 - 2|\mathcal{C}|\,(1 + 1/t)\varepsilon^{1/2}} \geq \sqrt{\xi^2 - 2|\mathcal{C}|\,(1 + 1/t)\varepsilon^{1/2}}$. Since this holds for any $c \in \mathcal{C}$, we conclude that $\xi' \geq \sqrt{\xi^2 - 2|\mathcal{C}|\,(1 + 1/t)\varepsilon^{1/2}}$. □

Lemma 10. *Consider the same setting as introduced in Theorem 5. Define also,*

$$X = \min_{c^-:c^-\neq c} \|\mathbf{v}_c - \mathbf{v}_{c-}\|^2, \qquad X^* = \min_{c^-:c^-\neq c} \left\|\mathbf{v}_c^* - \mathbf{v}_{c-}^*\right\|^2,$$

where $\mathbf{v}_c = f(x_c)$ *with* $x_c \in \arg\min_{\overline{\{x^+:h(x^+)=c\}}} \mathbb{E}_{x\sim p(\cdot|c)}\left\|f(x) - f(x^+)\right\|^2$ *for each* $c \in \mathcal{C}$. *We have,*

$$\mathbb{E}X^* - 2(1 + 1/t)\sqrt{\varepsilon} \leq \mathbb{E}X \leq \mathbb{E}X^*.$$

Proof. By Theorem 7 we know there is an f^* attaining the minimum $\inf_{\bar{f} \text{ measurable}} \mathcal{L}_\infty^*(\bar{f})$ and that this f^* attains $\mathcal{L}_{\text{align}}^*(f^*) = 0$, and also minimizes the uniformity term $\mathcal{L}_{\text{unif}}^*(f)$, taking the value $\mathcal{L}_{\text{unif}}^*(f^*) = \mathbb{E}_{c\sim\rho} \max_{c^-:c^-\neq c} \mathbf{v_c^*}^\top \mathbf{v}_{c^-}^*$. Because of this we find,

$$\begin{aligned}
\mathcal{L}_{\text{unif}}^*(f) &\leq \left(\mathcal{L}_\infty^*(f) - \mathcal{L}_\infty^*(f^*)\right) + \left(\mathcal{L}_{\text{align}}^*(f^*) - \mathcal{L}_{\text{align}}^*(f)\right) + \mathcal{L}_{\text{unif}}^*(f^*) \\
&\leq \left(\mathcal{L}_\infty^*(f) - \mathcal{L}_\infty^*(f^*)\right) + \mathcal{L}_{\text{unif}}^*(f^*) \\
&\leq \varepsilon + \mathcal{L}_{\text{unif}}^*(f^*) \\
&= \varepsilon + \mathbb{E}_{c\sim\rho} \max_{c^-:c^-\neq c} \mathbf{v_c^*}^\top \mathbf{v}_{c^-}^*.
\end{aligned}$$

Since we would like to bound $\mathbb{E}_{c\sim\rho}\max_{c^-:c^-\neq c}\mathbf{v_c}^\top\mathbf{v}_{c^-}$ in terms of $\mathbb{E}_{c\sim\rho}\max_{c^-:c^-\neq c}\mathbf{v_c^*}^\top\mathbf{v}^*_{c^-}$, this observation means that is suffices to bound $\mathbb{E}_{c\sim\rho}\max_{c^-:c^-\neq c}\mathbf{v_c}^\top\mathbf{v}_{c^-}$ in terms of $\mathcal{L}^*_{\text{unif}}(f)$. To this end, note that for a fixed c, and x such that $h(x)=c$ we have,

$$\begin{aligned}\sup_{x^-\nsim x} f(x)^\top f(x^-) &= \sup_{x^-\nsim x}\left\{\mathbf{v}_c{}^\top f(x^-) + (f(x)-\mathbf{v}_c)^\top f(x^-)\right\}\\ &= \sup_{x^-\nsim x}\mathbf{v}_c{}^\top f(x^-) - \left\|f(x)-\mathbf{v}_c\right\|/t\\ &\geq \max_{x^-\in\{x_c\}_{c\in\mathcal{C}}}\mathbf{v}_c{}^\top f(x^-) - \left\|f(x)-\mathbf{v}_c\right\|/t\\ &= \max_{c^-\neq c}\mathbf{v}_c{}^\top\mathbf{v}_{c^-} - \left\|f(x)-\mathbf{v}_c\right\|/t\end{aligned}$$

where the inequality follows since $\{x_c\}_{c\in\mathcal{C}}$ is a subset of the closure of $\{x^-: x^-\nsim x\}$. Taking expectations over c, x,

$$\begin{aligned}\mathcal{L}^*_{\text{unif}}(f) &= \mathbb{E}_{x,c}\sup_{x^-\nsim x} f(x)^\top f(x^-)\\ &\geq \mathbb{E}_{c\sim\rho}\max_{c^-\neq c}\mathbf{v}_c{}^\top\mathbf{v}_{c^-} - \mathbb{E}_{x,c}\left\|f(x)-\mathbf{v}_c\right\|/t\\ &\geq \mathbb{E}_{c\sim\rho}\max_{c^-\neq c}\mathbf{v}_c{}^\top\mathbf{v}_{c^-} - \sqrt{\mathbb{E}_{x,c}\left\|f(x)-\mathbf{v}_c\right\|^2}/t\\ &\geq \mathbb{E}_{c\sim\rho}\max_{c^-\neq c}\mathbf{v}_c{}^\top\mathbf{v}_{c^-} - \sqrt{\varepsilon}/t.\end{aligned}$$

So since $\varepsilon\leq\sqrt{\varepsilon}$, we have found that

$$\mathbb{E}_{c\sim\rho}\max_{c^-\neq c}\mathbf{v}_c{}^\top\mathbf{v}_{c^-}\leq\sqrt{\varepsilon}/t+\varepsilon+\mathbb{E}_{c\sim\rho}\max_{c^-:c^-\neq c}\mathbf{v_c^*}^\top\mathbf{v}^*_{c^-}\leq(1+1/t)\sqrt{\varepsilon}+\mathbb{E}_{c\sim\rho}\max_{c^-:c^-\neq c}\mathbf{v_c^*}^\top\mathbf{v}^*_{c^-}.$$

Of course we also have,

$$\mathbb{E}_{c\sim\rho}\max_{c^-:c^-\neq c}\mathbf{v_c^*}^\top\mathbf{v}^*_{c^-}=\mathcal{L}^*_{\text{unif}}(f^*)\leq\mathbb{E}_{c\sim\rho}\max_{c^-:c^-\neq c}\mathbf{v_c}^\top\mathbf{v}_{c^-}$$

since the embedding $f(x)=\mathbf{v}_c$ whenever $h(x)=c$ is also a feasible solution. Combining these two inequalities with the simple identity $\mathbf{x}^\top\mathbf{y}=1/t^2-\|\mathbf{x}-\mathbf{y}\|^2/2$ for all length $1/t$ vectors $\mathbf{x},\mathbf{y}$, we find,

$$\begin{aligned}1/t^2-\mathbb{E}_{c\sim\rho}\max_{c^-:c^-\neq c}\left\|\mathbf{v_c^*}-\mathbf{v}^*_{c^-}\right\|^2/2 &\leq 1/t^2-\mathbb{E}_{c\sim\rho}\max_{c^-:c^-\neq c}\|\mathbf{v_c}-\mathbf{v}_{c^-}\|^2/2\\ &\leq 1/t^2-\mathbb{E}_{c\sim\rho}\max_{c^-:c^-\neq c}\left\|\mathbf{v_c^*}-\mathbf{v}^*_{c^-}\right\|^2/2+(1+1/t)\sqrt{\varepsilon}.\end{aligned}$$

Subtracting $1/t^2$ and multiplying by -2 yields the result.

□

B Graph Representation Learning

We describe in detail the hard sampling method for graphs whose results are reported in Section 5.2. Before getting that point, in the interests of completeness we cover some required background details on the InfoGraph method of Sun et al. (2020). For further information see the original paper (Sun et al., 2020).

B.1 BACKGROUND ON GRAPH REPRESENTATIONS

We observe a set of graphs $\mathbf{G} = \{G_j \in \mathbb{G}\}_{j=1}^n$ sampled according to a distribution p over an ambient graph space $\mathbb{G}$. Each node u in a graph G is assumed to have features $h_u^{(0)}$ living in some Euclidean space. We consider a K-layer graph neural network, whose k-th layer iteratively computes updated embeddings for each node $v \in G$ in the following way,

$$h_v^{(k)} = \text{COMBINE}^{(k)}\left(h_v^{(k-1)}, \text{AGGREGATE}^{(k)}\left(\left\{\left(h_v^{(k-1)}, h_u^{(k-1)}, e_{uv}\right) : u \in \mathcal{N}(v)\right\}\right)\right)$$

where $\text{COMBINE}^{(k)}$ and $\text{AGGREGATE}^{(k)}$ are parameterized learnable functions and $\mathcal{N}(v)$ denotes the set of neighboring nodes of v. The K embeddings for a node u are collected together to obtain a single final summary embedding for u. As recommended by Xu et al. (2019) we use concatenation, $h^u = h^u(G) = \text{CONCAT}\left(\{h_u^{(k)}\}_{k=1}^K\right)$ to obtain an embedding in $\mathbb{R}^d$. Finally, the node representations are combined together into a length d graph level embedding using a readout function,

$$H(G) = \text{READOUT}\left(\{h^u\}_{u \in G}\right)$$

which is typically taken to be a simple permutation invariant function such as the sum or mean. The InfoGraph method aims to maximize the mutual information between the graph level embedding $H(G)$ and patch-level embeddings $h^u(G)$ using the following objective,

$$\max_h \mathbb{E}_{G \sim p} \frac{1}{|G|} \sum_{u \in G} I\left(h^u(G); H(G)\right)$$

In practice the population distribution p is replaced by its empirical counterpart, and the mutual information I is replaced by a variational approximation I_T. In line with Sun et al. (2020) we use the Jensen-Shannon mutual information estimator as formulated by Nowozin et al. (2016). It is defined using a neural network discriminator $T : \mathbb{R}^{2d} \to \mathbb{R}$ as,

$$I_T\left(h^u(G); H(G)\right) = \mathbb{E}_{G \sim p}\left[-\text{sp}(-T\left(h^u(G), H(G)\right))\right] - \mathbb{E}_{(G,G') \sim p \times p}\left[\text{sp}(T\left(h^u(G), H(G')\right))\right]$$

where $\text{sp}(z) = \log(1+e^z)$ denotes the softplus function. The finial objective is the joint maximization over h and T,

$$\max_{\theta, \psi} \mathbb{E}_{G \sim p} \frac{1}{|G|} \sum_{u \in G} I_T\left(h^u(G); H(G)\right)$$

B.2 HARD NEGATIVE SAMPLING FOR LEARNING GRAPH REPRESENTATIONS

In order to derive a simple modification of the NCE hard sampling technique that is appropriate for use with InfoGraph, we first provide a mildly generalized view of hard sampling. Recall that the NCE contrastive objective can be decomposed into two constituent pieces,

$$\mathcal{L}(f, q) = \mathcal{L}_{\text{align}}(f) + \mathcal{L}_{\text{unif}}(f, q)$$

where q is in fact a family of distributions $q(x^-; x)$ over x^- that is indexed by the possible values of the anchor x. $\mathcal{L}_{\text{align}}$ performs the role of "aligning" positive pairs (embedding near to one-another), while $\mathcal{L}_{\text{unif}}$ repels negative pairs. The hard sampling framework aims to solve,

$$\inf_f \sup_q \mathcal{L}(f, q).$$

In the case of NCE loss we take,

$$\mathcal{L}_{\text{align}}(f) = -\mathbb{E}_{\substack{x\sim p \\ x^+\sim p_x^+}} f(x)^T f(x^+),$$
$$\mathcal{L}_{\text{unif}}(f, q) = \mathbb{E}_{\substack{x\sim p \\ x^+\sim p_x^+}} \log \left\{ e^{f(x)^T f(x^+)} + Q\mathbb{E}_{x^-\sim q}[e^{f(x)^T f(x^-)}] \right\}.$$

View this view, we can easily adapt to the InfoGraph framework, taking

$$\mathcal{L}_{\text{align}}(h, T) = -\mathbb{E}_{G\sim p}\frac{1}{|G|}\sum_{u\in G} \text{sp}(-T\left(h^u(G), H(G)\right)),$$
$$\mathcal{L}_{\text{unif}}(h, T, q) = -\mathbb{E}_{G\sim p}\frac{1}{|G|}\sum_{u\in G} \mathbb{E}_{G'\sim q}\text{sp}(T\left(h^u(G), H(G')\right))$$

Denote by $\hat{p}$ the distribution over nodes $u \in \mathbb{R}^s$ defined by first sampling $G \sim p$, then sampling $u \in G$ uniformly over all nodes of G. Then these two terms can be simplified to

$$\mathcal{L}_{\text{align}}(h, T) = -\mathbb{E}_{u\sim\hat{p}}\text{sp}(-T\left(h^u(G), H(G)\right)),$$
$$\mathcal{L}_{\text{unif}}(h, T, q) = -\mathbb{E}_{(u,G')\sim\hat{p}\times q}\text{sp}(T\left(h^u(G), H(G')\right))$$

At this point it becomes clear that, just as with NCE, a distribution $q^* \in \arg\max_q \mathcal{L}(f, q)$ in the InfoGraph framework if it is supported on $\arg\max_{G'\in\mathbb{G}} \text{sp}(T\left(h^u(G), H(G')\right))$. Although this is still hard to compute exactly, it can be approximated by,

$$q_u^\beta(G') \propto \exp\left(\beta T(h^u(G), H(G))\right) \cdot p(G').$$

C Additional Experiments

C.1 Hard negatives with large batch sizes

The vision experiments in the main body of the paper are all based off the SimCLR framework (Chen et al., 2020a). They use a relatively small batch size (up to 512). In order to test whether our hard negatives sampling method can help when the negative batch size is very large, we also run experiments using MoCo-v2 with standard negative memory bank size $N = 65536$ (He et al., 2020; Chen et al., 2020c). We adopt the official MoCo-v2 code[2]. Embeddings are trained for 200 epochs, with batch size 128. Figure 6 summarizes the results. We find that hard negative sampling can still improve the generalization of embeddings trained on CIFAR10: MoCo-v2 attains linear readout accuracy of 88.08%, and MoCo-v2 with hard negatives ($\beta = 0.2$, $\tau^+ = 0$) attains 88.47%.

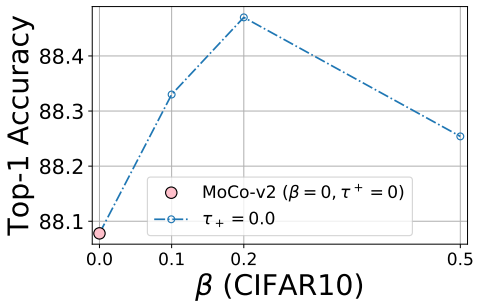

Figure 6: Hard negative sampling using MoCo-v2 framework. Results show that hard negative samples can still be useful when the negative memory bank is very large (in this case $N = 65536$).

C.2 Ablations

[2] https://github.com/facebookresearch/moco

To study the affect of varying the concentration parameter β on the learned embeddings Figure 9 plots cosine similarity histograms of pairs of similar and dissimilar points. The results show that for β moving from 0 through 0.5 to 2 causes both the positive and negative similarities to gradually skew left. In terms of downstream classification, an important property is the *relative* difference in similarity between positive and negative pairs. In this case $\beta = 0.5$ find the best balance (since it achieves the highest downstream accuracy). When β is taken very large ($\beta = 6$), we see a change in conditions. Both positive and negative pairs are assigned higher similarities in general. Visually it seems that the positive and negative histograms for $\beta = 6$ overlap a lot more than for smaller values, which helps explain why the linear readout accuracy is lower for $\beta = 6$.

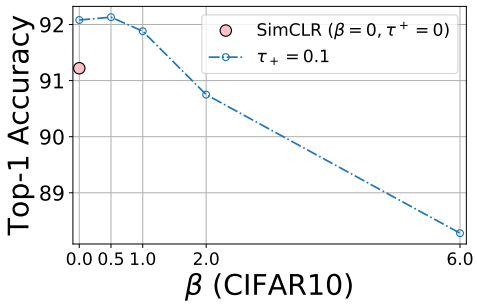

Figure 7: The effect of varying concentration parameter β on linear readout accuracy for CIFAR10. (Complements the left and middle plot from Figure 4.)

Figure 12 gives real examples of hard vs. uniformly sampled negatives. Given an anchor x (a monkey) and trained embedding f (trained on STL10 using standard SimCLR for 400 epochs), we sample a batch of 128 images. The top row shows the ten negatives x^- that have the largest inner product $f(x)^\top f(x^-)$, while the bottom row is a random sample from from the same batch. Negatives with the largest inner product with the anchor correspond to the items in the batch are the most important terms in the objective since they are given the highest weighting by q_β^-. Figure 12 shows that "real" hard negatives are conceptually similar to the idea as proposed in Figure 1: hard negatives are semantically similar to the anchor, possessing various similarities, including color (browns and greens), texture (fur), and objects (animals vs machinery).

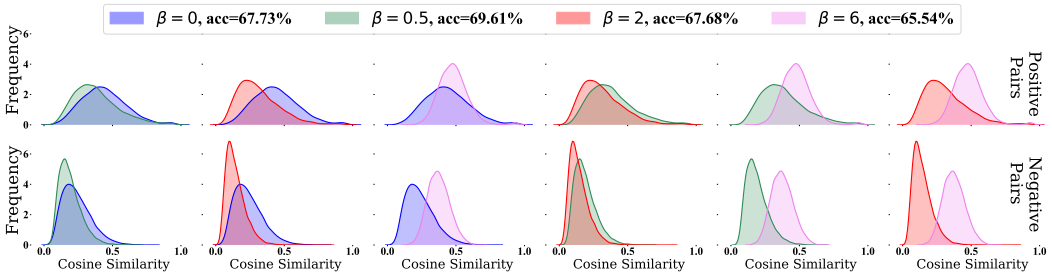

Figure 8: Histograms of cosine similarity of pairs of points with different label (bottom) and same label (top) for embeddings trained on CIFAR100 with different values of β. Histograms overlaid pairwise to allow for easy comparison.

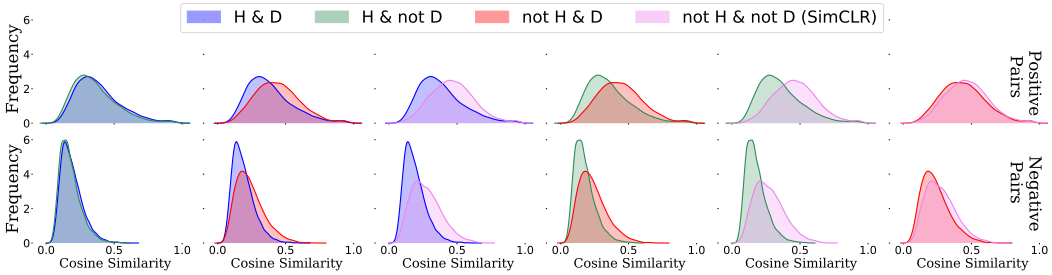

Figure 9: Histograms of cosine similarity of pairs of points with the same label (top) and different labels (bottom) for embeddings trained on CIFAR100 with four different objectives. H=Hard Sampling, D=Debiasing. Histograms overlaid pairwise to allow for convenient comparison.

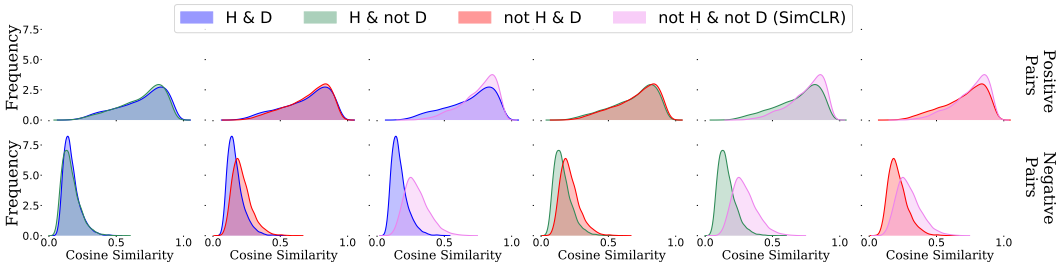

Figure 10: Histograms of cosine similarity of pairs of points with the same label (top) and different labels (bottom) for embeddings trained on CIFAR10 with four different objectives. H=Hard Sampling, D=Debiasing. Histograms overlaid pairwise to allow for convenient comparison.

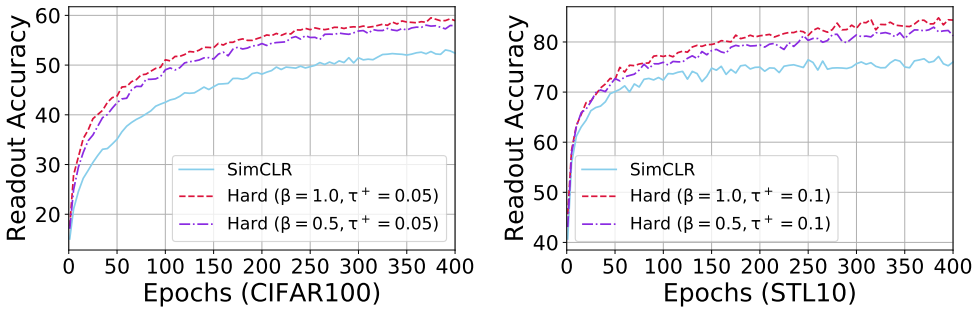

Figure 11: Hard sampling takes much fewer epochs to reach the same accuracy as SimCLR does in 400 epochs; for STL10 with $\beta = 1$ it takes only 60 epochs, and on CIFAR100 it takes 125 epochs (also with $\beta = 1$).

D Experimental Details

Figure 13 shows PyTorch-style pseudocode for the standard objective, the debiased objective, and the hard sampling objective. The proposed hard-sample loss is very simple to implement, requiring only two extra lines of code compared to the standard objective.

D.1 Visual Representations

We implement SimCLR in PyTorch. We use a ResNet-50 (He et al., 2016) as the backbone with embedding dimension 2048 (the representation used for linear readout), and projection head into the lower 128-dimensional space (the embedding used in the contrastive objective). We use the Adam optimizer (Kingma & Ba, 2015) with learning rate 0.001 and weight decay 10^{-6}. Official code will be released. Since we adopt the SimCLR framework, the number of negative samples $N = 2(\text{batch size} - 1)$. Since we always take the batch size to be a power of 2 $(16, 32, 64, 128, 256)$ the negative batch sizes are $30, 62, 126, 254, 510$ respectively.

Annealing β Method: We detail the annealing method whose results are given in Figure 4. The idea is to reduce the concentration parameter down to zero as training progresses. Specifically, suppose we have e number of total training epochs. We also specify a number ℓ of "changes" to the concentration parameter we shall make. We initialize the concentration parameter $\beta_1 = \beta$ (where this β is the number reported in Figure 4), then once every e/ℓ epochs we reduce β_i by β/ℓ. In other words, if we are currently on β_i, then $\beta_{i+1} = \beta_i - \beta/\ell$, and we switch from β_i to β_{i+1} in epoch number $i \cdot e/\ell$. The idea of this method is to select particularly difficult negative samples early on order to obtain useful gradient information early on, but later (once the embedding is already quite good) we reduce the "hardness" level so as to reduce the harmful effect of only approximately correcting for false negatives (negatives with the same labels as the anchor).

We also found the annealing in the opposite direction ("down") achieved similar performance.

Figure 12: Qualitative comparison of hard negatives and uniformly sampled negatives for embedding trained on STL10 for 400 epochs using SimCLR. Top row: selecting the 10 images with highest inner product with anchor in latent space from a batch of 128 inputs. Bottom row: a set of random samples from the same batch. Hard negatives are semantically much more similar to the anchor than uniformly sampled negatives - hard negatives possess many similar characteristics to the anchor, including texture, colors, animals vs machinery.

```
# pos      : exp of inner products for positive examples
# neg      : exp of inner products for negative examples
# N        : number of negative examples
# t        : temperature scaling
# tau_plus: class probability
# beta     : concentration parameter

#Original objective
standard_loss = -log(pos.sum() / (pos.sum() + neg.sum()))

#Debiased objective
Neg = max((-N*tau_plus*pos + neg).sum() / (1-tau_plus), e**(-1/t))
debiased_loss = -log(pos.sum() / (pos.sum() + Neg))

#Hard sampling objective (Ours)
reweight = (beta*neg) / neg.mean()
Neg = max((-N*tau_plus*pos + reweight*neg).sum() / (1-tau_plus), e**(-1/t))
hard_loss = -log( pos.sum() / (pos.sum() + Neg))
```

Figure 13: Pseudocode for our proposed new hard sample objective, as well as the original NCE contrastive objective, and debiased contrastive objective. In each case we take the number of positive samples to be $M = 1$. The implementation of our hard sampling method only requires two additional lines of code compared to the standard objective.

Bias-variance of empirical estimates in hard-negative objective: Recall the final hard negative samples objective we derive is,

$$\mathop{\mathbb{E}}_{\substack{x\sim p\\ x^+\sim p_x^+}}\left[-\log\frac{e^{f(x)^T f(x^+)}}{e^{f(x)^T f(x^+)}+\frac{Q}{\tau^-}\left(\mathbb{E}_{x^-\sim q_\beta}\left[e^{f(x)^T f(x^-)}\right]-\tau^+\mathbb{E}_{v\sim q_\beta^+}\left[e^{f(x)^T f(v)}\right]\right)}\right]. \tag{15}$$

This objective admits a practical counterpart by using empirical approximations to $\mathbb{E}_{x^-\sim q_\beta}[e^{f(x)^T f(x^-)}]$ and $\mathbb{E}_{v\sim q_\beta^+}[e^{f(x)^T f(v)}]$. In practice we use a fairly large number of samples (e.g. $N = 510$) to approximate the first expectation, and only $M = 1$ samples to approximate the second. Clearly in both cases the resulting estimator is unbiased. Further, since the first expectation is approximated using many samples, and the integrand is bounded, the resulting estimator is well concentrated (e.g. apply Hoeffding's inequality out-of-the-box). But what about the second expectation? This might seem uncontrolled since we use only one sample, however it turns out that the random variable $X = e^{f(x)^T f(v)}$ where $x \sim p$ and $v \sim q_\beta^+$ has variance that is bounded by $\mathcal{L}_{\text{align}}(f)$.

Lemma 11. *Consider the random variable $X = e^{f(x)^T f(v)}$ where $x \sim p$ and $v \sim q_\beta^+$. Then $Var(X) \leq \mathcal{O}\big(\mathcal{L}_{align}(f)\big)$.*

Recall that $\mathcal{L}_{\text{align}}(f) = \mathbb{E}_{x,x^+}\|f(x) - f(x^+)\|^2/2$ is termed *alignment*, and Wang & Isola (2020) show that the contrastive objective jointly optimize *alignment* and *uniformity*. Lemma 11 therefore

shows that as training evolves, the variance of the $X = e^{f(x)^T f(v)}$ where $x \sim p$ and $v \sim q_\beta^+$ is bounded by a term that we expect to see becoming small, suggesting that using a single sample ($M = 1$) to approximate this expectation is not unreasonable. We cannot, however, say more than this since we have no guarantee that $\mathcal{L}_{\text{align}}(f)$ goes to zero.

Proof. Fix an x and recall that we are considering $q_\beta^+(\cdot) = q_\beta^+(\cdot; x)$. First let X' be an i.i.d. copy of X, and note that, conditioning on x, we have $2\text{Var}(X|x) = \text{Var}(X|x) + \text{Var}(X'|x) = \text{Var}(X - X'|x) \leq \mathbb{E}\big[(X - X')^2|x\big]$. Bounding this difference,

$$\begin{aligned}
\mathbb{E}\big[(X - X')^2|x\big] &= \mathbb{E}_{v,v' \sim q_\beta^+}\left(e^{f(x)^\top f(v)} - e^{f(x)^\top f(v')}\right)^2 \\
&\leq \mathbb{E}_{v,v' \sim q_\beta^+}\left(e^{1/t^2}\left[f(x)^\top f(v) - f(x)^\top f(v')\right]\right)^2 \\
&\leq e^{1/t^4}\mathbb{E}_{v,v' \sim q_\beta^+}\left(\left[\|f(x)\|\|f(v) - f(v')\|\right]\right)^2 \\
&= \frac{e^{1/t^4}}{t^2}\mathbb{E}_{v,v' \sim q_\beta^+}\|f(v) - f(v')\|^2 \\
&\leq \mathcal{O}\left(\mathbb{E}_{v,v' \sim p^+}\|f(v) - f(v')\|^2\right)
\end{aligned}$$

where the first inequality follows since f lies on the sphere of radius $1/t$, the second inequality by Cauchy–Schwarz, the third again since f lies on the sphere of radius $1/t$, and the fourth since q_β^+ is absolutely continuous with respect to p^+ with bounded ratio.

Since $p^+(x^+) = p(x^+|h(x))$ only depends on $c = h(x)$, rather than x itself, taking expectations over $x \sim p$ is equivalent to taking expectations over $c \sim \rho$. Further, $\rho(c)p(v|c)p(v'|c) = p(v)p(v'|c) = p(v)p_v^+(v')$. So $\mathbb{E}_{c\sim\rho}\mathbb{E}_{v,v'\sim p^+}\|f(v) - f(v')\|^2 = \mathbb{E}_{x,x^+}\|f(x) - f(x^+)\|^2 = 2\mathcal{L}_{\text{align}}(f)$, where $x \sim p$ and $x^+ \sim p_x^+$. Thus we obtain the lemma. □

```
# pos     : exp of inner products for positive examples
# neg     : exp of inner products for negative examples
# N       : number of negative examples
# t       : temperature scaling
# tau_plus: class probability
# beta    : concentration parameter

#Clipping negatives trick before computing reweighting
reweight = 2*neg / max( neg.max().abs(), neg.min().abs() )
reweight = (beta*reweight) / reweight.mean()
Neg = max((-N*tau_plus*pos + reweight*neg).sum() / (1-tau_plus), e**(-1/t))
hard_loss = -log( pos.sum() / (pos.sum() + Neg))
```

Figure 14: In cases where the learned embedding is not normalized to lie on a hypersphere we found that clipping the negatives to live in a fixed range (in this case $[-2, 2]$) stabilizes optimization.

D.2 Graph Representations

All datasets we benchmark on can be downloaded at www.graphlearning.io from the TUDataset repository of graph classification problems (Morris et al., 2020). Information on basic statistics of the datasets is included in Tables 2 and 3. For fair comparison to the original InfoGraph method, we adopt the official code, which can be found at https://github.com/fanyun-sun/InfoGraph. We modify only the `gan_losses.py`

script, adding in our proposed hard sampling via reweighting. For simplicity we trained all models using the same set of hyperparameters: we used the GIN architecture (Xu et al., 2019) with $K = 3$ layers and embedding dimension $d = 32$. Each model is trained for 200 epochs with batch size 128 using the Adam optimizer (Kingma & Ba, 2015). with learning rate 0.001, and weight decay of 10^{-6}. Each embedding is evaluated using the average accuracy 10-fold cross-validation using an SVM as the classifier (in line with the approach taken by Morris et al. (2020)). Each experiment is repeated from scratch 10 times, and the distribution of results from these 10 runs is plotted in Figure 3.

Since the graph embeddings are not constrained to lie on a hypersphere, for a batch we clip all the inner products to live in the interval $[-2, 2]$ while computing the reweighting, as illustrated in Figure 14. We found this to be important for stabilizing optimization.

Dataset	**DD**	**PTC**	**REDDIT-B**	**PROTEINS**
No. graphs	1178	344	2000	1113
No. classes	2	2	2	2
Avg. nodes	284.32	14.29	429.63	39.06
Avg. Edges	715.66	14.69	497.75	72.82

Table 2: Basic statistics for graph datasets.

Dataset	**ENZYMES**	**MUTAG**	**IMDB-B**	**IMDB-M**
No. graphs	600	188	1000	1500
No. classes	6	2	2	3
Avg. nodes	32.63	17.93	19.77	13.00
Avg. Edges	62.14	19.79	96.53	65.94

Table 3: Basic statistics for graph datasets.

D.3 Sentence Representations

We adopt the official quick-thoughts vectors experimental settings, which can be found at `https://github.com/lajanugen/S2V`. We keep all hyperparameters at the default values and change only the `s2v-model.py` script. Since the official BookCorpus dataset Kiros et al. (2015) is not available, we use an unofficial version obtained using the following repository: `https://github.com/soskek/bookcorpus`. Since the sentence embeddings are also not constrained to lie on a hypersphere, we use the same clipping trick as for the graph embeddings, illustrated in Figure 14.

After training on the BookCorpus dataset, we evaluate the embeddings on six different classification tasks: paraphrase identification (MSRP) (Dolan et al., 2004), question type classification (TREC) (Voorhees & Harman, 2002), opinion polarity (MPQA) (Wiebe et al., 2005), subjectivity classification (SUBJ) (Pang & Lee, 2004), product reviews (CR) (Hu & Liu, 2004), and sentiment of movie reviews (MR) (Pang & Lee, 2005).

Comparison with Kalantidis et al. (2020): Kalantidis et al. (2020) also consider ways to sample negatives, and propose a mixing strategy for hard negatives, called MoCHi. The main points of difference are: 1) MoCHi considers the benefit of hard negatives, but does not consider the possibility of false negatives (Principle 1), which we found to be valuable. 2) MoCHi introduces three extra hyperparameters, while our method introduces only two (β, τ^+). If we discard Principle 1 (i.e. τ^+) then only β requires tuning. 3) our method introduces zero computational overhead, whereas MoCHi involves a small amount of extra computation.