Supplementary Material:

Dissecting Supervised Constrastive Learning

In this supplementary material, we provide all proofs (from §3 which were omitted in the main part of the manuscript. Results which are restated from the main manuscript have the same numbering, while Definitions, Lemmas, etc., which are only present in the supplementary material have their labels suffixed by an "S". Additionally, restatements are marked in blue.

In the following, we will frequently utilize standard inequalities (e.g., the Jensen inequality, or the Cauchy-Schwarz inequality) and analyzing their equality conditions will be key to get the desired results. In this context, the following notation has proved to be useful: If a symbol appears above an inequality sign in an equation, it denotes that the corresponding equality conditions will be discussed later on and referenced with the corresponding symbol. For example,

$$a \overset{(P)}{\geq} b \ ,$$

denotes $a \geq b$ and equality is attained if and only if the condition (P) is satisfied.

S1. Proofs for Section 3.2

In this section, we will prove Theorem 2 of the main manuscript (restated below). Throughout this section the following objects will appear repeatedly an thus are introduced one-off:

- $h, N, K \in \mathbb{N}$
- $\rho_{\mathcal{Z}} > 0$
- $\mathcal{Z} = \mathbb{S}^{h-1}_{\rho_{\mathcal{Z}}}$
- $\mathcal{Y} = \{1, \dots, K\} = [K]$

Further, we will consider batches $B \in \mathcal{B}$ of an arbitrary but fixed size $b \geq 3$. We additionally assume $|\mathcal{Y}| = K \leq h + 1$.

Theorem 2. *Let $\rho_{\mathcal{Z}} > 0$ and let $\mathcal{Z} = \mathbb{S}^{h-1}_{\rho_{\mathcal{Z}}}$. Further, let $Z = (z_1, \dots, z_N) \in \mathcal{Z}^N$ be an N point configuration with labels $Y = (y_1, \dots, y_N) \in [K]^N$. If the label configuration Y is balanced, it holds that*

$$\begin{aligned} &\mathcal{L}_{\mathrm{SC}}(Z; Y) \\ &\geq \sum_{l=2}^{b} l\, M_l \log\left(l - 1 + (b-l)\exp\left(-\frac{K\rho_{\mathcal{Z}}^2}{K-1}\right)\right) \ , \end{aligned}$$

where

$$M_l = \sum_{y \in \mathcal{Y}} \left|\{B \in \mathcal{B} : \ |B_y| = l\}\right| \ .$$

Equality is attained if and only if the following conditions are satisfied. There are $\zeta_1, \dots, \zeta_K \in \mathbb{R}^h$ such that:

(C1) $\forall n \in [N] : z_n = \zeta_{y_n}$
(C2) $\{\zeta_y\}_y$ *form a $\rho_{\mathcal{Z}}$-sphere-inscribed regular simplex*

S1.1. Definitions

First we will recall the definition of the supervised contrastive (SC) loss and introduce some necessary auxiliary definitions. The SC loss is given by

$$\mathcal{L}_{\mathrm{SC}}(Z;Y) = \sum_{B\in\mathcal{B}} \ell_{\mathrm{SC}}(Z;Y,B) \ , \tag{S1}$$

where $\ell_{\mathrm{SC}}(Z;Y,B)$ is the *batch-wise loss*

$$\ell_{\mathrm{SC}}(Z;Y,B) = -\sum_{\substack{i\in B\\ |B_{y_i}|>1}} \frac{1}{|B_{y_i}|-1} \sum_{j\in B_{y_i}\setminus\{\!\{i\}\!\}} \log\left(\frac{\exp\left(\langle z_i,z_j\rangle\right)}{\sum\limits_{k\in B\setminus\{\!\{i\}\!\}}\exp\left(\langle z_i,z_k\rangle\right)}\right) \ . \tag{S2}$$

We next introduce the *class-specific batch-wise loss*

$$\ell_{\mathrm{SC}}(Z;Y,B,y) = \begin{cases} -\sum\limits_{i\in B_y} \frac{1}{|B_{y_i}|-1} \sum\limits_{j\in B_{y_i}\setminus\{\!\{i\}\!\}} \log\left(\frac{\exp\left(\langle z_i,z_j\rangle\right)}{\sum\limits_{k\in B\setminus\{\!\{i\}\!\}}\exp\left(\langle z_i,z_k\rangle\right)}\right) & \text{if } |B_y|>1 \\ 0 & \text{else .} \end{cases} \ , \tag{S3}$$

This allows us to write

$$\ell_{\mathrm{SC}}(Z;Y,B) = -\sum_{\substack{i\in B\\ |B_{y_i}|>1}} \frac{1}{|B_{y_i}|-1} \sum_{j\in B_{y_i}\setminus\{\!\{i\}\!\}} \log\left(\frac{\exp\left(\langle z_i,z_j\rangle\right)}{\sum\limits_{k\in B\setminus\{\!\{i\}\!\}}\exp\left(\langle z_i,z_k\rangle\right)}\right) \tag{S4}$$

$$= -\sum_{\substack{y\in\mathcal{Y}\\ |B_y|>1}} \sum_{i\in B_y} \frac{1}{|B_{y_i}|-1} \sum_{j\in B_{y_i}\setminus\{\!\{i\}\!\}} \log\left(\frac{\exp\left(\langle z_i,z_j\rangle\right)}{\sum\limits_{k\in B\setminus\{\!\{i\}\!\}}\exp\left(\langle z_i,z_k\rangle\right)}\right) \tag{S5}$$

$$= \sum_{y\in\mathcal{Y}} \ell_{\mathrm{SC}}(Z;Y,B,y) \ , \tag{S6}$$

and so

$$\mathcal{L}_{\mathrm{SC}}(Z;Y) = \sum_{y\in\mathcal{Y}}\sum_{B\in\mathcal{B}} \ell_{\mathrm{SC}}(Z;Y,B,y) \ . \tag{S7}$$

We use the following notation: the multiplicity of an element x in the multiset M is denoted by $\mathrm{m}_M(x)$. Furthermore, we introduce the label configuration of a batch, i.e.,

$$\Upsilon(B) = \{\!\{y_i : \ i\in B\}\!\} \ , \tag{S8}$$

thus $\mathrm{m}_{\Upsilon(B)}(y) = |B_y|$.

For example, if $\mathcal{Y}=\{a,b\}$, $B=\{\!\{1,2,2,5,10\}\!\}$ and $a=y_1=y_2$, $b=y_5=y_{10}$, then $\mathrm{m}_B(2)=2$, $\Upsilon(B)=\{\!\{a,a,a,b,b\}\!\}$ and $\mathrm{m}_{\Upsilon(B)}(a)=3=|\{\!\{1,2,2\}\!\}|=|B_a|$. We will slightly abuse notation (Y is a tuple, not a multiset) and write $\mathrm{m}_Y(y)=\mathrm{m}_{\Upsilon([N])}(y)=|\{\!\{n\in[N]:y_n=y\}\!\}|$. For every batch $B\in\mathcal{B}$ and label $y\in\mathcal{Y}$, we will also write ${B_y}^C := \{\!\{i\in B:\ y_i\neq y\}\!\}$ for the complement of $B_y := \{\!\{i\in B:\ y_i=y\}\!\}$, which was already introduced in the Definition 2 of the supervised contrastive loss.

Definition 4 (Auxiliary functions S, S_{rep}, S_{att})**.** *Let $h,N\in\mathbb{N}$, $\rho_{\mathcal{Z}}>0$ and $\mathcal{Z}=\mathbb{S}^{h-1}_{\rho_{\mathcal{Z}}}$. For fixed label configuration $Y\in\mathcal{Y}^N$, batch $B\in\mathcal{B}$ and label $y\in\mathcal{Y}$ with $\mathrm{m}_{\Upsilon(B)}(y)>1$, we define*

$$S(\,\cdot\,;Y,B,y) : \mathcal{Z}^N \to \mathbb{R} \tag{S9}$$

$$Z \mapsto S_{att}(Z;Y,B,y) + S_{rep}(Z;Y,B,y) \ , \tag{S10}$$

where

$$S_{att}(Z;Y,B,y) = -\frac{1}{|B_y|\,(|B_y|-1)} \sum_{i\in B_y} \sum_{j\in B_y\setminus\{\{i\}\}} \langle z_i, z_j\rangle \tag{S11}$$

$$S_{rep}(Z;Y,B,y) = \begin{cases} \frac{1}{|B_y|\,|B_y{}^C|} \sum_{i\in B_y} \sum_{j\in B_y{}^C} \langle z_i, z_j\rangle & \text{if } \mathrm{m}_{\Upsilon(B)(y)} \neq b \\ 0 & \text{if } \mathrm{m}_{\Upsilon(B)}(y) = b \end{cases} . \tag{S12}$$

Definition 5 (Auxiliary partition of $\mathcal{B}$)**.** *For every $y \in \mathcal{Y}$ and every $l \in \{0,\dots,b\}$, we define*

$$\mathcal{B}_{y,l} := \{B \in \mathcal{B} :\ \mathrm{m}_{\Upsilon(B)}(y) = l\} \ . \tag{S13}$$

S1.2. Proof of Theorem 2

Proof. We first present the main steps of the proof of Theorem 2 and refer to subsequent technical lemmas when needed.

(**Step 1**) For each class $y \in \mathcal{Y}$ and each batch $B \in \mathcal{B}$ with $\mathrm{m}_{\Upsilon(B)}(y) > 1$, the class-specific batch-wise loss $\ell_{\mathrm{SC}}(Z;Y,B,y)$ (see Lemma S1) is bounded from below by

$$\ell_{\mathrm{SC}}(Z;Y,B,y) \geq |B_y| \log\left(|B_y| - 1 + |B_y{}^C| \exp\left(S(Z;Y,B,y)\right)\right) \ . \tag{S14}$$

(**Step 2**) We regroup the addends of the sum $\mathcal{L}_{\mathrm{SC}}(Z;Y)$, i.e.,

$$\mathcal{L}_{\mathrm{SC}}(Z;Y) = \sum_{B\in\mathcal{B}} \sum_{y\in\mathcal{Y}} \ell_{\mathrm{SC}}(Z;Y,B,y) \ , \tag{S15}$$

such that each group is defined by addends requiring $B \in \mathcal{B}_{y,l} = \{B \in \mathcal{B}|\ \mathrm{m}_{\Upsilon(B)}(y) = l\}$. As a result, we can leverage the bound of **(Step 1)** on each group, i.e.,

$$\mathcal{L}_{\mathrm{SC}}(Z;Y) = \sum_{B\in\mathcal{B}} \sum_{y\in\mathcal{Y}} \ell_{\mathrm{SC}}(Z;Y,B,y) \tag{S16}$$

$$= \sum_{l=0}^{b} \sum_{y\in\mathcal{Y}} \sum_{B\in\mathcal{B}_{y,l}} \ell_{\mathrm{SC}}(Z;Y,B,y) \tag{S17}$$

$$\geq \sum_{l=2}^{b} \sum_{y\in\mathcal{Y}} \sum_{B\in\mathcal{B}_{y,l}} l \log\left(l - 1 + (b-l)\exp\left(S(Z;Y,B,y)\right)\right) \ . \tag{S18}$$

Here the sum over the value of l starts at $l = 2$, because $\ell_{\mathrm{SC}}(Z;Y,B,y) = 0$ vanishes for batches $B \in \{\mathcal{B}_{y,0}, \mathcal{B}_{y,1}\}$.

(**Step 3**) Applying Jensen's inequality (see Lemma S2), then yields

$$\mathcal{L}_{\mathrm{SC}}(Z;Y) \geq \sum_{l=2}^{b} l M_l \log\left(l - 1 + (b-l)\exp\left(\frac{1}{M_l} \sum_{y\in\mathcal{Y}} \sum_{B\in\mathcal{B}_{y,l}} S(Z;Y,B,y)\right)\right) \ , \tag{S19}$$

where $M_l = \sum_{y\in\mathcal{Y}} |\mathcal{B}_{y,l}|$.

(**Step 4**) In Lemma S3, we characterize the equality case of the bound above. It is achieved if and only if all intra-class and inter-class inner products agree, respectively.

The next steps investigate, for each $l \in \{2,\dots,b-1\}$, the sum

$$\sum_{y\in\mathcal{Y}} \sum_{B\in\mathcal{B}_{y,l}} S(Z;Y,B,y) = \left(\sum_{y\in\mathcal{Y}} \sum_{B\in\mathcal{B}_{y,l}} S_{\text{att}}(Z;Y,B,y)\right) + \left(\sum_{y\in\mathcal{Y}} \sum_{B\in\mathcal{B}_{y,l}} S_{\text{rep}}(Z;Y,B,y)\right) \ . \tag{S20}$$

(**Step 5**) The sum of the attraction terms, S_{att}, is maximal if and only if all intra-class inner products are maximal, i.e., they are equal to ${\rho_{\mathcal{Z}}}^2$ (see Lemma S4). This implies

$$\sum_{y\in\mathcal{Y}}\sum_{B\in\mathcal{B}_{y,l}} S_{\text{att}}(Z;Y,B,y) \geq -M_l\,{\rho_{\mathcal{Z}}}^2 \ . \tag{S21}$$

(**Step 6**) If $|\mathcal{Y}|>2$, then the trivial bound on the repulsion term (inner products $=-{\rho_{\mathcal{Z}}}^2$) is not tight as this could only be achieved if the classes were on opposite poles of the sphere. Thus we need an additional step: instead of summing the repulsion terms, $S_{\text{rep}}(Z;Y,B,y)$, over all labels and batches, as done in **(Step 4)**, we re-write this summation as a sum over pairs of indices $(n,m)\in[N]^2$ of different classes $y_n\neq y_m$ (see Lemma S5). That is,

$$\sum_{y\in\mathcal{Y}}\sum_{B\in\mathcal{B}_{y,l}} S_{\text{rep}}(Z;Y,B,y) = \sum_{y\in\mathcal{Y}}\sum_{\substack{n\in[N]\\ y_n=y}}\sum_{\substack{m\in[N]\\ y_m\neq y}} K_{n,m}(y,l)\langle z_n,z_m\rangle \ , \tag{S22}$$

where $K_{n,m}(y,l) := \frac{1}{l(b-l)}\sum_{B\in\mathcal{B}_{y,l}} \mathrm{m}_{B_y}(n)\,\mathrm{m}_{B_y{}^C}(m)$.

(**Step 7**) As we assume the label configuration Y to be balanced, we get that

$$K_{n,m}(y,l) = \frac{M_l}{N^2}\frac{|\mathcal{Y}|}{|\mathcal{Y}|-1} \ , \tag{S23}$$

which only depends on l (and not on y), see Lemma S7 and Eq. (S98). Thus, it suffices to bound

$$\sum_{y\in\mathcal{Y}}\sum_{\substack{n\in[N]\\ y_n=y}}\sum_{\substack{m\in[N]\\ y_n\neq y}}\langle z_n,z_m\rangle \geq -{\rho_{\mathcal{Z}}}^2\sum_{y\in\mathcal{Y}}\mathrm{m}_Y(y)^2 = -\frac{N^2}{|\mathcal{Y}|}{\rho_{\mathcal{Z}}}^2 \ , \tag{S24}$$

where equality is attained if and only if (a) $\sum_{n\in[N]} z_n = 0$, and (b) $y_n=y_m \Rightarrow z_n=z_m$ (see Lemma S8).

(**Step 8**) Finally, we combine all results from (Steps 1-7) and obtain the asserted lower bound (see Lemma S11), i.e.,

$$\mathcal{L}_{\text{SC}}(Z;Y) \geq \sum_{l=2}^{b} M_l l \log\left(l-1+(b-l)\exp\left(-\frac{1}{M_l}\left(\frac{M_l}{N^2}\frac{|\mathcal{Y}|}{|\mathcal{Y}|-1}\frac{N^2}{|\mathcal{Y}|}+M_l\right){\rho_{\mathcal{Z}}}^2\right)\right) \tag{S25}$$

$$= \sum_{l=2}^{b} M_l l \log\left(l-1+(b-l)\exp\left(-\frac{|\mathcal{Y}|}{|\mathcal{Y}|-1}{\rho_{\mathcal{Z}}}^2\right)\right) \ , \tag{S26}$$

where equality is attained if and only if all instances z_n with equal label y_n collapse to a vertex ζ_{y_n} of a regular simplex, inscribed in a hypersphere of radius $\rho_{\mathcal{Z}}$, i.e., conditions (C1) and (C2) in Theorem 2.

□

S1.3. Technical lemmas

In the following, we provide proofs for all technical lemmas invoked throughout Steps 1-8 in the proof of Theorem 2.

Lemma S1. *Fix a class $y \in \mathcal{Y}$ and a batch $B \in \mathcal{B}$ with $\mathrm{m}_{\Upsilon(B)}(y) \in \{2, \dots, b\}$. For every $Z \in \mathcal{Z}^N$ and every $Y \in \mathcal{Y}^N$, the class-specific batch-wise loss $\ell_{\mathrm{SC}}(Z; Y, B, y)$ is bounded from below by*

$$\ell_{\mathrm{SC}}(Z; Y, B, y) \geq \mathrm{m}_{\Upsilon(B)}(y) \log\left(\mathrm{m}_{\Upsilon(B)}(y) - 1 + (b - \mathrm{m}_{\Upsilon(B)}(y)) \exp\left(S(Z; Y, B, y)\right)\right) \quad , \tag{S27}$$

where equality is attained if and only if all of the following hold:

(Q1) $\forall i \in B$ there is a $C_i(B, y)$ such that $\forall j \in B_y \setminus \{\!\{i\}\!\}$ all inner products $\langle z_i, z_j \rangle = C_i(B, y)$ are equal.

(Q2) $\forall i \in B$ there is a $D_i(B, y)$ such that $\forall j \in {B_y}^C$ all inner products $\langle z_i, z_j \rangle = D_i(B, y)$ are equal.

Proof. The lemma follows from an application of Jensen's inequality. In particular, we first need to bring the class-specific batch-wise loss in a form amenable to Jensen's inequality. Since $\mathrm{m}_{\Upsilon(B)}(y) > 1$, we have that

$$\ell_{\mathrm{SC}}(Z; Y, B, y) = -\frac{1}{|B_{y_i}| - 1} \sum_{j \in B_{y_i} \setminus \{\!\{i\}\!\}} \log\left(\frac{\exp\left(\langle z_i, z_j \rangle\right)}{\sum\limits_{k \in B \setminus \{\!\{i\}\!\}} \exp\left(\langle z_i, z_k \rangle\right)}\right) \tag{S28}$$

$$= \sum_{i \in B_y} \log\left(\frac{\sum\limits_{k \in B \setminus \{\!\{i\}\!\}} \exp\left(\langle z_i, z_k \rangle\right)}{\prod\limits_{j \in B_{y_i} \setminus \{\!\{i\}\!\}} \exp\left(\langle z_i, z_j \rangle\right)^{1/|B_{y_i}| - 1}}\right) \tag{S29}$$

$$= \sum_{i \in B_y} \log\left(\frac{\sum\limits_{k \in B \setminus \{\!\{i\}\!\}} \exp\left(\langle z_i, z_k \rangle\right)}{\exp\left(\frac{1}{|B_y \setminus \{\!\{i\}\!\}|} \sum\limits_{j \in B_{y_i} \setminus \{\!\{i\}\!\}} \langle z_i, z_j \rangle\right)}\right) \quad . \tag{S30}$$

We will focus on the sum in the numerator: For every $i \in B_y$, we write

$$\sum_{k \in B \setminus \{\!\{i\}\!\}} \exp(\langle z_i, z_k \rangle) = \sum_{k \in B_y \setminus \{\!\{i\}\!\}} \exp(\langle z_i, z_k \rangle) + \sum_{k \in {B_y}^C} \exp(\langle z_i, z_k \rangle) \quad . \tag{S31}$$

First, assume that $\mathrm{m}_{\Upsilon(B)}(y) \neq b$. As the exponential function is convex, we can leverage Jensen's inequality on both sums, resulting in

$$\sum_{k \in B_y \setminus \{\!\{i\}\!\}} \exp(\langle z_i, z_k \rangle) \overset{(Q1)}{\geq} |B_y \setminus \{\!\{i\}\!\}| \exp\left(\frac{1}{|B_y \setminus \{\!\{i\}\!\}|} \sum_{k \in B_y \setminus \{\!\{i\}\!\}} \langle z_i, z_k \rangle\right) \tag{S32}$$

$$\sum_{k \in {B_y}^C} \exp(\langle z_i, z_k \rangle) \overset{(Q2)}{\geq} |{B_y}^C| \exp\left(\frac{1}{|{B_y}^C|} \sum_{k \in {B_y}^C} \langle z_i, z_k \rangle\right) \quad . \tag{S33}$$

Herein, equality is attained if and only if

(Q1) There is a $C_i(B, y)$ such that $\forall j \in B_y \setminus \{\!\{i\}\!\}$ all inner products $\langle z_i, z_j \rangle = C_i(B, y)$ are equal.

(Q2) There is a $D_i(B, y)$ such that $\forall j \in B_y \setminus \{\!\{i\}\!\}$ all inner products $\langle z_i, z_j \rangle = D_i(B, y)$ are equal.

Thus, using $\exp(a)/\exp(b) = \exp(a-b)$, we bound the argument of the $\log$ in Eq. (S30) by

$$\frac{\sum_{k\in B\setminus\{\!\{i\}\!\}} \exp(\langle z_i, z_k\rangle)}{\exp\left(\frac{1}{|B_y\setminus\{\!\{i\}\!\}|}\sum_{j\in B_y\setminus\{\!\{i\}\!\}}\langle z_i, z_j\rangle\right)} \tag{S34}$$

$$\geq |B_y \setminus \{\!\{i\}\!\}| + |{B_y}^C| \exp\left(\underbrace{\frac{1}{|{B_y}^C|}\sum_{k\in {B_y}^C}\langle z_i, z_k\rangle - \frac{1}{|B_y\setminus\{\!\{i\}\!\}|}\sum_{j\in B_y\setminus\{\!\{i\}\!\}}\langle z_i, z_j\rangle}_{S(Z;Y,B,y)}\right) \quad . \tag{S35}$$

Hence, using $|B_y| = \mathrm{m}_B(y)$ and the definition of $S(Z;Y,B,y)$, we obtain the claimed bound

$$\ell_{\mathrm{SC}}(Z;Y,B) \geq \mathrm{m}_{\Upsilon(B)}(y)\log\left(\mathrm{m}_{\Upsilon(B)}(y) - 1 + (b - \mathrm{m}_{\Upsilon(B)}(y))\exp\left(S(Z;Y,B,y)\right)\right) \quad . \tag{S36}$$

Note that equality is attained, if and only if the above conditions hold for every $i \in B_y$. Also, note that the respective constants, $C_i(B,y)$ and $D_i(B,y)$, depend indeed on the batch B and the label y.

The case of $\mathrm{m}_{\Upsilon(B)}(y) = b$ follows from a analogous argument starting from Eq. (S35) under the observation that, in this case, $B_y = B$ and ${B_y}^C = \emptyset$. This leads to the inequality

$$\ell_{\mathrm{SC}}(Z;Y,B) \geq b\log(b-1) \quad , \tag{S37}$$

with equality condition (Q1). Note that the statement of the lemma is phrased such that the results from this case are automatically included. □

Lemma S2. *Let $l \in \{2,\dots,b\}$. For every $Y \in \mathcal{Y}^N$ and every $Z \in \mathcal{Z}^N$, we have that*

$$\frac{1}{M_l}\sum_{y\in\mathcal{Y}}\sum_{B\in\mathcal{B}_{y,l}}\log\left(l - 1 + (b-l)\exp\left(S(Z;Y,B,y)\right)\right) \tag{S38}$$

$$\geq \log\left(l - 1 + (b-l)\exp\left(\frac{1}{M_l}\sum_{y\in\mathcal{Y}}\sum_{B\in\mathcal{B}_{y,l}} S(Z;Y,B,y)\right)\right) \quad ,$$

where $M_l = \sum_{y\in\mathcal{Y}}|\mathcal{B}_{y,l}|$ and equality is attained if and only if:

(Q3) $l = b$ or there is a constant $D(l)$ such that for every $y \in \mathcal{Y}$ and for every $B \in \mathcal{B}_{y,l}$ the values of $S(Z;Y,B,y) = D(l)$ agree.

Proof. Let $\alpha, \beta > 0$ and $f : \mathbb{R} \to \mathbb{R}$, $x \mapsto \log(\alpha + \beta\exp(x))$. The function f is smooth with second derivative $f''(x) = \frac{\alpha\beta e^x}{(\alpha+\beta e^x)^2} > 0$ and therefore it is convex. Thus, by Jensen' inequality, for every finite sequence $(x_{B,y})_{(y\in\mathcal{Y}, B\in\mathcal{B}_{y,l})}$ it holds that

$$\frac{1}{\sum_{y\in\mathcal{Y}}|\mathcal{B}_{y,l}|}\sum_{y\in\mathcal{Y}}\sum_{B\in\mathcal{B}_{y,l}} f(x_{B,y}) \overset{(Q3)}{\geq} f\left(\frac{1}{\sum_{y\in\mathcal{Y}}|\mathcal{B}_{y,l}|}\sum_{y\in\mathcal{Y}}\sum_{B\in\mathcal{B}_{y,l}} x_{B,y}\right) \quad . \tag{S39}$$

Setting $\alpha = l-1$, $\beta = b-l$ and $x_{B,y} = S(Z;Y,B,y)$, we obtain the bound from the statement of the lemma. Furthermore, equality is attained if and only if:

(Q3) There is a constant $D(l)$ such that for every $y \in \mathcal{Y}$ and for every $B \in \mathcal{B}_{y,l}$ the values of $S(Z;Y,B,y) = D(l)$ agree.

□

Next, we combine Lemma S1 with Lemma S2, which implies a bound with more tangible equality conditions.

Lemma S3. *For every $Y \in \mathcal{Y}^N$ and every $Z \in \mathcal{Z}^N$ the supervised contrastive loss $\mathcal{L}_{\mathrm{SC}}$ is bounded from below by*

$$\mathcal{L}_{\mathrm{SC}}(Z;Y) \geq \sum_{l=2}^{b} l\, M_l \log\left(l - 1 + (b-l)\exp\left(\frac{1}{M_l}\sum_{y\in\mathcal{Y}}\sum_{B\in\mathcal{B}_{y,l}} S(Z;Y,B,y)\right)\right) \quad , \tag{S40}$$

where M_l is defined as in Lemma S2 and equality is attained if and only if:

(A1) There exists a constant α, such that $\forall n, m \in [N]$, $y_n = y_m$ implies $\langle z_n, z_m\rangle = \alpha$.
(A2) There exists a constant β, such that $\forall n, m \in [N]$, $y_n \neq y_m$ implies $\langle z_n, z_m\rangle = \beta$.

Proof. First, observe that $\ell_{\mathrm{SC}}(Z;Y,B,y) = 0$ if $B \in \{\mathcal{B}_{y,0}, \mathcal{B}_{y,1}\}$. Leveraging Lemma S1 and Lemma S2, we get

$$\begin{aligned}
\mathcal{L}_{\mathrm{SC}}(Z;Y) &= \sum_{B\in\mathcal{B}}\sum_{y\in\mathcal{Y}} \ell_{\mathrm{SC}}(Z;Y,B,y) && \text{(S41)}\\
&= \sum_{l=2}^{b}\sum_{y\in\mathcal{Y}}\sum_{B\in\mathcal{B}_{y,l}} \ell_{\mathrm{SC}}(Z;Y,B,y) && \text{(S42)}\\
&\overset{\text{Lem. S1}}{\geq} \sum_{l=2}^{b}\sum_{y\in\mathcal{Y}}\sum_{\mathcal{B}_{y,l}} l\, \log\left(l - 1 + (b-l)\exp\left(S(Z;Y,B,y)\right)\right) && \text{(S43)}\\
&\overset{\text{Lem. S2}}{\geq} \sum_{l=2}^{b} l\, M_l \log\left(l - 1 + (b-l)\exp\left(\frac{1}{M_l}\sum_{y\in\mathcal{Y}}\sum_{B\in\mathcal{B}_{y,l}} S(Z;Y,B,y)\right)\right) \quad . && \text{(S44)}
\end{aligned}$$

Here equality is attained if and only if all of the following conditions hold:

(Q1) $\forall l \in \{2,\dots,b\}$, $\forall y \in \mathcal{Y}$, $\forall B \in \mathcal{B}_{y,l}$ and $\forall i \in B$ there is a $C_i(B,y)$ such that $\forall j \in B_y \setminus \{\!\{i\}\!\}$ all inner products $\langle z_i, z_j\rangle = C_i(B,y)$ are equal.
(Q2) $\forall l \in \{2,\dots,b\}$, $\forall y \in \mathcal{Y}$, $\forall B \in \mathcal{B}_{y,l}$ and $\forall i \in B$ there is a $D_i(B,y)$ such that $\forall j \in {B_y}^C$ all inner products $\langle z_i, z_j\rangle = D_i(B,y)$ are equal.
(Q3) $\forall l \in \{2,\dots,b-1\}$, there is a constant $D(l)$ such that for every $y \in \mathcal{Y}$ and for every $B \in \mathcal{B}_{y,l}$ the values of $S(Z;Y,B,y) = D(l)$ agree.

It remains to show that (Q1) & (Q2) & (Q3) is equivalent to:

(A1) There exists a constant α, such that $\forall n, m \in [N]$, $y_n = y_m$ implies $\langle z_n, z_m\rangle = \alpha$.
(A2) There exists a constant β, such that $\forall n, m \in [N]$, $y_n \neq y_m$ implies $\langle z_n, z_m\rangle = \beta$.

Recall the definition of the auxiliary function S, i.e

$$S(Z;Y,B,y) = S_{\text{att}}(Z;Y,B,y) + S_{\text{rep}}(Z;Y,B,y) \quad \text{, where} \tag{S45}$$

$$S_{\text{att}}(Z;Y,B,y) = -\frac{1}{|B_y|\,(|B_y|-1)}\sum_{i\in B_y}\sum_{j\in B_y\setminus\{\!\{i\}\!\}}\langle z_i, z_j\rangle \tag{S46}$$

$$S_{\text{rep}}(Z;Y,B,y) = \frac{1}{|B_y|\,|{B_y}^C|}\sum_{i\in B_y}\sum_{j\in {B_y}^C}\langle z_i, z_j\rangle \quad . \tag{S47}$$

We start with the direction (A1) & (A2) $\implies$ (Q1) & (Q2) & (Q3).

(Q1) Fix $l \in \{2,\dots,b\}$, $y \in Y$, $B \in \mathcal{B}_{y,l}$ and $i \in B$. Let $j \in B_y \setminus \{\!\{i\}\!\}$, i.e., $y_j = y = y_i$. Therefore condition (A1) implies $\langle z_i, z_j\rangle = \alpha$, i.e., condition (Q1) is fulfilled with $C_i(B,y) = \alpha$.

(Q2) Fix $l \in \{2, \dots, b\}$, $y \in Y$, $B \in \mathcal{B}_{y,l}$ and $i \in B$. Let $j \in {B_y}^C$, i.e., $y_j \neq y = y_i$. Therefore condition (A2) implies $\langle z_i, z_j \rangle = \beta$, i.e. condition (Q2) is fulfilled with $D_i(B, y) = \beta$.

(Q3) Fix $l \in \{2, \dots, b-1\}$. Let $y \in \mathcal{Y}$ and $B \in \mathcal{B}_{y,l}$. By condition (A1), $S_{\text{att}}(Z; Y, B, y) = -\alpha$ and by condition (A2), $S_{\text{rep}}(Z; Y, B, y) = \beta$, so $S(Z; Y, B, y) = S_{\text{rep}}(Z; Y, B, y) + S_{\text{att}}(Z; Y, B, y) = \beta - \alpha$ and condition (Q3) is fulfilled with $D(l) = \beta - \alpha$.

Next, we show (Q1) & (Q2) & (Q3) $\implies$ (A1) & (A2).

Let $y, y' \in \mathcal{Y}$ and $n, m, n', m' \in [N]$ with $y_n = y_m = y$ and $y_{n'} = y_{m'} = y'$. For brevity, we write multisets such that the multiplicity of each element is denoted as a superscript, i.e., $\{\!\{n^b\}\!\}$ denotes the multiset $\{\!\{n, \dots, n\}\!\}$ which contains n exactly b times. Recall, that we assume the $b \geq 3$.

(A1) We need to show that $\langle z_n, z_m \rangle = \langle z_{n'}, z_{m'} \rangle$. There are two cases: $y = y'$ and $y \neq y'$.
First, assume $y \neq y'$.
Choose $l = 2$ and pick the batch $B_1 = \{\!\{n, m, (n')^{b-2}\}\!\} \in \mathcal{B}_{y,2}$. Then

$$S(Z; Y, B_1, y) = S_{\text{att}}(Z; Y, B_1, y) + S_{\text{rep}}(Z; Y, B_1, y) = -\langle z_n, z_m \rangle + \frac{1}{2}\langle z_n, z_{n'} \rangle + \frac{1}{2}\langle z_m, z_{n'} \rangle \ .$$

Condition (Q2) implies that $\langle z_n, z_{n'} \rangle = \langle z_m, z_{n'} \rangle$, and so the above simplifies to $S(Z; Y, B_1, y) = -\langle z_n, z_m \rangle + \langle z_n, z_{n'} \rangle$. An analogous argument for the batch $B_2 = \{\!\{n', m', n^{b-2}\}\!\} \in \mathcal{B}_{y',2}$ implies that $S(Z; Y, B_2, y') = -\langle z_{n'}, z_{m'} \rangle + \langle z_n, z_{n'} \rangle$. Finally, by condition (Q3), we have that $S(Z; Y, B_1, y) = S(Z; Y, B_2, y')$ and thus $\langle z_n, z_m \rangle = \langle z_{n'}, z_{m'} \rangle$.

Now, assume $y = y'$.
Let $p \in [N]$ such that $y_p \neq y$. Again, choose $l = 2$ and pick batches $B_1 = \{\!\{n, m, p^{b-2}\}\!\} \in \mathcal{B}_{y,2}$ and $B_2 = \{\!\{n', m', p^{b-2}\}\!\} \in \mathcal{B}_{y,2}$. By the same argument as in the preceding case of $y \neq y'$, we have that $S(Z; Y, B_1, y) = -\langle z_n, z_m \rangle + \langle z_n, z_p \rangle$ and $S(Z; Y, B_2, y) = -\langle z_{n'}, z_{m'} \rangle + \langle z_{n'}, z_p \rangle$. Therefore, condition (Q3) implies that

$$-\langle z_n, z_m \rangle + \langle z_n, z_p \rangle = -\langle z_{n'}, z_{m'} \rangle + \langle z_{n'}, z_p \rangle \ .$$

Now, pick the batch $B_3 = \{\!\{z_n, z_m, p^{b-2}\}\!\}$. From condition (Q2) it follows that $\langle z_n, p \rangle = \langle z_m, p \rangle$ and thus $\langle z_{n'}, z_{m'} \rangle = \langle z_n, z_m \rangle$.

(A2) We need to show that $\langle z_n, z_{n'} \rangle = \langle z_m, z_{m'} \rangle$ if $y \neq y'$.
Choose $l = 2$ and pick the batches $B_1 = \{\!\{n^2, (n')^{b-2}\}\!\} \in \mathcal{B}_{y,2}$ and $B_2 = \{\!\{m^2, (m')^{b-2}\}\!\} \in \mathcal{B}_{y,2}$. We can already assume that condition (A1) holds, so for every batch $B \in \{B_{y,2}\}$, we have that $S_{\text{att}}(Z; Y, B, y) = -\alpha$ and thus

$$S(Z; Y, B, y) = -\alpha + S_{\text{rep}}(Z; Y, B, y) = -\alpha + \frac{1}{2(b-2)} \sum_{i \in B_y} \sum_{j \in {B_y}^C} \langle z_i, z_j \rangle \ . \tag{S48}$$

Therefore, $S(Z; Y, B_1, y) = \langle z_n, z_{n'} \rangle - \alpha$ and $S(Z; Y, B_2, y) = \langle z_m, z_{m'} \rangle - \alpha$. By condition (Q3), we have that $S(Z; Y, B_1, y) = S(Z; Y, B_2, y)$ and so $\langle z_n, z_{n'} \rangle = \langle z_m, z_{m'} \rangle$.

□

In the following, we address the two parts of the sum in the exponent in Eq. (S40), i.e.,

$$\sum_{y \in \mathcal{Y}} \sum_{B \in \mathcal{B}_{y,l}} S(Z; Y, B, y) = \underbrace{\left(\sum_{y \in \mathcal{Y}} \sum_{B \in \mathcal{B}_{y,l}} S_{\text{att}}(Z; Y, B, y) \right)}_{\text{Lem. } S4} + \underbrace{\left(\sum_{y \in \mathcal{Y}} \sum_{B \in \mathcal{B}_{y,l}} S_{\text{rep}}(Z; Y, B, y) \right)}_{\text{Lem. } S9} \ . \tag{S49}$$

While the first summand is handled easily via Lemma S4, handling the second summand needs further considerations, encapsulated in Lemmas S5, S6, S7, S8 and finally combined into Lemma S9.

Lemma S4 (Sum of attraction terms)**.** *Let $l \in \{2, \dots, b\}$ and let $\mathcal{Z} = \mathbb{S}_{\rho_\mathcal{Z}}$. For every $Y \in \mathcal{Y}^N$ and every $Z \in \mathcal{Z}^N$, it holds that*

$$\sum_{y \in \mathcal{Y}} \sum_{B \in \mathcal{B}_{y,l}} S_{att}(Z; Y, B, y) \geq -\left(\sum_{y \in \mathcal{Y}} |\mathcal{B}_{y,l}| \right) {\rho_\mathcal{Z}}^2 \ , \tag{S50}$$

where equality is attained if and only if:

(Q4) For every $n, m \in [N]$, $y_n = y_m$ implies $z_n = z_m$.

Proof. Recall the definition of $S_{\text{att}}(Z; Y, B, y)$ from Eq. (S11):

$$S_{\text{att}}(Z; Y, B, y) = -\frac{1}{|B_y| \, |B_y \setminus \{\!\{i\}\!\}|} \sum_{i \in B_y} \sum_{j \in B_y \setminus \{\!\{i\}\!\}} \langle z_i, z_j \rangle \ . \tag{S51}$$

Using the Cauchy-Schwarz inequality and the assumption that $\mathcal{Z}$ is a hypersphere of radius $\rho_\mathcal{Z}$, $S_{\text{att}}(Z; Y, B, y)$ is bounded from below by

$$S_{\text{att}}(Z; Y, B, y) \overset{(Q4)}{\geq} -\frac{1}{|B_y| \, |B_y \setminus \{\!\{i\}\!\}|} \sum_{i \in B_y} \sum_{j \in B_y \setminus \{\!\{i\}\!\}} \|z_i\| \, \|z_j\| = -{\rho_\mathcal{Z}}^2 \ , \tag{S52}$$

which already implies the bound in the statement of the lemma.

For fixed $l \in \{2, \dots, b\}$, equality is attained if and only if there is equality in the Cauchy-Schwarz inequality. This means, that for every $y \in \mathcal{Y}$, for every $B \in \mathcal{B}_{y,l}$ and for every $i, j \in B_y$ there exists $\lambda \geq 0$, such that $z_i = \lambda z_j$. Since the z_i and z_j are on a hypersphere, this is equivalent to $z_i = z_j$. Furthermore, for each pair of indices $n, m \in [N]$ with equal class $y_n = y_m = y$, there exists a batch $B \in \mathcal{B}_{y,l}$ containing both indices. Hence the equality condition is equivalent to

(Q4) For every $n, m \in [N]$, $y_n = y_m$ implies $z_n = z_m$.

□

Next, we consider the repulsion component. Recall the definition of $S_{\text{rep}}(Z; Y, B, y)$ from Eq. (S12). We want to bound

$$\sum_{y \in \mathcal{Y}} \sum_{B \in \mathcal{B}_{y,l}} S_{\text{rep}}(Z; Y, B, y) = \sum_{y \in \mathcal{Y}} \sum_{B \in \mathcal{B}_{y,l}} \frac{1}{|B_y| \, |{B_y}^C|} \sum_{i \in B_y} \sum_{j \in {B_y}^C} \langle z_i, z_j \rangle \ . \tag{S53}$$

Similarly to the case of the sum of the attraction terms in Lemma S4, we could bound each inner product by $\langle z_i, z_j \rangle \geq -{\rho_\mathcal{Z}}^2$. However, the obtained inequality will not be tight and thus useless for identifying the minimizer of the sum. This is due to the fact that, in this case, equality would be attained if and only if all points $z_n, z_m \in \mathcal{Z}$ of different class $y_n \neq y_m$ were on opposite poles of the sphere. Yet, this is impossible for $|\mathcal{Y}| > 2$, i.e., if there are more than two classes.

Therefore, the argumentation is a bit more complex and we split it in a sequence of lemmas.

Lemma S5. *Let $l \in \{2, \dots, b-1\}$ and let $y \in \mathcal{Y}$. For every $Y \in \mathcal{Y}^N$ and every $Z \in \mathcal{Z}^N$ the following identity holds:*

$$\sum_{B \in \mathcal{B}_{y,l}} S_{rep}(Z; Y, B, y) = \sum_{\substack{n \in [N] \\ y_n = y}} \sum_{\substack{m \in [N] \\ y_m \neq y}} K_{n,m}(y, l) \, \langle z_n, z_m \rangle \ , \tag{S54}$$

where for each $n, m \in [N]$ with $y_n = y$ and $y_m \neq y$ the combinatorial factor $K_{n,m}(y, l)$ is defined by

$$K_{n,m}(y, l) = \frac{1}{l(b-l)} \sum_{B \in \mathcal{B}_{y,l}} \mathrm{m}_{B_y}(n) \, \mathrm{m}_{{B_y}^C}(m) \ . \tag{S55}$$

Proof. The lemma follows from appropriately partitioning the sum:

$$\sum_{B\in\mathcal{B}_{y,l}} S_{\text{rep}}(Z;Y,B,y) = \sum_{B\in\mathcal{B}_{y,l}} \frac{1}{|B_y|\,|B_y{}^C|} \sum_{i\in B_y}\sum_{j\in B_y{}^C} \langle z_i, z_j\rangle \tag{S56}$$

$$= \sum_{n\in[N]}\sum_{m\in[N]} \frac{1}{l(b-l)} \sum_{B\in\mathcal{B}_{y,l}} \sum_{\substack{i\in B_y\\ i=n}} \sum_{\substack{j\in B_y{}^C\\ j=m}} \langle z_i, z_j\rangle \tag{S57}$$

$$= \sum_{\substack{n\in[N]\\ y_n=y}}\sum_{\substack{m\in[N]\\ y_m\neq y}} \langle z_n, z_m\rangle \frac{1}{l(b-l)} \sum_{B\in\mathcal{B}_{y,l}} \left(\sum_{\substack{i\in B_y\\ i=n}} 1\right)\left(\sum_{\substack{j\in B_y{}^C\\ j=m}} 1\right) \tag{S58}$$

$$= \sum_{\substack{n\in[N]\\ y_n=y}}\sum_{\substack{m\in[N]\\ y_m\neq y}} \langle z_n, z_m\rangle \frac{1}{l(b-l)} \sum_{B\in\mathcal{B}_{y,l}} \mathrm{m}_{B_y}(n)\,\mathrm{m}_{B_y^C}(m) \tag{S59}$$

$$= \sum_{\substack{n\in[N]\\ y_n=y}}\sum_{\substack{m\in[N]\\ y_m\neq y}} \langle z_n, z_m\rangle\, K_{n,m}(y,l) \ . \tag{S60}$$

□

In order to address the quantities $K_{n,m}(y,l)$, we will need the combinatorial identities of the subsequent Lemma S6.

Lemma S6. *Let $n, m \in \mathbb{N}$.*

1. *The number of m-multisets over $[n]$ is*

$$\left(\!\binom{n}{m}\!\right) = \binom{n+m-1}{m} \ . \tag{S61}$$

2.
$$\sum_{k=0}^{m} \left(\!\binom{n}{k}\!\right) = \left(\!\binom{n+1}{m}\!\right) \tag{S62}$$

3.
$$\left(\!\binom{n+1}{m}\!\right) = \left(\!\binom{n}{m}\!\right) + \left(\!\binom{n+1}{m-1}\!\right) \tag{S63}$$

4. *Let $m \geq 1$, then*
$$\sum_{k\in[m]} k \left(\!\binom{n-1}{m-k}\!\right) = \left(\!\binom{n+1}{m-1}\!\right) = \frac{m}{n}\left(\!\binom{n}{m}\!\right) \tag{S64}$$

Proof. The first three identities are well known and imply the last one.

Ad (1): Follows from the stars and bars representation of multisets. Therein, every m-mutliset over $[n]$ is uniquely determined by the position of m bars in an $n+m-1$ tuple of stars and bars. Hence, the cardinality of the set of such multisets is the number of m-element subsets of an $n+m-1$-element set, which is given by the binomial coefficient in Eq. (S61). More precisely, the multiplicity of a number $k \in [n]$ in the multiset is encoded by the number of stars between the $(k-1)$-th and the k-th bar. For example, for $n=5$ and $k=4$ the multiset $1,3,4,4$ is represented by $(*||*\,|\,**|)$.

Ad (2): Denote by $\mathcal{P}_{\{\{\,\}\}}(n,m)$ the set of all m-multisets over $[n]$. Then

$$\mathcal{P}_{\{\{\,\}\}}(n+1,m) = \bigsqcup_{k=0}^{m} \{M \in \mathcal{P}_{\{\{\,\}\}}(n+1,m) |\ \mathrm{m}_M(n+1) = k\} \ . \tag{S65}$$

Thinking of a m-multiset over $[n+1]$ containing the element $(n+1)$ exactly k-times as a $m-k$ multiset over $[n]$, we get from Eq. (S61)

$$\left(\!\binom{n+1}{m}\!\right) = |\mathcal{P}_{\{\{\}\}}(n+1,m)| \tag{S66}$$

$$= \sum_{k=0}^{m} |\{M \in \mathcal{P}_{\{\{\}\}}(n+1,m)| \, \mathrm{m}_M(n+1) = k\}| \tag{S67}$$

$$= \sum_{k=0}^{m} \left(\!\binom{n}{m-k}\!\right) = \sum_{k=0}^{m} \left(\!\binom{n}{k}\!\right) \quad . \tag{S68}$$

Ad (3): Follows directly from the previous argument. In particular,

$$\left(\!\binom{n+1}{m}\!\right) \overset{\text{(S62)}}{=} \sum_{k=0}^{m} \left(\!\binom{n}{k}\!\right) = \sum_{k=0}^{m-1} \left(\!\binom{n}{k}\!\right) + \left(\!\binom{n}{m}\!\right) \overset{\text{(S62)}}{=} \left(\!\binom{n+1}{m-1}\!\right) + \left(\!\binom{n}{m}\!\right) \quad . \tag{S69}$$

Ad (4): The second equality is obvious, once both sides are expanded to the level of factorials. For the the first equality, we prove by induction the equivalent formula

$$\sum_{k=0}^{m} (m-k) \left(\!\binom{n-1}{k}\!\right) = \left(\!\binom{n+1}{m-1}\!\right) \quad . \tag{S70}$$

First, consider the case $m = 1$. Then both

$$\sum_{k=0}^{1} (1-k) \left(\!\binom{n-1}{k}\!\right) = (1-0) \left(\!\binom{n-1}{0}\!\right) = 1 \;\text{ and }\; \left(\!\binom{n+1}{m-1}\!\right) = \left(\!\binom{n+1}{0}\!\right) = 1 \;. \tag{S71}$$

Secondly, assume that Eq. (S70) holds for m. We show that it then also holds for $m+1$, i.e.

$$\sum_{k=0}^{m+1} (m+1-k) \left(\!\binom{n-1}{k}\!\right) = \left(\!\binom{n+1}{m}\!\right) \quad . \tag{S72}$$

The proof is a simple application of the previously derived summation identities:

$$\sum_{k=0}^{m+1} (m+1-k) \left(\!\binom{n-1}{k}\!\right) = \underbrace{\sum_{k=0}^{m} (m-k) \left(\!\binom{n-1}{k}\!\right)}_{\text{(S70)}} + \underbrace{\sum_{k=0}^{m} \left(\!\binom{n-1}{k}\!\right)}_{\text{(S62)}} \tag{S73}$$

$$= \left(\!\binom{n+1}{m-1}\!\right) + \left(\!\binom{n}{m}\!\right) \tag{S74}$$

$$\overset{\text{(S63)}}{=} \left(\!\binom{n+1}{m}\!\right) \quad . \tag{S75}$$

□

Lemma S7. *Let $l \in \{1, \dots, b-1\}$, $Y \in \mathcal{Y}^N$ and $y \in \mathcal{Y}$. For every $n, m \in [N]$, the combinatorial factor $K_{n,m}(y,l)$ has value*

$$K_{n,m}(y,l) = \frac{|\mathcal{B}_{y,l}|}{\mathrm{m}_Y(y)(N - \mathrm{m}_Y(y))} \ . \tag{S76}$$

Proof. We have

$$K_{n,m}(y,l) = \frac{1}{l(b-l)} \sum_{B \in \mathcal{B}_{y,l}} \mathrm{m}_{B_y}(n)\, \mathrm{m}_{B_y^C}(m) = \frac{1}{l(b-l)} \sum_{p=1}^{l} \sum_{q=1}^{b-l} \sum_{\substack{B \in \mathcal{B}_{y,l} \\ \mathrm{m}_{B_y}(n)=p \\ \mathrm{m}_{B_y{}^C}(m)=q}} \mathrm{m}_{B_y}(n)\, \mathrm{m}_{B_y^C}(m) \tag{S77}$$

$$= \frac{1}{l(b-l)} \sum_{p=1}^{l} p \sum_{q=1}^{b-l} q \sum_{\substack{B \in \mathcal{B}_{y,l} \\ \mathrm{m}_{B_y}(n)=p \\ \mathrm{m}_{B_y{}^C}(m)=q}} 1 \ . \tag{S78}$$

Therefore, it is crucial to calculate the cardinality $|\{B \in \mathcal{B}_{y,l} : \mathrm{m}_{B_y}(n) = p,\ \mathrm{m}_{B_y{}^C}(m) = q\}|$.

We can think of each batch $B \in \mathcal{B}$ satisfying the condition

$$B \in \{B \in \mathcal{B}_{y,l},\ \mathrm{m}_{B_y}(n) = p,\ \mathrm{m}_{B_y{}^C}(m) = q\}$$

as a disjoint union of multisets $B = C_n \sqcup C_m \sqcup C_y \sqcup C_{y^C}$, where

- C_n is a p-multiset over the singleton $\{n\}$,
- C_m is a q-multiset over the singleton $\{m\}$,
- C_y is a $(l-p)$-set over the multiset $\{i \in [N] \setminus \{n\} |\ y_i = y\}$ of cardinality $\mathrm{m}_Y(y) - 1$ and
- C_m is a $(b-l-q)$-set over the multiset $\{j \in [N] \setminus \{n\} |\ y_j \neq y\}$ of cardinality $N - \mathrm{m}_Y(y) - 1$.

We write $\mathcal{C}_n, \mathcal{C}_m, \mathcal{C}_y$ and $\mathcal{C}_{y^C}$ for the respective sets of multisets. These sets are of cardinalities (see Eq. (S61))

$$|\mathcal{C}_n| = \left(\!\!\binom{1}{p}\!\!\right) = 1, \qquad |\mathcal{C}_y| = \left(\!\!\binom{\mathrm{m}_Y(y) - 1}{l-p}\!\!\right),$$
$$|\mathcal{C}_m| = \left(\!\!\binom{1}{q}\!\!\right) = 1, \qquad |\mathcal{C}_{y^C}| = \left(\!\!\binom{N - \mathrm{m}_Y(y) - 1}{b-l-q}\!\!\right),$$

and so

$$|\{B \in \mathcal{B}_{y,l} : \mathrm{m}_{B_y}(n) = p,\ \mathrm{m}_{B_y{}^C}(m) = q\}| = |\mathcal{C}_n|\,|\mathcal{C}_m|\,|\mathcal{C}_y|\,|\mathcal{C}_{y^C}|$$
$$= \underbrace{\left(\!\!\binom{1}{p}\!\!\right)\left(\!\!\binom{1}{q}\!\!\right)}_{=1} \left(\!\!\binom{\mathrm{m}_Y(y) - 1}{l-p}\!\!\right)\left(\!\!\binom{N - \mathrm{m}_Y(y) - 1}{b-l-q}\!\!\right) \ . \tag{S79}$$

By a similar argument,

$$|\{B \in \mathcal{B}_{y,l}\}| = \left(\!\!\binom{\mathrm{m}_Y(y)}{l}\!\!\right)\left(\!\!\binom{N - \mathrm{m}_Y(y)}{b-l}\!\!\right) \ . \tag{S80}$$

Therefore, the sum from Eq. (S78) simplifies to

$$K_{n,m}(y,l) = \frac{1}{l(b-l)} \sum_{p=1}^{l} p \sum_{q=1}^{b-l} q |\{B \in \mathcal{B}_{y,l},\ \mathrm{m}_{B_y}(n) = p,\ \mathrm{m}_{B_y{}^C}(m) = q\}| \tag{S81}$$

$$= \frac{1}{l(b-l)} \sum_{p=1}^{l} p \left(\!\!\binom{\mathrm{m}_Y(y) - 1}{l-p}\!\!\right) \sum_{q=1}^{b-l} q \left(\!\!\binom{N - \mathrm{m}_Y(y) - 1}{b-l-q}\!\!\right) \ . \tag{S82}$$

Leveraging Eq. (S64), we get the claimed result

$$K_{n,m}(y,l) = \frac{1}{l(b-l)} \frac{l}{\mathrm{m}_Y(y)} \left(\!\!\binom{\mathrm{m}_Y(y)}{l}\!\!\right) \frac{b-l}{N-\mathrm{m}_Y(y)} \left(\!\!\binom{N-\mathrm{m}_Y(y)}{b-l}\!\!\right) \tag{S83}$$

$$= \frac{|\mathcal{B}_{y,l}|}{\mathrm{m}_Y(y)(N-\mathrm{m}_Y(y))} \ . \tag{S84}$$

□

Lemma S8. *Let $\mathcal{Z} = \mathbb{S}_{\rho_{\mathcal{Z}}}$. For every $Z \in \mathcal{Z}^N$ and every $Y \in \mathcal{Y}^N$, we have that*

$$\sum_{y\in\mathcal{Y}} \sum_{\substack{n\in[N]\\ y_n=y}} \sum_{\substack{m\in[N]\\ y_n\neq y}} \langle z_n, z_m\rangle \geq -{\rho_{\mathcal{Z}}}^2 \sum_{y\in\mathcal{Y}} \mathrm{m}_Y(y)^2 \ , \tag{S85}$$

where equality is attained if and only if the following conditions hold:

(Q5) $\sum_{n\in[N]} z_n = 0$.
(Q6) For every $n, m \in [N]$, $y_n = y_m$ implies $z_n = z_m$.

Proof. We first rewrite the sum as

$$\sum_{y\in\mathcal{Y}} \sum_{\substack{n\in[N]\\ y_n=y}} \sum_{\substack{m\in[N]\\ y_n\neq y}} \langle z_n, z_m\rangle = \sum_{y\in\mathcal{Y}} \sum_{\substack{y'\in\mathcal{Y}\\ y\neq y'}} \sum_{\substack{n\in[N]\\ y_n=y}} \sum_{\substack{m\in[N]\\ y_m=y'}} \langle z_n, z_m\rangle \tag{S86}$$

$$= \sum_{y\in\mathcal{Y}} \sum_{y'\in\mathcal{Y}} \sum_{\substack{n\in[N]\\ y_n=y}} \sum_{\substack{m\in[N]\\ y_m=y'}} \langle z_n, z_m\rangle - \sum_{y\in\mathcal{Y}} \sum_{\substack{n\in[N]\\ y_n=y}} \sum_{\substack{m\in[N]\\ y_m=y}} \langle z_n, z_m\rangle \tag{S87}$$

$$= \sum_{n\in[N]} \sum_{m\in[N]} \langle z_n, z_m\rangle - \sum_{y\in\mathcal{Y}} \sum_{\substack{n\in[N]\\ y_n=y}} \sum_{\substack{m\in[N]\\ y_m=y}} \langle z_n, z_m\rangle \tag{S88}$$

$$= \langle \sum_{n\in[N]} z_n, \sum_{n\in[N]} z_n\rangle - \sum_{y\in\mathcal{Y}} \sum_{\substack{n\in[N]\\ y_n=y}} \sum_{\substack{m\in[N]\\ y_m=y}} \langle z_n, z_m\rangle \ , \tag{S89}$$

where, for the last step, we used the linearity of the inner product. Using that the norm is positive-definite and applying the Cauchy-Schwarz inequality, yields the following lower bound:

$$\sum_{y\in\mathcal{Y}} \sum_{\substack{n\in[N]\\ y_n=y}} \sum_{\substack{m\in[N]\\ y_n\neq y}} \langle z_n, z_m\rangle = \left\| \sum_{n\in[N]} z_n \right\|^2 - \sum_{y\in\mathcal{Y}} \sum_{\substack{n\in[N]\\ y_n=y}} \sum_{\substack{m\in[N]\\ y_m=y}} \langle z_n, z_m\rangle \tag{S90}$$

$$\overset{(Q5)}{\geq} 0 - \sum_{y\in\mathcal{Y}} \sum_{\substack{n\in[N]\\ y_n=y}} \sum_{\substack{m\in[N]\\ y_m=y}} \langle z_n, z_m\rangle \tag{S91}$$

$$\overset{(Q6)}{\geq} -\sum_{y\in\mathcal{Y}} \sum_{\substack{n\in[N]\\ y_n=y}} \|z_n\| \sum_{\substack{m\in[N]\\ y_m=y}} \|z_m\| \tag{S92}$$

$$= -\sum_{y\in\mathcal{Y}} \left(\mathrm{m}_Y(y)\rho_{\mathcal{Z}}\right)^2 = -{\rho_{\mathcal{Z}}}^2 \sum_{y\in\mathcal{Y}} \mathrm{m}_Y(y)^2 \tag{S93}$$

Equality is attained if and only if the following conditions hold:

(Q5) $\sum_n z_n = 0$
(Q6) We have equality in all applications of the Cauchy-Schwarz inequality, i.e., for every $y \in \mathcal{Y}$ and every $n, m \in [N]$ with $y_n = y_m = y$ there exists $\lambda(y, n, m) \geq 0$ such that $z_n = \lambda(y, n, m) z_m$. Since $\mathcal{Z}$ is a sphere $\lambda(y, n, m) = 1$ and so the above is equivalent to $y_n = y_m$ implies $z_n = z_m$.

□

Lemma S9 (Sum of repulsion terms)**.** *Let $l \in \{2, \dots, b-1\}$ and let $\mathcal{Z} = \mathbb{S}^{h-1}_{\rho_{\mathcal{Z}}}$. For every $Z \in \mathcal{Z}^N$ and every balanced $Y \in \mathcal{Y}^N$, we have that*

$$\sum_{y \in \mathcal{Y}} \sum_{B \in \mathcal{B}_{y,l}} S_{rep}(Z; Y, B, y) \geq -|\mathcal{B}_{y,l}| \frac{|\mathcal{Y}|}{|\mathcal{Y}|-1} {\rho_{\mathcal{Z}}}^2 \ , \tag{S94}$$

where equality is attained if and only if the conditions (Q5) & (Q6) from Lemma S8 are fulfilled.

Proof. Recall from Lemma S5 that

$$\sum_{B \in \mathcal{B}_{y,l}} S_{\text{rep}}(Z; Y, B, y) = \sum_{\substack{n \in [N] \\ y_n = y}} \sum_{\substack{m \in [N] \\ y_m \neq y}} K_{n,m}(y, l) \langle z_n, z_m \rangle \ , \tag{S95}$$

and from Lemma S7 that

$$K_{n,m}(y, l) = \frac{|\mathcal{B}_{y,l}|}{\mathrm{m}_Y(y)(N - \mathrm{m}_Y(y))} \ . \tag{S96}$$

Therefore,

$$\sum_{y \in \mathcal{Y}} \sum_{B \in \mathcal{B}_{y,l}} S_{\text{rep}}(Z; Y, B, y) \overset{\text{Lem. } S5}{=} \sum_{y \in \mathcal{Y}} \sum_{\substack{n \in [N] \\ y_n = y}} \sum_{\substack{m \in [N] \\ y_m \neq y}} \frac{|\mathcal{B}_{y,l}|}{\mathrm{m}_Y(y)(N - \mathrm{m}_Y(y))} \langle z_n, z_m \rangle \ . \tag{S97}$$

Since Y is balanced, i.e. $\mathrm{m}_Y(y) = {}^N\!/_{|\mathcal{Y}|}$ for every $y \in \mathcal{Y}$, the term

$$\frac{|\mathcal{B}_{y,l}|}{\mathrm{m}_Y(y)(N - \mathrm{m}_Y(y))} = \frac{|\mathcal{B}_{y,l}|}{N^2} \frac{|\mathcal{Y}|^2}{|\mathcal{Y}|-1} \tag{S98}$$

does not depend on the labels y as (1) $|\mathcal{B}_{y,l}|$ is independent from y due to symmetry and (2) so is $\mathrm{m}_Y(y)$. For brevity, we will still write $|\mathcal{B}_{y,l}|$ in the following, but keep in mind that it is constant w.r.t. y. Furthermore, by Lemma S8

$$\sum_{y \in \mathcal{Y}} \sum_{\substack{n \in [N] \\ y_n = y}} \sum_{\substack{m \in [N] \\ y_m \neq y}} \langle z_n, z_m \rangle \geq -{\rho_{\mathcal{Z}}}^2 \sum_{y \in \mathcal{Y}} \mathrm{m}_Y(y)^2 = -\frac{N^2}{|\mathcal{Y}|} {\rho_{\mathcal{Z}}}^2 \ , \tag{S99}$$

where equality is attained if and only if the conditions (Q5) & (Q6) are fulfilled. Therefore, we obtain the claimed bound

$$\sum_{y \in \mathcal{Y}} \sum_{B \in \mathcal{B}_{y,l}} S_{\text{rep}}(Z; Y, B, y) = \frac{|\mathcal{B}_{y,l}|}{N^2} \frac{|\mathcal{Y}|^2}{|\mathcal{Y}|-1} \sum_{y \in \mathcal{Y}} \sum_{\substack{n \in [N] \\ y_n = y}} \sum_{\substack{m \in [N] \\ y_m \neq y}} \langle z_n, z_m \rangle \tag{S100}$$

$$\geq -\frac{|\mathcal{B}_{y,l}|}{N^2} \frac{|\mathcal{Y}|^2}{|\mathcal{Y}|-1} \frac{N^2}{|\mathcal{Y}|} {\rho_{\mathcal{Z}}}^2 \tag{S101}$$

$$= -|\mathcal{B}_{y,l}| \frac{|\mathcal{Y}|}{|\mathcal{Y}|-1} {\rho_{\mathcal{Z}}}^2 \ . \tag{S102}$$

□

As we have lower-bounded the attraction and repulsion components in Lemmas S4 and S9, respectively, the following lemma, bounding the exponent in Eq. (S40) of Lemma S3, is an immediate consequence.

Lemma S10. *Let $l \in \{2, \dots, b-1\}$ and let $\mathcal{Z} = \mathbb{S}_{\rho_\mathcal{Z}}$. For every $Z \in \mathcal{Z}^N$ and every balanced $Y \in \mathcal{Y}^N$, we have that*

$$\frac{1}{M_l} \sum_{y \in \mathcal{Y}} \sum_{B \in \mathcal{B}_{y,l}} S(Z; Y, B, y) \geq \frac{|\mathcal{Y}|}{|\mathcal{Y}| - 1} {\rho_\mathcal{Z}}^2 \ , \tag{S103}$$

where equality is attained if and only if the following conditions hold:

(A3) For every $n, m \in [N]$, $y_n = y_m$ implies $z_n = z_m$.
(A4) $\sum_{n \in [N]} z_n = 0$.

Proof. Since Y is balanced, $|\mathcal{B}_{y,l}|$ does not depend on y, and so

$$M_l = \sum_{y \in Y} |\mathcal{B}_{y,l}| = |\mathcal{Y}||\mathcal{B}_{y,l}| \ . \tag{S104}$$

Leveraging the bounds on the sums of the attraction terms $S_{\text{att}}(Z; Y, B, y)$ and of the repulsion terms $S_{\text{rep}}(Z; Y, B, y)$ from Lemma S4 and Lemma S9, respectively, we get

$$\sum_{y \in \mathcal{Y}} \sum_{B \in \mathcal{B}_{y,l}} S(Z; Y, B, y) = \left(\sum_{y \in \mathcal{Y}} \sum_{B \in \mathcal{B}_{y,l}} S_{\text{att}}(Z; Y, B, y) \right) + \left(\sum_{y \in \mathcal{Y}} \sum_{B \in \mathcal{B}_{y,l}} S_{\text{rep}}(Z; Y, B, y) \right) \tag{S105}$$

$$\geq -|\mathcal{Y}||\mathcal{B}_{y,l}| {\rho_\mathcal{Z}}^2 - |\mathcal{B}_{y,l}| \frac{|\mathcal{Y}|}{|\mathcal{Y}| - 1} {\rho_\mathcal{Z}}^2 \tag{S106}$$

$$= -|\mathcal{Y}||\mathcal{B}_{y,l}| {\rho_\mathcal{Z}}^2 \left(1 + \frac{1}{|\mathcal{Y}| - 1} \right) \tag{S107}$$

$$= -|\mathcal{Y}||\mathcal{B}_{y,l}| {\rho_\mathcal{Z}}^2 \frac{|\mathcal{Y}|}{|\mathcal{Y}| - 1} \ . \tag{S108}$$

This is the bound as stated in the lemma. Herein, equality is attained if and only if equality is attained in Lemma S4 and Lemma S9. Since conditions (Q4) and (Q6) are the same as condition (A3) and, additionally, since condition (Q5) is the same as condition (A4), the lemma follows. □

Lemma S11. *Combining Lemma S3 and Lemma S10 implies that the supervised contrastive loss $\mathcal{L}_{\text{SC}}(Z; Y)$ is bounded from below by*

$$\mathcal{L}_{\text{SC}}(Z; Y) \geq \sum_{l=2}^{b} l M_l \log \left(l - 1 + (b - l) \exp \left(- \frac{|\mathcal{Y}|}{|\mathcal{Y}| - 1} {\rho_\mathcal{Z}}^2 \right) \right) \ , \tag{S109}$$

where equality is attained if and only if there are $\zeta_1, \dots, \zeta_{|\mathcal{Y}|} \in \mathbb{R}^h$ such that the following conditions hold:

(C1) $\forall n \in [N] : z_n = \zeta_{y_n}$
(C2) $\{\zeta_y\}_{y \in \mathcal{Y}}$ form a $\rho_\mathcal{Z}$-sphere-inscribed regular simplex

Proof. We have that

$$\mathcal{L}_{\text{SC}}(Z; Y) \overset{\text{Lem. } S3}{\geq} \sum_{l=2}^{b} l M_l \log \left(l - 1 + (b - l) \exp \left(\frac{1}{M_l} \sum_{y \in \mathcal{Y}} \sum_{\substack{B \in \mathcal{B} \\ \text{m}_{Y(B)}(y) = l}} S(Z; Y, B, y) \right) \right) \tag{S110}$$

$$\overset{\text{Lem. } S10}{\geq} \sum_{l=2}^{b} l M_l \log \left(l - 1 + (b - l) \exp \left(- \frac{|\mathcal{Y}|}{|\mathcal{Y}| - 1} {\rho_\mathcal{Z}}^2 \right) \right) \ . \tag{S111}$$

Equality holds if and only if the equality conditions of Lemma S3 and Lemma S10 are fulfilled, i.e. if and only if:

(A1) There exists a constant α, such that $\forall n, m \in [N]$, $y_n = y_m$ implies $\langle z_n, z_m \rangle = \alpha$.
(A2) There exists a constant β, such that $\forall n, m \in [N]$, $y_n \neq y_m$ implies $\langle z_n, z_m \rangle = \beta$.
(A3) For every $n, m \in [N]$, $y_n = y_m$ implies $z_n = z_m$.
(A4) $\sum_{n \in [N]} z_n = 0$

Note that Lemma S10 does not hold for $l = b$, so the exponent in Eq. (S110) might differ in this case. However, this is irrelevant as, in this case, the factor $(b - l)$ in front of the exponential function vanishes.

To finish the proof, we need to show under the assumption $\mathcal{Z} = \mathbb{S}_{\rho_{\mathcal{Z}}}$, that these conditions are equivalent to that there are $\zeta_1, \ldots, \zeta_{|\mathcal{Y}|}$ such that

(C1) $\forall n \in [N] : z_n = \zeta_{y_n}$ and
(C2) $\{\zeta_y\}_{y \in \mathcal{Y}}$ form a $\rho_{\mathcal{Z}}$-sphere-inscribed regular simplex, i.e.,
  (S1) $\sum_{y \in \mathcal{Y}} \zeta_y = 0$,
  (S2) $\|\zeta_y\| = \rho_{\mathcal{Z}}$ for $y \in \mathcal{Y}$,
  (S3) $\exists d \in \mathbb{R} : d = \langle \zeta_y, \zeta_{y'} \rangle$ for $y, y' \in \mathcal{Y}$ with $y \neq y'$.

Obviously, (A3) $\iff$ (C1), (S2) holds by assumption, (A4) $\iff$ (S1) and (A2) $\implies$ (S3).

Thus it remains only to show that (C1) & (C2) $\implies$ (A1). Let $n, m \in [N]$ such that $y = y_n = y_m$. By condition (C1), $z_n = z_m = \zeta_y$, so by condition (S2), $\langle z_n, z_m \rangle = \|\zeta_y\|^2 = {\rho_{\mathcal{Z}}}^2$, which does not depend on n and m. □

S2. Proofs for Section 3.1

In this section, we will prove Theorem 1 of the main manuscript and its corollaries. First, we recall the main definitions of the paper and introduce an auxiliary function.

Throughout this section the following objects will appear repeatedly an thus are introduced one-off:

- $h, N, K \in \mathbb{N}$
- $\mathcal{Z} = \mathbb{R}^h$
- $\mathcal{Y} = \{1, \dots, K\} = [K]$

We additionally assume $|\mathcal{Y}| = K \leq h + 1$.

Definition 1 (Cross-entropy loss). *Let $\mathcal{Z} \subseteq \mathbb{R}^h$ and let Z be an N point configuration, $Z = (z_1, \dots, z_N) \in \mathcal{Z}^N$, with labels $Y = (y_1, \dots, y_N) \in [K]^N$; let w_y be the y-th row of the linear classifiers weight matrix $W \in \mathbb{R}^{K \times h}$. The cross-entropy loss $\mathcal{L}_{\mathrm{CE}}(\cdot, W; Y) : \mathcal{Z}^N \to \mathbb{R}$ is defined as*

$$Z \mapsto \frac{1}{N} \sum_{n=1}^{N} \ell_{\mathrm{CE}}(Z, W; Y, n) \tag{2}$$

with $\ell_{\mathrm{CE}}(\cdot, W; Y, n) : \mathcal{Z}^N \to \mathbb{R}$ given by

$$\ell_{\mathrm{CE}}(Z, W; Y, n) = -\log \left(\frac{\exp(\langle z_n, w_{y_n} \rangle)}{\sum_{l=1}^{K} \exp(\langle z_n, w_l \rangle)} \right) . \tag{3}$$

Definition 3 (ρ-Sphere-inscribed regular simplex). *Let $h, K \in \mathbb{N}$ with $K \leq h + 1$. We say that $\zeta_1, \dots, \zeta_K \in \mathbb{R}^h$ form the vertices of a regular simplex inscribed in the hypersphere of radius $\rho > 0$, if and only if the following conditions hold:*
(S1) $\sum_{i \in [K]} \zeta_i = 0$
(S2) $\|\zeta_i\| = \rho$ *for* $i \in [K]$
(S3) $\exists d \in \mathbb{R} : d = \langle \zeta_i, \zeta_j \rangle$ *for* $1 \leq i < j \leq K$

Definition S1 (Auxiliary function S)**.** Let $\mathcal{Z} = \mathbb{R}^h$, then we define

$$\begin{aligned} S(\cdot, \cdot; Y) : \mathcal{Z}^N \times \mathcal{Z}^K &\to \mathbb{R} \\ (Z, W) &\mapsto \frac{1}{N} \frac{K}{K-1} \sum_{n \in [N]} \langle z_n, \bar{w} - w_{y_n} \rangle \ , \end{aligned}$$

where $\bar{w} = \frac{1}{|\mathcal{Y}|} \sum_{y \in \mathcal{Y}} w_y$.

Lemma S12. *Let $h, K, N \in \mathbb{N}$, $\mathcal{Z} = \mathbb{R}^h$. Further, let*

$$\begin{aligned} Z &= (z_n)_{i=n}^{N} \in \mathcal{Z}^N, \\ W &= (w_y)_{y=1}^{K} \in \mathcal{Z}^K, \\ Y &= (y_n)_{i=n}^{N} \in \mathcal{Y}^N \ . \end{aligned}$$

It holds that

$$\mathcal{L}_{\mathrm{CE}}(Z, W; Y) \geq \log \Big(1 + (K-1) \exp \big(S(Z, W; Y) \big) \Big) \ ,$$

where S is as in Definition S1. Equality is attained if and only if the following conditions hold:

(P1) $\forall n \in [N]\ \exists M_n$ *such that* $\forall y \in \mathcal{Y} \setminus \{y_n\}$ *all inner products* $\langle z_n, w_y \rangle = M_n$ *are equal.*
(P2) $\exists M \in \mathbb{R}$ *such that* $\forall n \in [N]$ *it holds that* $\sum_{\substack{y \in \mathcal{Y} \\ y \neq y_n}} \left(\langle z_n, w_y \rangle - \langle z_n, w_{y_n} \rangle \right) = M$.

Proof. Using the identities $\log(t) = -\log(1/t)$ and $\exp(a)/\exp(b) = \exp(a-b)$, rewrite the cross-entropy loss in the equivalent form

$$\mathcal{L}_{\mathrm{CE}}(Z, W; Y) = \frac{1}{N} \sum_{n \in [N]} \log \left(1 + \sum_{\substack{y \in \mathcal{Y} \\ y \neq y_n}} \exp\left(\langle z_n, w_y \rangle - \langle z_n, w_{y_n} \rangle \right) \right) \ . \tag{S112}$$

In order to bound $\mathcal{L}_{\mathrm{CE}}$ from below, we apply Jensen's inequality twice; first for the convex function $t \mapsto \exp(t)$ and then for the convex function $t \mapsto \log(1 + \exp(t))$:

$$\mathcal{L}_{\mathrm{CE}}(Z, W; Y) \overset{(P1)}{\geq} \frac{1}{N} \sum_{n \in [N]} \log \left(1 + (K-1) \exp \left(\frac{1}{K-1} \sum_{\substack{y \in \mathcal{Y} \\ y \neq y_n}} \left(\langle z_n, w_y \rangle - \langle z_n, w_{y_n} \rangle \right) \right) \right) \tag{S113}$$

$$\overset{(P2)}{\geq} \log \left(1 + (K-1) \exp \left(\frac{1}{N} \frac{1}{K-1} \sum_{n \in [N]} \sum_{\substack{y \in \mathcal{Y} \\ y \neq y_n}} \left(\langle z_n, w_y \rangle - \langle z_n, w_{y_n} \rangle \right) \right) \right) \ . \tag{S114}$$

By the linearity of the inner product and as the addend for $y = y_n$ equals zero, the exponent in Eq. (S114) is simply $S(Z, W; Y)$, which proves the bound.

The equality condition is obtained from the combination of the equality cases in both applications of Jensen's inequality. These are:

(P1) $\forall n \in [N]\ \exists M_n$ such that $\forall y \in \mathcal{Y} \setminus \{y_n\}$ all inner products $\langle z_n, w_y \rangle = M_n$ are equal.
(P2) $\exists M \in \mathbb{R}$ such that $\forall n \in [N]$ it holds that $\sum_{\substack{y \in \mathcal{Y} \\ y \neq y_n}} \left(\langle z_n, w_y \rangle - \langle z_n, w_{y_n} \rangle \right) = M$.

□

Lemma S13. *Let* $h, K, N \in \mathbb{N}$, $\rho_{\mathcal{Z}} > 0$ *and* $\mathcal{Z} = \{z \in \mathbb{R}^h : \|z\| \leq \rho_{\mathcal{Z}}\}$. *Further, let*

$$\begin{aligned} Z &= (z_n)_{n=1}^N \in \mathcal{Z}^N \ , \\ W &= (w_y)_{y=1}^K \in (\mathbb{R}^h)^K \ , \\ Y &= (y_n)_{n=1}^N \in \mathcal{Y}^N \ . \end{aligned}$$

If the class configuration Y *is balanced, i.e., for all* $y \in \mathcal{Y}$, $N_y = \left| \{i \in [N] : y_i = y\} \right| = {}^N\!/_K$, *then*

$$S(Z, W;\, Y) \geq -\rho_{\mathcal{Z}} \frac{\sqrt{K}}{K-1} \|W\|_F \ ,$$

where $\|\cdot\|_F$ *denotes the Frobenius norm. We get equality if and only if the following conditions hold:*

(P3) $\forall n \in [N]$ *there* $\exists \lambda_n \leq 0$ *such that* $z_n = \lambda_n(\bar{w} - w_{y_n})$
(P4) $\forall n : \|z_n\| = \rho_{\mathcal{Z}}$
(P5) $\forall y \in \mathcal{Y}$ *the terms* $\|\bar{w}\|^2 + \|w_y\|^2 - 2\langle \bar{w}, w_y \rangle$ *are equal*
(P6) $\bar{w} = 0$

Proof. We will bound the function S from Lemma S12, using the norm constraint on each $z_n \in \mathcal{Z}$. In particular,

$$
\begin{aligned}
S(Z, W; Y) &= \frac{1}{N}\frac{K}{K-1}\sum_{n\in[N]} \langle z_n, \bar{w} - w_{y_n}\rangle \\
&\overset{(P3)}{\geq} -\frac{1}{N}\frac{K}{K-1}\sum_{n\in[N]} \|z_n\|\,\|\bar{w} - w_{y_n}\| \\
&\overset{(P4)}{\geq} -\frac{1}{N}\frac{K}{K-1}\sum_{n\in[N]} \rho_{\mathcal{Z}}\,\|\bar{w} - w_{y_n}\| \\
&= -\frac{1}{N}\frac{K}{K-1}\rho_{\mathcal{Z}}\sum_{y\in\mathcal{Y}} \|\bar{w} - w_y\| \left(\sum_{\substack{n\in[N]\\ y_n = y}} 1\right) \\
&= -\frac{1}{N}\frac{K}{K-1}\rho_{\mathcal{Z}}\sum_{y\in\mathcal{Y}} \|\bar{w} - w_y\|\, N_y \\
&= -\frac{1}{N}\frac{K}{K-1}\rho_{\mathcal{Z}}\frac{N}{K}\sum_{y\in\mathcal{Y}} \|\bar{w} - w_y\| \qquad \text{(by assumption } N_y = \tfrac{N}{K}\text{)} \\
&= -\frac{1}{K-1}\rho_{\mathcal{Z}}\sum_{y\in\mathcal{Y}} \sqrt{\|\bar{w}\|^2 + \|w_y\|^2 - 2\langle \bar{w}, w_y\rangle} \\
&\overset{(P5)}{\geq} -\rho_{\mathcal{Z}}\frac{K}{K-1}\sqrt{\frac{1}{K}\sum_{y\in\mathcal{Y}} \left(\|\bar{w}\|^2 + \|w_y\|^2 - 2\langle \bar{w}, w_y\rangle\right)} \\
&= -\rho_{\mathcal{Z}}\frac{K}{K-1}\sqrt{\frac{1}{K}\left(K\,\|\bar{w}\|^2 + \sum_{y\in\mathcal{Y}} \|w_y\|^2 - 2\langle \bar{w}, \sum_{y\in\mathcal{Y}} w_y\rangle\right)} \\
&= -\rho_{\mathcal{Z}}\frac{K}{K-1}\sqrt{\frac{1}{K}\sum_{y\in\mathcal{Y}} \|w_y\|^2 - \|\bar{w}\|^2} \\
&\overset{(P6)}{\geq} -\rho_{\mathcal{Z}}\frac{K}{K-1}\sqrt{\frac{1}{K}\sum_{y\in\mathcal{Y}} \|w_y\|^2} \\
&= -\rho_{\mathcal{Z}}\frac{\sqrt{K}}{K-1}\|W\|_F \ ,
\end{aligned}
$$

where

- (P3) follows from the Cauchy-Schwarz inequality with equality if and only if $\forall n \in [N]$ there $\exists \lambda_n \leq 0$ such that $z_n = \lambda_n(\bar{w} - w_{y_n})$,
- (P4) follows from the assumption on the space $\mathcal{Z}$, with equality if and only if $\forall n$, $\|z_n\| = \rho_{\mathcal{Z}}$ is maximal,
- (P5) follows from Jensen's inequality for the convex function $t \mapsto -\sqrt{t}$ with equality if and only if $\forall y \in \mathcal{Y}$ the terms $\|\bar{w}\|^2 + \|w_y\|^2 - 2\langle \bar{w}, w_y\rangle$ are equal,
- (P6) follows from the positivity of the norm, with equality if and only if $\bar{w} = 0$, i.e. W is centered at the origin.

□

Theorem 1. *Let $\rho_{\mathcal{Z}} > 0$, $\mathcal{Z} = \{z \in \mathbb{R}^h : \|z\| \leq \rho_{\mathcal{Z}}\}$. Further, let $Z = (z_1, \ldots, z_N) \in \mathcal{Z}^N$ be an N point configuration with labels $Y = (y_1, \ldots, y_N) \in [K]^N$ and let $W \in \mathbb{R}^{K \times h}$ be the weight matrix of the linear classifier from Definition 1. If the label configuration Y is balanced,*

$$\mathcal{L}_{\mathrm{CE}}(Z, W; Y) \geq \log\left(1 + (K-1)\exp\left(-\rho_{\mathcal{Z}}\frac{\sqrt{K}}{K-1}\|W\|_F\right)\right) ,$$

holds, with equality if and only if there are $\zeta_1, \ldots, \zeta_K \in \mathbb{R}^h$ such that:

(C1) $\forall n \in [N] : z_n = \zeta_{y_n}$
(C2) $\{\zeta_y\}_y$ *form a $\rho_{\mathcal{Z}}$-sphere-inscribed regular simplex*
(C3) $\exists \rho_{\mathcal{W}} > 0 : \forall y \in \mathcal{Y} : w_y = \frac{\rho_{\mathcal{W}}}{\rho_{\mathcal{Z}}}\zeta_y$

Proof. To prove the bound, we consecutively leverage Lemma S12 and Lemma S13.

$$\mathcal{L}_{\mathrm{CE}}(Z, W; Y) \overset{\text{Lem. } S12}{\geq} \log\left(1 + (K-1)\exp\left(S(Z, W; Y)\right)\right) \overset{\text{Lem. } S13}{\geq} \log\left(1 + (K-1)\exp\left(-\rho_{\mathcal{Z}}\frac{\sqrt{K}}{K-1}\|W\|_F\right)\right) .$$

The application of Lemma S12 and S13 above also yields the sufficient and necessary conditions for equality, which are (P1), (P2), (P3), (P4), (P5) and (P6). It remains to prove that those conditions are equivalent to (C1), (C2), (C3). That is, we need to show that

(P1) $\forall n \in [N]$ $\exists M_n$ such that $\forall y \in \mathcal{Y} \setminus \{y_n\}$ all inner products $\langle z_n, w_y\rangle = M_n$ are equal,
(P2) $\exists M \in \mathbb{R}$ such that $\forall n \in [N]$ it holds that $\sum_{\substack{y \in \mathcal{Y} \\ y \neq y_n}} \langle z_n, w_y\rangle - \langle z_n, w_{y_n}\rangle = M$.
(P3) $\forall n \in [N]$ there $\exists \lambda_n \leq 0$ such that $z_n = \lambda_n(\bar{w} - w_{y_n})$,
(P4) $\forall n : \|z_n\| = \rho_{\mathcal{Z}}$,
(P5) $\forall y \in \mathcal{Y}$ the terms $\|\bar{w}\|^2 + \|w_y\|^2 - 2\langle \bar{w}, w_y\rangle$ are equal,
(P6) $\bar{w} = 0$

are equivalent to that the existence of $\zeta_1, \ldots, \zeta_{|\mathcal{Y}|} \in \mathbb{R}^h$ such that

(C1) $\forall n \in [N] : z_n = \zeta_{y_n}$
(C2) $\{\zeta_y\}_{y \in \mathcal{Y}}$ form a $\rho_{\mathcal{Z}}$-sphere-inscribed regular simplex, i.e., it holds that
  (S1) $\sum_{y \in \mathcal{Y}} \zeta_y = 0$,
  (S2) $\|\zeta_y\| = \rho_{\mathcal{Z}}$ for $y \in \mathcal{Y}$,
  (S3) $\exists d \in \mathbb{R} : d = \langle \zeta_y, \zeta_{y'}\rangle$ for $1 \leq y < y' \leq K$.
(C3) $\exists \rho_{\mathcal{W}} > 0 : \forall y \in \mathcal{Y} : w_y = \frac{\rho_{\mathcal{W}}}{\rho_{\mathcal{Z}}}\zeta_y$.

The arguments for the equivalencies are given below:

First, we show (P1) - (P6) $\implies$ (C1) - (C3):

Ad (C1): We need to show that $\forall n \in [N] : z_n = \zeta_{y_n}$.

Let $n \in [N]$. Conditions (P3) and (P6) yield $z_n = -\lambda_n w_{y_n}$ where $\lambda_n \leq 0$. If $w_{y_n} = 0$, this immediately implies (C1) with $\zeta_{y_n} = 0$. If $w_{y_n} \neq 0$, we have $|\lambda| = \|z_n\| / \|w_{y_n}\|$, and thus by (P4)

$$z_n = -\left(-\frac{\|z_n\|}{\|w_{y_n}\|}\right) w_{y_n} \overset{(P4)}{=} \frac{\rho_{\mathcal{Z}}}{\|w_{y_n}\|} w_{y_n} . \tag{S115}$$

Consequently, condition (C1) is fulfilled with $\zeta_{y_n} = \rho_{\mathcal{Z}} \frac{w_{y_n}}{\|w_{y_n}\|}$.

Ad (C3): We need to show that $\exists \rho_{\mathcal{W}} > 0$ such that $\forall y \in \mathcal{Y}$ we have $w_y = \frac{\rho_{\mathcal{W}}}{\rho_{\mathcal{Z}}} \zeta_y$.

Since Y is balanced, for every label $y \in \mathcal{Y}$ we have that $N_y = N/K > 0$ and so there exists $n \in [N]$ with $y_n = y$. Thus Eq. (S115) implies for every $y \in \mathcal{Y}$ that $\zeta_y = \rho_{\mathcal{Z}} \frac{w_y}{\|w_y\|}$. Hence, condition (C3) is fulfilled with $\rho_{\mathcal{W}} = \|w_y\|$ if all such norms $\|w_y\|$ agree. Indeed, by condition (P5), there is $C \in \mathbb{R}$ such that for each $y \in \mathcal{Y}$

$$C \overset{(P5)}{=} \|\bar{w}\|^2 + \|w_y\|^2 - 2\langle \bar{w}, w_y \rangle \overset{(P6)}{=} 0 + \|w_y\|^2 - 2 \cdot 0 = \|w_y\|^2 \quad .$$

Ad (C2): We need to show that $\{\zeta_y\}_{y \in \mathcal{Y}}$ fulfill the requirements (S1), (S2) and (S3) of the regular simplex from Definition 3.

From condition (C1) and condition (C3), we already know that

$$\frac{\rho_{\mathcal{Z}}}{\rho_{\mathcal{W}}} \cdot w_y = \zeta_y \text{ for } y \in \mathcal{Y} \quad , \tag{S116}$$

which we will use in the following.

Ad (S1): We need to show that $\sum_{y \in \mathcal{Y}} \zeta_y = 0$.

This follows directly from Eq. (S116) and condition (P6), because

$$\sum_{y \in \mathcal{Y}} \zeta_y \overset{\text{Eq. (S116)}}{=} \frac{\rho_{\mathcal{Z}}}{\rho_{\mathcal{W}}} \sum_{y \in \mathcal{Y}} w_y \overset{(P6)}{=} 0 \quad . \tag{S117}$$

Ad (S2): We need to show for every $y \in \mathcal{Y}$ that $\|\zeta_y\| = \rho_{\mathcal{Z}}$.

This follows directly from Eq. (S116) and the already proven condition (C3), because

$$\|\zeta_y\| \overset{\text{Eq. (S116)}}{=} \|\frac{\rho_{\mathcal{Z}}}{\rho_{\mathcal{W}}} \cdot w_{y_n}\| \overset{(C3)}{=} \frac{\rho_{\mathcal{Z}}}{\rho_{\mathcal{W}}} \cdot \rho_{\mathcal{W}} = \rho_{\mathcal{Z}} \quad . \tag{S118}$$

Ad (S3): We need to show that for every $y, y' \in \mathcal{Y}$ with $y \neq y'$ there $\exists d \in \mathbb{R} : d = \langle \zeta_y, \zeta_{y'} \rangle$.

Let $y, y' \in \mathcal{Y}$ with $y \neq y'$. Since Y is balanced, we have that $N_{y'} = {}^{N}/_{K} > 0$. Hence, there exists $n \in [N]$ with $y' = y_n$ and so

$$\frac{\rho_{\mathcal{W}}}{\rho_{\mathcal{Z}}} \langle \zeta_{y'}, \zeta_y \rangle = \langle \zeta_{y_n}, \frac{\rho_{\mathcal{W}}}{\rho_{\mathcal{Z}}} \zeta_y \rangle \overset{\text{Eq. (S116)}}{=} \langle \zeta_{y_n}, w_y \rangle \overset{(C1)}{=} \langle z_n, w_y \rangle \overset{(P1)}{=} M_n \quad . \tag{S119}$$

Similarly,

$$\langle z_n, w_{y_n} \rangle \overset{(C1)}{=} \langle \zeta_{y_n}, w_{y_n} \rangle \overset{\text{Eq. (S116)}}{=} \langle \zeta_{y_n}, \frac{\rho_{\mathcal{W}}}{\rho_{\mathcal{Z}}} \zeta_{y_n} \rangle = \frac{\rho_{\mathcal{W}}}{\rho_{\mathcal{Z}}} \|\zeta_{y_n}\|^2 \overset{(S2)}{=} \rho_{\mathcal{W}} \rho_{\mathcal{Z}} \quad . \tag{S120}$$

We leverage condition (P1) and condition (P2) to get that there exists $M \in \mathbb{R}$ such that

$$M \overset{(P2)}{=} \sum_{\substack{\hat{y} \in \mathcal{Y} \\ \hat{y} \neq y_n}} (\langle z_n, w_{\hat{y}} \rangle - \langle z_n, w_{y_n} \rangle) \tag{S121}$$

$$\overset{\text{(S120)}}{=} \Big(\sum_{\substack{\hat{y} \in \mathcal{Y} \\ \hat{y} \neq y_n}} \langle z_n, w_{\hat{y}} \rangle \Big) - (K-1) \rho_{\mathcal{W}} \rho_{\mathcal{Z}} \tag{S122}$$

$$\overset{(P1)}{=} (K-1)(M_n - \rho_{\mathcal{W}} \rho_{\mathcal{Z}}) \tag{S123}$$

$$\overset{\text{(S119)}}{=} (K-1)(\frac{\rho_{\mathcal{W}}}{\rho_{\mathcal{Z}}} \langle \zeta_{y'}, \zeta_y \rangle - \rho_{\mathcal{W}} \rho_{\mathcal{Z}}) \quad . \tag{S124}$$

Thus $\langle \zeta_{y'}, \zeta_y \rangle = d$ is constant, and d can be calculated by rearranging the equation above.

Next, we show (C1) - (C3) $\implies$ (P1) - (P6) :

We assume that there exist $\zeta_1, \ldots, \zeta_K \in \mathbb{R}^h$ such that conditions (C1) - (C3) are fulfilled.

Ad (P1): We need to show that $\forall n \in [N]\ \exists M_n$ such that $\forall y \in \mathcal{Y} \setminus \{y_n\}$ all inner products $\langle z_n, w_y \rangle = M_n$ are equal.

Let $n \in [N]$ and $y \in \mathcal{Y} \setminus \{y_n\}$, then

$$\langle z_n, w_y \rangle \overset{(C1)}{=} \langle \zeta_{y_n}, w_y \rangle \overset{(C3)}{=} \langle \zeta_{y_n}, \frac{\rho_{\mathcal{W}}}{\rho_{\mathcal{Z}}} \zeta_y \rangle \overset{(S3)}{=} \frac{\rho_{\mathcal{W}}}{\rho_{\mathcal{Z}}} d \ , \tag{S125}$$

so condition (P1) is fulfilled with $M_n = \frac{\rho_{\mathcal{W}}}{\rho_{\mathcal{Z}}} d$.

Ad (P2): We need to show that $\exists M$ such that $\forall n \in [N]$ it holds that $\sum_{\substack{y \in \mathcal{Y} \\ y \neq y_n}} (\langle z_n, w_y \rangle - \langle z_n, w_{y_n} \rangle) = M$. Let $n \in [N]$. From Eq. (S125), we already now that for $y \in \mathcal{Y} \setminus \{y\}$ it holds that $\langle z_n, w_y \rangle = \frac{\rho_{\mathcal{W}}}{\rho_{\mathcal{Z}}} d$. Similarly,

$$\langle z_n, w_{y_n} \rangle \overset{(C1)}{=} \langle \zeta_{y_n}, w_y \rangle \overset{(C3)}{=} \langle \zeta_{y_n}, \frac{\rho_{\mathcal{W}}}{\rho_{\mathcal{Z}}} \zeta_{y_n} \rangle \overset{(S2)}{=} \rho_{\mathcal{W}} \rho_{\mathcal{Z}} \ . \tag{S126}$$

Therefore,

$$\sum_{y \in \mathcal{Y} \setminus \{y_n\}} (\langle z_n, w_y \rangle - \langle z_n, w_{y_n} \rangle) = (K-1) \left(\frac{\rho_{\mathcal{W}}}{\rho_{\mathcal{Z}}} d - \rho_{\mathcal{W}} \rho_{\mathcal{Z}} \right) \tag{S127}$$

and condition (P2) is fulfilled with $M = (K-1) \left(\frac{\rho_{\mathcal{W}}}{\rho_{\mathcal{Z}}} d - \rho_{\mathcal{W}} \rho_{\mathcal{Z}} \right)$.

Ad (P4): We need to show that $\forall n : \|z_n\| = \rho_{\mathcal{Z}}$.

This follows immediately from condition (C1) and condition (S2):

$$\|z_n\| \overset{(C1)}{=} \|\zeta_{y_n}\| \overset{(S2)}{=} \rho_{\mathcal{Z}} \ . \tag{S128}$$

Ad (P6): We need to show that $\bar{w} = 0$.

This follows immediately from condition (C3) and condition (S1):

$$\bar{w} = \frac{1}{K} \sum_{y \in \mathcal{Y}} w_y \overset{(C3)}{=} \frac{1}{K} \frac{\rho_{\mathcal{W}}}{\rho_{\mathcal{Z}}} \sum_{y \in \mathcal{Y}} \zeta_y \overset{(S1)}{=} 0 \ . \tag{S129}$$

Ad (P5): We need to show that $\forall y \in \mathcal{Y}$ the terms $\|\bar{w}\|^2 + \|w_y\|^2 - 2\langle \bar{w}, w_y \rangle$.

This follows from conditions (C3) and (S2), such as the already proven condition (P6).

Let $y \in \mathcal{Y}$ then

$$\|\bar{w}\|^2 + \|w_y\|^2 - 2\langle \bar{w}, w_y \rangle \overset{(P6)}{=} 0 + \|w_y\|^2 - 2 \cdot 0 \overset{(C3)}{=} \|\frac{\rho_{\mathcal{W}}}{\rho_{\mathcal{Z}}} \zeta_{y_n}\|^2 \overset{(S2)}{=} \rho_{\mathcal{W}}^2 \ , \tag{S130}$$

which, indeed, does not depend on y.

Ad (P3): We need to show that $\forall n \in [N]$ there $\exists \lambda_n \leq 0$ such that $z_n = \lambda_n (\bar{w} - w_{y_n})$.

Let $n \in [N]$ and consider

$$z_n \overset{(C1)}{=} \zeta_{y_n} \overset{(C3)}{=} \frac{\rho_{\mathcal{Z}}}{\rho_{\mathcal{W}}} w_{y_n} \tag{S131}$$

Thus, from the already proven condition (P6), it follows that

$$\bar{w} - w_{y_n} \overset{(P6)}{=} -w_{y_n} = -\frac{\rho_{\mathcal{W}}}{\rho_{\mathcal{Z}}} z_n \tag{S132}$$

and condition (P3) is fulfilled with $\lambda_n = -\frac{\rho_{\mathcal{Z}}}{\rho_{\mathcal{W}}} \leq 0$. □

Corollary 1. *Let Z, Y, W be defined as in Theorem 1. Upon requiring that $\forall y \in [K] : \|w_y\| \leq r_{\mathcal{W}}$, it holds that*

$$\mathcal{L}_{\mathrm{CE}}(Z, W; Y) \geq \log\left(1 + (K-1)\exp\left(-\frac{K\,\rho_{\mathcal{Z}}\,r_{\mathcal{W}}}{K-1}\right)\right)$$

with equality if and only if (C1) and (C2) from Theorem 1 are satisfied and condition (C3) changes to (C3r) $\forall y \in \mathcal{Y} : w_y = \frac{r_{\mathcal{W}}}{\rho_{\mathcal{Z}}}\zeta_y$.

Proof. By leveraging Theorem 1, we get

$$\mathcal{L}_{\mathrm{CE}}(Z, W; Y) \overset{\text{Thm. 1}}{\geq} \log\left(1 + (K-1)\exp\left(-\rho_{\mathcal{Z}}\frac{\sqrt{K}}{K-1}\|W\|_F\right)\right) \tag{S133}$$

$$= \log\left(1 + (K-1)\exp\left(-\rho_{\mathcal{Z}}\frac{\sqrt{K}}{K-1}\sqrt{\sum_{y\in\mathcal{Y}}\|w_y\|^2}\right)\right) \tag{S134}$$

$$\geq \log\left(1 + (K-1)\exp\left(-\rho_{\mathcal{Z}}\frac{\sqrt{K}}{K-1}\sqrt{\sum_{y\in\mathcal{Y}}r_{\mathcal{W}}^2}\right)\right) \tag{S135}$$

$$= \log\left(1 + (K-1)\exp\left(-\rho_{\mathcal{Z}}\frac{K}{K-1}r_{\mathcal{W}}\right)\right) \quad , \tag{S136}$$

where equality is attained if and only if the bound from Theorem 1 is tight, i.e., conditions (C1), (C2), (C3) are fulfilled and, additionally,

$$r_{\mathcal{W}} = \|w_y\| \text{ for } y \in \mathcal{Y} \quad . \tag{S137}$$

It remains to show that if conditions (C1) (C2) are fulfilled, then the following equivalency holds:

$$\big(r_{\mathcal{W}} = \|w_y\| \text{ for } y \in \mathcal{Y} \;\wedge\; (C3)\big) \iff (C3r) \quad . \tag{S138}$$

"$\Longrightarrow$": We need to show that $\forall y \in \mathcal{Y} : w_y = \frac{r_{\mathcal{W}}}{\rho_{\mathcal{Z}}}\zeta_y$.

So, let $y \in \mathcal{Y}$. By condition (C3), there is $\rho_{\mathcal{W}} > 0$ such that

$$w_y = \frac{\rho_{\mathcal{W}}}{\rho_{\mathcal{Z}}}\zeta_y \quad . \tag{S139}$$

Thus (C3r) holds if $\rho_{\mathcal{W}} = r_{\mathcal{W}}$. Indeed,

$$r_{\mathcal{W}} \overset{\text{(S137)}}{=} \|w_y\| \overset{\text{(S139)}}{=} \frac{\rho_{\mathcal{W}}}{\rho_{\mathcal{Z}}}\|\zeta_y\| \overset{(C2)}{=} \frac{\rho_{\mathcal{W}}}{\rho_{\mathcal{Z}}}\rho_{\mathcal{Z}} = \rho_{\mathcal{W}} \quad . \tag{S140}$$

"$\Longleftarrow$": Follows immediately as we can choose $\rho_{\mathcal{W}} = r_{\mathcal{W}} > 0$.

□

Lemma S14. *Let λ, $\rho_{\mathcal{Z}} > 0$, $K, h \in \mathbb{N}$ and $W \in (\mathbb{R}^h)^K$. The function*

$$f(x) = \log\left(1 + (K-1)\exp\left(-\rho_{\mathcal{Z}}\frac{K}{K-1}x\right)\right) + \lambda K x^2 \tag{S141}$$

is minimized by $x_0 = r_{\mathcal{W}}(\rho_{\mathcal{Z}}, \lambda) > 0$, i.e., the unique *solution to*

$$0 = K\left(2\lambda x - \frac{\rho_{\mathcal{Z}}}{e^{\frac{K\rho_{\mathcal{Z}}x}{K-1}} + K - 1}\right) \ . \tag{S142}$$

Proof. The first derivative of f is given by

$$f'(x) = K\left(2\lambda x - \frac{\rho_{\mathcal{Z}}}{e^{\frac{K\rho_{\mathcal{Z}}x}{K-1}} + K - 1}\right) \ . \tag{S143}$$

Note that f' is strictly increasing. Thus f is strictly convex and has a unique minimum at the point x_0 where $f'(x_0) = 0$. As f' is continuous on $(0, \infty)$ with

$$f'(0) = -\rho_{\mathcal{Z}} < 0 \tag{S144}$$

and

$$\lim_{x\to\infty} f'(x) = \infty \ , \tag{S145}$$

the intermediate value theorem implies $0 < x_0 = r_{\mathcal{W}}(\rho_{\mathcal{Z}}, \lambda) < \infty$. □

Corollary 2. *Let Z, Y, W be defined as in Theorem 1. For the L_2-regularized objective $\mathcal{L}_{\mathrm{CE}}(Z, W; Y) + \lambda\|W\|_F^2$ with $\lambda > 0$, it holds that*

$$\begin{aligned}&\mathcal{L}_{\mathrm{CE}}(Z, W; Y) + \lambda\|W\|_F^2 \\ &\geq \log\left(1 + (K-1)\exp\left(-\rho_{\mathcal{Z}}\frac{K}{K-1}r_{\mathcal{W}}(\rho_{\mathcal{Z}}, \lambda)\right)\right) \\ &\quad + \lambda K r_{\mathcal{W}}(\rho_{\mathcal{Z}}, \lambda)^2 \ ,\end{aligned}$$

where $r_{\mathcal{W}}(\rho_{\mathcal{Z}}, \lambda) > 0$ denotes the unique solution, in x, of

$$0 = K\left(2\lambda x - \frac{\rho_{\mathcal{Z}}}{\exp(\frac{K\rho_{\mathcal{Z}}x}{K-1}) + K - 1}\right) \ .$$

Equality is attained in the bound if and only if (C1) and (C2) from Theorem 1 are satisfied and (C3) changes to

(C3wd) $\forall y \in \mathcal{Y} : w_y = \frac{r_{\mathcal{W}}(\rho_{\mathcal{Z}}, \lambda)}{\rho_{\mathcal{Z}}}\zeta_y$.

Proof. By leveraging Theorem 1 and Lemma S14 (with $x = \|W\|_F/\sqrt{K}$), we get

$$\begin{aligned}&\mathcal{L}_{\mathrm{CE}}(Z, W; Y) + \lambda\|W\|_F^2 \\ &\overset{\text{Thm. 1}}{\geq} \log\left(1 + (K-1)\exp\left(-\rho_{\mathcal{Z}}\frac{\sqrt{K}}{K-1}\|W\|_F\right)\right) + \lambda\|W\|_F^2 \\ &= \log\left(1 + (K-1)\exp\left(-\rho_{\mathcal{Z}}\frac{K}{K-1}x\right)\right) + \lambda K x^2 \qquad \text{(by setting } x = \|W\|_F/\sqrt{K}\text{)} \\ &\overset{\text{Lem. S14}}{\geq} \log\left(1 + (K-1)\exp\left(-\rho_{\mathcal{Z}}\frac{K}{K-1}r_{\mathcal{W}}(\rho_{\mathcal{Z}}, \lambda)\right)\right) + \lambda K r_{\mathcal{W}}(\rho_{\mathcal{Z}}, \lambda)^2 \ ,\end{aligned}$$

where equality is attained if and only if the bound from Theorem 1 is tight, i.e., conditions (C1), (C2), (C3) are fulfilled and, additionally,

$$\|W\|_F/\sqrt{K} = r_{\mathcal{W}}(\rho_{\mathcal{Z}}, \lambda) \ . \tag{S146}$$

It remains to show that if (C1) and (C2) are fulfilled, it holds that

$$\left(\|W\|_F/\sqrt{K} = r_{\mathcal{W}}(\rho_{\mathcal{Z}}, \lambda) \ \wedge \ (C3)\right) \iff (C3wd) \ . \tag{S147}$$

"$\Longrightarrow$": We need to show for every $y \in \mathcal{Y}$ that $w_y = \frac{r_{\mathcal{W}}(\rho_{\mathcal{Z}},\lambda)}{\rho_{\mathcal{Z}}}\zeta_y$.

So let $y \in \mathcal{Y}$. By condition (C3), there exists $\rho_{\mathcal{W}} > 0$ such that $w_y = {}^{\rho_{\mathcal{W}}}/_{\rho_{\mathcal{Z}}}\zeta_y$. Thus, condition (C3wd) is fulfilled if $\rho_{\mathcal{W}} = r_{\mathcal{W}}(\rho_{\mathcal{Z}}, \lambda)$. Indeed,

$$r_{\mathcal{W}}(\rho_{\mathcal{Z}}, \lambda) \overset{(S146)}{=} \frac{\|W\|_F}{\sqrt{K}} = \sqrt{\frac{1}{K}\sum_{y\in\mathcal{Y}}\|w_y\|^2} \overset{(C3)}{=} \sqrt{\sum_{y\in\mathcal{Y}}\|\frac{\rho_{\mathcal{W}}}{\rho_{\mathcal{Z}}}\zeta_y\|^2} \tag{S148}$$

$$\overset{(C2)}{=} \sqrt{\frac{1}{K}\sum_{y\in\mathcal{Y}}\rho_{\mathcal{W}}^2} = \rho_{\mathcal{W}} \ . \tag{S149}$$

"$\Longleftarrow$":

Condition (C3) Is fulfilled as we can choose

$$\rho_{\mathcal{W}} = r_{\mathcal{W}}(\rho_{\mathcal{Z}}, \lambda) \overset{\text{Lem. } S14}{>} 0 \ . \tag{S150}$$

Finally, $r_{\mathcal{W}}(\rho_{\mathcal{Z}}, \lambda) = \|W\|_F/\sqrt{K}$, as

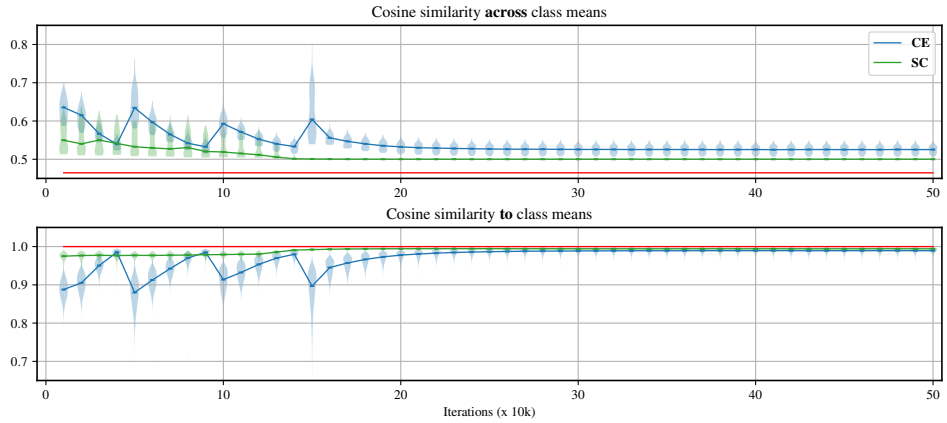
$$\|W\|_F = \sqrt{\sum_{y\in\mathcal{Y}}\|w_y\|^2} \overset{(C3wd)}{=} \sqrt{\sum_{y\in\mathcal{Y}}\|\frac{r_{\mathcal{W}}(\rho_{\mathcal{Z}},\lambda)}{\rho_{\mathcal{Z}}}\zeta_y\|^2}$$

$$\overset{(C2)}{=} \sqrt{\sum_{y\in\mathcal{Y}} r_{\mathcal{W}}(\rho_{\mathcal{Z}},\lambda)^2} = \sqrt{K}\, r_{\mathcal{W}}(\rho_{\mathcal{Z}},\lambda) \ .$$

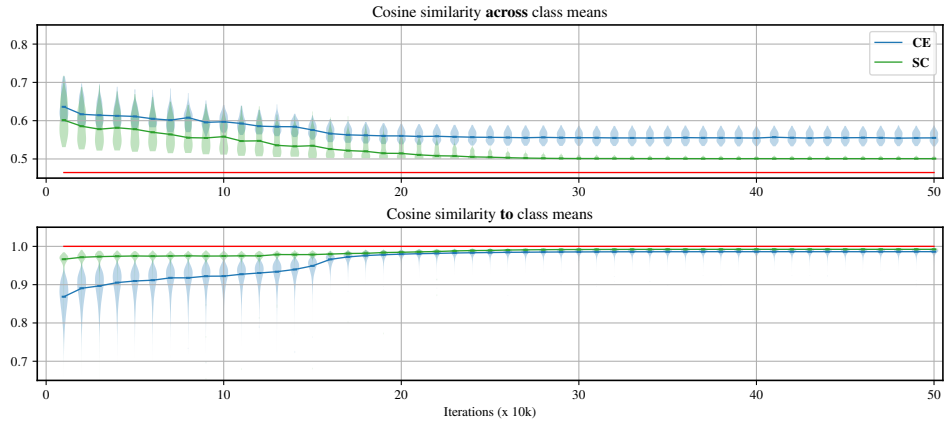
□

S3. Additional Experiments

The experiments in §5.2 suggest that representations learned by minimizing the **SC** loss might arrange closer to the (theoretically optimal) simplex configuration, compared to representations learned by minimizing the **CE** loss. To corroborate that this disparity is due to differing optimization dynamics of the loss functions, i.e., differing trajectories in the parameter space, and not an artifact of terminating the loss minimization to early, we repeat[6] these experiments when optimizing over **500k** SGD iterations instead of 100k. After every 10k iterations, we freeze the model, compute the class means of representation of the training data and evaluate two geometric properties on all of the training data: (1) the *cosine similarity **across** class means* and (2) the *cosine similarity **to** class means*, [7], as illustrated in Figs. S1, S2.



(a) CIFAR10 (without augmentation)



(b) CIFAR10 (with augmentation)

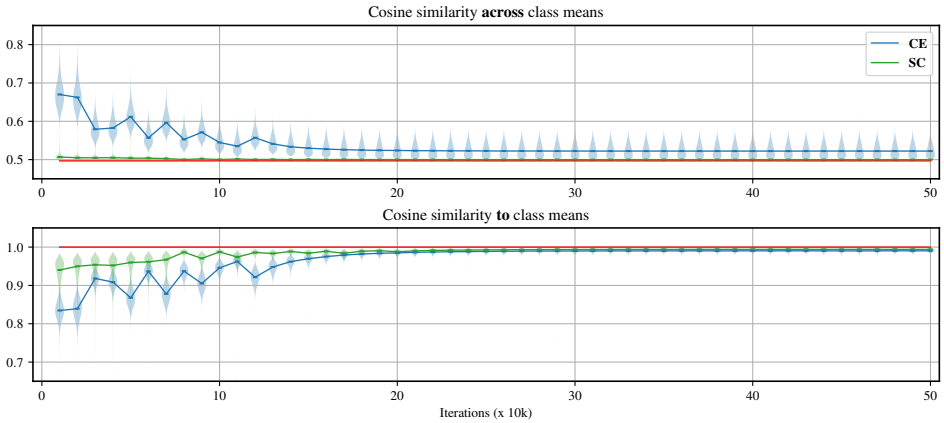
Figure S1: Distribution of geometric properties of representations, $\varphi_\theta(x_n)$, tracked during training. Representations are obtained from a ResNet-18 model trained (b) with and (a) without data augmentation on **CIFAR10**, with **CE** and **SC**, respectively. Blue and green lines indicate the evolution of the medians over the iterations; Red lines indicate the *sought-for* value at a regular simplex configuration.

The results reveal that (1) optimizing for 500k iterations improves convergence to the optimal state, yet at a very low speed, and (2) minimizing **SC** still yields representations closer to the simplex, compared to **CE**. The latter not only holds at the terminal stage of training, but at (almost) every evaluation step. Interestingly, on both datasets, the distributions of the computed properties obtained from the model trained via **CE** have notably more spread than the ones obtained from the model trained with **SC**.

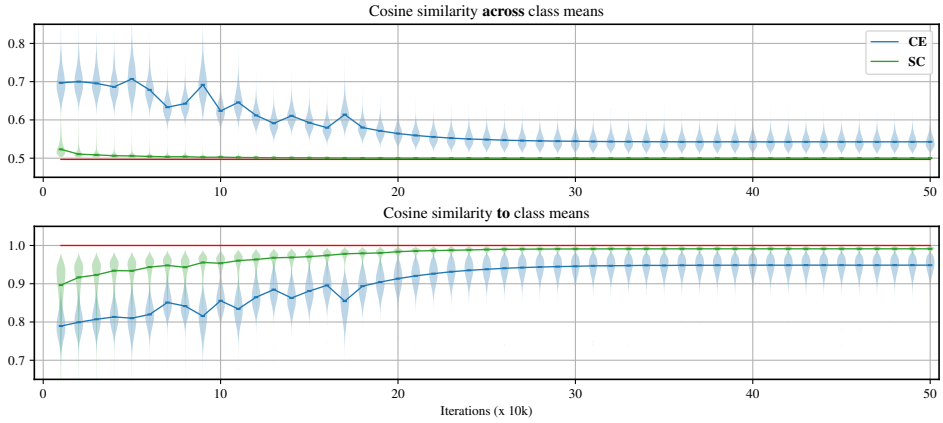
[6] i.e., the same setup and hyperparameters as in §5.2, except for the *number of training iterations*

[7] We omit the *cosine similarity across weights*, as, for **SC**, this requires to train an additional linear classifier each time.

Finally, we compare the geometric properties after training for 500k iteration with the ones from training over 100k iterations, i.e., Fig. 7 in §5.2. In case of **SC**, the distributions are roughly the same, whereas for **CE**, the distributions after 500k iterations are notably closer to the theoretical optimum than the ones after 100k iterations, particularly on the more complex CIFAR100 dataset. Once more, this highlights the faster convergence to the simplex arrangement via minimizing **SC**.



(a) CIFAR100 (without augmentation)



(b) CIFAR100 (with augmentation)

Figure S2: Distribution of geometric properties of representations, $\varphi_\theta(x_n)$, tracked during training. Representations are obtained from a ResNet-18 model trained (b) with and (a) without data augmentation on **CIFAR10**, with **CE** and **SC**, respectively. Blue and green lines indicate the evolution of the medians over the iterations; Red lines indicate the *sought-for* value at a regular simplex configuration.