Which Features are Learnt by Contrastive Learning? On the Role of Simplicity Bias in Class Collapse and Feature Suppression

Yihao Xue [1] **Siddharth Joshi** [1] **Eric Gan** [1] **Pin-Yu Chen** [2] **Baharan Mirzasoleiman** [1]

Abstract

Contrastive learning (CL) has emerged as a powerful technique for representation learning, with or without label supervision. However, supervised CL is prone to collapsing representations of subclasses within a class by not capturing all their features, and unsupervised CL may suppress harder class-relevant features by focusing on learning easy class-irrelevant features; both significantly compromise representation quality. Yet, there is no theoretical understanding of *class collapse* or *feature suppression* at *test* time. We provide the first unified theoretically rigorous framework to determine *which* features are learnt by CL. Our analysis indicate that, perhaps surprisingly, bias of (stochastic) gradient descent towards finding simpler solutions is a key factor in collapsing subclass representations and suppressing harder class-relevant features. Moreover, we present increasing embedding dimensionality and improving the quality of data augmentations as two theoretically motivated solutions to feature suppression. We also provide the first theoretical explanation for why employing supervised and unsupervised CL together yields higher-quality representations, even when using commonly-used stochastic gradient methods.

1. Introduction

Learning high-quality representations that generalize well to a variety of downstream prediction tasks has been a long-standing goal of machine learning (Hinton et al., 2006; Ranzato et al., 2006). Contrastive learning (CL) has emerged as an effective approach for solving this problem, both with and without supervision (Chen et al., 2020; Chuang et al., 2020; Grill et al., 2020; Khosla et al., 2020). Unsupervised CL learns representations of training examples by maximizing agreement between augmented views of the same example. Similarly, supervised CL maximizes agreement between augmented views of examples in the same class. Despite their empirical success, both supervised and unsupervised contrastive learning fail to capture *all* semantically relevant features in the data. In particular, *supervised* CL can fall prey to *class collapse* (Graf et al., 2021; Chen et al., 2022), where representations of *subclasses* within a class may no longer be distinguishable from each other; thus, yielding a poor classification performance at the subclass level. Similarly, *unsupervised* CL can be afflicted with *feature suppression* (Chen et al., 2021; Robinson et al., 2021) where easy but class-irrelevant features suppress the learning of harder class-relevant ones; deteriorating the generalizability of the obtained representations.

In spite of the significance of these failure modes, there is no clear theoretical understanding of them and consequently, no rigorous solution. Feature suppression has not been studied theoretically by prior work and the only theoretical work on class collapse (Graf et al., 2021) cannot explain why we observe class collapse at *test time*.

Addressing class collapse and feature suppression requires a theoretical understanding of *which* features CL learns. However, existing CL theory (Wang & Isola, 2020; Graf et al., 2021; Lee et al., 2021; Tosh et al., 2021a;b; Arora et al., 2019b; Tsai et al., 2020; HaoChen et al., 2021; Wen & Li, 2021; Ji et al., 2021) only explains how semantically relevant features are learned. The implicit assumption is that *all* semantically relevant features are learned, but the occurrence of class collapse and feature suppression proves otherwise. We propose the first unified (i.e. for both supervised and unsupervised CL) framework to answer *which* semantically relevant features are learned. We then leverage this framework to characterize class collapse and feature suppression. Table 1 summarizes the main findings in this paper, which are detailed below.

Class Collapse in Supervised CL. We prove that, perhaps surprisingly and in contrast to the current understanding (Graf et al., 2021), global minimizers of the supervised contrastive loss do not necessarily collapse the representations

[*]Equal contribution [1]Department of Computer Science, University of California, Los Angeles, USA [2]IBM Research. Correspondence to: Yihao Xue <yihaoxue@g.ucla.edu>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

Table 1. A concise overview of the key findings in this research. In the table, 'CC' and 'FS' refers to class collapse and feature suppression, respectively. 'Thm' and 'Exp' refers to theorem and experiment, respectively.

Loss	Finding	Thm/Exp	Implication
SCL	min loss $\not\Rightarrow$ CC	Thm 4.3	Simplicity bias of (S)GD contributes to CC
	(min loss & min norm) $\Rightarrow$CC	Thm 4.4 & 4.7	
	(S)GD learns subclasses early in training	Thm 4.5 & Exp	
	(S)GD eventually unlearns subclasses, leading to CC	Exp	
UCL	With insufficient embedding size, (min loss & min norm) $\Rightarrow$ FS	Thm 5.1 & Exp	Simplicity bias of (S)GD contributes to FS; Larger embedding size/better augmentation alleviates FS
	With imperfect data augmentation, (min loss & min norm) $\Rightarrow$ FS, even with sufficient embedding size	Thm 5.4	
Joint	Joint loss can avoid both CC and FS	Thm 6.1 & Exp	Justification of joint loss

of the subclasses at *test* time. We find, however, that the *minimum norm* global minimizer does suffer from class collapse on test data.

We then study minimizing the supervised contrastive loss using (S)GD and show that, interestingly, subclass features are learned early in training. However, we verify empirically, that as training proceeds, (S)GD forgets the learned subclass features and collapses class representations.

Altogether, our findings indicate that the bias of SGD towards finding simpler solutions (Lyu et al., 2021) is the main deriving factor in collapsing class representations.

Feature Suppression in Unsupervised CL. We provide the first theoretical characterization of feature suppression in unsupervised CL. In particular, we show that the *minimum norm* global minimizer of the unsupervised contrastive loss results in feature suppression, when the embedding dimensionality is small *or* when data augmentations preserve class-irrelevant features better than class-relevant features. Again, our results identify the simplicity bias of (S)GD as a key factor in suppressing features of the input data. In addition, our findings suggest practical solutions to the problem of feature suppression: increasing embedding dimensionality and/or improving the quality of data augmentations.

Theoretical Justification for Combining Supervised and Unsupervised CL to Obtain Superior Representations. Finally, we prove that the *minimum norm* global minimizer of the joint loss (weighted sum of the supervised and unsupervised contrastive loss) does not suffer from class collapse or feature suppression, explaining why Chen et al. (2022); Islam et al. (2021) observes this empirically (i.e. even when using SGD).

2. Related Work

Theory of CL. While there has been much progress in theoretically understanding CL, most prior work (Wang & Isola, 2020; Graf et al., 2021; Lee et al., 2021; Tosh et al., 2021a;b; Arora et al., 2019b; Tsai et al., 2020; HaoChen et al., 2021) are focused on understanding how CL clusters examples using semantically meaningful information or providing generalization guarantees on downstream tasks. Feature learning has only been studied by (Wen & Li, 2021; Ji et al., 2021) which show that CL learns semantically meaningful features from the data. In contrast, we show that CL may not learn *all* semantically relevant features.

Other important recent work (Saunshi et al., 2022; HaoChen & Ma, 2022) studied the role of inductive bias of the function class in the success of CL. Our analysis, however, is focused on understanding failure modes of CL i.e. class collapse and feature suppression.

Class Collapse in Supervised CL. Chen et al. (2022) empirically demonstrates *class collapse* on test data, but does not offer any rigorous theoretical explanation. Graf et al. (2021) proves that optimizing the supervised contrastive loss leads to class-collapsed training set representations. However, we show that there exist many minimizers with such class-collapsed training set representations and not all of them suffer from class collapse at *test time*. We also present the first theoretical characterization of class collapse at test time.

Feature Suppression in Unsupervised CL. Feature suppression has been empirically observed by Tian et al. (2020); Chen et al. (2021); Robinson et al. (2021) but we lack a theoretical formulation of this phenomenon. Li et al. (2020) shows that InfoNCE has local minimums that exhibit feature suppression, thus attributing this phenomenon to failure of optimizing the loss. However, Robinson et al. (2021) shows that the InfoNCE loss can be minimized by many models, some of which learn all task-relevant features, while others do not. We put forth the only theoretical characterization of feature suppression and consequently, use this understanding to suggest practical solutions to remedy this problem.

Joint Supervised and Unsupervised Contrastive Loss. Recently, several versions of loss functions that combine supervised and unsupervised contrastive losses (Islam et al., 2021; Chen et al., 2022) have been empirically observed to have superior transfer learning performance, by avoiding

class collapse. We provide the first theoretically rigorous analysis of which features the *minimum norm* global minimizer of the joint loss learns, provably demonstrating that it can avoid class collapse and feature suppression. To the best of our knowledge, this is the only theoretical result that can be used to understand the empirical success of joint losses.

3. Problem Formulation

3.1. Data distribution

We define data distribution $\mathcal{D}_{\text{orig}}$ below. Each example $(\boldsymbol{x}, y, y_{\text{sub}}) \in \mathcal{D}_{\text{orig}}$ is generated as follows:

$$\boldsymbol{x} = \boldsymbol{u} + \boldsymbol{\xi}, \qquad \text{where}$$
$$\boldsymbol{u} = (y\phi_1 + \mu_1)\boldsymbol{v}_1 + (y_{\text{sub}}\phi_2 + \mu_2)\boldsymbol{v}_2 + (\rho_k\phi_k + \mu_k)\boldsymbol{v}_k,$$

and k is uniformly selected from $3, \ldots, K$; and $y, y_{\text{sub}}, \rho_k$ are uniformly sampled from $\{-1, 1\}$.

Features and Noise. We assume features and noise form an orthonormal basis of $\mathbb{R}^d$, i.e., a set of unit orthogonal vectors $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_d\}$ in $\mathbb{R}^d$. W.l.o.g., one can let $\boldsymbol{v}$'s be the standard basis, where the first K basis are feature vectors. $\{\phi_1, \ldots, \phi_K\}$ are constants indicating the strength of each feature, and $\{\mu_1, \ldots, \mu_K\}$ are the means of the corresponding entries in the feature vectors. In particular:

- Class Feature: $\boldsymbol{v}_1$.
- Subclass Feature: $\boldsymbol{v}_2$.
- (Class and subclass) irrelevant features:[1] $\boldsymbol{v}_3, \ldots, \boldsymbol{v}_K$.
- Noise $\boldsymbol{\xi} \sim \mathcal{D}_\xi$: $\mathcal{D}_\xi$ is a uniform distribution over features $\sigma_\xi \boldsymbol{v}_1, \ldots, \sigma_\xi \boldsymbol{v}_d$, where σ_ξ indicates the variance of the noise.[2]

We sample n examples from $\mathcal{D}_{\text{orig}}$ to form the original dataset $\hat{\mathcal{D}}_{\text{orig}}$.

Assumption 3.1 (Balanced Dataset)**.** All combinations of $(y_i, y_{\text{sub},i}, k_i, \rho_i)$ are equally represented in $\hat{\mathcal{D}}_{\text{orig}}$.[3]

A Concrete Example of the Above Data Distribution. Let $y = 1$ be dogs and $y = -1$ be cats, $y_{\text{sub}} = 1$ if they are fluffy and $y_{\text{sub}} = -1$ if they are not-fluffy. Then $(\phi_1 + \mu_1)\boldsymbol{v}_1 + (\phi_2 + \mu_2)\boldsymbol{v}_2$ denotes a fluffy dog. Here, the background can be interpreted as an irrelevant feature: let $\rho_3 = 1$ for grass and $\rho_3 = -1$ for forest. Then $(\phi_1 + \mu_1)\boldsymbol{v}_1 + (\phi_2 + \mu_2)\boldsymbol{v}_2 + (\phi_3 + \mu_3)v_3$ represents a fluffy dog on grass. Note that each example only selects one irrelevant feature, which mimics the real world, where examples do not necessarily have all types of objects in the background i.e. many examples have neither grass or forests as their background.

Rationale for Including Feature Means μ_i. In general, it is unreasonable to expect all features to have 0 expectation over entire data, thus we introduce μ to further generalize our analysis. We find that considering a non-zero mean for the subclass feature is sufficient to provide novel insights into class collapse (Theorem 4.5). Therefore, for clarity, we set all the μ's except μ_2 to zero.

Relation to Sparse Coding Model. This data distribution is a variant of the sparse coding model which is usually considered as a provision model for studying the feature learning process in machine learning (e.g., (Zou et al., 2021; Wen & Li, 2021; Liu et al., 2021)). It naturally fits into many settings in machine learning, and in general mimics the outputs of intermediate layers of neural networks which have been shown to be sparse (Papyan et al., 2017). It is also used to model the sparse occurrences of objects in image tasks (Olshausen & Field, 1997; Vinje & Gallant, 2000; Foldiak, 2003; Protter & Elad, 2008; Yang et al., 2009; Mairal et al., 2014) and polysemy of words in language tasks (Arora et al., 2018).

3.2. Data Augmentation $\mathscr{A}(\cdot)$

For each example in $\hat{\mathcal{D}}_{\text{orig}}$, we generate m augmentations to form $\hat{\mathcal{D}}_{\text{aug}}$. We consider the following augmentation strategy: given an example $\boldsymbol{x} = \boldsymbol{u} + \boldsymbol{\xi}$, its augmentation is given by $\mathscr{A}(\boldsymbol{x}) = \boldsymbol{u} + \boldsymbol{\xi}'$, where $\boldsymbol{\xi}'$ is a new random variable from $\mathcal{D}_\xi$ independent of $\boldsymbol{\xi}$. This is an abstract of augmentations used in practice where two augmentations from the same example share certain parts of the features and have the correlation between their noise parts removed or weakened.

Assumption 3.2 (High dimensional regime)**.** d is at least $\omega(n^2 m^2)$.

Assumption 3.3 (Sufficient sample size)**.** The noise-to-sample-size ratio is not too large $\frac{\sigma_\xi^2}{mn} = o(1)$.

3.3. Linear Model

We consider a linear model with p outputs. The model has weights $\boldsymbol{W} \in \mathbb{R}^{p \times d}$ and bias $\boldsymbol{b} \in \mathbb{R}^p$ where $p \geq 3$. The function represented by the model is $f_{\boldsymbol{\Theta}}(\boldsymbol{x}) = \boldsymbol{W}\boldsymbol{x} + \boldsymbol{b}$, where we define $\boldsymbol{\Theta} \in \mathbb{R}^{p \times (d+1)}$ as the concatenated parameter $[\boldsymbol{W} \;\; \boldsymbol{b}]$. We establish theoretical proofs of class collapse and feature suppression for linear model, and also empirically verified them for (non-linear) deep neural networks.

[1] In the rest of the paper, we use irrelevant features to refer to features that may have semantic meaning but are irrelevant to class and subclass.

[2] This definition of noise is nearly identical to Gaussian noise $\mathcal{N}(0, \frac{\sigma_\xi^2}{d}\mathbf{I}_d)$ in the high-dimensional regime but keeps the analysis clear. Our results can be extended to the Gaussian noise setting.

[3] This can be approximately achieved when n is sufficiently larger than K. While our analysis can be generalized to consider imbalanced data, this is outside the scope of this work.

3.4. Loss function

For unsupervised contrastive learning, we use the unsupervised spectral contrastive loss popular in prior theoretical and empirical work (HaoChen et al., 2021; Saunshi et al., 2022; HaoChen & Ma, 2022) and for supervised contrastive learning, we consider the natural generalization of this loss to incorporate supervision. Let $\mathcal{A}_i$ denote the set of augmentations in $\hat{\mathcal{D}}_{\text{aug}}$ generated from the i-th original example with $\mathscr{A}(\cdot)$. Let $\mathcal{S}_{+1}$ and $\mathcal{S}_{-1}$ denote the set of augmentations in $\hat{\mathcal{D}}_{\text{aug}}$ with class labels $+1$ and -1, respectively. Let $\hat{\mathbb{E}}$ denote the empirical expectation. Then we have the following loss functions:

$$\begin{aligned}\mathcal{L}_{\text{UCL}}(\boldsymbol{\Theta}) = &-2\hat{\mathbb{E}}_{i\in[n],\boldsymbol{x}\in\mathcal{A}_i,\boldsymbol{x}^+\in\mathcal{A}_i}\left[f_\Theta(\boldsymbol{x})^\top f_\Theta(\boldsymbol{x}^+)\right] \\ &+\hat{\mathbb{E}}_{\boldsymbol{x}\in\hat{\mathcal{D}}_{\text{aug}},\boldsymbol{x}^-\in\hat{\mathcal{D}}_{\text{aug}}}\left[(f_\Theta(\boldsymbol{x})^\top f_\Theta(\boldsymbol{x}^-))^2\right] \quad (1)\\ \mathcal{L}_{\text{SCL}}(\boldsymbol{\Theta}) = &-2\hat{\mathbb{E}}_{c\in\{-1,1\},\boldsymbol{x}\in\mathcal{S}_c,\boldsymbol{x}^+\in\mathcal{S}_c}\left[f_\Theta(\boldsymbol{x})^\top f_\Theta(\boldsymbol{x}^+)\right] \\ &+\hat{\mathbb{E}}_{\boldsymbol{x}\in\hat{\mathcal{D}}_{\text{aug}},\boldsymbol{x}^-\in\hat{\mathcal{D}}_{\text{aug}}}\left[(f_\Theta(\boldsymbol{x})^\top f_\Theta(\boldsymbol{x}^-))^2\right]. \quad (2)\end{aligned}$$

4. Simplicity Bias Contributes to Class Collapse in Supervised CL

We make two key observations through our theoretical analysis and experiments (henceforth we refer to class collapse at *test time* simply as 'class collapse'):

1. Theoretically, not all global minimizers exhibit class collapse, but the *minimum norm* minimizer does.
2. Theoretically and empirically, when the model is trained using (S)GD, some subclasses are *provably* learned early in training. Empirically, however, those subclasses will eventually be unlearned i.e. S(GD) converges to minimizers that exhibit class collapse.

Altogether, these observations suggest that class collapse, which has been observed in practice when certain gradient-based algorithms are used to minimize the loss, cannot be explained by simply analyzing the loss function. This highlights the importance of studying the dynamics and inductive bias of training algorithms in contrastive learning.

4.1. What Minimizers Have Class Collapse?

We first define class collapse in terms of the alignment between the model weights and the subclass feature.

Definition 4.1 (Exact class collapse)**.** We say exact class collapse happens at test time when:

$$\forall \boldsymbol{\beta}\in\mathbb{R}^p, \Pr_{(\boldsymbol{x},y,y_{\text{sub}})\sim\mathcal{D}_{\text{orig}}}(y_{\text{sub}}\boldsymbol{\beta}^\top f_{\boldsymbol{\Theta}}(\boldsymbol{x})>0)=1/2.$$

The definition means that no linear classifier on the embeddings of examples drawn from $\mathcal{D}_{\text{orig}}$ can predict the subclass label with accuracy beyond random guess.[4]

This is different from class collapse on the training set which is not defined on the population set $\mathcal{D}_{\text{orig}}$ but on the training samples $\hat{\mathcal{D}}_{\text{orig}}$.

Proposition 4.2. *For any $\boldsymbol{\Theta}^*\in\min_{\boldsymbol{\Theta}}\mathcal{L}_{SCL}(\boldsymbol{\Theta})$, we have $f_{\boldsymbol{\Theta}^*}(\boldsymbol{x}_i)=f_{\boldsymbol{\Theta}^*}(\boldsymbol{x}_j)$ for all $\boldsymbol{x}_i,\boldsymbol{x}_j$ in the training set $\hat{\mathcal{D}}_{aug}$ such that $y_i=y_j$.*

This directly implies that minimizing the loss results in class collapse on the training set. However, the following theorem 4.3 shows that minimizing the loss does not necessarily lead to class collapse on the test set. To determine whether class collapse occurs, we need to determine whether the model learns the subclass feature. With a linear model, this exactly corresponds to constant alignment between weights and the subclass feature.

Theorem 4.3 (Minimizing $\mathcal{L}_{\text{SCL}}\not\Rightarrow$ Class Collapse)**.** *With high probability i.e. at least $1-O(\frac{m^2n^2}{d})=1-o(1)$, there exists $\boldsymbol{\Theta}^*=[\boldsymbol{W}^*\ \ \boldsymbol{b}^*]$ such that $\boldsymbol{\Theta}^*\in\min_{\boldsymbol{\Theta}}\mathcal{L}_{SCL}(\boldsymbol{\Theta})$ $\boldsymbol{W}^*$ has constant alignment with subclass feature $\boldsymbol{v}_2$ i.e.*

$$\|\boldsymbol{W}^*\boldsymbol{v}_2\|=\Omega(1).$$

Hence, there exists a linear classifier in the embedding space that can predict subclass labels almost perfectly. I.e.,

$$\exists\boldsymbol{\beta},\ s.t.\ \Pr_{(\boldsymbol{x},y,y_{sub})\sim\mathcal{D}_{orig}}(y_{sub}\boldsymbol{\beta}^\top\boldsymbol{W}^*\boldsymbol{x}>0|y)=1-o(1).$$

We prove the theorem in Appendix D. The proof utilizes Lemma C.1 which implies that, due to the high-dimensionality, the noise vectors have non-trivial effects on the empirical covariance matrix by rotating its kernel space. This results in the kernel space to have a $\Theta(\frac{\sigma_\xi}{\sqrt{mn}})$ alignment with the subclass feature. Since minimizers of the loss can behave arbitrarily on this kernel space, without any additional restriction, they can have any alignment with the subclass feature.

Next, we show that, the *minimum norm* minimizer exhibits class collapse.

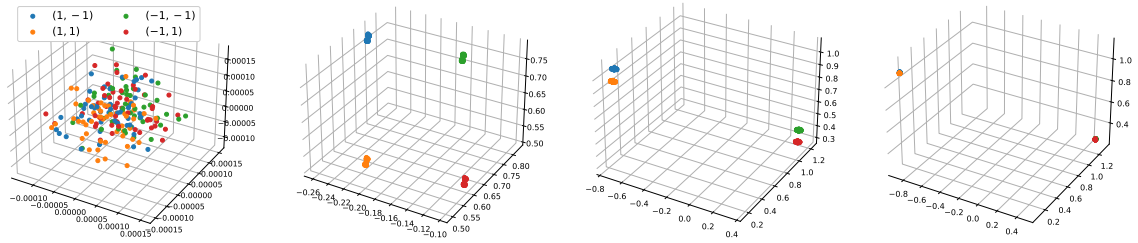
Theorem 4.4 (Minimizing $\mathcal{L}_{\text{SCL}}$ + Minimum Norm $\Longrightarrow$ Class Collapse)**.** *Assume $\mu_2=0$. Let $\boldsymbol{\Theta}^{**}=[\boldsymbol{W}^{**}\ \ \boldsymbol{b}^{**}]$ be the minimum norm minimizer of $\mathcal{L}_{SCL}$, i.e.,*

$$\boldsymbol{\Theta}^{**}=\arg\min_{\boldsymbol{\Theta}^*}\|\boldsymbol{\Theta}^*\|_F\ s.t.\ \boldsymbol{\Theta}^*\in\arg\min_{\boldsymbol{\Theta}}\mathcal{L}_{SCL}(\boldsymbol{\Theta}).$$

Then with high probability i.e. at least $1-O(\frac{m^2n^2}{d})=1-o(1)$, $\boldsymbol{W}^{**}$ has no alignment with subclass feature $\boldsymbol{v}_2$ i.e.

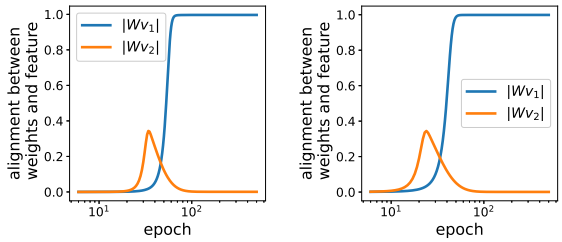
$$\|\boldsymbol{W}^{**}\boldsymbol{v}_2\|=0.$$

[4] Actually we are able to analyze a stronger version of class collapse: $\Pr_{(\boldsymbol{x},y,y_{\text{sub}})\sim\mathcal{D}_{\text{orig}}}(f_{\boldsymbol{\Theta}}(\boldsymbol{x})|y_{\text{sub}})=\Pr_{(\boldsymbol{x},y,y_{\text{sub}})\sim\mathcal{D}_{\text{orig}}}(f_{\boldsymbol{\Theta}}(\boldsymbol{x}))$, which means the distributions of embeddings given and not given the subclass label are exactly the same. Nonetheless, we present this simpler formulation for clarity.



(a) epoch 0 (b) epoch 45 (c) epoch 60 (d) epoch 100

Figure 1. Visualization of the embedding space at different epochs. We let $p = 3$ so that we can see the whole embedding space from a 3D plot. Other parameters: $n = 1000, m = 5, d = 2000, K = 4, \phi_1 = \phi_2 = \phi_3 = \phi_4 = 1, \mu = 1, \sigma = 2, \sigma_0 = 0.001, \eta = 0.05$. Colors represent combinations of class and subclass labels (y, y_{sub}). We use test examples for the plots. At epoch 45, the four groups of examples are well separated in the embedding space. However groups in the same classes are merged afterwards.



(a) $p = 3$ (b) $p = 500$

Figure 2. $\|\boldsymbol{W}_t\boldsymbol{v}_1\|$ and $\|\boldsymbol{W}_t\boldsymbol{v}_2\|$ at different epochs. Both features are learned early in training, but $\boldsymbol{v}_2$ is unlearned later.

This means class collapse occurs at test time (Definition 4.1), and no linear classifier does better than random guess for predicting subclass labels.

Theorems 4.3 and 4.4 show that minimizing the training loss does not necessarily lead to class collapse on test data, but does with additional constraint on the weights of the model. This is not due to a degenerate solution, as we show that both minimizers learn the class feature $\boldsymbol{v}_1$ (see corollary C.5).

4.2. Intriguing Properties of GD

We now further our theoretical characterization of class collapse by investigating the setting where $\mathcal{L}_{\text{SCL}}$ is minimized by GD. This is an important step toward understanding class collapse in practice, where similar optimization algorithms are used to minimize the loss. Our findings indicate that it is likely the simplicity bias of commonly used optimization algorithms that eventually leads to class collapse.

We consider GD with a constant learning rate η. The weights are initialized from a Gaussian distribution, i.e., the initial weight $\boldsymbol{\Theta}_0$ has each of its element drawn from $\mathcal{N}(0, \frac{\sigma_0^2}{d})$. And the weights at training epoch t are given by:

$$\boldsymbol{\Theta}_t = \boldsymbol{\Theta}_{t-1} - \eta \nabla_{\boldsymbol{\Theta}} \mathcal{L}_{\text{SCL}}(\boldsymbol{\Theta}_{t-1}).$$

Early in Training Some Subclasses are Provably Learned. By analyzing the training dynamics of GD, we find that subclasses are learned early in training.

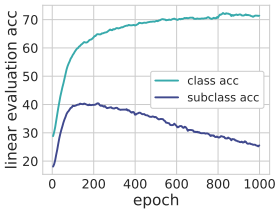
Theorem 4.5 (Early in training subclass features are learned)**.** *Assume $\sigma_0\sqrt{\frac{p}{d}} = o(1)$ and $\sigma_\xi = o(1)$. If the subclass feature has a constant non-zero mean such that $1 + \mu^2 > \phi_1^2$, then with probability at least $1 - O(\frac{m^2n^2}{d} + \frac{1}{\text{poly}(p)}) = 1 - o(1)$ the following holds:*

- $\|\boldsymbol{W}_0\boldsymbol{v}_2\| = o(1)$.
- $\exists t = O(\ln(\frac{1}{\sigma_0}\sqrt{\frac{d}{p}}))$, *s.t.* $\|\boldsymbol{W}_t\boldsymbol{v}_2\| = \Omega(1)$, *and*
- $\exists \boldsymbol{\beta}$, *s.t.* $\Pr_{(\boldsymbol{x},y,y_{sub})\sim\mathcal{D}_{orig}}(y_{sub}\boldsymbol{\beta}^\top \boldsymbol{W}_t\boldsymbol{x} > 0|y) = 1 - o(1)$.

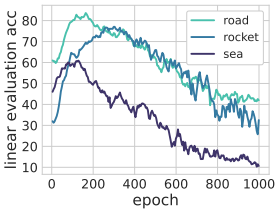
The above theorem shows that there exists an epoch where the weights have constant alignment with the subclass feature and produce distinguishable subclass embeddings (proof in Appendix G).

The key step of our analysis is showing that early in training, GD aligns the weights with the first eigenvector of the covariance matrix of class centers. This alignment grows exponentially faster than alignments with any other directions. When $1 + \mu^2 > \phi_1^2$, the subclass feature has a constant projection onto the first eigenvector and is therefore learned by the model.

More importantly, the same phenomenon can be observed in *neural networks*. We use SGD to train a ResNet18 (He et al., 2016) on CIFAR-100 (Krizhevsky et al., 2009) with supervised CL loss (Khosla et al., 2020) with 20 class (superclass) labels, and perform linear evaluation on embeddings of test data with 100 subclass (class) labels (see details in Appendix H). We observe that the subclass accuracy increases during the first 200 epochs before it starts to drop (Figure 3(a)). Some subclasses can even achieve a high accuracy around 80% (Figure 3(b)). This is surprising as it confirms that models trained with commonly used loss functions *do* learn subclass features early in training.



(a) Average subclass accuracy and class accuracy



(b) Subclasses are initially learned well but later unlearned

Figure 3. (a) Average subclass accuracy and class accuracy. (b) Accuracy in subclasses 'road', 'rocket' and 'sea'. In both plots, the subclass accuracy increases and then decreases, which confirms that subclasses are learned early in training before class collapse happens. The class accuracy only increases during training.

Empirical Evidence Showing that Class Collapse Eventually Happens in (S)GD. We simulate our theoretical analysis using numerical experiments to show that gradient descent converges to a minimizer that exhibits class collapse, despite learning subclasses early in training. We visualize the embeddings of test data at different epochs in Figure 1, and plot the alignment between weights and class/subclass features in Figure 2. Subclasses are perfectly separated and the weights align with both $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ after around 100 epochs of training. The model then starts unlearning $\boldsymbol{v}_2$ which causes the alignment to drop, thus subclasses are merged in the embedding space. We also confirm that same conclusion holds for neural networks in realistic settings. In Figure 3, we see that the subclass accuracy drops after around 200 epochs of training and eventually reaches a low value. In contrast, the class accuracy does not drop during training.

Minimum Norm Minimizer Exhibits Class Collapse. Note that in Theorem 4.5, assuming $\mu \neq 0$ leads us to discovering that subclasses are learned early in training. Here, we extend Theorem 4.4 to this setting under asymptotic class collapse.

Definition 4.6 (Asymptotic Class Collapse)**.** We say asymptotic class collapse happens when $\|\boldsymbol{W}\boldsymbol{v}_2\| = O(\frac{\sigma_\xi}{\sqrt{mn}}) = o(1)$.

This definition implies that: (1) representations of subclasses are not well separated, hence it is nearly impossible to distinguish between them, and (2) the distinguishability of subclasses is at odds with generalization, which improves as number of augmented views per example m and size of training data n increase. Thus, while this definition is a relaxation of Definition 4.1, practically, this results in equally severe class collapse.

Theorem 4.7 (Extension of Theorem 4.4 for $\mu_2 \neq 0$)**.** *Let* $\boldsymbol{\Theta}^{**} = [\boldsymbol{W}^{**}\ \boldsymbol{b}^{**}]$ *be the minimum norm minimizer of* $\mathcal{L}_{SCL}$*:*

$$\boldsymbol{\Theta}^{**} = \arg\min_{\boldsymbol{\Theta}^*} \|\boldsymbol{\Theta}^*\|_F \ s.t.\ \boldsymbol{\Theta}^* \in \arg\min_{\boldsymbol{\Theta}} \mathcal{L}_{SCL}(\boldsymbol{\Theta}).$$

Then with probability at least $1 - O(\frac{m^2n^2}{d}) = 1 - o(1)$*, asymptotic class collapse happens, i.e.,*

$$\|\boldsymbol{W}^{**}\boldsymbol{v}_2\| = O(\frac{\sigma_\xi}{\sqrt{mn}}) = o(1).$$

4.3. Simplicity Bias of (S)GD

We reiterate our main findings:

1. Minimizing the supervised contrastive loss *does not* necessarily lead to class collapse.
2. However, *simpler* minimizers of the supervised contrastive loss (e.g. *minimum norm*) do suffer from class collapse.
3. Optimizing with (S)GD does learn the subclass features early in training, but eventually unlearns them, resulting in class collapse.

These coupled with the fact that (S)GD is known to have a bias towards simpler solutions (Kalimeris et al., 2019) prompt us to conjecture:

The simplicity bias of (S)GD leads it to unlearn subclass features, thus causing class collapse.

The simplicity bias of (S)GD has not been rigorously studied for CL, and our results indicate the surprising role it may play in class collapse. Note that, the supervised contrastive loss is different than common supervised objectives, where the role of such bias of (S)GD is understood better (Gunasekar et al., 2018; Soudry et al., 2018; Ji & Telgarsky, 2019; Wu et al., 2019; Lyu et al., 2021). Rather, the supervised CL objective can be reformulated as a matrix factorization objective (Eq. 39), where the debate on the bias of (S)GD (e.g., minimum norm (Gunasekar et al., 2017) or rank (Arora et al., 2019a; Razin & Cohen, 2020)) is still ongoing.

5. Understanding Feature Suppression in Unsupervised CL

Empirically, feature suppression can be observed due to a variety of reasons (Li et al., 2020; Chen et al., 2021; Robinson et al., 2021). Easy features for unsupervised CL are those that allow the model to discriminate between examples (highly discriminative). Here, we consider different ways irrelevant features can be easy (highly discriminative) and characterize how this can lead to feature suppression. We show that the types of feature suppression we consider can be largely attributed to insufficient embedding dimensionality and/or poor data augmentations. Surprisingly, we find again that the minimum norm simplicity bias is critical in explaining this phenomenon.

5.1. Feature Suppression due to Easy Irrelevant Features and Limited Embedding Space

In Theorem 5.1, we show that easy (discriminative) irrelevant features can suppress the class feature when the embedding dimensionality is limited. For clarity, we let $\mu_2 = 0$.

Theorem 5.1 (Feature Suppression 1)**.** *Assume $p \leq K$. Let L be the $(K+1)$-element tuple $\left[1, \phi_1^2, \phi_2^2, \frac{\phi_3^2}{K-2}, \ldots, \frac{\phi_K^2}{K-2}\right]$ whose last K elements are the variances of features. If ϕ_1^2 is not among the p largest elements in L, then with probability at least $1-O(\frac{m^2n^2}{d}) = 1-o(1)$: (1) there exists a global minimizer $\boldsymbol{\Theta}^*$ of $\mathcal{L}_{UCL}$ such that $\|\boldsymbol{W}^*\boldsymbol{v}_1\| = \Omega(1)$, (2) However, the minimum norm minimizer $\boldsymbol{\Theta}^{**}$ satisfies $\|\boldsymbol{W}^{**}\boldsymbol{v}_1\| = 0$.*

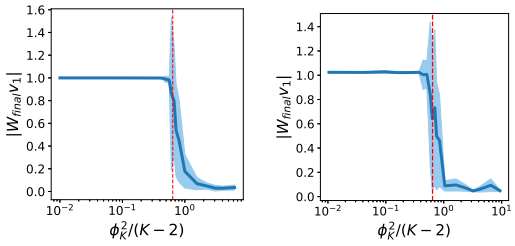
We prove the theorem in Appendix E. The elements except the first one in tuple L can be interpreted as the variance of examples at each coordinate $\boldsymbol{v}_k, k = 1, 2, \ldots, K$, which indicates how much the examples are discriminated by each feature. The theorem shows that when the embedding space is not large enough to represent all the K features (which requires $K+1$ dimensions), the minimum norm minimizer only picks the most discriminative ones. In practice, the embedding space in unsupervised CL is relatively low-dimensional (compared to input dimensionality) and thus the model cannot fit all the information about inputs into the embedding space. As is suggested by Theorem 5.1, if the training algorithm prefers functions with certain simple structures, only the easiest (most discriminative) features that can be mapped into the embedding space by less complex functions (e.g., smaller norm) are learned. The class features are suppressed if they are not amongst the easiest ones.

Remark 5.2. Following the same analysis we can also show that when ϕ_1 is among the p largest elements in L, i.e., the class feature is among the easiest (most discriminative) ones, the class feature $\boldsymbol{v}_1$ is learned by the minimum norm minimizer; when ϕ_1 is exactly on par with some other element as the p-th largest, there exist both minimum norm minimizers that learn and do not learn the class feature $\boldsymbol{v}_1$.

Numerical Experiments with GD. Our theory for the minimum norm minimizer matches the experimental results for models trained with GD. We let $p = K$ and let $1 \geq \phi_2^2 \geq \frac{\phi_3^2}{K-2} \geq \cdots \geq \frac{\phi_{K-1}^2}{K-2} > \phi_1^2$ so that ϕ_1^2 must be among the smallest two variances i.e. $\boldsymbol{v}_1$ is among the two most difficult features. Then we vary ϕ_K and see how the trained weights align with $\boldsymbol{v}_1$. Consistent with Theorem 1, Figure 4 shows that $\boldsymbol{v}_1$ is suppressed when $\frac{\phi_K^2}{K-2} > \phi_1^2$. Interestingly, we also see that the result at $\frac{\phi_K^2}{K-2} = \phi_1^2$ diverges, indicating that GD can find both minimizers that learn and do not learn $\boldsymbol{v}_1$ when the variances at $\boldsymbol{v}_1$ and $\boldsymbol{v}_K$ are the same.

Empirically Verifying Benefits of Larger Embedding Size. Theorem 5.1 also provides one practical solution for feature suppression due to limited embedding size: increasing the embedding size so that every feature can be learned by the model. To provide empirical evidence for this, we conduct two sets of experiments:

First, we train 5-layer convolutional networks on the RandomBit dataset with the same setup as in (Chen et al., 2021), but we vary the embedding size (see details in Appendix H). Varying the # bits in the extra channel intuitively controls how discriminative the irrelevant feature are, i.e., how easy-to-learn it is for CL. In this setting, the



(a) $K = 3, p = 3$ (b) $K = 50, p = 50$

Figure 4. **The irrelevant feature suppresses the class feature when its variance is beyond the variance of the class feature (the red vertical line).** We let $d = 2000, p = K, \phi_1 = 0.8, \phi_2 = 1, \mu = 0, \frac{\phi_k^2}{K-2} > \phi_1, \forall k \in [K-1]$ and vary ϕ_K. Thus whether ϕ_1^2 is among the p largest variances only depends on ϕ_K. We train the linear model to convergence. Plots show that the alignment between the trained weights and $\boldsymbol{v}_1$ drops when ϕ_K increases. We report the average of 10 runs. The result diverges at $\frac{\phi_K^2}{K-2} = \phi_1^2$ indicating that the model can learn either $\boldsymbol{v}_1$ or $\boldsymbol{v}_K$ in this case.

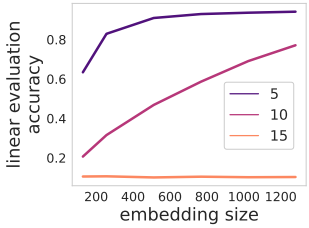


Figure 5. Effect of embedding size on feature suppression in MNIST RandBit(Chen et al., 2021). Legends show the number of bits in the extra channel which indicates how easy (discriminative) the irrelevant features are. We observe that (1) increasing the easiness of irrelevant features exacerbates feature suppression; (2) increasing the embedding size alleviates feature suppression.

Table 2. Effect of embedding size on feature suppression in CIFAR-10/100 RandBit. 'Acc' refers to class accuracy and 'Sub Acc' refers to subclass accuracy. We see that increasing embedding size alleviates feature suppression, improving class/subclass accuracy.

w	CIFAR-10 RandBit		CIFAR-100 RandBit	
	Sub Acc	Acc	Sub Acc	Acc
4	34.38	86.73	11.67	23.53
64	71.96	96.82	34.11	52.32
128	76.69	97.65	38.51	57.40

random bit can suppress the MNIST digits. We make two observations in Figure 5.1: (1) with a fixed embedding size, increasing easiness (number of random bits) of the irrelevant features exacerbates feature suppression; (2) with a fixed easiness of irrelevant features, increasing the embedding size alleviates feature suppression.

Second, we train ResNet18 (He et al., 2016) on the CIFAR-10/100 RandBit Dataset, constructed similarly to the MNIST RandBit dataset but with images from CIFAR-10/100 (Krizhevsky et al., 2009) (see Appendix H.1). For CIFAR-10, we use 2 random bits, and for CIFAR-100, we use one random bit as the class irrelevant features. Table 2 presents the test performance for different values of the model width w, where a larger w indicates a larger embedding size (see Appendix H.3 for details). On both datasets, increasing the embedding size alleviates feature suppression, leading to improvements in both class and subclass accuracies. We also provide additional experiments and discussion in Appendix H.3. Both experimental results confirm the conclusion drawn from the theoretical analysis.

5.2. Feature Suppression due to High-dimensional Irrelevant Features and Imperfect Augmentation

Empirically, another form of feature suppression has been observed that cannot be remedied by larger embedding dimensionality (Li et al., 2020). We characterize this form of feature suppression by defining easy irrelevant features as being: (1) drawn from a high dimensional space so that the collection of irrelevant features is large and discriminating based on irrelevant features is easier, (2) less altered by data augmentation compared to the class feature.

For (1), formally we assume $K = \omega(n^2)$, as opposed to assumption 3.1 which implies that K is smaller than n. A consequence of this assumption is that with high probability the n original examples each have a unique irrelevant feature. For (2) we consider the following imperfect data augmentation:

Definition 5.3 (Imperfect data augmentation $\mathscr{A}'(\cdot)$)**.** For a given example $\boldsymbol{x} = \boldsymbol{\mu} + \boldsymbol{\xi} \in \hat{\mathcal{D}}_{\text{orig}}$,

$$\mathscr{A}'(\boldsymbol{x}) = \boldsymbol{u} + \zeta' \boldsymbol{v}_1 + \zeta'' \boldsymbol{v}_2 + \boldsymbol{\xi}',$$

where $\zeta' \sim \mathcal{N}(0, \sigma_\zeta'^2)$, $\zeta'' \sim \mathcal{N}(0, \sigma_\zeta''^2)$, $\sigma_\zeta'^2, \sigma_\zeta''^2 \neq 0$ and $\boldsymbol{\xi}'$ is a new random variable drawn from $\mathcal{N}(\boldsymbol{\xi}, \boldsymbol{\Sigma}_{\boldsymbol{\xi}})$ with $\text{rank}(\boldsymbol{\Sigma}_{\boldsymbol{\xi}}) \leq \frac{m}{2}$.

In the definition, the data augmentation adds small perturbations (ζ' and ζ'') to class and subclass features, weakly alters the noise, but preserves the irrelevant features. For example, on Colorful-Moving-MNIST (Tian et al., 2020) constructed by assigning each MNIST digit a background object image selected randomly from STL-10, the colorful background objects are high-dimensional and the colors are invariant to data augmentations without color distortion.

Theorem 5.4 (Feature Suppression 2)**.** *If $K = \omega(n^2)$ and augmentation is $\mathscr{A}'(\cdot)$, with probability $\geq 1 - o(\frac{n^2 m^2}{d} + \frac{1}{n}) = 1 - o(1)$, the minimum norm minimizer $\boldsymbol{\Theta}^* = [\boldsymbol{W}^*, \boldsymbol{b}^*]$ satisfies $\|\boldsymbol{W}^* \boldsymbol{v}_1\| = 0$.*

This theorem shows that feature suppression can happen even when embedding dimensionality p is arbitrarily large and helps understand empirical observations made both in our work (Figure 5.1, the line with 15 bits) and previous work. For example Li et al. (2020) showed that on Colorful-Moving-MNIST, the colorful background can suppress learning the digits especially when color distortion is not used in augmentation, and increasing embedding size does not address the issue.

In conclusion, Theorem 5.4 highlights that designing data augmentations that disrupt the highly-discriminative irrelevant features is a key to addressing feature suppression.

6. Combining Supervised and Unsupervised CL Losses Can Avoid Both Class Collapse and Feature Suppression

We now consider the following loss which is a weighted sum of the supervised and unsupervised CL loss functions:

$$\mathcal{L}_{\text{joint},\beta}(\boldsymbol{\Theta}) = \beta \mathcal{L}_{\text{SCL}}(\boldsymbol{\Theta}) + (1-\beta)\mathcal{L}_{\text{UCL}}(\boldsymbol{\Theta}).$$

Similar loss functions have been proposed recently with notable empirical success. For example, Chen et al. (2022) put forth a weighted sum of supervised CL loss and class-conditional InfoNCE (which has similar effect as $\mathcal{L}_{\text{UCL}}$ in our setting) to avoid class collapse. Islam et al. (2021) empirically observed that the joint objective of supervised and unsupervised contrastive loss leads to better transferability of the learned models than their supervised counterparts. However, we still lack a theoretical understanding of why this weighted sum of losses can outperform both losses.

From our investigation of class collapse and feature suppression, the benefit of the joint objective $\mathcal{L}_{\text{joint}}$ becomes evident: the unsupervised term in $\mathcal{L}_{\text{joint}}$ increases the chance of learning features that do not appear relevant to the labels but might be useful for downstream tasks, while the supervised term in $\mathcal{L}_{\text{joint}}$ ensures that even hard-to-learn class features are learnt. Thus, $\mathcal{L}_{\text{joint}}$ can learn rich representations capturing more task relevant information than either $\mathcal{L}_{\text{UCL}}(\boldsymbol{\Theta})$ or $\mathcal{L}_{\text{SCL}}(\boldsymbol{\Theta})$. We show below that with an appropriate choice of β, $\mathcal{L}_{\text{joint}}$ can provably succeed where $\mathcal{L}_{\text{SCL}}$ fails due to collapse and $\mathcal{L}_{\text{UCL}}$ fails due to feature suppression (for clarity, we let $\mu = 0$).

Theorem 6.1. *W.L.O.G., assume $\phi_3 \geq \phi_4 \geq \cdots \geq \phi_K$. If $p \leq K$, $\phi_2^2 > \frac{\phi_{p-2}^2}{K-2}$ and $\phi_1^2 < \frac{\phi_{p-1}^2}{K-2}$, then by Theorem 4.4 the minimum norm minimizer of $\mathcal{L}_{SCL}$ suffers from class collapse and by Theorem 5.1 the minimum norm minimizer*

Table 3. Joint loss alleviates class collapse on CIFAR-100.

Loss	Subclass Acc
SCL	26.11
Joint loss ($\beta = 0.8$)	41.37

Table 4. Joint loss alleviates feature suppresion on MNIST RandBit.

Loss	Class Acc
UCL	61.21
Joint loss ($\beta = 0.5$)	79.37

Table 5. Joint loss alleviates both class collapse and feature suppresion on CIFAR-100 RandBit.

Loss	Subclass Acc	Class Acc
SCL	28.13	61.10
UCL	34.11	52.32
Joint loss ($\beta = 0.8$)	35.72	63.94

of $\mathcal{L}_{UCL}$ suffers from feature suppresion. However, for constant $\beta \in (0, 1)$, the minimum norm minimizer of $\mathcal{L}_{joint,\beta}$, denoted by $\boldsymbol{\Theta}^* = [\boldsymbol{W}^* \ \ \boldsymbol{b}^*]$, satisfies $\|\boldsymbol{W}^*\boldsymbol{v}_1\| = \Omega(1)$ and $\|\boldsymbol{W}^*\boldsymbol{v}_2\| = \Omega(1)$.

Empirically Verifying Benefits of the Joint Loss. We empirically examine the impact of the joint loss on MNIST RandBit, CIFAR-100, and CIFAR-100 RandBit. The training details are in Appenidx H.2. The results indicate that the joint loss significantly improves performance in scenarios where SCL suffers from class collapse (Table 3) and UCL suffers from feature suppression (Table 4). Furthermore, on CIFAR-100 RandBit dataset, where both phenomena can occur simultaneously, the joint loss effectively alleviates both issues (Table 5).

7. Discussion

Negative Impact of Simplicity Bias in Deep Learning. The simplicity bias of optimization algorithms has been studied as a key beneficial factor in achieving good generalization (Gunasekar et al., 2017; 2018; Soudry et al., 2018; Ji & Telgarsky, 2019; Wu et al., 2019; Lyu et al., 2021). However, our study reveals the negative impact of simplicity bias in CL. In fact, it has also been conjectured to lead to undesirable outcomes in other scenarios, such as learning spurious correlations (Sagawa et al., 2020) and shortcut solutions (Robinson et al., 2021). We hope our study can inspire further theoretical characterization of the negative role of simplicity bias in these scenarios, thereby deepening our understanding and fostering potential solutions.

Connection to Neural Collapse. Neural collapse (NC) (Papyan et al., 2020) refers to the collapse of representations within each class in supervised learning. Similar to the rationale in this study, overparameterized models that exhibit NC on training data can demonstrate different behaviors on test data due to their capacity to implement training set NC in various ways, and it is worth considering whether current theoretical frameworks (Han et al., 2021; Zhu et al., 2021; Zhou et al., 2022b;a; Lu & Steinerberger, 2022; Fang et al., 2021) can effectively capture NC on test data. In fact, the empirical results in (Hui et al., 2022) emphasize the distinction between NC on training and test data, as there can be an inverse correlation between the two. Our results suggest that analyzing the learned features and considering the inductive bias of training algorithms can aid in this distinction.

Theoretical Characterization of Class Collapse in (S)GD. The results in Section 4.2 highlight the need for theoretical characterization of class collapse in (S)GD. We provide two potential approaches for future investigation. (1) Given that the objective can be reformulated as matrix factorization (Eq. 39), and our Theorems 4.4 and 4.7 on minimum norm minimizer, it is reasonable to investigate whether the implicit bias of (S)GD is to seek the minimum norm solution. We note that understanding the implicit bias in matrix factorization is a longstanding pursuit in the machine learning community, with no consensus reached thus far (see Appendix I.1). Hence, further effort is still needed. (2) As elaborated in Appendix I.2, the gradient consists of two terms with distinct roles. One promotes alignment with the subclass feature, while the other counteracts its influence. The relative scale of these two terms undergoes a phase transition (Figure 6), and analyzing this can provide insights into class collapse.

8. Conclusion

To conclude, we present the first theoretically rigorous characterization of the failure modes of CL: class collapse and feature suppression at test time. We explicitly construct minimizers of supervised contrastive loss to show that optimizing this loss does not necessarily lead to class collapse. Then we show that the minimum norm minimizer does exhibit class collapse. Our analysis also reveals a peculiar phenomenon for supervised CL, when optimized with (S)GD: subclass features are learned early in training and then unlearned. To analyze feature suppression, we consider two formalisms of easy features that can prevent learning of class features and provably attribute feature suppression to insufficient embedding space and/or imperfect data augmentations; thus, motivating practical solutions to this problem. The unified framework we develop to determine which features are learnt by CL allows us to also offer the only theoretical justification for recent empirical proposals to combine unsupervised and supervised contrastive losses. Perhaps, most surprisingly, our findings from this theoretical study indicate that simplicity bias of (S)GD is likely the driving factor behind class collapse and feature suppression.

Acknowledgment. This research was supported by the National Science Foundation CAREER Award 2146492.

References

Arora, S., Li, Y., Liang, Y., Ma, T., and Risteski, A. Linear algebraic structure of word senses, with applications to polysemy. *Transactions of the Association for Computational Linguistics*, 6:483–495, 2018.

Arora, S., Cohen, N., Hu, W., and Luo, Y. Implicit regularization in deep matrix factorization. *Advances in Neural Information Processing Systems*, 32, 2019a.

Arora, S., Khandeparkar, H., Khodak, M., Plevrakis, O., and Saunshi, N. A theoretical analysis of contrastive unsupervised representation learning. *arXiv preprint arXiv:1902.09229*, 2019b.

Chen, M., Fu, D. Y., Narayan, A., Zhang, M., Song, Z., Fatahalian, K., and Ré, C. Perfectly balanced: Improving transfer and robustness of supervised contrastive learning. In *International Conference on Machine Learning*, pp. 3090–3122. PMLR, 2022.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A Simple Framework for Contrastive Learning of Visual Representations. February 2020. doi: 10.48550/arXiv.2002.05709. URL https://arxiv.org/abs/2002.05709v3.

Chen, T., Luo, C., and Li, L. Intriguing properties of contrastive losses. *Advances in Neural Information Processing Systems*, 34:11834–11845, 2021.

Chuang, C.-Y., Robinson, J., Lin, Y.-C., Torralba, A., and Jegelka, S. Debiased Contrastive Learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 8765–8775. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/hash/63c3ddcc7b23daa1e42dc41f9a44a873-Abstract.html.

Fang, C., He, H., Long, Q., and Su, W. J. Exploring deep neural networks via layer-peeled model: Minority collapse in imbalanced training. *Proceedings of the National Academy of Sciences*, 118(43):e2103091118, 2021.

Foldiak, P. Sparse coding in the primate cortex. *The handbook of brain theory and neural networks*, 2003.

Graf, F., Hofer, C., Niethammer, M., and Kwitt, R. Dissecting supervised constrastive learning. In *International Conference on Machine Learning*, pp. 3821–3830. PMLR, 2021.

Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P. H., Buchatskaya, E., Doersch, C., Pires, B. A., Guo, Z. D., Azar, M. G., Piot, B., Kavukcuoglu, K., Munos, R., and Valko, M. Bootstrap your own latent: A new approach to self-supervised Learning, September 2020. URL http://arxiv.org/abs/2006.07733. arXiv:2006.07733 [cs, stat].

Gunasekar, S., Woodworth, B. E., Bhojanapalli, S., Neyshabur, B., and Srebro, N. Implicit regularization in matrix factorization. *Advances in Neural Information Processing Systems*, 30, 2017.

Gunasekar, S., Lee, J. D., Soudry, D., and Srebro, N. Implicit bias of gradient descent on linear convolutional networks. *Advances in Neural Information Processing Systems*, 31, 2018.

Han, X., Papyan, V., and Donoho, D. L. Neural collapse under mse loss: Proximity to and dynamics on the central path. *arXiv preprint arXiv:2106.02073*, 2021.

HaoChen, J. Z. and Ma, T. A theoretical study of inductive biases in contrastive learning. *arXiv preprint arXiv:2211.14699*, 2022.

HaoChen, J. Z., Wei, C., Gaidon, A., and Ma, T. Provable guarantees for self-supervised deep learning with spectral contrastive loss. *Advances in Neural Information Processing Systems*, 34:5000–5011, 2021.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Hinton, G. E., Osindero, S., and Teh, Y.-W. A fast learning algorithm for deep belief nets. *Neural computation*, 18 (7):1527–1554, 2006.

Hui, L., Belkin, M., and Nakkiran, P. Limitations of neural collapse for understanding generalization in deep learning. *arXiv preprint arXiv:2202.08384*, 2022.

Islam, A., Chen, C.-F. R., Panda, R., Karlinsky, L., Radke, R., and Feris, R. A broad study on the transferability of visual representations with contrastive learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8845–8855, 2021.

Ji, W., Deng, Z., Nakada, R., Zou, J., and Zhang, L. The power of contrast for feature learning: A theoretical analysis. *arXiv preprint arXiv:2110.02473*, 2021.

Ji, Z. and Telgarsky, M. The implicit bias of gradient descent on nonseparable data. In *Conference on Learning Theory*, pp. 1772–1798. PMLR, 2019.

Kalimeris, D., Kaplun, G., Nakkiran, P., Edelman, B., Yang, T., Barak, B., and Zhang, H. Sgd on neural networks learns functions of increasing complexity. *Advances in neural information processing systems*, 32, 2019.

Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., and Krishnan, D. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33:18661–18673, 2020.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. 2009.

Laurent, B. and Massart, P. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, pp. 1302–1338, 2000.

Lee, J. D., Lei, Q., Saunshi, N., and Zhuo, J. Predicting what you already know helps: Provable self-supervised learning. *Advances in Neural Information Processing Systems*, 34:309–323, 2021.

Li, T., Fan, L., Yuan, Y., He, H., Tian, Y., Feris, R., Indyk, P., and Katabi, D. Addressing feature suppression in unsupervised visual representations, 2020. URL https://arxiv.org/abs/2012.09962.

Liu, H., HaoChen, J. Z., Gaidon, A., and Ma, T. Self-supervised learning is more robust to dataset imbalance. *arXiv preprint arXiv:2110.05025*, 2021.

Lu, J. and Steinerberger, S. Neural collapse under cross-entropy loss. *Applied and Computational Harmonic Analysis*, 59:224–241, 2022.

Lyu, K., Li, Z., Wang, R., and Arora, S. Gradient descent on two-layer nets: Margin maximization and simplicity bias. *Advances in Neural Information Processing Systems*, 34:12978–12991, 2021.

Mairal, J., Bach, F., Ponce, J., et al. Sparse modeling for image and vision processing. *Foundations and Trends® in Computer Graphics and Vision*, 8(2-3):85–283, 2014.

Olshausen, B. A. and Field, D. J. Sparse coding with an overcomplete basis set: A strategy employed by v1? *Vision research*, 37(23):3311–3325, 1997.

Papyan, V., Romano, Y., and Elad, M. Convolutional neural networks analyzed via convolutional sparse coding. *The Journal of Machine Learning Research*, 18(1):2887–2938, 2017.

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

Protter, M. and Elad, M. Image sequence denoising via sparse and redundant representations. *IEEE transactions on Image Processing*, 18(1):27–35, 2008.

Ranzato, M., Poultney, C., Chopra, S., and Cun, Y. Efficient learning of sparse representations with an energy-based model. *Advances in neural information processing systems*, 19, 2006.

Razin, N. and Cohen, N. Implicit regularization in deep learning may not be explainable by norms. *Advances in neural information processing systems*, 33:21174–21187, 2020.

Robinson, J., Sun, L., Yu, K., Batmanghelich, K., Jegelka, S., and Sra, S. Can contrastive learning avoid shortcut solutions?, 2021. URL https://arxiv.org/abs/2106.11230.

Sagawa, S., Raghunathan, A., Koh, P. W., and Liang, P. An investigation of why overparameterization exacerbates spurious correlations. In *International Conference on Machine Learning*, pp. 8346–8356. PMLR, 2020.

Saunshi, N., Ash, J., Goel, S., Misra, D., Zhang, C., Arora, S., Kakade, S., and Krishnamurthy, A. Understanding contrastive learning requires incorporating inductive biases. *arXiv preprint arXiv:2202.14037*, 2022.

Shorack, G. R. and Shorack, G. *Probability for statisticians*, volume 951. Springer, 2000.

Soudry, D., Hoffer, E., Nacson, M. S., Gunasekar, S., and Srebro, N. The implicit bias of gradient descent on separable data. *The Journal of Machine Learning Research*, 19(1):2822–2878, 2018.

Tian, Y., Sun, C., Poole, B., Krishnan, D., Schmid, C., and Isola, P. What makes for good views for contrastive learning? *Advances in Neural Information Processing Systems*, 33:6827–6839, 2020.

Tosh, C., Krishnamurthy, A., and Hsu, D. Contrastive estimation reveals topic posterior information to linear models. *J. Mach. Learn. Res.*, 22:281–1, 2021a.

Tosh, C., Krishnamurthy, A., and Hsu, D. Contrastive learning, multi-view redundancy, and linear models. In *Algorithmic Learning Theory*, pp. 1179–1206. PMLR, 2021b.

Tsai, Y.-H. H., Wu, Y., Salakhutdinov, R., and Morency, L.-P. Self-supervised learning from a multi-view perspective. *arXiv preprint arXiv:2006.05576*, 2020.

Vinje, W. E. and Gallant, J. L. Sparse coding and decorrelation in primary visual cortex during natural vision. *Science*, 287(5456):1273–1276, 2000.

Wang, T. and Isola, P. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, pp. 9929–9939. PMLR, 2020.

Wen, Z. and Li, Y. Toward understanding the feature learning process of self-supervised contrastive learning. In *International Conference on Machine Learning*, pp. 11112–11122. PMLR, 2021.

Wu, Y., Poczos, B., and Singh, A. Towards understanding the generalization bias of two layer convolutional linear classifiers with gradient descent. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 1070–1078. PMLR, 2019.

Yang, J., Yu, K., Gong, Y., and Huang, T. Linear spatial pyramid matching using sparse coding for image classification. In *2009 IEEE Conference on computer vision and pattern recognition*, pp. 1794–1801. IEEE, 2009.

Zhou, J., Li, X., Ding, T., You, C., Qu, Q., and Zhu, Z. On the optimization landscape of neural collapse under mse loss: Global optimality with unconstrained features. In *International Conference on Machine Learning*, pp. 27179–27202. PMLR, 2022a.

Zhou, J., You, C., Li, X., Liu, K., Liu, S., Qu, Q., and Zhu, Z. Are all losses created equal: A neural collapse perspective. *arXiv preprint arXiv:2210.02192*, 2022b.

Zhu, Z., Ding, T., Zhou, J., Li, X., You, C., Sulam, J., and Qu, Q. A geometric analysis of neural collapse with unconstrained features. *Advances in Neural Information Processing Systems*, 34:29820–29834, 2021.

Zou, D., Cao, Y., Li, Y., and Gu, Q. Understanding the generalization of adam in learning neural networks with proper regularization. *arXiv preprint arXiv:2108.11371*, 2021.

A. Preliminaries

A.1. Effective dataset

Analyzing training a linear model with bias on the data is equivalent to analyzing trainng a linear model without bias on: $\{\begin{bmatrix}1\\ \boldsymbol{x}_i\end{bmatrix} : \boldsymbol{x}_i \in \mathcal{D}_{\text{aug}}\}$. Equivalently we can consider a dataset distribution where

$$\begin{aligned}\boldsymbol{x} =& \boldsymbol{u} + \boldsymbol{\xi},\\ \text{where } \boldsymbol{u} =& \boldsymbol{v}_0 + y\phi_1\boldsymbol{v}_1 + (y_{\text{sub}}\phi_2 + \mu)\boldsymbol{v}_2 + \rho\phi_k\boldsymbol{v}_k.\end{aligned}$$

The definitions are identical to the one in Section 3.1 except that each data now is in $\mathbb{R}^{d+1}$ and has one constant feature $\boldsymbol{v}_0$ orthogonal to other $\boldsymbol{v}$'s. We train a linear model $f(\boldsymbol{x}) = \boldsymbol{W}\boldsymbol{x}$ on such data. The definition of other notations such as $\hat{\mathcal{D}}_{\text{aug}}$ in the following analysis are also adapted to this dataset accordingly. Other notations such as $\hat{\mathcal{D}}_{\text{aug}}$ in the subsequent analysis are adjusted accordingly to accommodate this dataset.

A.2. Loss functions

The loss functions can be rewritten as follows

$$\begin{aligned}\mathcal{L}_{\text{SCL}} =& -2\hat{\mathbb{E}}_{i\in[n],\boldsymbol{x}\in\mathcal{A}_i,\boldsymbol{x}^+\in\mathcal{A}_i}\left[\boldsymbol{x}^\top\boldsymbol{W}^\top\boldsymbol{W}\boldsymbol{x}^+\right] + \hat{\mathbb{E}}_{\boldsymbol{x}\in\hat{\mathcal{D}}_{\text{aug}},\boldsymbol{x}^-\in\hat{\mathcal{D}}_{\text{aug}}}\left[\left(\boldsymbol{x}^\top\boldsymbol{W}^\top\boldsymbol{W}\boldsymbol{x}^-\right)^2\right] && (3)\\ =& -\operatorname{Tr}(2\boldsymbol{M}^+\boldsymbol{W}\boldsymbol{W}^\top) + \operatorname{Tr}(\boldsymbol{M}\boldsymbol{W}^\top\boldsymbol{W}\boldsymbol{M}\boldsymbol{W}^\top\boldsymbol{W})\\ \mathcal{L}_{\text{UCL}} =& -2\hat{\mathbb{E}}_{c\in\{-1,1\},\boldsymbol{x}\in\mathcal{S}_c,\boldsymbol{x}^+\in\mathcal{S}_c}\left[\boldsymbol{x}^\top\boldsymbol{W}^\top\boldsymbol{W}\boldsymbol{x}^+\right] + \hat{\mathbb{E}}_{\boldsymbol{x}\in\hat{\mathcal{D}}_{\text{aug}},\boldsymbol{x}^-\in\hat{\mathcal{D}}_{\text{aug}}}\left[\left(\boldsymbol{x}^\top\boldsymbol{W}^\top\boldsymbol{W}\boldsymbol{x}^-\right)^2\right] && (4)\\ =& -\operatorname{Tr}(2\tilde{\boldsymbol{M}}\boldsymbol{W}\boldsymbol{W}^\top) + \operatorname{Tr}(\boldsymbol{M}\boldsymbol{W}^\top\boldsymbol{W}\boldsymbol{M}\boldsymbol{W}^\top\boldsymbol{W})\\ \mathcal{L}_{\text{joint}} =& (1-\beta)\mathcal{L}_{\text{SCL}} + \beta\mathcal{L}_{\text{UCL}} && (5)\\ =& -\operatorname{Tr}(2\bar{\boldsymbol{M}}\boldsymbol{W}\boldsymbol{W}^\top) + \operatorname{Tr}(\boldsymbol{M}\boldsymbol{W}^\top\boldsymbol{W}\boldsymbol{M}\boldsymbol{W}^\top\boldsymbol{W}),\end{aligned}$$

where we define the following

Definition A.1. $\boldsymbol{M}, \boldsymbol{M}^+, \tilde{\boldsymbol{M}}$ are the covariance matrices of training examples, class centers and augmentation centers, respectively

$$\begin{aligned}\boldsymbol{M} =& \frac{1}{mn}\sum_{i=1}^{mn}\boldsymbol{x}_i\boldsymbol{x}_i^\top\\ \boldsymbol{M}^+ =& \frac{1}{2}\sum_{c\in\{-1,1\}}(\frac{2}{mn}\sum_{\boldsymbol{x}\in\mathcal{S}_c}\boldsymbol{x})(\frac{2}{mn}\sum_{\boldsymbol{x}\in\mathcal{S}_c}\boldsymbol{x})^\top\\ \tilde{\boldsymbol{M}} =& \frac{1}{n}\sum_{i=1}^{n}(\frac{1}{m}\sum_{\boldsymbol{x}\in\mathcal{A}_i}\boldsymbol{x})(\frac{1}{m}\sum_{\boldsymbol{x}\in\mathcal{A}_i}\boldsymbol{x})^\top,\end{aligned}$$

and

$$\bar{\boldsymbol{M}} = (1-\beta)\boldsymbol{M}^+ + \beta\tilde{\boldsymbol{M}}.$$

B. Minimizers of Loss Functions

We start with a technical lemma which we will need:

Lemma B.1. *The product of two positive semidefinite matrices is diagonalizable.*

Next, we present a lemma that facilitates the analysis of minimizers for various contrastive loss functions. To apply the lemma, simply substitute the respective covariance matrices ($\boldsymbol{M}$, $\boldsymbol{M}^+$, $\tilde{\boldsymbol{M}}$) into $\boldsymbol{P}$ and $\boldsymbol{Q}$ as indicated.

Lemma B.2. *Let $\boldsymbol{P}, \boldsymbol{Q} \in \mathbb{R}^{d+1}$ be positive semidefinite matrices such that **colsp**$(\boldsymbol{P}) \subset$ **colsp**$(\boldsymbol{Q})$. Consider the function $\mathcal{L} : \mathbb{R}^{p\times(d+1)} \to \mathbb{R}$ given by*

$$\mathcal{L}(\boldsymbol{W}) = \operatorname{Tr}[-2\boldsymbol{W}^\top\boldsymbol{W}\boldsymbol{P} + \boldsymbol{W}^\top\boldsymbol{W}\boldsymbol{Q}\boldsymbol{W}^\top\boldsymbol{W}\boldsymbol{Q}] \tag{6}$$

Then $\boldsymbol{W}$ is a global minimizer of $\mathcal{L}$ if and only if

$$\boldsymbol{W}^\top \boldsymbol{W}\boldsymbol{Q} = [\boldsymbol{Q}^\dagger \boldsymbol{P}]_p$$

*where notation $[\boldsymbol{A}]_p$ represents the matrix composed of the first p eigenvalues and eigenvectors of a positive semidefinite $\boldsymbol{A}$ (if $p \geq$ **rank** $\boldsymbol{A}$ then $[\boldsymbol{A}]_p = \boldsymbol{A}$).*

*Moreover, if $p \geq$ **rank**$(\boldsymbol{P})$, then $\boldsymbol{W}^{**}$ is a minimum norm global minimizer if and only if*

$$\boldsymbol{W}^{**\top}\boldsymbol{W}^{**} = \boldsymbol{Q}^\dagger \boldsymbol{P}\boldsymbol{Q}^\dagger$$

Proof. First consider points that satisfy the first order condition

$$\nabla_{\boldsymbol{W}}(\mathcal{L}) = -4\boldsymbol{W}\boldsymbol{P} + 4\boldsymbol{W}\boldsymbol{Q}\boldsymbol{W}^\top\boldsymbol{W}\boldsymbol{Q} = 0 \tag{7}$$

where $\nabla_{\boldsymbol{W}}(\mathcal{L})$ is the matrix of partial derivatives of $\mathcal{L}$ with respect to each entry of $\boldsymbol{W}$.

Since $\boldsymbol{Q}$ is positive semidefinite, it decomposes $\mathbb{R}^{d+1}$ into the orthogonal direct sum $\ker(\boldsymbol{Q}) \oplus$ **colsp**$(\boldsymbol{Q})$. Observe that both subspaces are invariant under both $\boldsymbol{P}$ and $\boldsymbol{Q}$.

Now let $\boldsymbol{v} \in$ **colsp**$(\boldsymbol{Q}) \cap \ker(\boldsymbol{W}\boldsymbol{Q})$. Note that $\boldsymbol{P}\boldsymbol{v} \in$ **colsp**$(\boldsymbol{Q})$, so $\boldsymbol{P}v = \boldsymbol{Q}\boldsymbol{Q}^\dagger\boldsymbol{P}v$. Then from equation 7,

$$\boldsymbol{0} = (\boldsymbol{W}\boldsymbol{P} - \boldsymbol{W}\boldsymbol{Q}\boldsymbol{W}^\top\boldsymbol{W}\boldsymbol{Q})\boldsymbol{v} = \boldsymbol{W}\boldsymbol{Q}(\boldsymbol{Q}^\dagger\boldsymbol{P} - \boldsymbol{W}^\top\boldsymbol{W}\boldsymbol{Q})\boldsymbol{v} \tag{8}$$

If in addition we assume $\boldsymbol{v} \in \ker(\boldsymbol{W}\boldsymbol{Q})$, then

$$\boldsymbol{0} = \boldsymbol{W}\boldsymbol{Q}(\boldsymbol{Q}^\dagger\boldsymbol{P}\boldsymbol{v})$$

namely $\boldsymbol{Q}^\dagger\boldsymbol{P}\boldsymbol{v} \in \ker(\boldsymbol{W}\boldsymbol{Q})$. But **colsp**$(\boldsymbol{Q})$ is also $\boldsymbol{Q}^\dagger$-invariant, so $\boldsymbol{Q}^\dagger\boldsymbol{P}\boldsymbol{v} \in$ **colsp**$(\boldsymbol{Q})$. We conclude that $\ker(\boldsymbol{W}\boldsymbol{Q}) \cap$ **colsp**$(\boldsymbol{Q})$ is $\boldsymbol{Q}^\dagger\boldsymbol{P}$-invariant. Since $\boldsymbol{Q}^\dagger$ and $\boldsymbol{P}$ are positive semidefinite, by B.1 $\boldsymbol{Q}^\dagger\boldsymbol{P}$ is diagonalizable. The only invariant subspaces of a diagonalizable operator are spans of its eigenvectors, so $\ker(\boldsymbol{W}\boldsymbol{Q}) \cap$ **colsp**$(\boldsymbol{Q})$ is the span of eigenvectors of $\boldsymbol{Q}^\dagger\boldsymbol{P}$.

Let **colsp**$(\boldsymbol{Q}) = \ker(\boldsymbol{W}\boldsymbol{Q}) \cap$ **colsp**$(\boldsymbol{Q}) \oplus U$, where U is the span of the remaining eigenvectors of $\boldsymbol{Q}^\dagger\boldsymbol{P}$ in **colsp**$(\boldsymbol{Q})$. Then by equation 8, $\boldsymbol{W}^\top\boldsymbol{W}\boldsymbol{Q} = \boldsymbol{Q}^\dagger\boldsymbol{P}$ on U.

Thus we have a basis $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_r, \ldots, \boldsymbol{v}_{\boldsymbol{s}}, \ldots, \boldsymbol{v}_{\boldsymbol{d}}$ s.t. $Span(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_r) = U, Span(\boldsymbol{v}_{r+1}, \ldots, \boldsymbol{v}_s) = \ker(\boldsymbol{W}\boldsymbol{Q}) \cap$

colsp$(\boldsymbol{Q}), Span(v_{s+1}, \ldots, v_{d+1}) = \ker \boldsymbol{Q}$, and in this basis

$$\boldsymbol{Q}^\dagger \boldsymbol{P} = \begin{pmatrix} \lambda_1 & & & & & & & \\ & \ddots & & & & & & \\ & & \lambda_r & & & & & \\ & & & \lambda_{r+1} & & & & \\ & & & & \ddots & & & \\ & & & & & \lambda_s & & \\ & & & & & & 0 & \\ & & & & & & & \ddots \\ & & & & & & & & 0 \end{pmatrix}$$

$$\boldsymbol{W}^\top \boldsymbol{W} \boldsymbol{Q} = \begin{pmatrix} \lambda_1 & & & & & & & \\ & \ddots & & & & & & \\ & & \lambda_r & & & & & \\ & & & 0 & & & & \\ & & & & \ddots & & & \\ & & & & & 0 & & \\ & & & & & & 0 & \\ & & & & & & & \ddots \\ & & & & & & & & 0 \end{pmatrix}$$

$$\boldsymbol{W}^\top \boldsymbol{W} \boldsymbol{P} = \begin{pmatrix} \lambda_1^2 & & & & & & & \\ & \ddots & & & & & & \\ & & \lambda_r^2 & & & & & \\ & & & 0 & & & & \\ & & & & \ddots & & & \\ & & & & & 0 & & \\ & & & & & & 0 & \\ & & & & & & & \ddots \\ & & & & & & & & 0 \end{pmatrix}$$

with $\lambda_1, \ldots, \lambda_r, \ldots, \lambda_q \neq 0$ for some $r \leq q \leq s$, where $r = \textbf{rank}\boldsymbol{W} \leq p, q = \textbf{rank}(\boldsymbol{P})$.

Then for all such $\boldsymbol{W}$,

$$\begin{aligned} \mathcal{L} &= \mathrm{Tr}[-2\boldsymbol{W}^\top \boldsymbol{W} \boldsymbol{P} + \boldsymbol{W}^\top \boldsymbol{W} \boldsymbol{Q} \boldsymbol{W}^\top \boldsymbol{W} \boldsymbol{Q}] \\ &= -2 \sum_{i=1}^{r} \lambda_i^2 + \sum_{i=1}^{r} \lambda_i^2 \\ &= -\sum_{i=1}^{r} \lambda_i^2 \end{aligned}$$

It is clear from the above expression that the minimum among critical points is achieved if and only if

$$\boldsymbol{W}^\top \boldsymbol{W} \boldsymbol{Q} = [\boldsymbol{Q}^\dagger \boldsymbol{P}]_p$$

(note that if matching anything beyond the qth eigenvalue is trivial since all such eigenvalues are zero).

It remains to check the behavior as $\|\boldsymbol{W}\|_F$ grows large. Equivalently, $\boldsymbol{W}^\top \boldsymbol{W}$ has a large eigenvalue λ. Let $\boldsymbol{w}$ be a corresponding eigenvector. If $\boldsymbol{w} \in \ker \boldsymbol{Q}$, then $\boldsymbol{Q}\boldsymbol{w} = \boldsymbol{P}\boldsymbol{w} = 0$, so we see that the loss is unchanged. Otherwise, $\boldsymbol{w}$ has some nonzero alignment with **colsp**$(\boldsymbol{W})$. But then $\mathrm{Tr}[\boldsymbol{W}^\top \boldsymbol{W} \boldsymbol{Q} \boldsymbol{W}^\top \boldsymbol{W} \boldsymbol{Q}]$ grows quadratically in λ, but $\mathrm{Tr}[-2\boldsymbol{W}^\top \boldsymbol{W} \boldsymbol{P}]$

grows at most linearly in λ, hence the loss is large. We conclude that the previously found condition in fact specifies the global minimizers of $\mathcal{L}$.

From now on, assume that $p \geq q$. Then the global minimum is achieved if and only if

$$\boldsymbol{W}^\top \boldsymbol{W} \boldsymbol{Q} = \boldsymbol{Q}^\dagger \boldsymbol{P} \tag{9}$$

Let us now consider the minimum norm solution, i.e. the one that minimizes $\mathrm{Tr}(\boldsymbol{W}^\top \boldsymbol{W})$. Note that $\boldsymbol{W}^\top \boldsymbol{W}$ and $\boldsymbol{Q}^\dagger \boldsymbol{P} \boldsymbol{Q}^\dagger$ are positive semidefinite. Let $\mathcal{B}$ be an orthonormal basis of eigenvectors for $\textbf{colsp}(\boldsymbol{Q}), \mathcal{C}$ an orthonormal basis for $\ker \boldsymbol{Q}$. Then in the orthonormal basis $\mathcal{B} \cup \mathcal{C}$, we have the following block form of $\boldsymbol{Q}^\dagger \boldsymbol{P} \boldsymbol{Q}^\dagger$

$$\boldsymbol{Q}^\dagger \boldsymbol{P} \boldsymbol{Q}^\dagger = \begin{pmatrix} \boldsymbol{A} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{pmatrix} \tag{10}$$

where $\boldsymbol{A}$ is positive semidefinite.

Now equation 9 implies that $\boldsymbol{W}\boldsymbol{W}^\top$ has the form

$$\boldsymbol{W}^\top \boldsymbol{W} = \begin{pmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{B}^\top & \boldsymbol{C} \end{pmatrix} \tag{11}$$

where $\boldsymbol{C}$ is also positive semidefinite matrix. Then $\|\boldsymbol{W}\|_F = \mathrm{Tr}[\boldsymbol{W}^\top \boldsymbol{W}]$ is minimized exactly when $\mathrm{Tr}[\boldsymbol{C}] = 0$. But this holds if and only if $\boldsymbol{C} = \boldsymbol{0}$. Now suppose for the sake of contradiction $\boldsymbol{B} \neq \boldsymbol{0}$, say $b_{ij} \neq 0$ for some i, j. Then $\boldsymbol{W}^\top \boldsymbol{W}$ contains a submatrix

$$\begin{pmatrix} a_{ii} & b_{ij} \\ b_{ij} & 0 \end{pmatrix} \tag{12}$$

which has negative determinant. But this implies that $\boldsymbol{W}^\top \boldsymbol{W}$ is not positive semidefinite, a contradiction. We conclude that $\boldsymbol{B} = \boldsymbol{0}$ so that the minimum norm solution is precisely

$$\boldsymbol{W}^{**\top} \boldsymbol{W}^{**} = \boldsymbol{Q}^\dagger \boldsymbol{P} \boldsymbol{Q}^\dagger.$$

This completes the proof.

□

C. Some Properties of The Covariance Matrices

We assume $\frac{\sigma_\xi^2}{mn} = o(1)$.

With probability $\geq 1 - O(\frac{m^2 n^2}{d})$, we have that $\boldsymbol{\xi}_i^\top \boldsymbol{v}_k = 0,\ \forall k, i$ and $\boldsymbol{\xi}_i^\top \boldsymbol{\xi}_j = 0,\ \forall i, j$. The following discussion focuses on the properties of $\boldsymbol{M}$, $\boldsymbol{M}^+$, and $\tilde{\boldsymbol{M}}$ when this condition is met.

Write $\boldsymbol{X} = \boldsymbol{V} \begin{bmatrix} \boldsymbol{S} \\ \sigma_\xi \mathbf{I}_{mn} \end{bmatrix}$ where $\boldsymbol{V} = [\boldsymbol{v}_0, \boldsymbol{v}_1\ \dots\ \boldsymbol{v}_K \dots \boldsymbol{v}_{K+1} \dots \boldsymbol{v}_{mn+K}]$ where $\boldsymbol{v}_{K+i}$ is the noise vector selected by example $\boldsymbol{x}_i$, and

$$\begin{aligned} \boldsymbol{S} = & \begin{bmatrix} 1 & 1 & \dots & 1 \\ y_1 \phi_1 & y_2 \phi_1 & \dots & y_{mn} \phi_1 \\ \mu + y_{\text{sub},1} \phi_2 & \mu + y_{\text{sub},2} \phi_2 & \dots & y_{\text{sub},mn} \phi_2 \\ \rho_1 \mathbb{1}_{k_1=3} \phi_3 & \rho_2 \mathbb{1}_{k_2=3} \phi_3 & \dots & \rho_{mn} \mathbb{1}_{k_{mn}=3} \phi_3 \\ \rho_1 \mathbb{1}_{k_1=4} \phi_4 & \rho_2 \mathbb{1}_{k_2=4} \phi_4 & \dots & \rho_{mn} \mathbb{1}_{k_{mn}=4} \phi_4 \\ \vdots & \vdots & \ddots & \vdots \\ \rho_1 \mathbb{1}_{k_1=K} \phi_K & \rho_2 \mathbb{1}_{k_2=K} \phi_K & \dots & \rho_{mn} \mathbb{1}_{k_{mn}=K} \phi_K \end{bmatrix} \\ = & \boldsymbol{S}' \bar{\boldsymbol{Y}}, \end{aligned} \tag{13}$$

where

$$\boldsymbol{S}' := \begin{bmatrix} \sqrt{mn} & 0 & 0 & 0 & \dots & 0 \\ 0 & \sqrt{mn}\phi_1 & 0 & 0 & \dots & 0 \\ \sqrt{mn}\mu & 0 & \sqrt{mn}\phi_2 & 0 & \dots & 0 \\ 0 & 0 & 0 & \sqrt{\frac{mn}{K-2}}\phi_3 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \dots & \sqrt{\frac{mn}{K-2}}\phi_K \end{bmatrix},$$

and

$$\bar{\boldsymbol{Y}} := \begin{bmatrix} \frac{1}{\sqrt{mn}} & \frac{1}{\sqrt{mn}} & \dots & \frac{1}{\sqrt{mn}} \\ y_1 \frac{1}{\sqrt{mn}} & y_2 \frac{1}{\sqrt{mn}} & \dots & y_{mn} \frac{1}{\sqrt{mn}} \\ y_{\text{sub},1} \frac{1}{\sqrt{mn}} & y_{\text{sub},2} \frac{1}{\sqrt{mn}} & \dots & y_{\text{sub},mn} \frac{1}{\sqrt{mn}} \\ \rho_1 \mathbb{1}_{k_1=3} \sqrt{\frac{K-2}{mn}} & \rho_2 \mathbb{1}_{k_2=3} \sqrt{\frac{K-2}{mn}} & \dots & \rho_{mn} \mathbb{1}_{k_{mn}=3} \sqrt{\frac{K-2}{mn}} \\ \rho_1 \mathbb{1}_{k_1=4} \sqrt{\frac{K-2}{mn}} & \rho_2 \mathbb{1}_{k_2=4} \sqrt{\frac{K-2}{mn}} & \dots & \rho_{mn} \mathbb{1}_{k_{mn}=4} \sqrt{\frac{K}{mn}} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_1 \mathbb{1}_{k_1=K} \sqrt{\frac{K-2}{mn}} & \rho_2 \mathbb{1}_{k_2=K} \sqrt{\frac{K-2}{mn}} & \dots & \rho_{mn} \mathbb{1}_{k_{mn}=K} \sqrt{\frac{K-2}{mn}} \end{bmatrix}$$

It should be noted that the rows of $\bar{\boldsymbol{Y}}$ are orthonormal due to the assumption of a balanced dataset. Consequently, to obtain the singular value decomposition (SVD) of $\boldsymbol{S}$, it suffices to find the SVD of $\boldsymbol{S}' = \boldsymbol{P}'\boldsymbol{\Lambda}'\boldsymbol{Q}'^\top$. Moreover, the right singular vectors of $\boldsymbol{S}$ with non-zero singular values are given by the rows of $\boldsymbol{Q}'^\top \bar{\boldsymbol{Y}}$.

We write $\boldsymbol{M}$ as $\boldsymbol{V}\boldsymbol{G}\boldsymbol{V}^\top$ where $\boldsymbol{G}$ is given by

$$\begin{bmatrix} \frac{1}{mn}\boldsymbol{S}\boldsymbol{S}^\top & \frac{\sigma_\xi}{mn}\boldsymbol{S} \\ \frac{\sigma_\xi}{mn}\boldsymbol{S}^\top & \frac{\sigma_\xi^2}{mn}\mathbf{I}_{mn} \end{bmatrix}.$$

Now we are ready to show the following lemma which describes the SVD of $\boldsymbol{G}$.

Lemma C.1. *Let $\boldsymbol{S} \in \mathbb{R}^{K\times nm}$ be a rank-K matrix with SVD $\boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{Q}^\top$, where $\boldsymbol{P} \in \mathbb{R}^{K\times K}, \boldsymbol{\Lambda} \in \mathbb{R}^{K\times mn}$ and $\boldsymbol{Q} \in \mathbb{R}^{mn\times mn}$. The mn none-zero eigenvalues of the following matrix $\boldsymbol{G}$*

$$\begin{bmatrix} \frac{1}{mn}\boldsymbol{S}\boldsymbol{S}^\top & \frac{\sigma_\xi}{mn}\boldsymbol{S} \\ \frac{\sigma_\xi}{mn}\boldsymbol{S}^\top & \frac{\sigma_\xi^2}{mn}\mathbf{I}_{mn} \end{bmatrix}$$

are given by $\frac{\sigma_\xi^2}{mn} + \frac{\lambda_1^2}{mn}, \frac{\sigma_\xi^2}{mn} + \frac{\lambda_2^2}{mn}, \dots, \frac{\sigma_\xi^2}{mn} + \frac{\lambda_K^2}{mn}, \frac{\sigma_\xi^2}{mn}, \dots, \frac{\sigma_\xi^2}{mn}$, with the corresponding eigenvectors $\begin{bmatrix} \frac{1}{\sqrt{1+r_1^2}}\boldsymbol{p}_1 \\ \frac{r_1}{\sqrt{1+r_1^2}}\boldsymbol{q}_1 \end{bmatrix}, \begin{bmatrix} \frac{1}{\sqrt{1+r_2^2}}\boldsymbol{p}_2 \\ \frac{r_2}{\sqrt{1+r_2^2}}\boldsymbol{q}_2 \end{bmatrix}, \dots, \begin{bmatrix} \frac{1}{\sqrt{1+r_K^2}}\boldsymbol{p}_K \\ \frac{r_K}{\sqrt{1+r_K^2}}\boldsymbol{q}_K \end{bmatrix}, \begin{bmatrix} \mathbf{0}_K \\ \boldsymbol{q}_{K+1} \end{bmatrix}, \dots, \begin{bmatrix} \mathbf{0}_K \\ \boldsymbol{q}_{mn} \end{bmatrix}$, where $r_k = \frac{\sigma_\xi}{\lambda_k}$.

Proof. Let $\begin{bmatrix} \boldsymbol{P}\boldsymbol{a} \\ \boldsymbol{Q}\boldsymbol{b} \end{bmatrix}$ where $\boldsymbol{a} \in \mathbb{R}^K$ and $\boldsymbol{b} \in \mathbb{R}^{mn}$ be an eigenvector of $\boldsymbol{G}$. By the definition of eigenvector there should exist α such that $\boldsymbol{G}\begin{bmatrix} \boldsymbol{P}\boldsymbol{a} \\ \boldsymbol{Q}\boldsymbol{b} \end{bmatrix} = \alpha \begin{bmatrix} \boldsymbol{P}\boldsymbol{a} \\ \boldsymbol{Q}\boldsymbol{b} \end{bmatrix}$, i.e.,

$$\begin{cases} \frac{1}{mn}\boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top\boldsymbol{a} + \frac{\sigma_\xi}{mn}\boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{b} = \alpha\boldsymbol{P}\boldsymbol{a} \\ \frac{\sigma_\xi}{mn}\boldsymbol{Q}\boldsymbol{\Lambda}^\top\boldsymbol{a} + \frac{\sigma_\xi^2}{mn}\boldsymbol{Q}\boldsymbol{b} = \alpha\boldsymbol{Q}\boldsymbol{b}, \end{cases}$$

which reduces to

$$\begin{cases} (\alpha\mathbf{I}_K - \frac{1}{mn}\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top)\boldsymbol{a} = \frac{\sigma_\xi}{mn}\boldsymbol{\Lambda}\boldsymbol{b} \\ \frac{\sigma_\xi}{mn}\boldsymbol{\Lambda}^\top\boldsymbol{a} = (\alpha - \frac{\sigma_\xi^2}{mn})\boldsymbol{b}. \end{cases}$$

Firstly, we observe that the rank of $\mathbf{G}$ is at most mn because $\boldsymbol{G} = \frac{1}{mn}\begin{bmatrix}\boldsymbol{S}\\ \sigma_\xi \mathbf{I}_{mn}\end{bmatrix}\begin{bmatrix}\boldsymbol{S}\\ \sigma_\xi \mathbf{I}_{mn}\end{bmatrix}^\top$. Then it is easy to check that the eigenvalues and eigenvectors in Lemma C.1 satisfy the above conditions and the eigenvectors are indeed orthonormal, which completes the proof. □

Corollary C.2. *The projection of $\boldsymbol{v}_2$ onto* $\ker \boldsymbol{M}$ *has magnitude* $\Theta(\frac{\sigma_\xi}{\sqrt{mn}})$.

Corollary C.3. . *Assuming the dataset is balanced, then*

$$\sqrt{\boldsymbol{v}_2^\top \boldsymbol{M}^\dagger \boldsymbol{M}^+ \boldsymbol{M}^\dagger \boldsymbol{v}_2} = \begin{cases} 0, if \;\; \mu = 0 \\ O(\frac{\sigma_\xi}{\sqrt{mn}}), if \;\; \mu \neq 0 \;\; and \;\; \mu = \Theta(1). \end{cases}$$

Proof. Let $\boldsymbol{LAL}^\top$ be the eigendecomposition of $\boldsymbol{G}$. Then

$$\boldsymbol{M}^\dagger \boldsymbol{v}_2 = \boldsymbol{VLA}^\dagger \boldsymbol{L}^\top \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

When $\mu = 0$, we can express the SVD of $\boldsymbol{S}$ (equation 13) and apply Lemma C.1 to obtain the following result.

$$\lambda_3 = \sqrt{mn}\phi_2,\; a_3 = \frac{\sigma_\xi^2}{mn} + \phi_2^2,\;\; r_3 = \frac{\sigma_\xi}{\sqrt{mn}\phi_2}$$

$$\boldsymbol{p}_k = \boldsymbol{e}_k,\; \forall k \in [K] \;\text{ and }\; \boldsymbol{q}_3 = \begin{bmatrix} \frac{1}{\sqrt{mn}} y_{\text{sub},1} \\ \frac{1}{\sqrt{mn}} y_{\text{sub},2} \\ \vdots \\ \frac{1}{\sqrt{mn}} y_{\text{sub},mn} \end{bmatrix}, \;\text{ and }\; \boldsymbol{l}_3 = \begin{bmatrix} \frac{1}{\sqrt{1+r_3^2}} \boldsymbol{p}_3 \\ \frac{r_3}{\sqrt{1+r_3^2}} \boldsymbol{q}_3 \end{bmatrix}.$$

Thus

$$\begin{aligned} \boldsymbol{M}^\dagger \boldsymbol{v}_2 =& \frac{1}{a_3\sqrt{1+r_3^2}} \boldsymbol{V}\boldsymbol{l}_3 \\ =& \frac{1}{a_3\sqrt{1+r_3^2}} \Big(\frac{1}{\sqrt{1+r_3^2}} \boldsymbol{v}_2 + \frac{r_3}{\sqrt{1+r_3^2}} \sum_{i=1}^{mn} \frac{1}{\sqrt{mn}} y_{\text{sub},i} \boldsymbol{v}_{K+i} \Big). \end{aligned}$$

Let $\bar{\boldsymbol{x}}_y$ be the average of examples with label y and let $\mathcal{S}_y$ collects indices of examples with label y. Then

$$\bar{\boldsymbol{x}}_y = \boldsymbol{v}_0 + \boldsymbol{v}_1 + \frac{2\sigma_\xi}{mn} \sum_{i \in \mathcal{S}_y} \boldsymbol{v}_{K+i}, \tag{14}$$

and

$$\bar{\boldsymbol{x}}_y^\top \boldsymbol{M}^\dagger \boldsymbol{v}_2 = \frac{r_3}{a_3(1+r_3^2)} \frac{2\sigma_\xi}{mn} \sum_{i \in \mathcal{S}_y} \frac{1}{\sqrt{mn}} y_{\text{sub},i} = 0.$$

Write $\boldsymbol{M}^+$ as

$$\boldsymbol{M}^+ = \frac{1}{2}(\bar{\boldsymbol{x}}_{+1}\bar{\boldsymbol{x}}_{+1}^\top + \bar{\boldsymbol{x}}_{-1}\bar{\boldsymbol{x}}_{-1}^\top).$$

Then

$$\boldsymbol{v}_2^\top \boldsymbol{M}^\dagger \boldsymbol{M}^+ \boldsymbol{M}^\dagger \boldsymbol{v}_2 = \frac{1}{2}((\boldsymbol{v}_2^\top \boldsymbol{M}^\dagger \bar{\boldsymbol{x}}_{+1})^2 + (\boldsymbol{v}_2^\top \boldsymbol{M}^\dagger \bar{\boldsymbol{x}}_{-1})^2) = 0.$$

When $\mu \neq 0$, then there are at most two of $\boldsymbol{p}_k$'s that are not orthogonal to $\boldsymbol{e}_3$ (say $\boldsymbol{p}_1$ and $\boldsymbol{p}_3$). Additionally, all of their elements, except for the first one, are zero. The remaining corresponding quantities satisfy.

$$\begin{aligned}
&\lambda_1, \lambda_3 = \Theta(\sqrt{mn}),\\
&a_1 = \frac{\lambda_1^2}{mn} + \frac{\sigma_\xi^2}{mn}, \;\; a_3 = \frac{\lambda_3^2}{mn} + \frac{\sigma_\xi^2}{mn}\\
&r_1 = \frac{\sigma_\xi}{\lambda_1}, r_3 = \frac{\sigma_\xi}{\lambda_3},
\end{aligned}$$

and $\boldsymbol{q}_1$ and $\boldsymbol{q}_3$ are just linear combinations of $\bar{\boldsymbol{y}}_{\text{sub}}$ and $\frac{1}{\sqrt{mn}}\mathbf{1}$, where $\bar{\boldsymbol{y}}_{\text{sub}}$ is a vector whose i-th element is $\frac{1}{\sqrt{mn}} y_{\text{sub},i}$. Then

$$\boldsymbol{M}^\dagger \boldsymbol{v}_2 = \boldsymbol{V}\left[\frac{1}{a_1}\frac{1}{\sqrt{1+r_1^2}} c_{3,1}\boldsymbol{l}_1 + \frac{1}{a_3}\frac{1}{\sqrt{1+r_3^2}} c_{3,3}\boldsymbol{l}_3\right]$$

where $c_{i,j} = \boldsymbol{p}_j^\top \boldsymbol{e}_i$ are constants. For $i = 0, 2$

$$\begin{aligned}
\boldsymbol{v}_0^\top \boldsymbol{M}^\dagger \boldsymbol{v}_2 =& \boldsymbol{e}_1^\top \left[\frac{1}{a_1}\frac{1}{\sqrt{1+r_1^2}} c_{3,1}\boldsymbol{l}_1 + \frac{1}{a_3}\frac{1}{\sqrt{1+r_3^2}} c_{3,3}\boldsymbol{l}_3\right]\\
=& \frac{1}{a_1}\frac{1}{1+r_1^2} c_{3,1}c_{1,1} + \frac{1}{a_3}\frac{1}{1+r_3^2} c_{3,3}c_{1,3}\\
=& (\frac{mn}{\lambda_1^2} - \Theta(\frac{\sigma_\xi^2}{mn}))(1 - \Theta(\frac{\sigma_\xi}{\lambda_1}))c_{3,1}c_{1,1} + (\frac{mn}{\lambda_3^2} - \Theta(\frac{\sigma_\xi^2}{mn}))(1 - \Theta(\frac{\sigma_\xi}{\lambda_3}))c_{3,3}c_{1,3}\\
=&
\end{aligned}$$

where $|\epsilon_1| = O(\frac{\sigma_\xi}{\sqrt{mn}})$. Similarly,

$$\boldsymbol{v}_2^\top \boldsymbol{M}^\dagger \boldsymbol{v}_2 = \frac{mn}{\lambda_1^2} c_{3,1}c_{3,1} + \frac{mn}{\lambda_3^2} c_{3,3}c_{3,3} + \epsilon_2,$$

where $|\epsilon_2| = O(\frac{\sigma_\xi}{\sqrt{mn}})$. For $i > K$

$$\begin{aligned}
\boldsymbol{v}_i^\top \boldsymbol{M}^\dagger \boldsymbol{v}_2 =& \boldsymbol{v}_i \boldsymbol{V}\left[\frac{1}{a_1}\frac{1}{1+r_1^2} c_{3,1}\boldsymbol{l}_1 + \frac{1}{a_3}\frac{1}{1+r_3^2} c_{3,3}\boldsymbol{l}_3\right]\\
=& \boldsymbol{e}_i^\top \left[\frac{1}{a_1}\frac{1}{1+r_1^2} c_{3,1}\boldsymbol{l}_1 + \frac{1}{a_3}\frac{1}{1+r_3^2} c_{3,3}\boldsymbol{l}_3\right]\\
=& \epsilon_3,
\end{aligned}$$

where $|\epsilon_3| = O(\frac{\sigma_\xi}{mn})$. Additionally,

$$\bar{\boldsymbol{x}}_y = \boldsymbol{v}_0 + \boldsymbol{v}_1 + \mu \boldsymbol{v}_2 + \frac{2\sigma_\xi}{mn}\sum_{i \in \mathcal{S}_y} \boldsymbol{v}_{K+i}.$$

Then

$$\bar{\boldsymbol{x}}_y^\top \boldsymbol{M}^\dagger \boldsymbol{v}_2 = \frac{mn}{\lambda_1^2} c_{3,1}c_{1,1} + \frac{mn}{\lambda_3^2} c_{3,3}c_{1,3} + \frac{mn}{\lambda_1^2} c_{3,1}c_{3,1} + \frac{mn}{\lambda_3^2} c_{3,3}c_{3,3} + O(\frac{\sigma_\xi}{\sqrt{mn}}).$$

By straightforward calculation, we can verify that $\frac{mn}{\lambda_1^2} c_{3,1}c_{1,1} + \frac{mn}{\lambda_3^2} c_{3,3}c_{1,3} + \frac{mn}{\lambda_1^2} c_{3,1}c_{3,1} + \frac{mn}{\lambda_3^2} c_{3,3}c_{3,3} = 0$. This equation can be equivalently examined as the satisfaction of the following condition:

$$[1 \;\; \mu] \begin{bmatrix} 1 & \mu \\ \mu & \mu^2 + \phi_2^2 \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ 1 \end{bmatrix} = 0.$$

Therefore $|\bar{\boldsymbol{x}}_y^\top \boldsymbol{M}^\dagger \boldsymbol{v}_2| = O(\frac{\sigma_\xi}{\sqrt{mn}})$, and consequently $\sqrt{\boldsymbol{v}_2^\top \boldsymbol{M}^\dagger \boldsymbol{M}^+ \boldsymbol{M}^\dagger \boldsymbol{v}_2} = O(\frac{\sigma_\xi}{\sqrt{mn}})$. □

Corollary C.4. *Similar to Corollary C.3, we also have* $\sqrt{\boldsymbol{v}_k^\top \boldsymbol{M}^\dagger \boldsymbol{M}^+ \boldsymbol{M}^\dagger \boldsymbol{v}_k} = 0$, $k = 3, 4, \ldots, K$.

Corollary C.5. $\sqrt{\boldsymbol{v}_1^\top \boldsymbol{M}^\dagger \boldsymbol{M}^+ \boldsymbol{M}^\dagger \boldsymbol{v}_1} = \Theta(1)$. *It can be proved using the same strategy as in Corollary C.3.*

Lemma C.6. *(1) The first K eigenvectors/eigenvalues of $\tilde{\boldsymbol{M}}$ match those of $\boldsymbol{M}$. (2) $\boldsymbol{M}^\dagger \tilde{\boldsymbol{M}}$ is identity on* ***colsp***$(\tilde{\boldsymbol{M}})$ *and null on* $\ker(\tilde{\boldsymbol{M}})$*, i.e., $\boldsymbol{M}^\dagger \tilde{\boldsymbol{M}} = \tilde{\boldsymbol{M}}^\dagger \tilde{\boldsymbol{M}}$.*

Proof. We assign indices to the training examples such that the augmented examples from the same original example are indexed from $(l-1) \times m + 1$ to $l \times m$, where l ranges from 1 to n. Next, we define matrix $\tilde{\boldsymbol{V}} = [\tilde{\boldsymbol{v}}_1, \tilde{\boldsymbol{v}}_2, \ldots, \tilde{\boldsymbol{v}}_n] \in \mathbb{R}^{d \times n}$, where

$$\begin{aligned}
\tilde{\boldsymbol{v}}_i =& \boldsymbol{v}_i, \quad \forall 1 \le i \le K, \\
\tilde{\boldsymbol{v}}_i =& \frac{1}{\sqrt{m}} \sum_{j=1}^{m} \boldsymbol{v}_{K+(i-1)\times m+j}, \quad \forall K+1 \le i \le n.
\end{aligned}$$

In other words, $\tilde{\boldsymbol{V}}$ can be written as

$$\tilde{\boldsymbol{V}} = \boldsymbol{V}\boldsymbol{T},$$

where

$$\boldsymbol{T} = \begin{bmatrix} \boldsymbol{I}_K & \boldsymbol{0}_{K \times n} \\ \boldsymbol{0}_{mn \times K} & \begin{bmatrix} \frac{1}{\sqrt{m}}\boldsymbol{1}_{m\times 1} & 0 & 0 & \ldots & 0 \\ 0 & \frac{1}{\sqrt{m}}\boldsymbol{1}_{m\times 1} & 0 & \ldots & 0 \\ 0 & 0 & \frac{1}{\sqrt{m}}\boldsymbol{1}_{m\times 1} & \ldots & 0 \end{bmatrix} \end{bmatrix}$$

Note that, by the definition of our augmentation, the center of augmentations of the i-th original example, i.e., $\tilde{\boldsymbol{x}}_i = \frac{1}{m}\sum_{j=1}^{m} \boldsymbol{x}_{K+(i-1)\times m+j}$, can be considered as an example with the same features as $\boldsymbol{x}_i$ but with an added noise term of $\frac{\sigma_\xi}{\sqrt{m}}\tilde{\boldsymbol{v}}_i$. Therefore we can change the basis to $\tilde{\boldsymbol{V}}$ and express $\tilde{\boldsymbol{M}}$ as

$$\tilde{\boldsymbol{M}} = \tilde{\boldsymbol{V}}\tilde{\boldsymbol{G}}\tilde{\boldsymbol{V}}^\top,$$

where

$$\tilde{\boldsymbol{G}} = \begin{bmatrix} \frac{1}{n}\tilde{\boldsymbol{S}}\tilde{\boldsymbol{S}}^\top & \frac{1}{n}\frac{\sigma_\xi}{\sqrt{m}}\tilde{\boldsymbol{S}} \\ \frac{1}{n}\frac{\sigma_\xi}{\sqrt{m}}\tilde{\boldsymbol{S}}^\top & \frac{1}{n}\frac{\sigma_\xi^2}{m}\boldsymbol{I}_n \end{bmatrix}$$

and

$$\tilde{\boldsymbol{S}} = \tilde{\boldsymbol{S}}'\bar{\boldsymbol{Y}}_{\text{orig}}, \tag{15}$$

where

$$\tilde{\boldsymbol{S}}' := \begin{bmatrix} \sqrt{n} & 0 & 0 & 0 & \ldots & 0 \\ 0 & \sqrt{n}\phi_1 & 0 & 0 & \ldots & 0 \\ \sqrt{n}\mu & 0 & \sqrt{n}\phi_2 & 0 & \ldots & 0 \\ 0 & 0 & 0 & \sqrt{\frac{n}{K-2}}\phi_3 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \ldots & \sqrt{\frac{n}{K-2}}\phi_K \end{bmatrix},$$

and

$$\bar{\boldsymbol{Y}}_{\text{orig}} := \begin{bmatrix} \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \cdots & \frac{1}{\sqrt{n}} \\ y_1 \frac{1}{\sqrt{n}} & y_2 \frac{1}{\sqrt{n}} & \cdots & y_n \frac{1}{\sqrt{n}} \\ y_{\text{sub},1} \frac{1}{\sqrt{n}} & y_{\text{sub},2} \frac{1}{\sqrt{n}} & \cdots & y_{\text{sub},n} \frac{1}{\sqrt{n}} \\ \rho_1 \mathbb{1}_{k_1=3} \sqrt{\frac{K-2}{n}} & \rho_2 \mathbb{1}_{k_2=3} \sqrt{\frac{K-2}{n}} & \cdots & \rho_n \mathbb{1}_{k_n=3} \sqrt{\frac{K-2}{n}} \\ \rho_1 \mathbb{1}_{k_1=4} \sqrt{\frac{K-2}{n}} & \rho_2 \mathbb{1}_{k_2=4} \sqrt{\frac{K-2}{n}} & \cdots & \rho_n \mathbb{1}_{k_n=4} \sqrt{\frac{K}{n}} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_1 \mathbb{1}_{k_1=K} \sqrt{\frac{K-2}{n}} & \rho_2 \mathbb{1}_{k_2=K} \sqrt{\frac{K-2}{n}} & \cdots & \rho_n \mathbb{1}_{k_n=K} \sqrt{\frac{K-2}{n}} \end{bmatrix}_{\text{orig}}. \tag{16}$$

We note that we use the subscript 'orig' of a matrix to indicate that its elements represent the corresponding quantities on the original dataset (e.g., y_i is the label of the i-th original example). Let $\tilde{\boldsymbol{P}}'\tilde{\boldsymbol{\Lambda}}'\tilde{\boldsymbol{Q}}'^\top$ be the SVD of $\tilde{\boldsymbol{S}}$. Similar to equation 13, we observe that $\tilde{\boldsymbol{P}}'\tilde{\boldsymbol{\Lambda}}'(\tilde{\boldsymbol{Q}}'^\top \bar{\boldsymbol{Y}}_{\text{orig}})$ serves as an eigendecomposition of $\tilde{\boldsymbol{S}}$.

Now we make the following observations:

1. By Lemma C.1 (with $\boldsymbol{G}$ replaced by $\tilde{\boldsymbol{G}}$) and the fact that $\tilde{\boldsymbol{\Lambda}}'$ collects the eigenvalues of $\tilde{\boldsymbol{S}}$, the eigenvalues of $\tilde{\boldsymbol{G}}$ are $\frac{\sigma_\xi^2}{mn} + \frac{\lambda_1'^2}{n}, \frac{\sigma_\xi^2}{mn} + \frac{\lambda_2'^2}{n}, \ldots, \frac{\sigma_\xi^2}{mn} + \frac{\lambda_K'^2}{n}, \frac{\sigma_\xi^2}{n}, \ldots, \frac{\sigma_\xi^2}{n}$, which are also the eigenvalues of $\tilde{\boldsymbol{M}}$ because $\tilde{\boldsymbol{V}}$ has orthonormal columns. With the observation that $\tilde{\boldsymbol{S}}' = \frac{1}{\sqrt{m}}\boldsymbol{S}'$ ($\boldsymbol{S}'$ is defined in equation 13), we further conclude that the above eigenvalues equal eigenvalues of $\boldsymbol{G}$ and therefore $\boldsymbol{M}$.

2. Let $\tilde{\boldsymbol{q}}_i$ be the i-th column of $\bar{\boldsymbol{Y}}_{\text{orig}}^\top \tilde{\boldsymbol{Q}}'$. By Lemma C.1 (substitute $\boldsymbol{G}$ with $\tilde{\boldsymbol{G}}$), the i-th ($i \le K$) eigenvector of $\tilde{\boldsymbol{G}}$ is given by $\begin{bmatrix} \frac{1}{\sqrt{1+\tilde{r}_i^2}}\tilde{\boldsymbol{p}}_i' \\ \frac{\tilde{r}_i}{1+\tilde{r}_i^2}\tilde{\boldsymbol{q}}_i \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{1+\tilde{r}_i^2}}\boldsymbol{p}_i \\ \frac{\tilde{r}_i}{1+\tilde{r}_i^2}\tilde{\boldsymbol{q}}_i \end{bmatrix}$, where $\tilde{r}_i = \frac{\sigma_\xi}{\sqrt{m}\lambda_i'} = \frac{\sigma_\xi}{\lambda_i}$. The corresponding eigenvector of $\tilde{\boldsymbol{M}}$ is $\boldsymbol{V}\boldsymbol{T}\begin{bmatrix} \frac{1}{\sqrt{1+\tilde{r}_i^2}}\boldsymbol{p}_i \\ \frac{\tilde{r}_i}{1+\tilde{r}_i^2}\tilde{\boldsymbol{q}}_i \end{bmatrix}$. Observe that $\boldsymbol{T}\bar{\boldsymbol{Y}}_{\text{orig}}^\top = \bar{\boldsymbol{Y}}^\top$, therefore $\boldsymbol{V}\boldsymbol{T}\begin{bmatrix} \frac{1}{\sqrt{1+\tilde{r}_i^2}}\boldsymbol{p}_i \\ \frac{\tilde{r}_i}{1+\tilde{r}_i^2}\tilde{\boldsymbol{q}}_i \end{bmatrix} = \boldsymbol{V}\begin{bmatrix} \frac{1}{\sqrt{1+r_i^2}}\boldsymbol{p}_i \\ \frac{r_i}{1+r_i^2}\boldsymbol{q}_i \end{bmatrix}$ which is the i-th eigenvector of $\boldsymbol{M}$.

Combining the above two leads to the conclusion that the first K eigenvectors/eigenvalues of $\tilde{\boldsymbol{M}}$ and $\boldsymbol{M}$ match. Additionally, we observe that $\textbf{colsp}(\tilde{\boldsymbol{M}}) \subseteq \textbf{colsp}(\boldsymbol{M})$. Therefore the span of the last $n-K$ eigenvectors of $\tilde{\boldsymbol{M}}$ is a subspace of the span of the last $mn-K$ eigenvectors of $\boldsymbol{M}$. Since Lemma C.1 tells us that the remaining $mn-K$ eigenvalues of $\boldsymbol{M}$ are equal, $\boldsymbol{M}$ is identity on the span of the last $mn-K$ eigenvectors. Thus $\boldsymbol{M}$ is identity on the span of the last $n-K$ eigenvectors of $\tilde{\boldsymbol{M}}$. Now we can conclude that $\boldsymbol{M}^\dagger\tilde{\boldsymbol{M}} = \tilde{\boldsymbol{M}}^\dagger\tilde{\boldsymbol{M}}$. □

Lemma C.7. *Suppose that the first $\frac{mn}{2}$ examples have class label $+1$ and the others have class label -1. Let $\boldsymbol{L}^+\boldsymbol{A}^+\boldsymbol{L}^{+\top}$ (where $\boldsymbol{L}^+ \in \mathbb{R}^{d\times 2}$) be the eigendecomposition of $\boldsymbol{M}^+$, then*

$$\boldsymbol{l}_1^+ = \boldsymbol{V}\begin{bmatrix} \frac{1}{\sqrt{1+\mu^2+\frac{\sigma_\xi^2}{mn}}} \\ 0 \\ \frac{\mu}{\sqrt{1+\mu^2+\frac{\sigma_\xi^2}{mn}}} \\ \boldsymbol{0}_{(K-2)\times 1} \\ \frac{\sigma_\xi}{mn\sqrt{1+\mu^2+\frac{\sigma_\xi^2}{mn}}}\boldsymbol{1}_{\frac{mn}{2}\times 1} \\ \frac{\sigma_\xi}{mn\sqrt{1+\mu^2+\frac{\sigma_\xi^2}{mn}}}\boldsymbol{1}_{\frac{mn}{2}\times 1} \end{bmatrix}, \quad \boldsymbol{l}_2^+ = \boldsymbol{V}\begin{bmatrix} 0 \\ \frac{\phi_1}{\sqrt{\phi_1^2+\frac{\sigma_\xi^2}{mn}}} \\ 0 \\ \boldsymbol{0}_{(K-2)\times 1} \\ \frac{\sigma_\xi}{mn\sqrt{\phi_1^2+\frac{\sigma_\xi^2}{mn}}}\boldsymbol{1}_{\frac{mn}{2}\times 1} \\ \frac{-\sigma_\xi}{mn\sqrt{\phi_1^2+\frac{\sigma_\xi^2}{mn}}}\boldsymbol{1}_{\frac{mn}{2}\times 1} \end{bmatrix}, \quad a_1 = 1+\mu^2+\frac{\sigma_\xi^2}{mn}, \quad a_2 = \phi_1^2+\frac{\sigma_\xi^2}{mn} \tag{17}$$

D. Class Collapse in Supervised CL

D.1. Proof of Theorem 4.3

Let $\boldsymbol{l}_\perp$ be the projection of $\boldsymbol{v}_2$ onto $\ker \boldsymbol{M}$. By Corollary C.2, $\|\boldsymbol{l}_\perp\| = \Theta(\frac{\sigma_\xi}{\sqrt{mn}})$. Let $\boldsymbol{a} = \frac{mn}{\sigma_\xi^2}\boldsymbol{l}_\perp$. We can construct a $\boldsymbol{W}^*$ that satisfies the following

$$\boldsymbol{W}^{*\top}\boldsymbol{W}^* = \boldsymbol{M}^\dagger \boldsymbol{M}^+ \boldsymbol{M}^\dagger + \boldsymbol{a}\boldsymbol{a}^\top,$$

which, by Lemma B.2, satisfies the condition for being a minimizer of the loss. In the meantime, $\boldsymbol{W}^*$ also satisfies $\|\boldsymbol{W}^*\boldsymbol{v}_2\| = \Theta(1)$ by Corollary C.2. Note that both $\boldsymbol{v}_2$ and the projection of $\boldsymbol{v}_2$ onto $\textbf{colsp}(\boldsymbol{M})$ is orthogonal to $\boldsymbol{v}_k$ $(1 \leq k \leq K, k \neq 2)$ as well as $\boldsymbol{v}_k$ $(k > mn)$ by Lemma C.1, therefore

$$\boldsymbol{l}_\perp \text{is also orthogonal to } \boldsymbol{v}_k, \text{ for any } k \ s.t. \ 1 \leq k \leq K, k \neq 2 \text{ and } k > mn. \tag{18}$$

Then, for $\boldsymbol{x}$ from $\mathcal{D}_{\text{orig}}$ the following holds true

$$\boldsymbol{W}^*\boldsymbol{x} = c_0\boldsymbol{v}_0 + c_1 y \boldsymbol{v}_1 + y_{\text{sub}} c_2 \boldsymbol{v}_2 + \boldsymbol{h}_{\boldsymbol{x}} + \boldsymbol{W}^*\boldsymbol{\xi},$$

where c_1, c_2 are $\Theta(1)$, and $\boldsymbol{h}_{\boldsymbol{x}}$ is orthogonal to $\boldsymbol{v}_k, k = 0, \ldots, K$ and $\boldsymbol{h}_{\boldsymbol{x}} \in \textbf{colsp}(\boldsymbol{M})$ (by Lemmas C.1, C.5, C.4, equation 18 and that $\|\boldsymbol{W}^*\boldsymbol{v}_2\| = \Theta(1)$). Let $\boldsymbol{\beta} = c_2\boldsymbol{v}_2$, then

$$\boldsymbol{\beta}^\top \boldsymbol{W}^*\boldsymbol{x} = y_{\text{sub}} c_2^2 + \boldsymbol{\beta}^\top \boldsymbol{W}^*\boldsymbol{\xi}.$$

With probability $\geq 1 - \frac{mn}{d}$, $\boldsymbol{\xi} \notin \{\boldsymbol{v}_k\}_{k=1}^{mn}$, which indicates that $\boldsymbol{W}^*\boldsymbol{\xi} = 0$ by Lemma C.1 and equation 18. Therefore we can conclude

$$\Pr(y_{\text{sub}}\boldsymbol{\beta}^\top \boldsymbol{W}^*\boldsymbol{x} > 0 | y) \geq 1 - \frac{mn}{d} = 1 - o(1).$$

D.2. Proof of Theorems 4.4 and 4.7

Theorems 4.4 and 4.7 and can be proved by invoking Lemma B.2 and Corollary C.3.

E. Feature Suppression in Unsupervised CL

E.1. Feature Suppression 1

By Lemmas B.2 and C.6, when $p < K$, any global minimizer of $\mathcal{L}_{\text{UCL}}$ satisfies

$$\boldsymbol{W}^\top \boldsymbol{W} \boldsymbol{M} = \sum_{i=1}^{p} \boldsymbol{r}_i \boldsymbol{r}_i^\top, \tag{19}$$

where $\{\boldsymbol{r}_i\}_{i=1}^p$ can be an orthonormal basis of any p-dimensional subspace of $\textbf{colsp}(\tilde{\boldsymbol{M}})$. By equation 13 and Lemmas C.1 and C.6, $\boldsymbol{M}$ and $\tilde{\boldsymbol{M}}$ each have an eigenvector $\boldsymbol{c}_1$ with eigenvalue $\frac{\sigma_\xi^2}{mn} + \phi_1^2$ and a $\frac{1}{\sqrt{1+\frac{\sigma_\xi^2}{mn\phi_1^2}}}$ alignment with $\boldsymbol{v}_1$, with the other eigenvectors having no alignment with $\boldsymbol{v}_1$. Thus if we include $\boldsymbol{c}_1$ in $\{\boldsymbol{r}_i\}_{i=1}^p$ and let $\boldsymbol{W}^\top\boldsymbol{W}$ be null on $\ker \boldsymbol{M}$, then the constructed $\boldsymbol{W}$ is a minimizer of $\mathcal{L}_{\text{UCL}}$ with $\Theta(1)$ alignment with $\boldsymbol{v}_1$. Now let's look at the minimum norm minimizer, which should satisfy

$$\boldsymbol{W}^\top \boldsymbol{W} = \sum_{i=1}^{p} \boldsymbol{r}_i \boldsymbol{r}_i^\top \boldsymbol{M}^\dagger,$$

where $\{\boldsymbol{r}_i\}_{i=1}^p$ is selected such that $\boldsymbol{W}$ has the smallest norm. By Lemma C.6, $\{\boldsymbol{r}_i\}_{i=1}^p$ should be the p-eigenvectors of $\boldsymbol{M}$ with largest eigenvalues (so that the inverse of the eigenvalues are among the smallest). If among $\frac{(1+\mu^2+\phi_2^2)+\sqrt{(1+\mu^2+\phi_2)^2-4\phi_2^2}}{2}, \frac{(1+\mu^2+\phi_2^2)-\sqrt{(1+\mu^2+\phi_2)^2-4\phi_2^2}}{2}, \frac{\phi_3}{\sqrt{K-2}}, \ldots, \frac{\phi_K}{\sqrt{K-2}}$ there are p elements larger than ϕ_1, then $\frac{\sigma_\xi^2}{mn} + \phi_1^2$ is not among the p largest eigenvalues of $\boldsymbol{M}$. Thus $\boldsymbol{c}_1$ is not included in $\{\boldsymbol{r}_i\}_{i=1}^p$ and the corresponding $\boldsymbol{W}$ is orthogonal to $\boldsymbol{v}_1$.

E.2. Feature Suppression 2

We first present our result under slightly technical conditions.

Lemma E.1. *Let $\boldsymbol{v_1}, \dots, \boldsymbol{v_C} \in \mathcal{R}^d$ be nonzero and orthogonal, U, A are subspaces that are orthogonal to each other and all the $\boldsymbol{v_i}$. Suppose we have a data distribution $\mathcal{D} = \{(\boldsymbol{v_{y_i}} + \boldsymbol{u_{y_i}} + \boldsymbol{a_i}, y_i)\}_{i=1}^n \subset \mathbb{R}^d \times \{1, \dots, C\}$, where $\boldsymbol{u}_i \in U, \boldsymbol{a}_i \in A$ for all $i \in \{1, \dots, n\}$ (namely all examples in the same class c share the same $\boldsymbol{v}_c$ and $\boldsymbol{u}_c$).*

Denote $\boldsymbol{z}_{y_i} = \boldsymbol{v}_{y_i} + \boldsymbol{u}_{y_i}$, and let $\boldsymbol{M}, \boldsymbol{M}^+$ be the matrices defined for this dataset, and let $\boldsymbol{Z}, \boldsymbol{Z}^+$ and $\boldsymbol{A}, \boldsymbol{A}^+$ be the corresponding matrices when the data is $\{(\boldsymbol{z}_{y_i}, y_i)\}$ and $\{(\boldsymbol{a}_i, y_i)\}$, respectively. Suppose that $(\boldsymbol{A} - \boldsymbol{A}^+)\boldsymbol{v} \neq \boldsymbol{0}$ for all $\boldsymbol{v} \in \mathbb{R}^d$ s.t. $\boldsymbol{A}\boldsymbol{v} \neq 0$ and the output dimension $p \geq C$. Then $\boldsymbol{W}^\top \boldsymbol{W} = \boldsymbol{Z}^\dagger$ is the minimum norm solution to the contrastive learning objective on $\mathcal{D}$.

Proof. In this proof, we will use $\mathbb{E}$ to represent the empirical expectation over the dataset $\mathcal{D}$. Also, let n_c denote the number of examples in class c.

We first derive the following expression for $\boldsymbol{A}^+$:

$$\boldsymbol{A}^+ = \mathbb{E}[\boldsymbol{a_i} \boldsymbol{z}_{y_i}^\top] \boldsymbol{Z}^\dagger \mathbb{E}[\boldsymbol{z}_{y_i} \boldsymbol{a_i}^\top] \tag{20}$$

Define $\boldsymbol{B} = [\sqrt{n_1}\boldsymbol{z_1} \dots, \sqrt{n_C}\boldsymbol{z_C}] \in \mathbb{R}^{d \times C}, \boldsymbol{C} = [\boldsymbol{a}_1^*, \dots, \boldsymbol{a}_C^*] \in \mathbb{R}^{d \times C}$, where n_c is the number of examples in class c and $\boldsymbol{a}_c^* = \frac{1}{\sqrt{n_c}} \sum_{y_i = c} \boldsymbol{a_i}$. Then

$$\mathbb{E}[\boldsymbol{a_i} \boldsymbol{z}_{y_i}^\top] \boldsymbol{Z}^\dagger \mathbb{E}[\boldsymbol{z}_{y_i} \boldsymbol{a_i}^\top] = \frac{1}{n} \boldsymbol{C}\boldsymbol{B}^\top \left(\frac{1}{n} \boldsymbol{B}\boldsymbol{B}^\top \right)^\dagger \frac{1}{n} \boldsymbol{B}\boldsymbol{C}^\top \tag{21}$$

Now $\boldsymbol{B}$ has full column rank, so $\boldsymbol{B}^\top (\boldsymbol{B}\boldsymbol{B}^\top)^\dagger \boldsymbol{B} = \boldsymbol{I}$. Thus

$$\begin{aligned}
\mathbb{E}[\boldsymbol{a_i} \boldsymbol{z}_{y_i}^\top] \boldsymbol{Z}^\dagger \mathbb{E}[\boldsymbol{z}_{y_i} \boldsymbol{a_i}^\top] &= \frac{1}{n} \boldsymbol{C}\boldsymbol{C}^\top && (22)\\
&= \sum_{c=1}^{C} \frac{n_c}{n} \mathbb{E}_{y_i = c}[\boldsymbol{a_i}] \mathbb{E}_{y_i = c}[\boldsymbol{a_i}]^\top && (23)\\
&= \boldsymbol{A}^+ && (24)
\end{aligned}$$

Now we show that $\boldsymbol{W}^\top \boldsymbol{W} = \boldsymbol{Z}^\dagger$ is a global minimizer. It suffices to show that $\boldsymbol{M}\boldsymbol{W}^\top \boldsymbol{W}\boldsymbol{M} = \boldsymbol{M}^+$. Note that by assumption, we have $\langle \boldsymbol{z}_i, \boldsymbol{a}_j \rangle = 0$ for all $i \in \{1, \dots, C\}, j \in \{1, \dots, n\}$, so we have

$$\begin{aligned}
\boldsymbol{M}\boldsymbol{Z}^\dagger \boldsymbol{M} &= \mathbb{E}[(\boldsymbol{z}_{y_i} + \boldsymbol{a_i})(\boldsymbol{z}_{y_i} + \boldsymbol{a_i})^\top] \boldsymbol{M}_*^\dagger \mathbb{E}[(\boldsymbol{z}_{y_i} + \boldsymbol{a_i})(\boldsymbol{z}_{y_i} + \boldsymbol{a_i})^\top] && (25)\\
&= (\boldsymbol{Z} + \mathbb{E}[\boldsymbol{z}_{y_i} \boldsymbol{a_i}^\top] + \mathbb{E}[\boldsymbol{a_i} \boldsymbol{z}_{y_i}^\top] + \boldsymbol{A}) \boldsymbol{Z}^\dagger (\boldsymbol{Z} + \mathbb{E}[\boldsymbol{z}_{y_i} \boldsymbol{a_i}^\top] + \mathbb{E}[\boldsymbol{w_i} \boldsymbol{z}_{y_i}^\top] + \boldsymbol{A}) && (26)\\
&= \boldsymbol{Z}\boldsymbol{Z}^\dagger \boldsymbol{Z} + \boldsymbol{Z}\boldsymbol{Z}^\dagger \mathbb{E}[\boldsymbol{z}_{y_i} \boldsymbol{a_i}^\top] + \mathbb{E}[\boldsymbol{a_i} \boldsymbol{z}_{y_i}^\top] \boldsymbol{Z}^\dagger \boldsymbol{Z} + \mathbb{E}[\boldsymbol{a_i} \boldsymbol{z}_{y_i}^\top] \boldsymbol{Z}^\dagger \mathbb{E}[\boldsymbol{z}_{y_i} \boldsymbol{a_i}^\top] && (27)\\
&= \boldsymbol{Z} + \mathbb{E}[\boldsymbol{z}_{y_i} \boldsymbol{a_i}^\top] + \mathbb{E}[\boldsymbol{a_i} \boldsymbol{z}_{y_i}^\top] + \boldsymbol{A}^+ && (28)\\
&= \boldsymbol{Z}^+ + \mathbb{E}[\boldsymbol{z}_{y_i} \boldsymbol{a_i}^\top] + \mathbb{E}[\boldsymbol{a_i} \boldsymbol{z}_{y_i}^\top] + \boldsymbol{A}^+ && (29)\\
&= \boldsymbol{M}^+ && (30)
\end{aligned}$$

We now want to show that this is the minimum norm solution. It is sufficient to show that $\mathrm{im}(\boldsymbol{W}^\top \boldsymbol{W}) = \mathrm{im}(\boldsymbol{Z}^\dagger) = \mathrm{im}(\boldsymbol{Z}) \subset \mathrm{im}(\boldsymbol{M})$. Note that $\mathrm{im}(\boldsymbol{M}) \subset \mathrm{im}(\boldsymbol{A}) \oplus \mathrm{im}(\boldsymbol{Z})$, so we can restrict $\boldsymbol{M}$ to this subspace. We will show that $\boldsymbol{M}$ is invertible on $\mathrm{im}(\boldsymbol{A}) \oplus \mathrm{im}(\boldsymbol{Z})$. Suppose $\boldsymbol{v} = \boldsymbol{z} + \boldsymbol{a}$ with $\boldsymbol{z} \in \mathrm{im}(\boldsymbol{Z}), \boldsymbol{a} \in \mathrm{im}(\boldsymbol{A}), \boldsymbol{M}\boldsymbol{v} = 0$. This implies that

$$\boldsymbol{Z}\boldsymbol{z} + \mathbb{E}[\boldsymbol{z}_{y_i} \boldsymbol{a}_i^\top] \boldsymbol{a} = \boldsymbol{0} \tag{31}$$

$$\mathbb{E}[\boldsymbol{a}_i \boldsymbol{z}_{y_i}^\top] \boldsymbol{z} + \boldsymbol{A}\boldsymbol{a} = \boldsymbol{0} \tag{32}$$

Left-multiplying the first equation by $\mathbb{E}[\frac{n}{n_{y_i}\|\boldsymbol{v}_{y_i}\|^2}\boldsymbol{a_i}\boldsymbol{v_i}^\top]$, by orthogonality we have

$$\begin{aligned}
\mathbf{0} &= \mathbb{E}\left[\frac{n}{n_{y_i}\|\boldsymbol{v}_{y_i}\|^2}\boldsymbol{a_i}\boldsymbol{v}_{y_i}^\top\right]\left(\mathbb{E}[\boldsymbol{z}_{y_i}\boldsymbol{z}_{y_i}^\top]\boldsymbol{z} + \mathbb{E}[\boldsymbol{z}_{y_i}\boldsymbol{a}_i^\top]\boldsymbol{a}\right)\\
&= \mathbb{E}\left[\frac{n}{n_{y_i}\|\boldsymbol{v}_{y_i}\|^2}\boldsymbol{a_i}\boldsymbol{v}_{y_i}^\top\right]\left(\mathbb{E}[(\boldsymbol{v}_{y_i}+\boldsymbol{u}_{y_i})\boldsymbol{z}_{y_i}^\top]\boldsymbol{z} + \mathbb{E}[(\boldsymbol{v}_{y_i}+\boldsymbol{u}_{y_i})\boldsymbol{a}_i^\top]\boldsymbol{a}\right)\\
&= \mathbb{E}\left[\frac{n}{n_{y_i}\|\boldsymbol{v}_{y_i}\|^2}\boldsymbol{a_i}\boldsymbol{v}_{y_i}^\top\right]\left(\mathbb{E}[\boldsymbol{u}_{y_i}(\boldsymbol{z}_{y_i}^\top\boldsymbol{z}+\boldsymbol{a}_i^\top\boldsymbol{a})] + \mathbb{E}[\boldsymbol{v}_{y_i}(\boldsymbol{z}_{y_i}^\top\boldsymbol{z}+\boldsymbol{a}_i^\top\boldsymbol{a})]\right)\\
&= \mathbb{E}\left[\frac{n}{n_{y_i}\|\boldsymbol{v}_{y_i}\|^2}\boldsymbol{a_i}\boldsymbol{v}_{y_i}^\top\right]\mathbb{E}[\boldsymbol{v}_{y_i}(\boldsymbol{z}_{y_i}^\top\boldsymbol{z}+\boldsymbol{a}_i^\top\boldsymbol{a})]\\
&= \sum_{c=1}^{C}\frac{1}{nn_c\|\boldsymbol{v}_c\|^2}\left(\sum_{y_i=c}\boldsymbol{a_i}\right)\boldsymbol{v}_c^\top\boldsymbol{v_c}\left(n_c\boldsymbol{z_c}^\top\boldsymbol{z} + \sum_{y_i=c}\boldsymbol{a_i}^\top\boldsymbol{a}\right)\\
&= \sum_{c=1}^{C}\frac{1}{nn_c}\left(\sum_{y_i=c}\boldsymbol{a_i}\right)\left(n_c\boldsymbol{z_c}^\top\boldsymbol{z} + \sum_{y_i=c}\boldsymbol{a_i}\boldsymbol{a}\right)\\
&= \frac{1}{n^2}\sum_{c=1}^{C}\frac{1}{n}\left(\sum_{y_i=c}\boldsymbol{a_i}\right)\boldsymbol{z_c}^\top\boldsymbol{z} + \frac{1}{nn_c}\left(\sum_{y_i=c}\boldsymbol{a_i}\right)\left(\sum_{y_i=c}\boldsymbol{a_i}\right)^\top\boldsymbol{a}\\
&= E[\boldsymbol{a}_i\boldsymbol{z}_{y_i}^\top]\boldsymbol{z} + \boldsymbol{A}^+\boldsymbol{a}
\end{aligned}$$

Now substituting into the second equation, we find that

$$(\boldsymbol{A}-\boldsymbol{A}^+)\boldsymbol{a} = \mathbf{0} \tag{33}$$

But our assumptions imply that $\boldsymbol{a}=0$. Returning to the first equation, we now have $\boldsymbol{Z}\boldsymbol{z}=0$. But since $\boldsymbol{Z}$ is diagonalizable, $\boldsymbol{Z}$ must be invertible on its image, hence $\boldsymbol{z}=0$. We conclude that $\boldsymbol{v}=0$. This completes the proof.

□

We now want to show that we can simplify some of the conditions of the previous lemma to linear independence.

Lemma E.2. *Suppose $d \geq 3n-2$ and $\boldsymbol{x_1},\ldots,\boldsymbol{x_n} \in \mathbb{R}^d$ are linearly independent. Then there exists a set of nonzero orthogonal vectors $\boldsymbol{v_1},\ldots,\boldsymbol{v_n}$ s.t. $\boldsymbol{x_i}=\boldsymbol{v_i}+\boldsymbol{u_i}$ and $\boldsymbol{v_i},\boldsymbol{u_j}$ are orthogonal for all $i,\in\{1,\ldots,n\}$.*

Proof. WLOG assume the $\boldsymbol{x_i}$ are contained in the span of the first n basis vectors. The lemma amounts to finding an orthonormal matrix $\boldsymbol{\Omega}=\begin{pmatrix}\boldsymbol{A} & \boldsymbol{B}\\ \boldsymbol{C} & \boldsymbol{D}\end{pmatrix}$ s.t.

$$\begin{pmatrix}\boldsymbol{A} & \boldsymbol{B}\\ \boldsymbol{C} & \boldsymbol{D}\end{pmatrix}\begin{pmatrix}\boldsymbol{X}\\ \mathbf{0}\end{pmatrix} = \begin{pmatrix}\boldsymbol{AX}\\ \boldsymbol{CX}\end{pmatrix} = \begin{pmatrix}\boldsymbol{\Sigma}\\ \boldsymbol{F}\end{pmatrix} \tag{34}$$

where $\boldsymbol{\Sigma}$ is diagonal. Since the $\boldsymbol{x_i}$ are linearly independent, $\boldsymbol{X}$ is invertible, so there exists $\boldsymbol{A}'$ s.t. $\boldsymbol{A}'\boldsymbol{X}$ is diagonal.

We now want to construct a matrix $\boldsymbol{C}$ such that $\begin{pmatrix}\boldsymbol{A}'\\ \boldsymbol{C}'\end{pmatrix}$ has orthogonal columns, all with norm $l>0$. Note that $\boldsymbol{C}'$ has at least $2n-2$ rows. Set $\boldsymbol{C}'_{11}=1$, and the remaining entries in the first row so that when considering $\boldsymbol{A}$ and the first row of $\boldsymbol{C}'$, the first column is orthogonal to every other column. Now leave $\boldsymbol{C}'_{21}=0$, set $\boldsymbol{C}'_{22}=1$, and fill out the remaining entries in the second row so that when considering $\boldsymbol{A}$ and the first two rows of $\boldsymbol{C}'$, the second column is orthogonal to the remaining columns. Note that the first column remains orthogonal to all other columns. Continuing in this fashion, we can use the first $n-1$ rows of $\boldsymbol{C}'$ to guarantee that all n columns are orthogonal. Finally, suppose without loss of generality that whhen considering the $\boldsymbol{A}'$ and the first $n-1$ rows of $\boldsymbol{C}'$, the first column has the largest norm l. For each of the remaining $n-1$ rows, set the jth row to have all zero entries except possibly in the $(j+1)$-th column, which is set so that the jth column will also have norm l. Note that the columns remain orthogonal under this construction.

Now $\frac{1}{l}\begin{pmatrix}\boldsymbol{A}'\\ \boldsymbol{C}'\end{pmatrix}$ has orthonormal columns and $\frac{1}{l}\boldsymbol{A}'\boldsymbol{X}$ is still diagonal. By Gram-Schmidt, we can fill out the remaining columns of $\boldsymbol{\Omega}$ to construct an orthonormal matrix. □

We now present the feature result with simplified assumptions.

Lemma E.3. *Let Z, A be orthogonal subspaces. Suppose we have a data distribution $\mathcal{D} = \{(\boldsymbol{z}_{\boldsymbol{y_i}} + \boldsymbol{a_i}, y_i)\}_{i=1}^n \subset \mathbb{R}^d \times \{1, \dots, C\}$, where $\boldsymbol{z}_i \in Z, \boldsymbol{a}_i \in A$ for all $i \in \{1, \dots, n\}$, and the $\boldsymbol{z}_i$ are linearly independent.*

Let $\boldsymbol{M}, \boldsymbol{M}^+$ be the matrices defined for this dataset, and let $\boldsymbol{Z}, \boldsymbol{Z}^+$ and $\boldsymbol{A}, \boldsymbol{A}^+$ be the corresponding matrices when the data is $\{(\boldsymbol{z}_{y_i}, y_i)\}$ and $\{(\boldsymbol{a}_i, y_i)\}$, respectively. Suppose that $(\boldsymbol{A} - \boldsymbol{A}^+)\boldsymbol{v} \neq \boldsymbol{0}$ for all $\boldsymbol{v} \in \mathbb{R}^d$ s.t. $\boldsymbol{A}\boldsymbol{v} \neq 0$ and the output dimension $p \geq C$. Then $\boldsymbol{W}^\top \boldsymbol{W} = \boldsymbol{Z}^\dagger$ is the minimum norm solution to the contrastive learning objective on $\mathcal{D}$.

Proof. Assume that $d \geq 3C - 2$, otherwise embed the distribution in a space of sufficiently large dimension. By Lemma B.2, the minimum norm minimizer is unaffected by adding extra dimensions. Then Lemma E.2 applies, so linear independence of the $\boldsymbol{z}_{y_i}$ is sufficient to be able to construct $\boldsymbol{v}_1, \dots, \boldsymbol{v}_C, \boldsymbol{y}_1, \dots, \boldsymbol{y}_C$ satisfying Lemma E.1, from which the conclusion follows.

□

F. Minimizer of The Joint Loss

For simplicity we assume $\mu = 0$. Same strategy can be applied to prove the theorem when $\mu \neq 0$ but a more detailed discussion on the selection of β may be required.

By Lemmas C.7 and C.1 and the expression of $\boldsymbol{S}$ (equation 13), we observe that the two eigenvectors of $\boldsymbol{M}^+$ match two of the eigenvectors of $\boldsymbol{M}$. By combining this with Lemma C.6, we obtain that $\beta\boldsymbol{M}^\dagger\boldsymbol{M}^+ + (1-\beta)\boldsymbol{M}^\dagger\tilde{\boldsymbol{M}} = \boldsymbol{l}_1^+\boldsymbol{l}_1^{+\top} + \boldsymbol{l}_2^+\boldsymbol{l}_2^{+\top}$ on $\textbf{span}(\{\boldsymbol{l}_1^+, \boldsymbol{l}_2^+\})$ and $\beta\boldsymbol{M}^\dagger\boldsymbol{M}^+ + (1-\beta)\boldsymbol{M}^\dagger\tilde{\boldsymbol{M}} = (1-\beta)\tilde{\boldsymbol{M}}^\dagger\tilde{\boldsymbol{M}}$ on $\textbf{span}(\{\boldsymbol{l}_1^+, \boldsymbol{l}_2^+\})^\perp$. Thus the eigenvalues of $\beta\boldsymbol{M}^\dagger\boldsymbol{M}^+ + (1-\beta)\boldsymbol{M}^\dagger\tilde{\boldsymbol{M}}$ are $1, 1, 1-\beta, 1-\beta, \dots, 1-\beta$. When $\beta \in (0,1)$, $\boldsymbol{l}_1^+$ and $\boldsymbol{l}_2^+$ are the two eigenvectors of $\beta\boldsymbol{M}^\dagger\boldsymbol{M}^+ + (1-\beta)\boldsymbol{M}^\dagger\tilde{\boldsymbol{M}}$ with largest eigenvalues. For the remaining eigenvectors, since they have equally large eigenvalues (same as analyzed in E), the minimum norm minimizer will select the largest $p-2$ of them. In the setting of Theorem 6.1 $(1-\beta)(\phi_2^2 + \frac{\sigma_\xi^2}{mn})$ is one of the $p-2$ largest of the remaining. As a result, both components aligned with $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ are selected by the minimum norm minimizer of the joint loss.

G. Early in Training Subclasses Are Learned

We assume $\sigma_\xi = O(1)$.

G.1. Lemmas

Lemma G.1 (Laurent-Massart (Laurent & Massart, 2000) Lemma 1, page 1325)**.** *Let $v_1, \dots, v_d$ be i.i.d. Gaussian variables drawn from $\mathcal{N}(0,1)$. Let $\boldsymbol{a} = (a_1, \dots, a_d)$ be a vector with non-negative components. Let $Z = \sum_{i=1}^d a_i(v_i^2 - 1)$. The following inequalities hold for any positive t:*

$$\begin{aligned}&\Pr(Z \geq 2\|\boldsymbol{a}\|_2\sqrt{t} + 2\|\boldsymbol{a}\|_\infty t) \leq e^{-t},\\ &\Pr(Z \leq -2\|\boldsymbol{a}\|_2\sqrt{t}) \leq e^{-t}.\end{aligned} \tag{35}$$

Lemma G.2 (Mills' ratio. Exercise 6.1 in (Shorack & Shorack, 2000).)**.** *Let v be a Gaussian random variable drawn from $\mathcal{N}(0,1)$. Then for all $\lambda > 0$,*

$$\frac{\lambda}{\lambda^2+1}\frac{1}{\sqrt{2\pi}}e^{-\frac{\lambda^2}{2}} < \Pr(v \geq \lambda) < \frac{1}{\lambda}\frac{1}{\sqrt{2\pi}}e^{-\frac{\lambda^2}{2}}.$$

Corollary G.3. *Given a vector $\boldsymbol{q}$, and a random vector $\boldsymbol{z}$ drawn from $\mathcal{N}(0, \frac{\sigma}{d}\mathbf{I}_d)$, w.p. $\geq 1 - O(\frac{\delta}{\sqrt{\log 1/\delta}})$, $|\boldsymbol{z}^\top\boldsymbol{q}| = O(\frac{\|\boldsymbol{q}\|\sigma\sqrt{\log\frac{1}{\delta}}}{\sqrt{d}})$.*

Proof. This can be proven by considering the fact that $\boldsymbol{q}^\top \boldsymbol{z}$ is a Gaussian variable and applying Lemma G.2. □

Lemma G.4. *Let each element of $\boldsymbol{W}_0 \in \mathbb{R}^{p\times d}$ be randomly drawn from $\mathcal{N}(0, \frac{\sigma_0^2}{d}\mathbf{I}_d)$. Let $\boldsymbol{u} \in \mathbb{R}^d$ be a unit vector. With probability at least $1-\delta$, we have*

$$\|\boldsymbol{W}_0\boldsymbol{u}\| \geq \sigma_0\sqrt{\frac{p}{d}}\sqrt{1-2\sqrt{\frac{\ln 2/\delta}{p}}}$$
$$\|\boldsymbol{W}_0\boldsymbol{u}\| \leq \sigma_0\sqrt{\frac{p}{d}}\sqrt{1+2\sqrt{\frac{\ln 2/\delta}{p}}+2\frac{\ln 2/\delta}{p}}.$$

Proof. Firstly rewrite $\|\boldsymbol{W}_0\boldsymbol{u}\|$ as

$$\|\boldsymbol{W}_0\boldsymbol{u}\| = \sqrt{\sum_{i=1}^{p}(\boldsymbol{w}_0^{(i)\top}\boldsymbol{u})^2} = \sigma_0\sqrt{\frac{p}{d}}\sqrt{\frac{1}{p}\sum_{i=1}^{p}(\frac{\sqrt{d}}{\sigma_0}\boldsymbol{w}_0^{(i)\top}\boldsymbol{u})^2}.$$

By spherical symmetric, each $\frac{\sqrt{d}}{\sigma_0}\boldsymbol{w}_0^{(i)\top}\boldsymbol{u}$ is a random Gaussian variable drawn from $\mathcal{N}(0,1)$. By lemma G.1 we have

$$\Pr\left(\frac{1}{p}\sum_{i=1}^{p}(\frac{\sqrt{d}}{\sigma_0}\boldsymbol{w}_0^{(i)\top}\boldsymbol{u})^2 \leq 1-2\sqrt{\frac{\ln 2/\delta}{p}}\right) \leq \delta/2$$
$$\Pr\left(\frac{1}{p}\sum_{i=1}^{p}(\frac{\sqrt{d}}{\sigma_0}\boldsymbol{w}_0^{(i)\top}\boldsymbol{u})^2 \geq 1+2\sqrt{\frac{\ln 2/\delta}{p}}+2\frac{\ln 2/\delta}{p}\right) \leq \delta/2$$

which completes the proof. □

G.2. Proof of Theorem 4.5

We assume the dataset satisfies the condition in Section C (wich holds with probability $1-O(\frac{m^2n^2}{d})$). Let $\boldsymbol{L}\boldsymbol{A}\boldsymbol{L}^\top$ (where $\boldsymbol{C} \in \mathbb{R}^{d\times mn}$) be the eigendecomposition of $\boldsymbol{M}$. By equation 13 and Lemma C.7 and Lemma C.1, we observe that when $\mu \neq 0$ all but three of $\boldsymbol{M}$'s eigenvectors are orthogonal to $\boldsymbol{l}_1^+, \boldsymbol{l}_2^+$. W.L.O.G., let $\boldsymbol{l}_1, \boldsymbol{l}_2$ and $\boldsymbol{l}_3$ be those three eigenvectors. The corresponding three eigenvalues are all constants. Let $\boldsymbol{l}_3^+$ be a unit vector in $\textbf{span}(\{\boldsymbol{l}_1, \boldsymbol{l}_2, \boldsymbol{l}_3\}) - \textbf{span}(\{\boldsymbol{l}_1^+, \boldsymbol{l}_2^+\})$. Decompose $\boldsymbol{v}_2$ as $\frac{\mu}{\sqrt{1+\mu^2+\frac{\sigma_\xi^2}{mn}}}\boldsymbol{l}_1^+ + \frac{\sqrt{1+\frac{\sigma_\xi^2}{mn}}}{\sqrt{1+\mu^2+\frac{\sigma_\xi^2}{mn}}}\boldsymbol{l}_\perp$ where $\boldsymbol{l}_\perp$ is a unit vector that is orthogonal to $\boldsymbol{l}_1^+$. Since $\boldsymbol{v}_2 \perp \boldsymbol{l}_2^+$, we have $\boldsymbol{l}_\perp \perp \boldsymbol{l}_2^+$ thus $\boldsymbol{M}^+\boldsymbol{l}_\perp = 0$.

Define

$$
\begin{aligned}
\sqrt{\boldsymbol{M}} :=& \boldsymbol{L}\sqrt{\boldsymbol{A}} \\
\Gamma_i(t) :=& \|\boldsymbol{W}_t \boldsymbol{l}_i^+\|, \quad i = 1, 2, 3 \\
\Gamma_\perp(t) :=& \|\boldsymbol{W}_t \boldsymbol{l}_\perp\| \\
\Gamma_{:3}(t) :=& \sqrt{\sum_{i=1}^{3} \|\boldsymbol{W}_t \boldsymbol{l}_i^+\|^2} \\
\boldsymbol{B} :=& [\sqrt{a}_4 \boldsymbol{l}_4 \ \ \sqrt{a}_5 \boldsymbol{l}_5 \ \ \dots \ \ \sqrt{a}_{mn} \boldsymbol{l}_{mn}] \\
\Gamma_B(t) :=& \|\boldsymbol{W}_t \boldsymbol{B}\|_F \\
s :=& \|\sqrt{\boldsymbol{M}}\| = O(1) \\
h :=& \|\sqrt{\boldsymbol{M}}^\top \boldsymbol{B}\| = \sqrt{\sum_{i=4}^{mn} a_i^2} = \sqrt{\sum_{i=3}^{K} (\frac{\sigma_\xi^2}{mn} + \frac{\phi_i^2}{(K-2)})^2 + (mn-K)\frac{\sigma_\xi^4}{m^2 n^2}} \\
=& O(\sqrt{\frac{\sigma_\xi^2}{mn} + \frac{1}{K} + \frac{\sigma_\xi^4}{mn}}) = O(1)
\end{aligned}
$$

Then we bound $\|\boldsymbol{W}_t \sqrt{\boldsymbol{M}}\|_F$

$$
\begin{aligned}
\|\boldsymbol{W}_t \sqrt{\boldsymbol{M}}\|_F =& \|\boldsymbol{W}_t \boldsymbol{L}\sqrt{\boldsymbol{A}}\|_F \\
=& \|[\boldsymbol{W}_t \sqrt{a}_1 \boldsymbol{l}_1 \ \ \boldsymbol{W}_t \sqrt{a}_2 \boldsymbol{l}_2 \ \ \dots \ \ \boldsymbol{W}_t \sqrt{a}_{mn} \boldsymbol{l}_{mn}]\|_F \\
=& \sqrt{\sum_{i=1}^{3} \|\boldsymbol{W}_t \sqrt{a}_i \boldsymbol{l}_i\|^2 + \sum_{i=4}^{mn} \|\boldsymbol{W}_t \sqrt{a}_i \boldsymbol{l}_i\|^2} \\
\leq& \sqrt{c\Gamma_{:3}(t)^2 + \Gamma_B(t)^2},
\end{aligned}
$$

where c is a constant because a_1, a_2, a_3 are all $O(1)$ (by Lemma C.1) and each $\boldsymbol{l}_i$ $(i = 1, 2, 3)$ is a linear combination of $\boldsymbol{l}_1^{++}, \boldsymbol{l}_2^{++}, \boldsymbol{l}_3^+$ with $O(1)$ coefficients, with $\boldsymbol{l}_1^{++}, \boldsymbol{l}_2^{++}$ representing the projections of $\boldsymbol{l}_1^+, \boldsymbol{l}_2^+$ onto **span**$(\{\boldsymbol{l}_i\}_{i=1}^3)$.

By the rule of gradient descent we have

$$
\begin{aligned}
\boldsymbol{W_{t+1}} =& \boldsymbol{W_t} + \eta(4\boldsymbol{W_t}\boldsymbol{M}^+ - 4\boldsymbol{W_t}\boldsymbol{M}\boldsymbol{W_t}^\top \boldsymbol{W_t}\boldsymbol{M}) \\
=& \boldsymbol{W_t} + 4\eta \boldsymbol{W_t}\boldsymbol{M}^+ - 4\eta \boldsymbol{W_t}\boldsymbol{M}\boldsymbol{W_t}^\top \boldsymbol{W_t}\boldsymbol{M}m
\end{aligned} \tag{36}
$$

This is followed by Lemma G.5.

Lemma G.5. *By the update rule of GD we have the following recurrence relations*

$$
\begin{aligned}
\Gamma_1(t+1) \geq& (1 + 4\eta a_1^+)\Gamma_1(t) - 4\eta(c\Gamma_{:3}^2(t) + \Gamma_B(t)^2)^{3/2} s \\
\Gamma_1(t+1) \leq& (1 + 4\eta a_1^+)\Gamma_1(t) + 4\eta(c\Gamma_{:3}^2(t) + \Gamma_B(t)^2)^{3/2} s \\
\Gamma_2(t+1) \leq& (1 + 4\eta a_2^+)\Gamma_2(t) + 4\eta(c\Gamma_{:3}^2(t) + \Gamma_B(t)^2)^{3/2} s \\
\Gamma_3(t+1) \leq& \Gamma_3(t) + 4\eta(c\Gamma_{:3}^2(t) + \Gamma_B(t)^2)^{3/2} s \\
\Gamma_\perp(t+1) \leq& \Gamma_\perp(t) + 4\eta(c\Gamma_{:3}^2(t) + \Gamma_B(t)^2)^{3/2} s \\
\Gamma_B(t+1) \leq& \Gamma_B(t) + 4\eta(c\Gamma_{:3}^2(t) + \Gamma_B(t)^2)^{3/2} h.
\end{aligned}
$$

Then we prove the following Lemma

Lemma G.6. *At initialization the following holds with probability* $\geq 1 - O(\frac{1}{\text{poly}(p)})$

$$\begin{aligned}
\frac{\Gamma_B(0)}{\Gamma_1(0)} =& O(1) \\
\Gamma_i(0) =& \sigma_0 \sqrt{\frac{p}{d}} \left(1 \pm O(\sqrt{\frac{\log p}{p}})\right), \quad i = 1, 2, 3 \\
\Gamma_\perp(0) =& \sigma_0 \sqrt{\frac{p}{d}} \left(1 \pm O(\sqrt{\frac{\log p}{p}})\right)
\end{aligned}$$

Proof. We first bound $\Gamma_B(0)$

$$\begin{aligned}
\Gamma_B(0) =& \sqrt{\sum_{i=4}^{mn} a_i \|\boldsymbol{W}_0 \boldsymbol{l}_i\|^2} \\
=& \sqrt{\sum_{i=4}^{mn} a_i \sum_{j=1}^{p} \|\boldsymbol{w}_{0,j}^\top \boldsymbol{l}_i\|^2} \\
\leq& \sqrt{\frac{\phi_{\max}^2}{K-2} \sum_{i=4}^{K+1} \sum_{j=1}^{p} \|\boldsymbol{w}_{0,j}^\top \boldsymbol{l}_i\|^2 + \frac{\sigma_\xi^2}{mn} \sum_{i=K+2}^{mn} \sum_{j=1}^{p} \|\boldsymbol{w}_{0,j}^\top \boldsymbol{l}_i\|^2} \\
=& O(\sqrt{\frac{p\sigma_0^2}{d} + \sigma_\xi^2 \frac{p\sigma_0^2}{d}}) \quad ① \\
=& O(\sigma_0 \sqrt{\frac{p}{d}}).
\end{aligned}$$

Inequality ① holds with probability $\geq 1 - O(\frac{1}{\text{poly}(mnp)})$. It is obtained by obsreving that $\boldsymbol{w}_{0,j}^\top \boldsymbol{l}_i$'s are independent Gaussian variables (by the orthogonality of $\boldsymbol{l}_i$'s) and applying Lemma G.1 to the sum of $\|\boldsymbol{w}_{0,j}^\top \boldsymbol{l}_i\|^2$'s.

By Lemma G.4 and the above, at initialization the following holds with probability $\geq 1 - O(\frac{1}{\text{poly}(p)} + \frac{1}{\text{poly}(mnp)})$

$$\begin{aligned}
\frac{\Gamma_B(0)}{\Gamma_1(0)} =& O(1) \\
\Gamma_i(0) =& \sigma_0 \sqrt{\frac{p}{d}} \left(1 \pm O(\sqrt{\frac{\log p}{p}})\right), \quad i = 1, 2, 3 \\
\Gamma_\perp(0) =& \sigma_0 \sqrt{\frac{p}{d}} \left(1 \pm O(\sqrt{\frac{\log p}{p}})\right).
\end{aligned}$$

□

Let ψ, ψ_B be constants. Define

$$\begin{aligned}
\pi :=& \frac{\Gamma_B(0)}{\Gamma_1(0)} = O(1) \text{ by Lemma G.6} \\
\tau :=& (c(1 + 2(1+\psi)^2) + (\pi + \psi_B)^2)^{3/2} = \Theta(1).
\end{aligned}$$

Let γ be a constant satisfying the following

$$\gamma \leq \min \left\{ \sqrt{\frac{(a_1^+ - a_2^+)\psi}{\tau(s + s\psi)}}, \sqrt{\frac{a_1^+ \psi}{\tau(s + s\psi)}}, \sqrt{\frac{a_1^+ \psi_B}{\tau(h + s\psi_B)}}, \sqrt{\frac{a_1^+ - a_2^+}{\tau s}} \right\}.$$

Note that $a_1^+ - a_2^+ > 0$ because $\mu^2 + 1 > \phi_1^2$. Additionally, we define the following shorthand

$$\begin{aligned}
\epsilon :=& 4\eta\tau\gamma^2 s, \\
\epsilon_B :=& 4\eta\tau\gamma^2 h \\
\alpha :=& 1 + 4\eta a_1^+ - \epsilon \\
\hat{\alpha} :=& 1 + 4\eta a_1^+ + \epsilon \\
\kappa_2 :=& \frac{1 + 4\eta a_2^+}{\alpha} < 1 \text{ because } \mu^2 + 1 > \phi_1^2 \\
\kappa_3 :=& \frac{1}{\alpha} \\
\kappa_\perp :=& \frac{1}{\alpha} \\
\kappa_B :=& \frac{1}{\alpha}.
\end{aligned}$$

Now we are ready to prove the following Lemma.

Lemma G.7. *If $\forall t \leq T$, $\Gamma_1(t) \leq \gamma$. For any constants ψ, ψ_B, the following holds $\forall t \leq T+1$ with probability $1 - O(\frac{1}{\text{poly}(p)})$,*

- $\Gamma_1(t) \geq \alpha^t \Gamma_1(0)$
- $\Gamma_1(t) \leq \hat{\alpha}^t \Gamma_1(0)$.
- $\Gamma_i(t) \leq (\kappa_i^t + \psi)\Gamma_1(t), \;\; i = 2, 3.$
- $\Gamma_\perp(t) \leq (\kappa_\perp^t + \psi)\Gamma_1(t)$.

Proof. Let $S(k)$ be the following statement: $\forall t'$ such that $0 \leq t' \leq k$, the following holds

- $\Gamma_1(t') \geq \alpha^{t'} \Gamma_1(0)$,
- $\Gamma_1(t) \leq \hat{\alpha}^t \Gamma_1(0)$,
- $\Gamma_i(t') \leq (\kappa_i^{t'} + \psi)\Gamma_1(t'), \;\; i = 2, 3,$
- $\Gamma_\perp(t') \leq (\kappa_\perp^{t'} + \psi)\Gamma_1(t')$,
- $\Gamma_B(t') \leq (\kappa_B^{t'} \pi + \psi_B)\Gamma_1(t')$.

By Lemma G.6, $S(0)$ holds with high probability. Next we show that, $\forall t \in [0, T+1]$, if $S(t-1)$ holds then $S(t)$ also holds. By Lemma G.5, the induction hypothesis and $\kappa_2, \kappa_3, \kappa_\perp, \kappa_B < 1$, $\Gamma_1(t-1) \leq \gamma$, we have the following

$$\Gamma_1(t) \geq \alpha \Gamma_1(t-1) \tag{37}$$

$$\Gamma_1(t) \leq \hat{\alpha} \Gamma_1(t-1) \tag{38}$$

$$\begin{aligned}
\Gamma_2(t) \leq& \left((1 + 4\eta a_2^+)(\kappa_2^t + \psi) + \epsilon\right) \Gamma_1(t) \\
\Gamma_3(t) \leq& \left((\kappa_3^t + \psi) + \epsilon\right) \Gamma_1(t) \\
\Gamma_\perp(t) \leq& \left((\kappa_\perp^t + \psi) + \epsilon\right) \Gamma_1(t) \\
\Gamma_B(t) \leq& \left((\kappa_B^t \pi + \psi_B) + \epsilon_B\right) \Gamma_1(t).
\end{aligned}$$

By the construction of our κ's, α's, ϵ's and ψ's, the last three items in statement $S(t)$ hold. Combining the induction hypothesis with equations 37 and 38 yields the first two items in $S(t)$, which completes the proof. □

Now we are ready to prove the theorem.

Theorem G.8. *If $\sigma_0\sqrt{\frac{p}{d}} = o(1)$ and $\sigma_\xi = o(1)$, with probability at least $1 - O(\frac{m^2n^2}{d} + \frac{1}{\text{poly}(p)}) = 1 - o(1)$, the following holds*

- $\|\boldsymbol{W}_0\boldsymbol{v}_2\| = o(1)$.
- $\exists t = O(\ln(\frac{1}{\sigma_0}\sqrt{\frac{d}{p}}))$, *s.t.* $\|\boldsymbol{W}_t\boldsymbol{v}_2\| = \Omega(1)$.

Proof. $\|\boldsymbol{W}_0\boldsymbol{v}_3\| = o(1)$ follows Lemma G.4 and the assumption that $\sigma_0\sqrt{\frac{p}{d}} = o(1)$. Select a constant ψ such that $\psi < \frac{\mu}{\sqrt{1+\frac{\sigma_\xi^2}{mn}}}$. Note that $\frac{\mu}{\sqrt{1+\frac{\sigma_\xi^2}{mn}}} - \psi = \Theta(1)$. Let $T = \lfloor \frac{\ln(\gamma/\Gamma_1(0))}{\ln\alpha} \rfloor = \Theta(\ln\frac{1}{\sigma_0}\sqrt{\frac{d}{p}})$. There are two cases to consider.

- If $\forall t \le T, \Gamma_1(t) \le \gamma$, by Lemma G.7 we have $\Gamma_1(T+1) \ge \gamma$ and $\Gamma_\perp(T+1) \le (o(1)+\psi)\Gamma_1(T+1)$. Then

$$\begin{aligned}\|\boldsymbol{W}_{T+1}\boldsymbol{v}_2\| \ge& \frac{\mu}{\sqrt{1+\mu^2+\frac{\sigma_\xi^2}{mn}}}\|\boldsymbol{W}_{T+1}\boldsymbol{l}_1^+\| - \frac{\sqrt{1+\frac{\sigma_\xi^2}{mn}}}{\sqrt{1+\mu^2+\frac{\sigma_\xi^2}{mn}}}\|\boldsymbol{W}_{T+1}\boldsymbol{l}_\perp\| \\ \ge& (\frac{\mu}{\sqrt{1+\mu^2+\frac{\sigma_\xi^2}{mn}}} - \frac{\sqrt{1+\frac{\sigma_\xi^2}{mn}}}{\sqrt{1+\mu^2+\frac{\sigma_\xi^2}{mn}}}\psi - o(1))\gamma \\ =& \Omega(1).\end{aligned}$$

- If $\exists t \le T\ s.t.\ \Gamma_1(t) > \gamma$, we define $T^* = \frac{\ln(\frac{\gamma}{\Gamma_1(0)})}{\ln\hat{\alpha}}$ and $t^* = \min t\ s.t.\ \Gamma_1(t) > \gamma$. It follows that $\forall t \le t^*-1, \Gamma_1(t) \le \gamma$. Then we can apply Lemma G.7 to obtain $\Gamma_1(t^*) \le \hat{\alpha}^{t^*}\Gamma_1(0)$. If $t < T^*$, the above yields $\Gamma_1(t^*) < \gamma$, which contradicts the definition of t^*. Therefore we conclude $t^* \ge T^*$. Lemma G.7 also tells that $\Gamma_\perp(t^*) \le (\kappa_\perp^{t^*} + \psi)\Gamma_1(t^*)$. Since $t^* \ge T^*$ and $\kappa_\perp < 1$, we have $\kappa_\perp^{t^*} \le (\frac{\Gamma_1(0)}{\gamma})^{\frac{\ln(1/\kappa_3)}{\ln\hat{\alpha}}} = o(1)$. Therefore $\Gamma_\perp(t^*) \le (o(1)+\psi)\Gamma_1(t^*)$. By the definition of t^*, $\Gamma_1(t^*) > \gamma$. Then we can lower bound $\|\boldsymbol{W}_{t^*}\boldsymbol{v}_2\|$ in the same way as in the previous case

$$\begin{aligned}\|\boldsymbol{W}_{t^*}\boldsymbol{v}_2\| \ge& \frac{\mu}{\sqrt{1+\mu^2+\frac{\sigma_\xi^2}{mn}}}\|\boldsymbol{W}_{t^*}\boldsymbol{l}_1^+\| - \frac{\sqrt{1+\frac{\sigma_\xi^2}{mn}}}{\sqrt{1+\mu^2+\frac{\sigma_\xi^2}{mn}}}\|\boldsymbol{W}_{t^*}\boldsymbol{l}_\perp\| \\ \ge& (\frac{\mu}{\sqrt{1+\mu^2+\frac{\sigma_\xi^2}{mn}}} - \frac{\sqrt{1+\frac{\sigma_\xi^2}{mn}}}{\sqrt{1+\mu^2+\frac{\sigma_\xi^2}{mn}}}\psi - o(1))\gamma \\ =& \Omega(1).\end{aligned}$$

□

Table 6. increasing k improves both subclass and class accuracies on CIFAR-10 RandBit.

k	Sub Acc	Acc
1	34.38	86.73
16	58.12	94.09

H. Experimental Setup and Additional Experimental Results

H.1. Datasets

CIFAR-10/100. The two datasets each consist of 60000 32x32 colour images (Krizhevsky et al., 2009). In the case of CIFAR-10, the 'classes' refer to the original 10 classes defined in the dataset, while we define 'subclasses' as two subclasses: vehicles (airplane, automobile, ship, truck) and animals (bird, cat, deer, dog, frog, horse). On CIFAR-100, we refer to the 10 super-classes (e.g. aquatic mammals, fish, flowers) as our 'classes' and the 100 classes as our 'sub-classes'. These two datasets illustrate a natural setting where class collapse is extremely harmful, as it results in learning representations that do not capture much of the semantically relevant information from the data.

MNIST RandBit. The MNIST RandBit dataset Chen et al. (2021) is created by setting n, the # of bits that specifies how easy the useless feature will be. Larger n makes the feature more discriminative, thus 'easier' and more problematic for feature suppression. An extra channel is concatenated to MNIST images where each value in the feature map corresponds to a random integer between 0 and 2^n.

CIFAR-10/100 RandBit. The two datasets are constructed in a similar way as MNIST RandBit, but with images from CIFAR-10/100.

H.2. Training details

For the experiments on CIFAR-10/100 or CIFAR-100 RandBit, we use a ResNet-18 trained with (Momentum) SGD using learning rate = 0.01 and momentum = 0.9. We train with batch size set to 512 for 1000 epochs. For data augmentations, we consider the standard data augmentations from Chen et al. (2020).

For the feature suppression experiments on MNIST RandBit, we directly use the code provided by Chen et al. (2021). We consider a 5-Layer convolutional network. For our data augmentations, we consider the standard set of data augmentations for images and do not alter the useless feature (extra channel concatenated of RandBits).

H.3. Details and additional experiments on varying embedding size

In the experiments presented in Table 2, we vary the width, denoted by w, of the ResNet, which is controlled by the number of convolutional layer filters. For width w, there are w, $2w$, $4w$, $8w$ filters in each layer of the four ResNet blocks.

In addition, we explore an alternative way of varying the embedding size, which isolates the effect of the last layer's embedding size from the size of the lower layers. Specifically, we set the width parameter $w = 4$ and multiply the width of only the last ResNet block by a factor k. It is worth noting that doing this requires a much smaller total number of parameters. Table 6 presents the results on CIFAR-10 RandBit. We observe that increasing k also effectively improves the accuracy. Although the improvement is not as substantial as in the previous case where we increase w, it confirms the same trend predicted by the theory, supporting the conclusion that increasing the embedding size alleviates feature suppression.

I. Potential Approaches to Theoretical Characterization of Class Collapse in (S)GD

The most crucial aspect that remains to be tackled is how (S)GD unlearns subclass features that have already been learned early in training. We offer two potential approaches that could help in achieving this goal.

I.1. Through implicit bias of (S)GD in matrix factorization

The contrastive loss we are considering can be reformulated as a matrix factorization objective:

$$\min f(\boldsymbol{W}^\top \boldsymbol{W}) = \frac{1}{n^2} \sum_{i,j} (\boldsymbol{x}_i^\top \boldsymbol{W}^\top \boldsymbol{W} \boldsymbol{x}_j - a_{ij})^2, \tag{39}$$

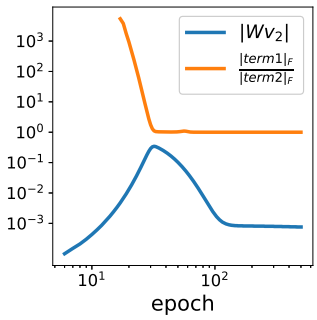


Figure 6. We plot the ratio between norms of term 1 and term 2 in orange, and the projection of the weights $\boldsymbol{W}$ onto the subclass feature ($\|\boldsymbol{W}\boldsymbol{v}_2\|$) in blue. The ratio between the two norms initially starts at a very large value, then decreases until it reaches a plateau around 1. The point at which the ratio dropped to around 1 coincided almost precisely with the peak of $\|\boldsymbol{W}\boldsymbol{v}_2\|$.

where

$$a_{ij} := \begin{cases} 2, & if \;\; x_i \text{ and } x_j \text{ are from the same class} \\ 0. & else \end{cases}$$

This opens up the possibility of leveraging the rich literature on matrix factorization to find a solution. Since we have already proven in Theorems 4.4 and 4.7 that the minimum norm minimizer of the loss function exhibits class collapse, and our experiments confirm that (S)GD does converge to a minimizer that exhibits class collapse, it is reasonable to investigate whether the implicit bias of (S)GD in our setting, specifically matrix factorization, is to seek the minimum norm solution.

We note that understanding the implicit bias in matrix factorization is a longstanding pursuit in the machine learning community. (Gunasekar et al., 2017) have provided empirical and theoretical evidence that under certain conditions, gradient descent converges to the minimum nuclear norm solution. Therefore, one can examine whether similar existing results can be applied to our setting and then combine that with our Theorems 4.4 or 4.7 to show class collapse in GD. However, (Arora et al., 2019a) and (Razin & Cohen, 2020) suggested that the implicit bias may be explained by rank rather than norm when the depth of a network ≥ 2.

I.2. Through analyzing the two terms in the gradient

Let's take a closer look at the update of the weights in GD, i.e., learning rate times minus gradient $4\boldsymbol{W}\boldsymbol{M}^+ - 4\boldsymbol{W}\boldsymbol{M}\boldsymbol{W}^\top\boldsymbol{W}\boldsymbol{M}$ (see Equation). There are two terms $4\boldsymbol{W}\boldsymbol{M}^+$ and $-4\boldsymbol{W}\boldsymbol{M}\boldsymbol{W}^\top\boldsymbol{W}\boldsymbol{M}$ which play different roles in the high level. Here $\boldsymbol{M}^+$ is the covariance of class centers, and $\boldsymbol{M}$ is the covariance of all training examples, as defined in Definition A.1.

Term 1 ($4\boldsymbol{W}\boldsymbol{M}^+$): The first term aligns the weights with $\boldsymbol{M}^+$ which has alignment with the subclass feature. This aligning effect of term 1 has already been theoretically characterized in our proof (Appendix G) for Theorem 4.5.

Term 2 ($-4\boldsymbol{W}\boldsymbol{M}\boldsymbol{W}^\top\boldsymbol{W}\boldsymbol{M}$): Although the effect of the second term is not entirely straightforward, we can gain some intuition by considering the simplest case where the embedding is one-dimensional. In this case, the second term takes the form of a negative scalar times WM, which can be seen as trying to 'discourage' alignment with $\boldsymbol{M}$, the covariance of all training examples.

In our numerical experiment, we observe that the ratio between norms of term 1 and term 2 initially starts at a very large value, then decreases until it reaches a plateau around 1, as shown in Figure 6. Interestingly, the point at which the ratio dropped to around 1 coincided almost precisely with the peak of the projection of the weights $\boldsymbol{W}$ onto the subclass feature. This leads us to the following intuition, which may serve as a proof sketch for showing class collapse at the end of training. In the following, we first describe what happens in the early phase (which we have already proven in the paper), then outline the high-level idea of how subclasses are eventually unlearned.

Phase I where the model learns the subclass feature: We have already proved this part in Appendix G. In summary, the intuition is that early in training, the scale of term 1 dominates over term 2, aligning the model with $\boldsymbol{M}^+$, which in turn aligns with the subclass feature. Therefore, the model learns the subclass feature during this phase.

Phase II where the model unlearns the subclass feature but the class feature remains: Note that the scale of the second term also increases during Phase I as $\boldsymbol{M}$ and $\boldsymbol{M}^+$ share certain components. Once the scale of term 2 reaches that of term 1, the effect of term 2 becomes more pronounced and Phase II begins. Since $\boldsymbol{M}$ exhibits a stronger correlation with the subclass feature than $\boldsymbol{M}^+$ does, the overall effect of the sum of term 1 and term 2 is to reduce alignment with the subclass feature. Thus, over time, the model unlearns the subclass feature. In contrast, for the class feature, $\boldsymbol{M}^+$ has a stronger correlation, causing GD to continue aligning the model with the class feature.