Dissecting Supervised Constrastive Learning

Florian Graf [1] **Christoph D. Hofer** [1] **Marc Niethammer** [2] **Roland Kwitt** [1]

Abstract

Minimizing cross-entropy over the softmax scores of a linear map composed with a high-capacity encoder is arguably the most popular choice for training neural networks on supervised learning tasks. However, recent works show that one can *directly* optimize the encoder instead, to obtain equally (or even more) discriminative representations via a supervised variant of a contrastive objective. In this work, we address the question whether there are fundamental differences in the sought-for representation geometry in the output space of the encoder at minimal loss. Specifically, we prove, under mild assumptions, that both losses attain their minimum once the representations of each class collapse to the vertices of a regular simplex, inscribed in a hypersphere. We provide empirical evidence that this configuration is attained in practice and that reaching a close-to-optimal state typically indicates good generalization performance. Yet, the two losses show remarkably different optimization behavior. The number of iterations required to perfectly fit to data scales *superlinearly* with the amount of randomly flipped labels for the supervised contrastive loss. This is in contrast to the approximately *linear* scaling previously reported for networks trained with cross-entropy.

1. Introduction

In modern machine learning, neural networks have become the prevalent choice to parametrize maps from a complex input space $\mathcal{X}$ to some target space $\mathcal{Y}$. In supervised learning tasks, where the output space is a set of discrete labels, $\mathcal{Y} = \{1, \dots, K\}$, it is common to implement predictors of the form

$$f = \operatorname{argmax} \circ W \circ \varphi \ . \tag{1}$$

In this construction, f is realized as the composition of an encoder $\varphi : \mathcal{X} \to \mathcal{Z} \subseteq \mathbb{R}^h$, a linear map/classifier $W : \mathbb{R}^h \to \mathbb{R}^K$ and the argmax operation which handles the transition from continuous output to discrete label space.

Despite myriad advances in designing networks that implement φ, such as (Krizhevsky et al., 2012; He et al., 2016a; Zagoruyko & Komodakis, 2016; Huang et al., 2017), the training routine rarely deviates from minimizing the *cross-entropy (CE)* between softmax scores of $W \circ \varphi$ and one-hot encoded discrete labels. Assuming sufficient encoder capacity, it is clear that, at minimal loss, the representations of training instances, i.e., their images under φ, are in a linearly separable configuration (as the classifier is implemented as a linear map). Remarkably, this behavior is not only observed on real data with semantically meaningful labels, but also on real data with randomly flipped labels (Zhang et al., 2017).

Alternatively, one could aim for *directly* learning an encoder that is compatible with a linear classifier and the argmax decision rule. Recent works (Khosla et al., 2020; Han et al., 2020) have shown that this is indeed possible via a supervised variant of a contrastive loss (Chopra et al., 2005; Hadsell et al., 2006) that has full access to label information. Informally, this *supervised contrastive (SC)* loss comprises two competing dynamics: an attraction and a repulsion force. The former pulls representations from the same class (positives) closer together, the latter pushes representations from different classes (negatives) away from each other. A similar mechanic underpins the triplet loss (Weinberger & Saul, 2009), the N-pairs loss (Sohn, 2016), or the soft nearest-neighbor loss (Salakhutdinov & Hinton, 2007; Frosst et al., 2019) and contributes to the the success of self-supervised learning, framed as an instance discrimination task (van den Oord et al., 2018; Chen et al., 2020; Hénaff et al., 2020). In the context of the latter, positives are typically defined as different views of the *same* instance.

Most notably, predictors obtained by first learning φ via the supervised contrastive loss, followed by a composition with a linear map, not only yield state-of-the-art results on popular benchmarks, but show increased robustness towards input corruptions and hyperparameter choices (Khosla et al., 2020). This warrants a closer analysis of the underlying effects. While we focus on the formulation of Khosla et al. (2020), a similar analysis most likely holds for related vari-

[1]Department of Computer Science, University of Salzburg, Austria [2]UNC Chapel Hill. Correspondence to: Florian Graf <florian.graf@sbg.ac.at>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021. Copyright 2021 by the author(s).

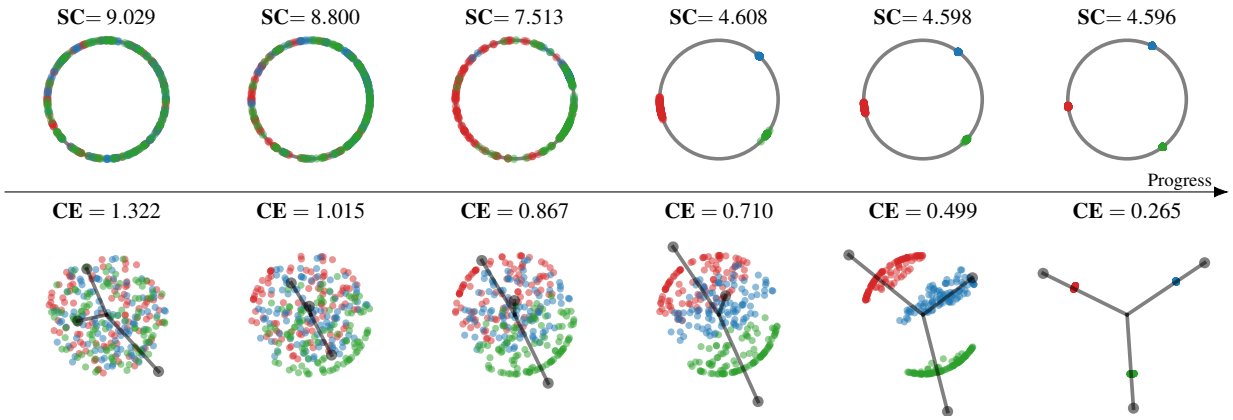


Figure 1: Loss comparison on a three-class toy problem in 2D with 100 points (z_i) per class. *Left to right* indicates optimization progress. The *top* row shows the point configurations while minimizing the supervised contrastive (SC) loss, w.r.t. z, on points drawn uniformly on $\mathbb{S}^1$. The *bottom* row shows the point configurations when minimizing cross-entropy (CE) over $\mathrm{softmax}(Wz)$ scores, w.r.t. W and z (and an L_2 penalty $\lambda\|W\|_F^2$), on points drawn uniformly within the unit disc. For the CE loss, gray discs (●) indicate the weights, the rays show the direction of the weights. In both cases, the z_i with equal label collapse to the vertices of a regular simplex.

ants. Specifically, we take a first step toward understanding potential differences in the output space of the encoder, induced either by (1) minimizing (softmax) cross-entropy over $W \circ \varphi$, or (2) minimizing the supervised contrastive loss directly over the outputs of φ. Characterizing the sought-for *geometric arrangement* of representations of training instances, at minimal loss, is an immediate starting point. Our analysis yields *two insights*, summarized below:

Insight 1 (theoretical). Under the assumption of an encoder φ that is powerful enough to realize any geometric arrangement of the representations in $\mathcal{Z}$, we analyze all loss minimizing configurations of the supervised contrastive and cross-entropy loss, respectively. More precisely, we prove (see §3.2) that the supervised contrastive loss (see Definition 2) attains its minimum if and only if the representations of each class collapse to the vertices of an origin-centered *regular* $K-1$ *simplex*, cf. Fig. 3. For the cross-entropy loss, we prove a similar, but more nuanced result (see §3.1) which is supplemental to an existing line of research. In particular, under a norm constraint on the outputs of φ, we show that (1) representations also collapse to the vertices of an origin-centered regular $K-1$ simplex and (2) the classifier weights are (positive) scalar multiples of the simplex vertices. Additionally, when subject to L_2 penalization, the weights attain equal norm, characterized by a function of the regularization strength. Fig. 1 visualizes the convergence to such a configuration on a toy example. In §4, we link these results to recent prior work, where an evenly spaced arrangement of classifier weights on the unit hypersphere is either *prescribed* or *explicitly* enforced.

Insight 2 (empirical). While our theoretical results assume an *ideal* encoder, we provide empirical evidence on popular vision benchmarks, that the sought-for regular simplex configurations can be attained in practice. Yet, networks trained with the supervised contrastive loss (1) tend to converge to a state closer to the loss minimizing configuration and (2) empirically yield better generalization performance. Hence, as loss minimization strives for a similar geometry of the encoder output for both loss functions (cf. Insight 1), we conjecture that differing optimization dynamics are the primary cause for obtaining solutions of different quality. One striking difference is observed when training on data with an increasing fraction of randomly flipped labels, illustrated in Fig. 2 for a ResNet-18 (CIFAR10), trained with (1) cross-entropy and (2) the supervised contrastive loss (with a subsequently optimized linear classifier W).

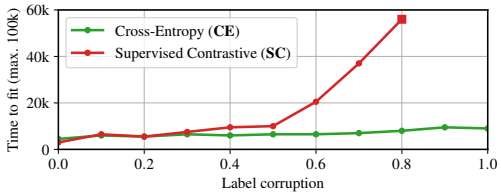


Figure 2: *Time to fit* of a ResNet-18 (on CIFAR10) as a function of increasing label corruption. The red square (■) marks the point at which zero training error can no longer be achieved.

While Zhang et al. (2017) report an approximately *linear* increase in the time to fit[1] for networks trained with cross-entropy, training with the supervised contrastive loss exhibits a clearly *superlinear* behavior. In fact, for a given iteration budget, fitting becomes impossible beyond a certain level of label corruption. This suggests that the supervised contrastive loss exerts some form of implicit regularization during optimization, yielding a parameter incarnation of the network which effectively prevents fitting to random labels.

[1] i.e., the number of iterations to reach zero training error.

Overview. §2 and §3 provide the technical details that underpin Insight 1. §4 draws connections to prior work and §5 presents further experiments along the lines of Insight 2. §6 concludes with a discussion of the main points.

2. Preliminaries

Consider a supervised learning task with $N \in \mathbb{N}$ training samples, i.e., the learner has access to data $X = (x_1, \dots, x_N) \in \mathcal{X}^N$, drawn i.i.d. from some distribution, and labels, $\{1, \dots, K\} = [K] \ni y_n = \mathfrak{c}(x_n)$, assigned to each x_n by an unknown function $\mathfrak{c} : \mathcal{X} \to [K]$.

We denote the unit-hypersphere (in $\mathbb{R}^h$) of radius $\rho > 0$ by $\mathbb{S}_\rho^{h-1} = \{x \in \mathbb{R}^h : \|x\| = \rho\}$; in case of $\rho = 1$, we write $\mathbb{S}^{h-1}$. The map $\varphi_\theta : \mathcal{X} \to \mathcal{Z} \subseteq \mathbb{R}^h$ identifies an encoder (see §1), parametrized by a neural network with parameters θ; we write $Z_\theta = \big(\varphi_\theta(x_1), \dots \varphi_\theta(x_N)\big)$ as the image of X under φ_θ. When required (e.g., in §3.2), we denote a batch by B, and identify the batch as the multi-set of indices $\{\!\{n_1, \dots, n_b\}\!\}$ with $n_i \in [N]$. For our analysis of the supervised contrastive loss, in §3.2, we assume $b \geq 3$.

Under the assumption of a *powerful enough encoder*, i.e., a map φ_θ that can realize every possible geometric arrangement, Z_θ, of the representations, we can decouple the loss formulations from the encoder. This facilitates to interpret Z_θ as a *free configuration* $Z = (z_1, \dots, z_N)$ of N labeled points (hence, we can omit the dependency on θ).

2.1. Definitions

For our purposes, we define the CE and SC loss, resp., as the loss over all N instances in Z. In case of the CE loss, this is the average over all instance losses; in case of the SC loss, we sum over *all* batches of size $b \in \mathbb{N}$. While the normalizing constant is irrelevant for our results, we point out that normalizing the SC loss would depend on the cardinality of *all* multi-sets of size b.

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 2 (Supervised contrastive loss). *Let $\mathcal{Z} = \mathbb{S}_{\rho_\mathcal{Z}}^{h-1} \subseteq \mathbb{R}^h$ and let Z be an N point configuration, $Z = (z_1, \dots, z_N) \in \mathcal{Z}^N$, with labels $Y = (y_1, \dots, y_N) \in [K]^N$. For a fixed batch size $b \in \mathbb{N}$, we define*

$$\mathcal{B} = \{\{\!\{n_1, \dots, n_b\}\!\} : n_1, \dots, n_b \in [N]\} \tag{4}$$

as the set of index multi-sets of size b. The supervised contrastive loss $\mathcal{L}_{\mathrm{SC}}(\cdot\,; Y) : \mathcal{Z}^N \to \mathbb{R}$ is defined as

$$Z \mapsto \sum_{B \in \mathcal{B}} \ell_{\mathrm{SC}}(Z; Y, B) \tag{5}$$

with $\ell_{\mathrm{SC}}(\cdot\,; Y, B) : \mathcal{Z}^N \to \mathbb{R}$ given by[2]

$$-\sum_{i \in B} \frac{\mathbb{1}_{\{|B_{y_i}| > 1\}}}{|B_{y_i}| - 1} \sum_{j \in B_{y_i} \setminus \{\!\{i\}\!\}} \log \left(\frac{\exp\big(\langle z_i, z_j \rangle\big)}{\sum\limits_{k \in B \setminus \{\!\{i\}\!\}} \exp\big(\langle z_i, z_k \rangle\big)} \right) \tag{6}$$

where $B_{y_i} = \{\!\{j \in B : y_j = y_i\}\!\}$ denotes the multi-set of indices in a batch $B \in \mathcal{B}$ with label equal to y_i. [3]

As the regular simplex, inscribed in a hypersphere, will play a key role in our results, we formally define this object next:

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* $1 \leq i < j \leq K$

Fig. 3 shows such configurations (for $K = 2, 3, 4$) on $\mathbb{S}^2$.

Remark 1. *The assumption $K \leq h + 1$ is crucial, as it is a necessary and sufficient condition for the existence of the regular simplex. In our context, K denotes the number of classes and $K \leq h + 1$ is typically satisfied, as the output spaces of encoders in contemporary neural networks are*

[2] for notational reasons, we set $\frac{0}{0} = 0$ when $|B_y| = 1$

[3] Definition 2 differs from the original definition by Khosla et al. (2020) in the following aspects: First, we do not explicitly duplicate batches (e.g., by augmenting each instance). For fixed index n, this does not guarantee that at least one other instance with label equal to y_n exists. However, this is formally irrelevant, as the contribution to the summation is zero in that case. Nevertheless, batch duplication is subsumed in our definition. Second, we adapt the definition to multi-sets, allowing for instances to occur more than once. If batches are drawn with replacement, this could indeed happen in practice. Third, we omit scaling the inner products $\langle \cdot, \cdot \rangle$ in Eq. (6) by a temperature parameter $1/\tau, \tau > 0$, as this complicates the notation. Instead we implicitly subsume this scaling into the radius $\rho_\mathcal{Z}$ of $\mathbb{S}_{\rho_\mathcal{Z}}^{h-1}$.

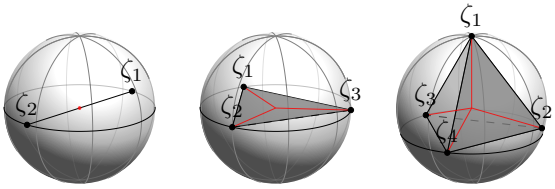


Figure 3: Regular simplices inscribed in $\mathbb{S}^2$.

high-dimensional, e.g., 512-dimensional for a ResNet-18 on CIFAR10/100. If it is violated, then the bounds derived in §3 still hold, but are not tight. Studying the loss minimizing configurations in this regime is much harder. Even for the related and more studied Thomson problem of minimizing the potential energy of K equally charged particles on the 2-dimensional sphere, the minimizers are only known for $K \in \{2, 3, 4, 5, 6, 12\}$ (Borodachov et al., 2019).

3. Analysis

We recap that we aim to address the following question: which N point configurations $Z = (z_1, \dots, z_N)$ yield minimal CE and SC loss? §3.1 and §3.2 answer this question, assuming a sufficiently high dimensional representation space $\mathcal{Z} \subseteq \mathbb{R}^h$, i.e., $K \leq h + 1$, and *balanced* class labels Y, i.e., $\big|\{i \in [N] : y_i = y\}\big| = {}^N\!/_K$, irrespective of the class y. *For detailed proofs we refer to the supplementary material.*

3.1. Cross-Entropy Loss

We start by providing a lower bound, in Theorem 1, on the CE loss, under the constraint of norm-bounded points.

Theorem 1. *Let $\rho_{\mathcal{Z}} > 0$, $\mathcal{Z} = \{z \in \mathbb{R}^h : \|z\| \leq \rho_{\mathcal{Z}}\}$. Further, let $Z = (z_1, \dots, z_N) \in \mathcal{Z}^N$ be an N point configuration with labels $Y = (y_1, \dots, y_N) \in [K]^N$ 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 $\zeta_1, \dots, \zeta_K \in \mathbb{R}^h$ 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*

(C3) $\exists \rho_{\mathcal{W}} > 0 : \forall y \in \mathcal{Y} : w_y = \frac{\rho_{\mathcal{W}}}{\rho_{\mathcal{Z}}}\zeta_y$

Importantly, Theorem 1 states that the bound is tight, if and only if all instances with the same label collapse to points and these points form the vertices of a regular simplex, inscribed in a hypersphere of radius $\rho_{\mathcal{Z}}$. Additionally, all weights, w_y, have to attain equal norm and have to be scalar multiples of the simplex vertices, thus also forming a regular simplex (inscribed in a hypersphere of radius $\rho_{\mathcal{W}}$).

Remark 2. *Our result complements recent work by Papyan et al. (2020), where it is empirically observed that training neural predictors as in Eq. (1)[4] leads to a within-class covariance collapse of the representations as we continue to minimize the CE loss beyond zero training error. By assuming representations to be Gaussian distributed around each class mean and taking the covariance collapse into account, the regular simplex arrangements of Theorem 1 arise. Specifically, this is the optimal configuration from the perspective of recovering the correct class labels. While the analysis in (Papyan et al., 2020) is decoupled from the loss function and hinges on a probabilistic argument, we study what happens as the CE loss attains its lower bound; our result, in fact, implies the covariance collapse.*

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

Notably, a special case of Corollary 1 appears in Proposition 2 of Wang et al. (2017), covering the case where $\forall n : z_n = w_{y_n}$ and $\forall y \in \mathcal{Y} : \|w_y\| = l$, i.e., equinorm weights and already collapsed classes. Corollary 1 obviates these constraints and provides a more general result, only assuming that $\forall n : \|z_n\| \leq \rho_{\mathcal{Z}}$ and $\forall y : \|w_y\| \leq r_{\mathcal{W}}$. However, constraining the norm of the weights seems artificial as, in practice, the weights are typically subject to an additional L_2 penalty. Corollary 2 directly addresses this connection, showing that applying an L_2 penalty of the form $\lambda\|W\|_F^2$ eliminates the necessity of an explicit norm constraint.

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*

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

[4] also including bias terms, i.e., $Wx + b$

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

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

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

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

Corollary 2 differs from Corollary 1 in that the characterization of w_y depends on $r_{\mathcal{W}}(\rho_{\mathcal{Z}}, \lambda)$, i.e., a function of the norm constraint, $\rho_{\mathcal{Z}}$, on the points and the regularization strength λ. While $r_{\mathcal{W}}(\rho_{\mathcal{Z}}, \lambda)$ has, to the best of our knowledge, no closed-form solution, it can be solved numerically. Fig. 1 illustrates the attained regular simplex configuration, on a toy example, in case of added L_2 regularization.

It is important to note that the assumed norm-constraint on points in $\mathcal{Z}$ is not purely theoretical. In fact, such a constraint often arises[5], e.g., via batch normalization (Ioffe & Szegedy, 2015) at the last layer of a network implementing φ_θ. While one could, in principle, derive a normalization dependent bound for the CE loss, it is unclear (to the best of our knowledge) if a regular simplex solution satisfying the corresponding equality conditions always exists.

Numerical Simulation

To empirically assess our theoretical results, we take the toy example from Fig. 1, where we minimize (via gradient descent) the L_2 regularized CE loss over W and Z with $\forall n : \|z_n\| \leq 1$. This setting corresponds to having an *ideal* encoder, φ, that can realize any configuration of points and matches the assumptions of Corollary 2. Fig. 4 (*right*) shows that the lower bound, for varying values of the regularization strength λ, closely matches the empirical loss. Additionally, Fig. 4 (*left*) shows a direct comparison of the empirical weight average, $\|\overline{w}\|$, *vs.* the corresponding theoretical value of $\|w_y\|$ (which is equal for all y in case of minimal loss). These experiments empirically confirm that conditions (C1) and (C2), as well as the adapted condition (C3wd) from Corollary 2 are satisfied. In §5, we will see that the sought-for regular simplex configurations actually arise (with varying quality) when minimizing the L_2 regularized CE loss for a ResNet-18 trained on popular vision benchmarks.

3.2. Supervised Contrastive Loss

An analysis of the SC loss, similar to §3.1, is less straightforward. In fact, as the loss is defined over *batches*, we can not simply sum up per-instance losses to characterize the ideal N point configuration. Instead, we need to consider *all* batch configurations of a specific size $b \in \mathbb{N}$.

[5] although it might not be explicitly enforced

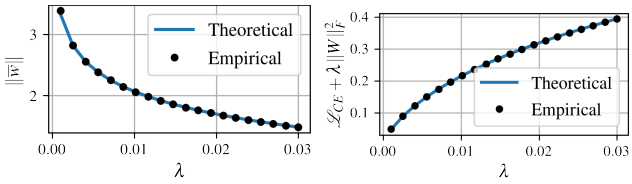


Figure 4: Numerical simulation for Corollary 2 (on the toy data of Fig. 1), as a function of the L_2 regularization strength λ. The *left* plot shows the theoretical norm of w_y (which is equal for all y at minimal loss) vs. the observed mean norm of the three weights. The *right* plot shows the theoretical bound vs. the empirical L_2 regularized CE loss.

We next state our lower bound for the SC loss with the corresponding equality conditions.

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

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

where

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

Equality is attained if and only if the following conditions are satisfied. 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*

Theorem 2 characterizes the geometric configuration of points in Z at minimal loss. We see that the equality conditions (C1) and (C2) from Theorem 1 equally appear in Theorem 2. This implies that, at minimal loss, each class collapses to a point and these points form a regular simplex.

Considering the guiding principle of the SC loss, i.e., separating instances from distinct classes and attracting instances from the same class, it seems plausible that constraining instances to the hypersphere would yield an evenly distributed arrangement of classes. However, a closer look at the SC loss reveals that this is not obvious by any means. In contrast to the physical (electrostatic) intuition, the involved attraction and repulsion forces are not pairwise, but depend on groups of samples, i.e., batches. Naively, one could try to characterize the loss minimizing configuration of points for each batch separately. Yet, this is destined to fail, as the minimizing arrangement of points in each batch depends on the label configuration; an example is visualized in Fig. 5. Hence, there is no simultaneous minimizer for all

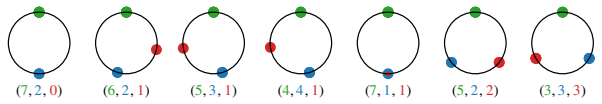


Figure 5: Illustration of *loss minimizing* point configurations of the batch-wise SC loss for varying label configurations and a batch size $b = 9$. Colored numbers indicate the *multiplicity* of each class in the batch.

	$\mathcal{L}_{\mathrm{SC}}(Z;Y)$
Empirical	12.12016
Theory	12.12015

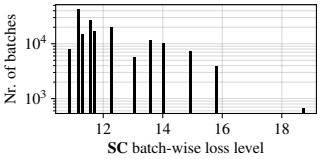


Figure 6: Numerical optimization of the SC loss for toy data on $\mathbb{S}^1$. *Left*: Comparison of mean batch-wise loss with the lower bound from Theorem 2. *Right*: Histogram (over all ≈170k batches) of the batch-wise loss values at convergence, showing the inhomogeneity of minimal loss values across batch configurations (cf. Fig. 5).

batch-wise losses. It is therefore crucial to understand the interaction of the attraction and repulsion forces across different batches. We sketch the argument of the proof below and refer to the supplementary material for details.

Proof Idea for Theorem 2

The key idea is to decouple the attraction and repulsion effects from the batch-wise formulation of the loss. Since each batch-wise loss contribution is actually a sum of label-wise contributions, the supervised contrastive loss can be considered as a sum over the Cartesian product of the set of all batches with the set of all labels. We partition this Cartesian product into appropriately constructed subsets, i.e., by label multiplicity. This allows to apply Jensen's inequality to each sum over such a subset. In the resulting lower bound, the repulsion and attraction effects are still allocated to the batches, but encoded more tangibly, i.e., linearly, as sums of inner products. Therefore, their interactions can be analyzed by a combinatorial argument which hinges on the balanced class label assumption. Minimality of the respective sums is attained if and only if (1) all classes are collapsed and (2) the mean of all instances (i.e., ignoring the class label) is zero. The simplex arrangement arises as consequence of (1) & (2) and, additionally, the equality conditions yielded by the previous application of Jensen's inequality, i.e., all intra-class and inter-class inner products are equal.

Numerical Simulation

For a large number of points, numerical computation of the bound in Theorem 2 is infeasible due to the combinatorial growth of the number of batches (even for the toy-example of Fig. 1 with 300 points). Hence, we consider a smaller setup. In particular, we take $K = 3$ classes, each consisting of 4 points on the unit circle $\mathbb{S}^1$, i.e., $Z = (z_1, \ldots, z_{12})$, $h = 2$ and $\rho = 1$. For a batch size of $b = 9$, this setup yields a total of 167,960 batches, i.e., the number of combinations with replacement. We initialize the z_i as the projection of points sampled from a standard Gaussian distribution and then minimize the SC loss (by stochastic gradient descent for 100k iterations) over the points in Z. Fig. 6 (*left*) shows that, at convergence, the lower bound on $\mathcal{L}_{\mathrm{SC}}(Z;Y)$ closely matches the empirical loss. Fig. 6 (*right*) shows the SC loss over *all* batches, highlighting the different loss levels depending on the label configuration in the batch (cf. Fig. 5).

4. Related work

We focus on works closely linked to our theoretical results of §3; we refer the reader to (Khosla et al., 2020) (and references therein) for additional background on the supervised contrastive loss and to (Le-Khac et al., 2020) for a general survey on contrastive learning.

Our results on the cross-entropy loss from §3.1 are partially related to a recent stream of research (Soudry et al., 2018; Nacson et al., 2018; Gunasekar et al., 2018) on characterizing the convergence speed and structure of homogeneous linear predictors (W) when minimizing cross-entropy via gradient descent on linearly separable data (i.e., no preceding learned encoder φ). In particular, Soudry et al. (2018) show that such predictors converge to the L_2 max margin separator. In our setting, the geometric structure of W (and the outputs of φ) becomes even more explicit, i.e., the weights reside at the vertices of a *regular simplex*. This is in line with the special case of equinorm representations and weights, presented in (Wang et al., 2017), and complements a recent optimality result by Papyan et al. (2020) (cf. Remark 2 for details).

Along another line of research, several works focus on controlling geometric properties of the classifier weights. In (Hoffer et al., 2018), for instance, the classifier weights are fixed prior to training, with one choice of the weight matrix being a random orthonormal projection. In this setup, all weights have unit norm, are well separated on the hypersphere, but do not form the vertices of a regular simplex. Yet, this empirically yields fast convergence, reduces the number of learnable parameters and has no negative impact on performance. In (Liu et al., 2018), separation of the classifier weights is achieved via a regularization term based on a Riesz s-potential (Borodachov et al., 2019). While this regularization term can be added to all network layers, in the special case of the linear classifier weights, the sought-for minimal energy (for $K \leq h + 1$) is again attained once the weights form the vertices of a regular simplex. Recently, Mettes et al. (2019) presented an approach that a-priori positions so called *prototypes* (one for each class) on the unit hypersphere such that the largest cosine similarity among the prototypes is minimized. Training then reduces to attract-

ing representations towards their corresponding prototypes. Again, in case of $K \leq h+1$, this yields a geometric prototype arrangement at the vertices of a regular simplex. In the context of Eq. (1), the prototypes correspond to the classifier weights and are compatible with the argmax decision rule.

Overall, (Hoffer et al., 2018; Liu et al., 2018; Mettes et al., 2019) all control, in one way or the other, the geometric arrangement of classifier weights and thereby, implicitly, the arrangement of the representations. This is decisively different to supervised contrastive learning, where the arrangement of the classifier weights is a consequence of the regular simplex arrangement of the representations at minimal loss. More precisely, if representations are already in a regular simplex configuration, the cross-entropy loss of a subsequently trained linear classifier is minimized if and only if the classifier weights are equinorm and scalar multiples of the simplex vertices (cf. Corollary 1).

We additionally point out that several works have recently started to establish a solid theoretical foundation for using contrastive loss functions in the context of unsupervised representation learning.

Through the concept of *latent classes* (i.e., a construction formalizing the notion of semantic similarity), Arora et al. (2019) prove generalization bounds for downstream supervised classification, under the assumption that the supervised task is defined on a subset of the latent classes. Central to their analysis is the *mean classifier* which is determined by the means of representations of training inputs with equal label. Notably, they empirically observe that this mean classifier performs well on models trained under full supervision. In light of our theoretical results, this can be easily explained by the fact that, at optimality, representations collapse to the simplex vertices.

The *unsupervised* counterpart of the objective we study in this work is analyzed by Wang & Isola (2020) from a probabilistic perspective. It is shown that minimizing the (unsupervised) contrastive loss promotes *alignment* and *uniformity* of representations on the unit hypersphere, two properties that empirically correlate with good performance on downstream tasks. More precisely, the authors split the (unsupervised) contrastive loss into two summands and show that in the limit of infinite negative samples, asymptotically one is minimized by a *peferctly aligned* and the other by a *perfectly uniform* encoder. As pointed out by the authors, if the data is finite, then there is no encoder which is both, i.e., perfectly aligned and perfectly uniform. Hence, in this case, their analysis does not provide an explicit characterization of the loss minimizer. Complementary to that, our analysis restricts to this very case of finite (training) data, but is able to characterize the loss minimizer in the *supervised* setup.

5. Experiments

In any practical setting, we do not have an *ideal* encoder (as in §3), but an encoder parameterized as a neural network, φ_θ. Hence, in §5.2, we first assess whether the regular simplex configurations actually arise (and to which extent), given a fixed iteration budget during optimization. Second, in §5.3, we study the optimization behavior of models under different loss functions in a series of random label experiments.

5.1. Setup

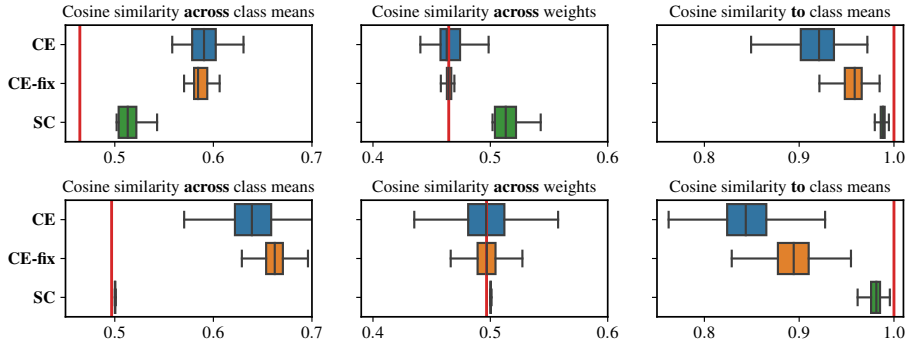
As our choice of φ_θ, we select a ResNet-18 (He et al., 2016a) model, i.e., all layers up to the linear classifier. Experiments are conducted on CIFAR10/100, for which this choice yields 512-dim. representations (and $K \leq h+1$ holds in all cases).

We either compose φ_θ with a linear classifier and train with the CE loss function (denoted as **CE**), or we directly optimize φ_θ via the SC loss function, then freeze the encoder parameters and train a linear classifier on top (denoted **SC**). In case of the latter, outputs of φ_θ are always projected onto a hypersphere of radius $\rho = 1/\sqrt{\tau}$ (with $\tau = 0.1$), which accounts for scaling the inner-products by the temperature parameter $1/\tau$ in the original formulation of Khosla et al. (2020). We want to stress that while Theorem 2 holds for every $\rho > 0$, the temperature crucially influences the optmization dynamics and needs to be tuned appropriately. For comparison, we also compose φ_θ with a *fixed* linear classifier, in particular, a classifier with weights a-priori optimized towards a regular simplex arrangement. This is similar to (Mettes et al., 2019), only that we minimize the CE loss (denoted as **CE-fix**) to learn predictors $W \circ \varphi_\theta$, as opposed to pulling outputs of φ_θ towards the fixed prototypes/weights.

Optimization is done via (mini-batch) stochastic gradient descent with L_2 regularization (10^{-4}) and momentum (0.9) for 100k iterations. The batch-size is fixed to 256 and the learning rate is annealed exponentially, starting from 0.1. When using data augmentation, we apply random cropping and random horizontal flipping, each with probability $1/2$.

5.2. Theory vs. Practice

To provide a first impression to which extend the representations of the training data achieve the loss minimizing geometric arrangement, we compare the empirical **CE** and **SC** loss values to the optima derived in §3, using ResNet-18 models trained on CIFAR10 (with data augmentation). The theoretical/empirical losses are (1) 7.64e-5 vs. 2.48e-4 (for **CE**) and (2) 824.487 vs. 824.731 (for **SC**), where we estimate the empirical **SC** loss over 1k training batches. Notably, when optimizing for 500k iterations (instead of 100k; see supplementary material), the loss values continue to move closer to the optimum, but at very low speed. In particular, the loss values change to 2.27e-4 (for **CE**) and



Loss	CIFAR10 err. [%] w aug.	w/o aug.
CE	6.3	16.4
CE-fix	6.2	14.7
SC	**5.7**	**13.8**

Loss	CIFAR100 err. [%] w aug.	w/o aug.
CE	27.0	41.8
CE-fix	26.3	**41.3**
SC	**24.9**	41.5

Figure 7: Geometric properties of representations, $\varphi_\theta(x_n)$, and weights, obtained from ResNet-18 models trained (*top*: **CIFAR10**; *bottom*: **CIFAR100**) with different losses (using data augmentation, i.e., w aug.). The *left* three panels show the distribution of cosine similarities across (1) distinct class means (quantifying class separation), (2) distinct class weights (quantifying classifier weight separation) and (3) representations with respect to their respective class means (quantifying within class spread). Red lines indicate the sought-for value at a regular simplex configuration. The *right-most* panel shows the testing error of all models with and without data augmentation.

824.523 (for **SC**), respectively. Overall, this suggests that out of these models, representations learned by minimizing the **SC** loss might arrange closer to the theoretically optimal configuration. As our results only cover the *loss minimizers* and not (close to optima) level sets, the latter hypothesis is more of a first guess and not predicted by the theory.

For a closer look at the geometric arrangement of the representations (and classifier weights), we compute three statistics, all based on the cosine similarity $\gamma : \mathbb{R}^h \times \mathbb{R}^h \to [0, 1]$, defined as

$$(x, y) \mapsto 1 - \cos^{-1}\left(\left\langle {}^{x}/_{\|x\|}, {}^{y}/_{\|y\|} \right\rangle\right) / \pi \quad . \tag{7}$$

First, we measure the separation of the class representations via the cosine similarity among the class means, $\mu_1, \ldots, \mu_K$, i.e., $\gamma(\mu_i, \mu_j)$ for $i \neq j$. Second, for the CE loss function, we compute the cosine similarity across the classifier weights, i.e., $\gamma(w_i, w_j), i \neq j$, quantifying their separation. Third, to quantify class collapse, we compute the cosine similarity among all representations and their respective class means, i.e., $\gamma(\varphi(x_n), \mu_{y_n})$. Note that our theoretical results imply that the classes should collapse and the pairwise similarities, as mentioned above, should be equal.

Fig. 7 illustrates the distribution of the cosine similarities for the ResNet-18 model trained with different loss functions (and using data augmentation). We observe that the SC loss leads to (1) an arrangement of the class means much closer to the ideal simplex configuration and (2) a tighter concentration of training representations around their class means. Furthermore, in case of the CE loss, the weight arrangement reaches, on average, a regular simplex configuration, while the representations slightly deviate. When using a-priori fixed weights in a simplex configuration, i.e., **CE-fix**, the situation is similar, but the within-class spread is smaller. In general, the statistics are comparable between CIFAR10 and CIFAR100, only that the distribution of all computed statistics widens for models trained with **CE** on CIFAR100. We conjecture that the increase in the number of classes, combined with the joint optimization of φ_θ and W complicates convergence to the loss minimizing state. Fig. 7 (*right*) further suggests that approaching this state positively correlates with generalization performance. Whether the latter is a general phenomenon, or may even have a theoretical foundation, is an interesting question for future work.

Finally, we draw attention to the comparatively large gap between the cosine similarities *across* the class means and their theoretical prediction in case of models trained on CIFAR10 (Fig. 1, *top left*). The aforementioned gap indicates that the chosen encoder might not be powerful enough to arrange the representations on a sphere-inscribed regular simplex. In fact, a standard ResNet (He et al., 2016a) utilizes a ReLU activation function after each block, including the last block before the linear classifier. Therefore, the coordinates of representations obtained by the encoder part of a standard ResNet are always *non-negative*, and so are the coordinates of the class means. Consequently, their inner products are non-negative as well, which corresponds to a minimal cosine similarity of 0.5 across the class means. Since the scalar products of vertices (considered as position vectors) of a unit sphere inscribed regular simplex with K vertices are $-1/(K-1)$, the deviation from the optimal class separation, resulting from the choice of encoder, is unnoticeable for models trained on CIFAR100 due to the large number of classes, i.e., $K = 100$, but becomes apparent in case of CIFAR10 where $K = 10$.

We suspect that architectures which do not implement the aforementioned non-negativity constraint in the encoder, e.g., the *pre-activation* variants of ResNets (He et al., 2016b), are capable of separating the classes to a larger extend and thus match the theoretical prediction more closely when trained on data with a small number of classes.

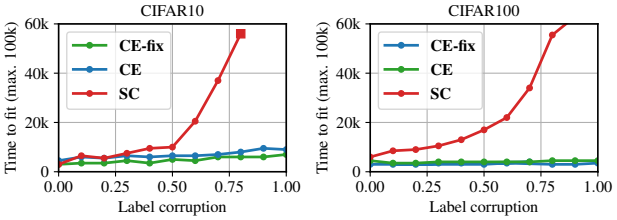


Figure 8: Time to fit for models of the form $W \circ \varphi_\theta$, based on ResNet-18 encoders, optimized under different loss functions.

5.3. Random Label Experiments

Despite the similarity of the loss minimizing geometric arrangements at the output of φ_θ, for both (CE, SC) losses, we have seen (in Fig. 7) that the extent to which this "optimal" state is achieved differs. These differences likely arise as a result of the underlying optimization dynamics, driven by the loss contribution of each batch. Notably, while the CE loss decomposes into independent instance-wise contributions, the SC loss does not (due to the interaction terms).

One way to explore this in greater detail, is to study optimization behavior as a function of label corruption. Specifically, as label corruption (i.e., the fraction of randomly flipped labels) increases, it is interesting to track the number of iterations (*time to fit*) to reach zero training error (Zhang et al., 2017), illustrated in Fig. 8.

On both datasets, **CE** and **CE-fix** show an approximately linear growth, while **SC** shows a remarkably *superlinear* growth. We argue that the latter primarily results from the profound interaction among instances in a batch. Intuitively, as the number of attraction terms for the SC loss function scales quadratically with the number of samples per class, increasing the number of semantically confounding labels equally increases the complexity of the optimization problem. In contrast, for **CE** and **CE-fix**, semantically confounding labels only impose per-instance constraints instead. This equally explains why, on CIFAR10, **SC** cannot achieve zero error beyond 80% corruption: fewer training instances per class (500 vs. 5,000) yield fewer pairwise intra-class constraints to be met.

6. Discussion

By focusing on predictors $\operatorname{argmax} \circ W \circ \varphi$, our results assert that the outputs of φ are *strikingly similar*, at minimal loss, irrespective of whether we train with cross-entropy or in the supervised contrastive regime.

Yet, from an optimization perspective, the choice of loss makes a profound difference, visible in the differing resilience to fit in the presence of corrupt label information. We argue that the advantages of supervised contrastive learning, reported in prior work, are rooted in the strong interaction terms among samples in a batch. While cross-entropy acts sample-wise, the supervised contrastive loss considers pair-wise sample relations, i.e., a *batch* is an *atomic* computational unit during stochastic optimization; in case of cross-entropy, the atomic unit is a *single* sample instead.

While we simplified the original setup of supervised contrastive learning, in particular, by detaching the commonly used *projective head*, we hope that our results provide a viable starting point for further analyses. Specifically, we think that a better theoretical understanding of the profound interaction between stochastic optimization and loss functions that capture pairwise constraints (rather than instance losses), could be a promising avenue to be explored in the context of the generalization puzzle.

Acknowledgements

This research was supported in part by the Austrian Science Fund (FWF): project FWF P31799-N38 and the Land Salzburg (WISS 2025) under project numbers 20102-F1901166-KZP and 20204-WISS/225/197-2019. We also like to thank the anonymous reviewers for the constructive feedback during the review process.

Source Code

Source code to reproduce experiments is publicly available: https://github.com/plus-rkwitt/py_supcon_vs_ce

References

Arora, S., Khandeparkar, H., Khodak, M., Plevrakis, O., and Saunshi, N. A theoretical analysis of contrastive unsupervised representation learning. In *ICML*, 2019.

Borodachov, S., Hardin, D., and Saff, E. *Discrete energy on rectifiable sets*. Springer, 2019.

Chen, X., Fan, H., Girshick, R., and He, K. Improved baselines with momentum contrastive learning. *arXiv preprint*, 2020. arXiv:2003.04297v1 [cs.CV].

Chopra, S., Hadsell, R., and LeCun, Y. Learning a similarity metric discriminatively, with application to face verification. In *CVPR*, 2005.

Frosst, N., Papernot, N., and Hinton, G. Analyzing and improving representations with the soft nearest neighbor loss. In *ICML*, 2019.

Gunasekar, S., Lee, J., Soudry, D., and Srebro, N. Characterizing implicit bias in terms of optimization geometry. In *ICML*, 2018.

Hadsell, R., Chopra, S., and LeCun, Y. Dimensionality reduction by learning an invariant mapping. In *CVPR*, 2006.

Han, T., Xie, W., and Zisserman, A. Self-supervised co-training for video representation learning. In *NeurIPS*, 2020.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *CVPR*, 2016a.

He, K., Zhang, X., Ren, S., and Sun, J. Identity mappings in deep residual networks. In *ECCV*, 2016b.

Hénaff, O., Srinivas, A., De Fauw, J., Razawi, A., Doersch, C., Ali Eslami, S., and van der Oord, A. Data-efficient image recognition with contrastive predictive coding. In *ICML*, 2020.

Hoffer, E., Hubara, I., and Soudry, D. Fix your classifier: the marginal value of training the last weight layer. In *ICLR*, 2018.

Huang, G., Liu, Z., van der Maaten, L., and Weinberger, K. Densely connected convolutional networks. In *CVPR*, 2017.

Ioffe, S. and Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, 2015.

Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., and Krishnan, D. Supervised contrastive learning. In *NeurIPS*, 2020.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

Le-Khac, P., Healy, G., and Smeaton, A. Contrastive representation learning: A framework and review. *IEEE Access*, 8:193907–193934, 2020.

Liu, W., Lin, R., Liu, Z., Liu, L., Yu, Z., Dai, B., and Song, L. Learning towards minimal hyperspherical energy. In *NeurIPS*, 2018.

Mettes, P., van der Pol, E., and Snoek, C. Hyperspherical prototype networks. In *NeurIPS*, 2019.

Nacson, M., Lee, J., Gunasekar, S., Savarese, P., Srebro, N., and Soudry, D. Convergence of gradient descent on separable data. In *AISTATS*, 2018.

Papyan, V., Han, X., and Donoho, D. Prevalence of neural collapse during the terminal phase of deep learning training. *PNAS*, 117(40):24652–24663, 2020.

Salakhutdinov, R. and Hinton, G. Learning a nonlinear embedding by preserving class neighbourhood structure. In *AISTATS*, 2007.

Sohn, K. Improved deep metric learning with multi-class N-pair loss objective. In *NIPS*, 2016.

Soudry, D., Hoffer, E., Nacson, M., Gunasekar, S., and Srebro, N. The implicit bias of gradient descent on separable data. *JMLR*, 19:1–57, 2018.

van den Oord, A., Li, Y., and Vinyals, O. Representation learning with contrastive predictive coding. *arXiv preprint*, 2018. arXiv:1807.03748v2 [cs.LG].

Wang, F., Xiang, X., Cheng, J., and Yuille, A. NormFace: l_2 hypersphere embedding for face verification. In *ACM Multimedia*, 2017.

Wang, T. and Isola, P. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *ICML*, 2020.

Weinberger, K. and Saul, L. Distance metric learning for large margin nearest neighbor classification. *JMLR*, 10: 207–244, 2009.

Zagoruyko, S. and Komodakis, N. Wide residual networks. In *BMVC*, 2016.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. Understanding deep learning requires rethinking generalization. In *ICLR*, 2017.