Rethinking Weak Supervision in Helping Contrastive Learning

Jingyi Cui [1 *] **Weiran Huang** [2 3 *] **Yifei Wang** [4 *] **Yisen Wang** [1 5]

Abstract

Contrastive learning has shown outstanding performances in both supervised and unsupervised learning, and has recently been introduced to solve weakly supervised learning problems such as semi-supervised learning and noisy label learning. Despite the empirical evidence showing that semi-supervised labels improve the representations of contrastive learning, it remains unknown if noisy supervised information can be directly used in training instead of after manual denoising. Therefore, to explore the mechanical differences between semi-supervised and noisy-labeled information in helping contrastive learning, we establish a unified theoretical framework of contrastive learning under weak supervision. Specifically, we investigate the most intuitive paradigm of jointly training supervised and unsupervised contrastive losses. By translating the weakly supervised information into a similarity graph under the framework of spectral clustering based on the posterior probability of weak labels, we establish the downstream classification error bound. We prove that semi-supervised labels improve the downstream error bound whereas noisy labels have limited effects under such a paradigm. Our theoretical findings here provide new insights for the community to rethink the role of weak supervision in helping contrastive learning.

1. Introduction

Contrastive learning has shown state-of-the-art empirical performances in unsupervised representation learning (Chen et al., 2020; He et al., 2020; Chen & He, 2021; Wang et al., 2021a). It learns good representations of high-dimensional observations from a large amount of unlabeled data, by pulling together an anchor and its augmented views in the embedding space. On the other hand, supervised contrastive learning (Khosla et al., 2020) uses same-class examples and their corresponding augmentations as positive labels, and achieves significantly better performance than both the unsupervised contrastive learning and the state-of-the-art classification losses, e.g. cross entropy loss. While accurate supervised signals are not always available, there is a lot of weak supervision accompanying the data. This makes us wonder: *can weak supervision help contrastive learning?*

There are two major types of weak supervision. The first is semi-supervised information, where the supervised labels are only available on a small fraction of samples. The second is noisy-labeled information, where the labels are available but unreliable, i.e. the labels can possibly be wrong. Empirical evidence has shown that semi-supervised information can directly be used in positive sample selection to help improve the representations of contrastive learning by jointly training the supervised and unsupervised contrastive losses (Assran et al., 2020; Acharya et al., 2022). By contrast, for noisy-labeled information, most methodological studies use contrastive learning as a tool to select confident samples based on the learned representations (Yao et al., 2021; Ortego et al., 2021; Li et al., 2022; Zhang et al., 2022a), whereas none of the existing literature demonstrates if noisy label information can help improve the representations of contrastive learning. Therefore, we are wondering *if the conclusion on noisy-labeled information is the same as the semi-supervised information. If not, what are the differences between semi-supervised and noisy-labeled information in helping contrastive learning?*

In this paper, we show an awkward fact that noisy labels have a limited effect on representations of contrastive learning under the joint training paradigm, and the winner of supervised and unsupervised contrastive learning itself serves as an embarrassingly strong baseline. As a preview of results, in Figure 1, we plot the result of unsupervised contrastive learning trained without labels (SimCLR (Chen et al., 2020)), supervised contrastive learning trained with noisy labels (SupCon (Khosla et al., 2020)), the winner of SupCon and SimCLR (Max), and joint training of SimCLR

[*]Equal contribution [1]National Key Lab of General Artificial Intelligence, School of Intelligence Science and Technology, Peking University [2]Qing Yuan Research Institute, Shanghai Jiao Tong University [3]Huawei Noah's Ark Lab [4]School of Mathematical Sciences, Peking University [5]Institute for Artificial Intelligence, Peking University. Correspondence to: Yisen Wang <yisen.wang@pku.edu.cn>.

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

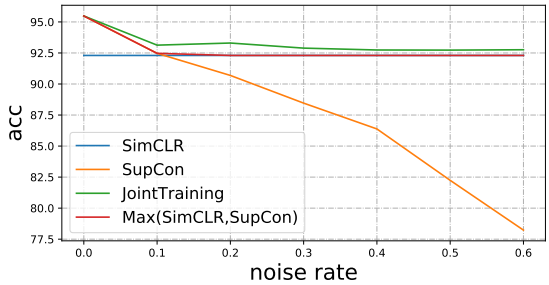


Figure 1: The winner of SupCon and SimCLR serves as a strong baseline of joint training under label noise on the CIFAR-10 dataset.

and SupCon under label noise (JointTraining), respectively. We conduct experiments under symmetric label noise with noise rates ranging from 0% to 60%, stepped by 10%. We evaluate the trained representations by linear probing on the clean testing data. The weight parameter θ of JointTraining is tuned in $\{0, 0.05, 0.1, 0.2, 0.4, 0.6, 0.8, 1.0\}$, and the best linear probing accuracy under the optimal θ is reported. We show that supervised training outperforms unsupervised training when the noise rate is low, and the situation reverses when the noise rate is high. Nonetheless, the joint training of SimCLR and SupCon (with finely tuned weights) has only limited advantage over the winner of SupCon and SimCLR. Moreover, its performance coincides with that of SupCon when noise rate is 0, and converges to that of SimCLR when noise rate increases. This indicates that noisy labels provide limited help to contrastive representation learning under the paradigm of joint training, and that the winner of SupCon and SimCLR itself serves as a strong baseline.

To explain the above phenomena in depth and exploit the mechanical differences between semi-supervised and noisy-labeled information in helping contrastive learning, we establish a unified theoretical framework of weakly supervised contrastive learning, applying to both semi-supervised and noisy-labeled settings. We investigate the joint training of supervised and unsupervised contrastive losses, and take spectral contrastive learning (HaoChen et al., 2021) as a performance proxy for (standard) contrastive learning to conduct the theoretical analysis. Based on the posterior probability of labeled samples, we translate the weakly supervised information (under symmetric label noise assumption) into a similarity graph under the framework of spectral clustering. This enables us to analyze the effect of the label information on the augmentation graph, and consequently on the derived error bound. Accordingly, we prove that under the semi-supervised setting, the label information helps improve the downstream error bound, whereas under the noisy-labeled setting, the joint training is no better than the winner of supervised and unsupervised contrastive learning in terms of an error bound.

The contributions of this paper are summarized as follows.

- We for the first time establish a theoretical framework for contrastive learning under weak supervision, including noisy label learning and semi-supervised learning.
- By formulating the label information into a similarity graph based on the posterior probability of labels, we derive the downstream error bound of jointly trained contrastive learning losses. We prove that semi-supervised labels improve the downstream error bound compared with unsupervised learning, whereas under the noisy-labeled setting, joint training fails to improve the error bound compared with the winner of supervised and unsupervised contrastive learning.
- We empirically verify that noisy labels have only limited help to contrastive representation learning under the paradigm of joint training. Thus, complex designs such as label denoising are still required for leveraging noisy labeled information in improving supervised contrastive learning.

2. Related Works

Theoretical Frameworks of Contrastive Learning. The theoretical frameworks of unsupervised contrastive learning can be divided into two major categories. The first category is devoted to building the relationship between unsupervised contrastive learning and supervised downstream classification. Arora et al. (2019) first introduce the concept of latent classes, hypothesizes that semantically similar points are sampled from the same latent class, and proves that the unsupervised contrastive loss serves as an upper bound of downstream supervised learning loss. Nozawa & Sato (2021); Ash et al. (2022); Bao et al. (2022) further investigate the effect of negative samples, and establish surrogate bounds for the downstream classification loss that better match the empirical observations on the negative sample size. However, studies in this category have to assume the existence of supervised latent classes, and that the positive pairs are conditionally independently drawn from the same latent class. This assumption fails to distinguish between supervised and unsupervised contrastive learning, and therefore cannot be used to analyze the weakly supervised setting.

Another major approach is to analyze contrastive learning by modeling the feature similarity. HaoChen et al. (2021) first introduce the concept of the *augmentation graph* to represent the feature similarity of the augmented samples, and analyzes contrastive learning from the perspective of spectral clustering. Shen et al. (2022) use a stochastic block model to analyze spectral contrastive learning for the problem of unsupervised domain adaption. Similarly, Wang et al. (2021b) propose the concept of *augmentation overlap* to

formulate how the positive samples are aligned. Moreover, contrastive learning is also understood through other existing theoretical frameworks of unsupervised learning, such as nonlinear independent component analysis (Zimmermann et al., 2021), neighborhood component analysis (Ko et al., 2022), variational autoencoder (Aitchison, 2021), stochastic neighbor embedding (Hu et al., 2023), geometric analysis of embedding space (Huang et al., 2023), and message passing (Wang et al., 2023).

In this paper, we follow the second category of contrastive learning approaches, and formulate the weakly supervised information into a similarity graph based on both label and feature information.

Contrastive Learning for Noisy Label Learning. Ghosh & Lan (2021) first find that pretraining with contrastive learning improves robustness to label noise through empirical evidence. Many methodological studies are carried out for noisy label learning with the help of contrastive learning. Yao et al. (2021); Ortego et al. (2021); Li et al. (2022) use representations learned from unsupervised contrastive learning to filter out confident samples from all noisy ones, and in turn use the confident samples to conduct supervised contrastive learning to generate better representations. Navaneet et al. (2022) introduce additional semantically similar supervision to contrastive representation learning by incorporating nearest neighbors under certain constraints as additional positive samples, which also adapts to noisy label learning. By contrast, Yan et al. (2022) follow the idea of negative learning (Kim et al., 2019; 2021b), and leverages the negative correlations from the noisy data to avoid same-class negatives in contrastive learning. Chuang et al. (2022) propose a robust contrastive loss function inspired by the symmetric losses that are proved to be noise tolerant. Very recently, Zhang et al. (2022a) use contrastive learning to handle noisy labels of long-tailed data. For theoretical studies, Cheng et al. (2021) analyze the robustness of cross-entropy with SSL features, and Xue et al. (2022) prove the robustness of downstream classifier in contrastive learning.

Contrastive Learning for Semi-supervised Learning. Lee et al. (2022); Yang et al. (2022) use contrastive regularization to enhance the reliability of pseudo-labeling in semi-supervised learning. Kim et al. (2021a) introduce a semi-supervised learning method that combines self-supervised contrastive pre-training and semi-supervised fine-tuning based on augmentation consistency regularization. Zhang et al. (2022b) use contrastive loss to model pairwise similarities among samples, generates pseudo labels from the cross entropy loss, and in turn calibrates the prediction distribution of the two branches.

To conclude, the existing studies of contrastive learning under weak supervision mainly focus on using contrastive learning as a tool to improve the weakly supervised learning performance, whereas to the best of our knowledge, none of the previous works reveals how weak supervision helps contrastive learning. To fill in the blank, in this paper, we establish a theoretical framework for contrastive learning under weak supervision, and show the effects of semi-supervised and noisy-labeled information on the error bounds of contrastive learning.

3. Preliminaries

Notations. Suppose that random variables $\bar{X} \in \bar{\mathcal{X}} := \mathbb{R}^d$, and $Y \in [r] := \{1, \dots, r\}$. Let the input natural data $\{(\bar{x}_i, y_i)\}_{i\in[N]}$ be i.i.d. sampled from the joint distribution $\mathrm{P}(\bar{X}, Y)$. Given a natural data $\bar{x} \in \bar{\mathcal{X}}$, we use $\mathcal{A}(\cdot|\bar{x})$ to denote the distribution of its augmentations and use $\mathcal{X}$ to denote the set of all augmented data, which is assumed to be finite but exponentially large. Denote $n = |\mathcal{X}|$.

3.1. Spectral Contrastive Learning

In HaoChen et al. (2021), an augmentation graph $\mathcal{G}$ is used to describe the distribution of augmented samples, where the edge weight $w_{xx'} := \mathbb{E}_{\bar{x}\sim\bar{\mathcal{P}}}[\mathcal{A}(x|\bar{x})\mathcal{A}(x'|\bar{x})]$ denotes the marginal probability of generating augmented views x and x' from the same natural data. Due to the total probability mass, $\sum_{x,x'\in\mathcal{X}} w_{xx'} = 1$. The adjacent matrix of the augmentation graph is denoted as $\boldsymbol{A} := (w_{xx'})_{x,x'\in\mathcal{X}} \in \mathbb{R}^{n\times n}$, and the normalized adjacent matrix is denoted as $\bar{\boldsymbol{A}} := D^{-1/2}\boldsymbol{A}D^{-1/2}$, where $D := \mathrm{diag}(w_x)_{x\in\mathcal{X}}$, and $w_x := \sum_{x'\in\mathcal{X}} w_{xx'}$.

In this paper, we consider the spectral contrastive loss $\mathcal{L}(f)$ proposed by HaoChen et al. (2021), that is, for an embedding function $f : \mathcal{X} \to \mathbb{R}^k$,

$$-2 \cdot \mathbb{E}_{x,x^+}[f(x)^\top f(x^+)] + \mathbb{E}_{x,x'}\left[\left(f(x)^\top f(x')\right)^2\right]. \quad (1)$$

Spectral contrastive loss is proved to be equivalent to the matrix factorization loss, i.e. for $F \in \mathbb{R}^{n\times k} := (u_x)_{x\in\mathcal{X}}$, $u_x := w_x^{1/2} f(x)$,

$$\mathcal{L}_{\mathrm{mf}}(F) := \|\bar{\boldsymbol{A}} - FF^\top\|_F^2 = \mathcal{L}(f) + const. \quad (2)$$

3.2. Noisy Label Learning

Recall that we denote the true label of a given instance $x \in \mathcal{X}$ is y. One common assumption of the generation procedure of label noise is as follows. Given the true labels, the noisy label is randomly flipped to another label $\tilde{y}$ with some probability. In this paper, we take the widely adopted symmetric label noise assumption as an example.

For notational simplicity, we write the symmetric label noise assumption in matrix form. Denote $\boldsymbol{Y} := (\eta_j(x_i))_{i\in[n],j\in[r]}$, $\eta_j(x) = \mathrm{P}(Y = j|x)$, as the posterior

probability matrix of the clean label distribution, and denote $\tilde{\boldsymbol{Y}} := (\tilde{\eta}_j(x_i))_{i\in[n],j\in[r]}$, $\tilde{\eta}_j(x) = \mathrm{P}(\tilde{Y} = j|x)$, as the noisy label distribution. In Assumption 3.1, we assume that the flipping probability is conditional independent of the input data, and that the flipping probability to all other classes is uniformly at random.

Assumption 3.1. For symmetric label noise with noise rate $\gamma \in (0, 1)$, we denote the transition matrix $\boldsymbol{T} = (t_{i,j})_{i\in[r],j\in[r]}$, where

$$t_{i,i} = 1 - \gamma, \text{ and } t_{i,j} = \frac{\gamma}{r-1}, \text{ for } j \neq i. \tag{3}$$

Then the noisy label posterior distribution is assumed to be

$$\tilde{\boldsymbol{Y}} = \boldsymbol{Y}\boldsymbol{T}. \tag{4}$$

Under Assumption 3.1, $\boldsymbol{T}$ is symmetric. Specifically, when $\gamma = 0$, $\boldsymbol{T}$ degenerates to the identity matrix $\boldsymbol{I}_{r\times r}$. Moreover, to guarantee PAC-learnability, we usually assume the true label is the dominating class, i.e. $\gamma < \frac{r-1}{r}$.

3.3. Semi-Supervised Learning

For $j \in [r]$, let n_j be the number of labeled samples of Class j. Let $n_L = \sum_{j\in[r]} n_{L,j}$ be the number of all labeled samples, and n_U be the number of unlabeled samples. Obviously, we have $n_L + n_U = n$. Usually, the number of labeled samples is much smaller than that of the unlabeled ones because human annotation is costly and labor-intensive. That is, we can naturally assume $n_L \ll n_U$.

In the following parts of the paper, we analyze the settings of noisy label learning and semi-supervised learning in a unified framework. Without loss of generality, we assume $(x_1, \ldots, x_{n_L})$ is labeled with noise rate $\gamma \in [0, \frac{r-1}{r})$, and denote the corresponding clean and noisy posterior probability matrices as $\boldsymbol{Y}_L$ and $\tilde{\boldsymbol{Y}}_L$, respectively. Then we have $\tilde{\boldsymbol{Y}}_L = \boldsymbol{Y}_L\boldsymbol{T}$. Specifically, when $\gamma = 0$, our analyzing framework degenerates to the standard setting of semi-supervised learning, and when $n_L = n$, our analyzing framework reduces to the standard noisy label learning.

4. Mathematical Formulations

We mention that our formulation of "similarity graph" is not a distributional assumption on the underlying similarity among data, but to formulate a possible probability of drawing positive samples in contrastive learning that takes both label and feature information into consideration. Specifically, in Sections 4.1 and 4.2, we only discuss the similarity graph induced by the weakly supervised labels and neglected feature similarity. Note that in Section 4.2 we investigate the setting of semi-supervised noisy labels, so as to include semi-supervised learning and noisy label learning in a unified framework. Then in Section 4.3, we take both label and feature similarity into consideration through the convex combination to describe the joint training loss.

4.1. Similarity Graph Describing Noisy Labels

To leverage the labeled information in the form of a similarity graph, we first consider a simple example where noise rate $\gamma = 0$ and the label distribution is deterministic, i.e. for a sample x with true label y, the posterior probability $\eta_y(x) = 1$ and $\eta_j(x) = 0$ for $j \neq y$. In this case, we can naturally assume that in the label similarity graph, the intra-class vertices are fully connected and the inter-class vertices are disconnected. That is, $w_{xx'} = 1$ if x and x' has the same label and otherwise $w_{xx'} = 0$.

Then we consider the more general stochastic label scenario. Recall that for unsupervised spectral contrastive learning, the edge weight $w_{xx'}$ in an augmentation graph $\mathcal{G}$ describes the marginal probability of generating x and x' from the same natural data. That is, $w_{xx'}$ describes the joint probability of a pair of positive samples. Similarly, since the positive samples for supervised contrastive learning (Khosla et al., 2020) are selected as all same-class samples, we can naturally define the edge weight $w_{xx'}$ as the probability of two views x and x' generating from the same class, i.e. $w_{xx'} = \sum_{j\in[r]} \eta_j(x)\eta_j(x')$, and therefore $\boldsymbol{A}_L := \boldsymbol{Y}_L\boldsymbol{Y}_L^\top$. Moreover, we denote $\bar{\boldsymbol{A}}$ as the normalized adjacent matrix. For the simplicity of notations, we consider the case where the data is class-balanced, i.e. $n_1 = \ldots = n_r = n_L/r$. Then we have $\bar{\boldsymbol{A}} = \frac{r}{n_L}\boldsymbol{A}$.

Next, we add label noise to our mathematical formulations. To be specific, when performing supervised contrastive learning based on noisy labeled data, we naturally select positive samples as the samples with the same *noisy* labeled data. According to Assumption 3.1, we have $\tilde{\boldsymbol{Y}}_L = \boldsymbol{Y}_L\boldsymbol{T}$, where $\boldsymbol{T}$ is symmetric. Then the adjacent matrix of the similarity graph describing noisy labels is formulated as

$$\begin{aligned}\boldsymbol{A}_L^\star :&= \tilde{\boldsymbol{Y}}_L\tilde{\boldsymbol{Y}}_L^\top = \boldsymbol{Y}_L\boldsymbol{T}(\boldsymbol{Y}_L\boldsymbol{T})^\top = \boldsymbol{Y}_L\boldsymbol{T}\boldsymbol{T}^\top\boldsymbol{Y}_L^\top \\ &= \boldsymbol{Y}_L\boldsymbol{T}^2\boldsymbol{Y}_L^\top.\end{aligned} \tag{5}$$

Similarly, when data is class balanced, we have the normalized adjacent matrix $\bar{\boldsymbol{A}}_L^\star = \frac{n_L}{r}\boldsymbol{A}_L^\star$.

4.2. Similarity Graph Describing Semi-Supervised Noisy Labels

Under the setting of semi-supervised learning, we have no prior knowledge about the label information of the unlabeled samples. Therefore, from the perspective of unsupervised contrastive learning, the unlabeled samples can be viewed as having unique class labels. Therefore, to construct the similarity graph, we attach sample-specific labels to the unlabeled samples. Thus, the posterior probability matrix of unlabeled samples $\boldsymbol{Y}_U$ is an identity matrix $\boldsymbol{I}_{n_U\times n_U}$. Note

that here we only discuss the similarity graph of supervised information, so the feature similarity between samples is not included in the similarity graph.

Combining both labeled and unlabeled samples, the posterior probability matrix of all semi-supervised samples can be denoted as

$$\tilde{\boldsymbol{Y}} = \begin{bmatrix} \tilde{\boldsymbol{Y}}_L & \boldsymbol{0} \\ \boldsymbol{0} & \tilde{\boldsymbol{Y}}_U \end{bmatrix} = \begin{bmatrix} \boldsymbol{Y}_L \boldsymbol{T} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}_{n_U \times n_U} \end{bmatrix}. \tag{6}$$

Therefore, the similarity graph of samples with n_L noisy labels can be denoted as

$$\boldsymbol{A}^\star = \tilde{\boldsymbol{Y}}\tilde{\boldsymbol{Y}}^\top = \begin{bmatrix} \boldsymbol{Y}_L \boldsymbol{T}^2 \boldsymbol{Y}_L^\top & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}_{n_U \times n_U}. \end{bmatrix} \tag{7}$$

In Lemma 4.1 we present the influence of symmetric label noise with noise rate γ on the similarity graph $\boldsymbol{A}^\star$.

Lemma 4.1. *Under Assumption 3.1, if the data is class balanced, i.e.* $n_1 = \ldots = n_r = \frac{n_L}{r}$*, then there holds*

$$\bar{\boldsymbol{A}}^\star = \begin{bmatrix} \alpha(\gamma)\bar{\boldsymbol{A}}_L + \beta(\gamma)\frac{r}{n_L}\vec{1}_{n_L}\vec{1}_{n_L}^\top & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}_{n_U \times n_U} \end{bmatrix}, \tag{8}$$

where $\alpha(\gamma) := \left(1 - \frac{r}{r-1}\gamma\right)^2$ *and* $\beta(\gamma) := \frac{\gamma}{r-1}\left(2 - \frac{r}{r-1}\gamma\right)$.

Note that without label noise, i.e. $\gamma = 0$, we have $\alpha(\gamma) = 1$ and $\beta(\gamma) = 0$. For the sake of simplicity, in the following, we write α and β instead of $\alpha(\gamma)$ and $\beta(\gamma)$ when no ambiguity is aroused.

In Lemma 4.1, we show that the effect of symmetric label noise is to add a uniform weight to the edges between all labeled samples. This uniform weight increases the confusion between intra- and inter-class similarities. For example, under the deterministic label scenario, we have $\boldsymbol{A}_L = \boldsymbol{I}_{n_L \times n_L}$, and under label noise, the original intra-class similarity is uniformly shrunk from 1 to α and the inter-class similarity increases from 0 to β. Moreover, as the noise rate γ increases, α decreases and β increases, which results in severer confusion between the intra- and inter-class similarities. Intuitively, as the similarity graph describes the sampling probability of positive pairs, α and β measure *how noisy that similarity is*. (The sampling probability of negative samples remains unaffected because they are assumed to be uniformly sampled regardless of labels.) Then under label noise, the probability of same-class samples being selected as positive pairs reduces from 1 to $\alpha + \frac{r}{n_L}\beta$, and the probability of different-class samples being selected as positive pairs raises from 0 to $\frac{r}{n_L}\beta$.

Note that our mathematical formulations can also be extended to generalized label noise assumptions of the noise transition matrix $\boldsymbol{T}$ (other than symmetric label noise). Specifically, under generalized assumptions, the term $\frac{r}{n_L}\vec{1}_{n_L}\vec{1}_{n_L}^\top$ in (8) will become a real symmetric matrix which depends on the specific form of the label noise assumption.

4.3. Similarity Graph Describing Joint Training

Recall that we want to investigate the effect of noisy labels in the joint training of supervised and unsupervised contrastive losses. Specifically, for $\theta \in (0, 1)$, we have the loss for joint training as

$$\mathcal{L}_{\text{JointTraining}} := (1-\theta)\mathcal{L}_{\text{unsup}} + \theta\mathcal{L}_{\text{sup}}, \tag{9}$$

where in $\mathcal{L}_{\text{unsup}}$, the positive samples are selected as augmentations of the anchor sample, and in $\mathcal{L}_{\text{sup}}$, the positive samples are selected as (possibly noisy) same-class samples. The specific algorithms of joint training for noisy label learning and semi-supervised learning are shown in Appendix B.1.

According to (2), the spectral contrastive loss is equivalent to the matrix factorization loss. Therefore, we can write (9) as a convex combination of matrix factorization losses with similarity graphs under label and feature information, i.e.

$$(1-\theta)\|\bar{\boldsymbol{A}}_0 - FF^\top\|_F^2 + \theta\|\bar{\boldsymbol{A}}^\star - FF^\top\|_F^2, \tag{10}$$

where we denote $\boldsymbol{A}_0$ as the augmentation graph of arbitrary unlabeled samples describing feature information, and $\boldsymbol{A}^\star$ is denoted in (7) describing the similarity graph induced by the semi-supervised noisy labeled (augmented) samples.

Note that (10) can be rewritten as

$$\|((1-\theta)\bar{\boldsymbol{A}}_0 + \theta\bar{\tilde{\boldsymbol{A}}}) - FF^\top\|_F^2 + c_0(\theta), \tag{11}$$

where $c_0(\theta) := (1-\theta)\|\bar{\boldsymbol{A}}_0\|_F^2 + \theta\|\bar{\tilde{\boldsymbol{A}}}\|_F^2 - \|(1-\theta)\bar{\boldsymbol{A}}_0 + \theta\bar{\tilde{\boldsymbol{A}}}\|_F^2$ is independent of F. As $c_0(\theta)$ does not affect the training procedure, optimizing (10) and (11) results in the same optimal F. Therefore, in the following, to analyze the joint training loss (9), we investigate the properties of the mixed similarity graph

$$\boldsymbol{A}_{\theta,\gamma,n_L} := (1-\theta)\bar{\boldsymbol{A}}_0 + \theta\bar{\boldsymbol{A}}^\star. \tag{12}$$

5. Theoretical Results

In this section, we first compute eigenvalues of the similarity graph induced by both label and feature information, which plays a key role in deriving the error bound of contrastive learning in Section 5.1. Then in Section 5.2, we show that (clean) label information in the semi-supervised setting can help improve the error bound, whereas in Section 5.3, we prove that the joint training of supervised and unsupervised contrastive learning fails to improve the error bound compared with purely supervised or purely unsupervised contrastive learning.

5.1. Eigenvalues of Similarity Graph Describing Joint Training

We first compute the eigenvalues of the similarity graph describing the weak labels (without describing feature information).

Proposition 5.1. *For arbitrary $\boldsymbol{Y}$, assume that the labeled data is class-balanced, i.e. $\sum_{i\in[n_L]}\eta_j(x_i) = n_L/r$ for $j\in[r]$. Assume that the eigenvalues of $\bar{\boldsymbol{A}}_L$ are $\mu_1,\ldots,\mu_n$ (in descending order). Then under Assumption 3.1, the eigenvalues of $\bar{\boldsymbol{A}}^\star$ are*

$$\tilde{\mu}_1 = \ldots = \tilde{\mu}_{n_U+1} = 1, \tag{13}$$

$$\tilde{\mu}_j = \mu_j\alpha, \text{ for } j = n_U+2,\ldots,n, \tag{14}$$

where $\alpha = \mu_j\left(1-\frac{r}{r-1}\gamma\right)^2$.

In Proposition 5.1, we show that the eigenvalues of $\bar{\boldsymbol{A}}^\star$ rely on the eigenvalues of $\bar{\boldsymbol{A}}$ and consequently rely on the posterior probabilities of clean labels. Specifically, if the true label has a higher posterior probability, i.e. $\max_{j\in[r]}\mathrm{P}(Y=j|x)$ is larger, then the eigenvalues of $\bar{\boldsymbol{A}}$ are larger. On the other hand, the existence of label noise uniformly shrinks the eigenvalues of $\bar{\boldsymbol{A}}^\star$ except for the largest ones, and larger noise rate γ results in smaller α and thus leads to smaller eigenvalues of $\bar{\boldsymbol{A}}^\star$. Moreover, the number of largest eigenvalues is decided by the number of unlabeled samples.

Note that $\mathrm{rank}(\boldsymbol{A}^\star)\le\mathrm{rank}(\boldsymbol{Y}_L)+n_U\le n_U+r$, and therefore we have $\tilde{\mu}_{n_U+r+1}=\ldots=\tilde{\mu}_n=0$. Specifically, under the deterministic label scenario, we have $\mu_{n_U+2}=\ldots=\mu_{n_U+r}=1$. Then the eigenvalues of $\bar{\boldsymbol{A}}^\star$ become

$$\tilde{\mu}_1 = \ldots = \tilde{\mu}_{n_U+1} = 1, \tag{15}$$

$$\tilde{\mu}_{n_U+2} = \ldots = \tilde{\mu}_{n_U+r} = \alpha = \left(1-\frac{r}{r-1}\gamma\right)^2, \tag{16}$$

$$\tilde{\mu}_{n_U+r+1} = \ldots = \tilde{\mu}_n = 0. \tag{17}$$

Then in the following proposition, we discuss the eigenvalues of the mixed similarity graph $\boldsymbol{A}_{\theta,\gamma,n_L}$ describing both weak labels and feature information.

Proposition 5.2. *Denote $\lambda_1,\ldots,\lambda_n$ as the eigenvalues of $\boldsymbol{A}_{\theta,\gamma,n_L}$. Then given the eigenvalues of $\bar{\boldsymbol{A}}_0$, i.e. $\nu_1,\ldots,\nu_n$ and the eigenvalues of $\bar{\boldsymbol{A}}_L$, i.e. $\mu_1,\ldots,\mu_{n_L}$(in descending order), under the deterministic scenarios, when $k\le n_U$, there holds*

$$\max\left\{\theta+(1-\theta)\nu_{n_L+k},(1-\theta)\nu_{k+1},\theta\alpha+(1-\theta)\nu_{n_L+k-r+1}\right\} \le \lambda_{k+1}\le\theta+(1-\theta)\nu_{k+1},$$

and for $k\ge n_U+r$,

$$(1-\theta)\nu_{k+1}\le\lambda_{k+1} \le\min\{\theta+(1-\theta)\nu_{k+1},\theta\alpha+(1-\theta)\nu_{k-n_U},(1-\theta)\nu_{k+1-r-n_U}\}.$$

In Proposition 5.2, we derive both the upper and lower bounds for the $k+1$-th largest eigenvalue of $\boldsymbol{A}_{\theta,\gamma,n_L}$. We see that under the deterministic scenario, the upper bound of the $k+1$-th largest eigenvalue of $\boldsymbol{A}_{\theta,\gamma,n_L}$ depends on at most three specific eigenvalues of the unsupervised augmentation graph $\boldsymbol{A}_0$. The value of λ_{k+1} is also affected by the weighting parameter θ. However, the specific dependence relies on the relative magnitudes of ν_{k+1}, ν_{k-n_U}, and ν_{k+1-n_U-r}. A perhaps somehow anti-intuitive conclusion is that when k is smaller than n_U, the upper bound of λ_{k+1} is unaffected by the noise rate. Similarly, we notice that for $k\ge n_U+r$, the lower bound of λ_{k+1} is unaffected by the noise rate.

5.2. Error Bound of Joint Training under Semi-supervised Setting

Recall that the goal of contrastive representation learning is to learn an embedding function $f:\mathcal{X}\to\mathbb{R}^k$. The quality of the learned embedding is often evaluated through linear evaluation. To be specific, denote $B\in\mathbb{R}^{k\times r}$ as the weights of the downstream linear classifier, and the linear predictor is denoted as $g_{f,B}(\bar{x}) = \arg\max_{i\in[r]}\mathrm{P}_{x\sim\mathcal{A}(\cdot|\bar{x})}(g_{f,B}(x)=i)$. In this paper, we focus on analyzing the error bound of the best possible downstream linear classifier $g_{f^*_{\mathrm{pop}},B^*}$, where $f^*_{\mathrm{pop}}\in\arg\min_{f:\mathcal{X}\to\mathbb{R}^k}$ is the minimizer of the population spectral contrastive loss $\mathcal{L}(f)$ defined in (1), and B^* is the optimal weight for the downstream linear classifier.

Following HaoChen et al. (2021), we assume that the labels are recoverable from augmentations, i.e. we assume there exists a classifier g that can predict $y(x)$ given x with error at most $\delta\in(0,1)$.

Assumption 5.3. Assume that for some $\delta_u,\delta_s>0$, there holds

$$\mathbb{E}_{\bar{x}\sim\mathcal{P}_{\bar{\mathcal{X}}},x\sim\mathcal{A}(\cdot|\bar{x})}\mathbf{1}[\hat{y}(x_i)\ne y(\bar{x})]\le\delta_u. \tag{18}$$

and

$$\frac{1}{n_L}\sum_{i\in[n_L]}\sum_{\ell\in[r]}\eta_\ell(x_i)\mathbf{1}[\hat{y}(x_i)\ne\ell]\le\delta_s. \tag{19}$$

Compared with Assumption 3.5 in HaoChen et al. (2021), Assumption 5.3 additionally assumes the recoverable of labels taking expectation under the posterior probability distribution. Intuitively, δ_u represents the error under unsupervised learning, and δ_s represents the error under supervised learning (with clean posterior distributions). Therefore, it is reasonable to assume that $\delta_s\le\delta_u$. Note that Assumption 5.3 is a minor revision of the original assumption. The additional assumption (19) does not change the nature of the original idea of label recovery, and will be used to bound the error term of learning from weakly supervised labels.

Then we derive the error bound of downstream linear evaluation learned by contrastive learning under semi-supervised setting.

Theorem 5.4. *Assume the assumptions in Theorem 5.5 hold. Then if*

$$\|B^*\|_F \leq 1/\max\Big\{(1-\theta)\nu_k, \theta + (1-\theta)\nu_{n_L+k-r}\Big\},$$

we have

$$\mathcal{E} \leq \frac{2\big[2\delta_u + [(1+\rho)\delta_s - 2\delta_u]\theta\big]}{1-\theta-(1-\theta)\nu_{k+1}} + 8\delta_u. \quad (20)$$

By Theorem 5.4, the error bound of linear probing is larger when the label recovery error δ_u and δ_s gets larger.

The error bound in Theorem 5.4 attains the minimum when $\theta = 1$ if $\rho \leq 2\delta_u/\delta_s - 1$. And therefore we have $\mathcal{E} \leq 2(1+\rho)\delta_s + 8\delta_u \leq 2(1+\rho)\delta_u + 8\delta_u$. Since $\nu_{k+1} \in [0,1]$, then when $\rho \leq (1+\nu_{k+1})/(1-\nu_{k+1})$, $2(1+\rho)\delta_u + 8\delta_u \leq \frac{4\delta_u}{1-\nu_{k+1}} + 8\delta_u$. Recall that in HaoChen et al. (2021), the error bound of purely unsupervised contrastive learning is $\frac{4\delta_u}{1-\nu_{k+1}} + 8\delta_u$. Our result indicates that semi-supervised information improves the error bound compared with purely unsupervised contrastive learning since by using all labeled samples. This theoretical point can also be verified by existing experimental research about semi-supervised contrastive learning, e.g. Assran et al. (2020).

5.3. Error Bound of Joint Training under Noisy-labeled Setting

In Theorem 5.5, we present the error bound of joint training contrastive learning under noisy label setting.

Theorem 5.5. *For arbitrary $\boldsymbol{Y}$, assume that the labeled data is class-balanced, i.e. $\sum_{i\in[n_L]} \eta_j(x_i) = n_L/r$ for $j \in [r]$. Denote $\nu_1, \ldots, \nu_n$ as the eigenvalues of $\bar{\boldsymbol{A}}_0$ (in descending order). Denote $\mathcal{E} := \mathrm{P}_{\bar{x}\sim\mathcal{P}_{\bar{\mathcal{X}}}, x\sim\mathcal{A}(\cdot|\bar{x})}\big(g_{f^*_{\mathrm{pop}},B^*}(x) \neq y(\bar{x})\big)$ as the linear evaluation error, where $B^* \in \mathbb{R}^{r\times k}$ with norm $\|B^*\|_F \leq 1/(1-\theta)\nu_k$. Assume there exists $\rho > 0$, such that $w_i/w_j < \rho$, for $i,j \in [n]$, and $k > r$. Then under the deterministic scenario and Assumptions 3.1 and 5.3, there holds*

$$\mathcal{E} \leq \frac{2\big[2\delta_u + [\alpha(1+\rho)\delta_s - 2\delta_u + (1-\alpha)]\theta\big]}{1-\lambda(\boldsymbol{\nu};\theta,\alpha)} + 8\delta_u, \quad (21)$$

where

$$\lambda(\boldsymbol{\nu};\theta,\alpha) = \min\{\theta + (1-\theta)\nu_{k+1}, \theta\alpha + (1-\theta)\nu_k, (1-\theta)\nu_{k+1-r}\}. \quad (22)$$

and $\alpha := \big(1 - \frac{r}{r-1}\gamma\big)^2$.

The bound in Theorem 5.5 gets larger when the noise rate γ and the label recovery error δ_u and δ_s gets larger. It is worth noting that Theorem 5.5 shows that joint training fails to improve the error bound. Specifically, since the numerator is positive, this bound is equivalent to the minimum of the following three terms

$$\frac{2\big[2\delta_u + [\alpha(1+\rho)\delta_s - 2\delta_u + (1-\alpha)]\theta\big]}{1-\theta-(1-\theta)\nu_{k+1}} + 8\delta_u, \quad (23)$$

$$\frac{2\big[2\delta_u + [\alpha(1+\rho)\delta_s - 2\delta_u + (1-\alpha)]\theta\big]}{1-\theta\alpha-(1-\theta)\nu_{k}} + 8\delta_u, \quad (24)$$

and

$$\frac{2\big[2\delta_u + [\alpha(1+\rho)\delta_s - 2\delta_u + (1-\alpha)]\theta\big]}{1-(1-\theta)\nu_{k+1-r}} + 8\delta_u. \quad (25)$$

Since each of the three terms is a monotonic function with respect to θ, the minimum must be attained at the end of the range of θ, i.e. $\theta = 0$ or $\theta = 1$. Therefore, the minimum of $\mathcal{E}$ must be attained at $\theta = 0$ or $\theta = 1$, too. Specifically, when the noise rate is relatively high, i.e. $\gamma > \gamma_{\text{threshold}} := \frac{r-1}{r}\Big(1 - \sqrt{\frac{1-\nu_{k+1}-2\delta_u}{(1-(1+\rho)\delta_s)(1-\nu_{k+1})}}\Big)$, $\mathcal{E}$ achieves its minimum at $\theta = 0$, whereas when the noise rate is relatively low, i.e. $\gamma < \gamma_{\text{threshold}}$, $\mathcal{E}$ achieves its minimum at $\theta = 1$. Either way, joint training of supervised and unsupervised contrastive learning ($\theta \in (0,1)$) does not improve the error bound compared with purely supervised or purely unsupervised contrastive learning.

Then we derive the finite sample bound for the linear probing error for contrastive learning under label noise, which explicitly depends on the training set size n.

Theorem 5.6. *Denote $\widehat{\mathcal{E}} := \mathrm{P}_{\bar{x}\sim\mathcal{P}_{\bar{\mathcal{X}}}, x\sim\mathcal{A}(\cdot|\bar{x})}\big(g_{\hat{f},\widehat{B}}(x) \neq y(\bar{x})\big)$ as the generalization bound for the linear probing error. For any labeling function $\hat{y} : \mathcal{X} \to [r]$, there exists a linear probe $\widehat{B} \in \mathbb{R}^{r\times k}$ such that with probability at least $1-\varepsilon$ over the randomness of data, we have*

$$\widehat{\mathcal{E}} \leq \min_{1\leq k'\leq k}\Bigg(\frac{2\big[2\delta_u + [\alpha(1+\rho)\delta_s - 2\delta_u + (1-\alpha)]\theta\big]}{1-\lambda(\boldsymbol{\nu};\theta,\alpha,k'+1)} + \frac{4k'\Big[c_1\cdot\widehat{R}_{n/2}(\mathcal{F}) + c_2\big(\sqrt{\frac{\log 2/\delta}{n}} + \varepsilon\big)\Big]}{\big((1-\theta)\nu_{k'} - \lambda(\boldsymbol{\nu};\theta,\alpha,k+1)\big)^2}\Bigg) + 8\delta_u, \quad (26)$$

where $\widehat{R}_{n/2}(\mathcal{F})$ is the maximal possible empirical Rademacher complexity of $\mathcal{F}$ over $n/2$ data, $\lambda(\boldsymbol{\nu};\theta,\alpha,k+1) = \min\{\theta+(1-\theta)\nu_{k+1}, \theta\alpha+(1-\theta)\nu_k, (1-\theta)\nu_{k+1-r}\}$, $\alpha := \big(1-\frac{r}{r-1}\gamma\big)^2$, $c_1 \lesssim k^2\kappa^2 + k\kappa$ and $c_2 \lesssim k\kappa^2 + k^2\kappa^4$.

Note that in Theorem 5.6, as the training size $n \to \infty$, the sample error term (second term) approximates 0, and therefore the bound in Theorem 5.6 degenerates to that in Theorem 5.5.

For a given sample size $n < \infty$, we observe a trade-off in the choice of k. Specifically, as k increases, the approximation

error (1st term in (26)) decreases, whereas the sample error (2nd term in (26)) increases. It will lead to the following two cases: 1) when k is small, the approximation error could be very large since λ_{k+1} is large; 2) when k is large, the eigen gap $\lambda_{k'} - \lambda_{k+1} \to 0$ since both $\lambda_{k'}$ and λ_{k+1} are very small, and accordingly the sample error goes to infinity. This suggests that we should choose a moderate feature dimension k, so that both approximation and sample error terms are relatively small.

5.4. Discussions

By comparing Theorems 5.4 and 5.5, we show that clean semi-supervised labels help improve the downstream linear error bound of contrastive representation learning, whereas jointly training supervised and unsupervised contrastive losses fails to improve the error bound under noisy labels. In other words, the winner of supervised and unsupervised contrastive learning itself serves as a strong baseline for noisy label contrastive learning. This theoretical finding partly explains why the intuitive joint-training method is not investigated by the community, and why complex algorithmic design such as label denoising is a popular approach to leveraging noisy labels in contrastive learning.

For technical contributions, although the theoretical analysis is based on HaoChen et al. (2021), our analysis is essentially different from existing works in the following aspects: 1) We for the first time establish a theoretical framework for weakly supervised contrastive learning, where we translate the label information into a similarity graph, whereas existing works analyzed pure unsupervised contrastive learning; 2) The main technical difficulty of our analysis is to discuss the eigenvalues of the mixed similarity graph containing both label and feature information (Proposition 5.2), rather than to utilize existing results about self-supervised contrastive learning.

The joint training of SimCLR and SupCon is also discussed in previous works (Islam et al., 2021; Chen et al., 2022), which focus on the empirical improvement of transfer performances whereas we focus on the theoretical properties of joint training. More details can be found in Appendix B.2.

6. Experiments

Recall that it is already empirically verified by Assran et al. (2020) that clean semi-supervised labels help improve over unsupervised contrastive learning. Therefore, in this section, we only empirically verify our theoretical results that noisy labels have limited effects in improving the performance of contrastive learning, and show that the winner of SupCon and SimCLR itself serves as a strong baseline for contrastive learning with noisy labels. Based on this, we discuss that complex designs are imperative for improving contrastive representation learning with noisy labels.

6.1. Experimental Setups

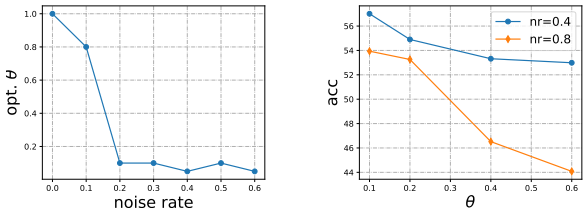
We conduct numerical comparisons on the CIFAR-10 and TinyImageNet-200 benchmark datasets. Because the standard supervised contrastive learning algorithm SupCon (Khosla et al., 2020) is adapted from the self-supervised contrastive learning framework SimCLR (Chen et al., 2020), for fair comparisons, we use SupCon as the supervised contrastive loss, and SimCLR as the self-supervised contrastive loss. We argue that we can to a large extent verify the theoretical insights discussed in the previous section, even if the theoretical parts consider the spectral contrastive loss. First of all, as shown in the original paper, spectral contrastive loss has comparative empirical performances with respect to that of SimCLR. Besides, the theoretical evidence can be found in Johnson et al. (2023), which proves that by interpreting the exponentiated dot product $e^{f(x)^\top f(x')}$ as the similarity and treating the exponential and temperature term τ as part of the model instead of part of the objective, InfoNCE loss and Spectral contrastive loss share the same population minimum. That means, by adopting similar kernel deriviations, our work also has the potential to extend to other contrastive losses including standard InfoNCE.

We follow the experimental setting of SimCLR and SupCon. Specifically, we use ResNet-50 as the encoder and a 2-layer MLP as the projection head. We set the batch size as 1024. We use 1000 epochs for training representations. We use the SGD optimizer with the learning rate 0.5 decayed at the 700-th, 800-th, and 900-th epochs with a weight decay 0.1. We run experiments on 4 NVIDIA Tesla V100 32GB GPUs. The data augmentations we use are random crop and resize (with random flip), color distortion, and color dropping. We evaluate the self-supervised learned representation by linear evaluation protocol, where a linear classifier is trained on the top of the encoder, and regard its test accuracy as the performance of the encoder. 100 epochs are used for linear probing on the clean data. The symmetric noisy labels are generated by flipping the labels of a given proportion of training samples uniformly to one of the other class labels. For the CIFAR-10 dataset, we run experiments with noise rate $\{0.1, 0.2, 0.3, 0.4, 0.5, 0.6\}$ and for TinyImageNet-200, we run experiments with noise rate $\{0.2, 0.4, 0.6, 0.8\}$.

6.2. Parameter Analysis of θ

We first conduct an analysis of the parameter θ for the joint training of SimCLR and SupCon. In Figure 2(a), we plot the optimal θ for $\mathcal{L}_{\mathrm{JointTraining}}$ on the CIFAR-10 dataset across various noise rates. We show that when the noise rate is relatively high, θ is relatively small, and $\mathcal{L}_{\mathrm{JointTraining}}$ relies more on unsupervised learning, whereas when the noise rate is relatively low, θ is relatively large, and $\mathcal{L}_{\mathrm{JointTraining}}$

relies more on supervised learning. Moreover, the optimal θ lies very close to either 0 or 1, which indicates that only one loss mainly contributes to the joint training. On the other hand, in Figure 2(b), we plot the performance changes of $\mathcal{L}_{\text{JointTraining}}$ with respect to the parameter θ under noise rates 0.4 and 0.8 on the TinyImageNet-200 dataset. We show that under both noise rates, as θ increases, the accuracy of $\mathcal{L}_{\text{JointTraining}}$ decreases. Moreover, the performance drop of $\mathcal{L}_{\text{JointTraining}}$ under high noise rates is more significant than that under low noise rates, indicating that too much noisy label information hurts the performance of joint training, especially when the noise rate is high.



(a) Optimal θ for various noise rates on CIFAR-10.

(b) Performance change w.r.t. θ on TinyImageNet-200.

Figure 2: Parameter analysis of θ.

6.3. Parameter Analysis of k

We empirically verify the discussions following Theorem 5.6 about the feature dimension k. We train the joint objective (12) with $\theta = 0.1$ using ResNet-50 on the CIFAR-10 dataset under 30% label noise with feature dimension varying in $\{128, 256, 512, 1024, 2048\}$, and report the linear probing accuracy in Table 1.

Table 1: Parameter analysis of k.

k	128	256	512	1024	2048
Acc (%)	91.89	**92.18**	92.12	91.77	90.9

We observe that as k increases, the linear probing accuracy first increases and then decreases, which validates the theoretical insight that we should choose a moderate feature dimension k in Theorem 5.6.

6.4. The Winner of SupCon and SimCLR Serves as a Strong Baseline

In this section, we show that under label noise, joint training of SupCon and SimCLR has limited improvement over a simple but strong baseline, that is, the winner of SupCon and SimCLR on the TinyImageNet-200 dataset. Specifically, in Figure 3, the box plot show the linear probing performances of $\mathcal{L}_{\text{JointTraining}}$ with $\theta \in \{0.1, 0.2, 0.4, 0.6\}$ under noise rates $\{0.2, 0.4, 0.6, 0.8\}$ respectively. And the curves respectively show the performance of SimCLR, SupCon, and

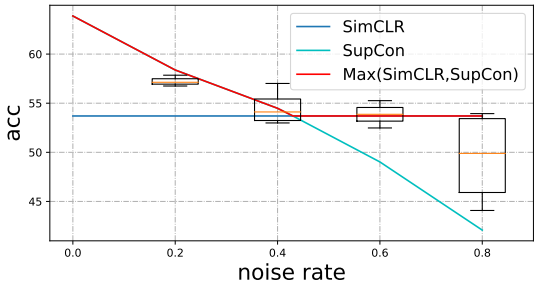


Figure 3: The winner of SupCon and SimCLR serves as a strong baseline of joint training under label noise on the TinyImageNet-200 dataset.

the winner of the two. We can see that the "Max" performance of SupCon and SimCLR lies within the box plots, indicating that the winner of SupCon and SimCLR has a comparative or sometimes better performance compared with the jointly trained model. Additionally, we observe the same trends on tasks like detection, segmentation, and fine-tuning in Appendix B.3. These experimental results verify the theoretical result in Theorem 5.5 that joint training does not improve the error bound.

7. Conclusion

In this paper, we establish a theoretical framework for weakly supervised contrastive learning, which is compatible with the settings of both noisy label learning and semi-supervised learning. We take spectral contrastive learning as a proxy for theoretical analysis. By formulating a mixed similarity graph induced by both weakly supervised label information and unsupervised feature information, we analyze the weakly supervised spectral contrastive learning based on the framework of spectral clustering, and derive the downstream linear evaluation error bound. Our theoretical results show that semi-supervised information improves the downstream error bound, whereas, under the setting of symmetric label noise, we prove that jointly training supervised and unsupervised contrastive losses fail to improve the error bound. Our theoretical findings here provide new insights for the community to rethink the role of weak supervision in helping the representation of contrastive learning. For future works, we will investigate the effect of more complex weak supervision, such as active learning and label-dependent label noise, on contrastive learning.

Acknowledgements

Yisen Wang is partially supported by the National Key R&D Program of China (2022ZD0160304), the National Natural Science Foundation of China (62006153), Open Research Projects of Zhejiang Lab (No. 2022RC0AB05), and Huawei Technologies Inc. Weiran Huang is funded by MSRA.

References

Acharya, A., Sanghavi, S., Jing, L., Bhushanam, B., Choudhary, D., Rabbat, M., and Dhillon, I. Positive unlabeled contrastive learning. *arXiv preprint arXiv:2206.01206*, 2022.

Aitchison, L. Infonce is a variational autoencoder. *arXiv preprint arXiv:2107.02495*, 2021.

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

Ash, J., Goel, S., Krishnamurthy, A., and Misra, D. Investigating the role of negatives in contrastive representation learning. In *AISTATS*, 2022.

Assran, M., Ballas, N., Castrejon, L., and Rabbat, M. Supervision accelerates pre-training in contrastive semi-supervised learning of visual representations. *arXiv preprint arXiv:2006.10803*, 2020.

Bao, H., Nagano, Y., and Nozawa, K. On the surrogate gap between contrastive and supervised losses. In *ICML*, 2022.

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 *ICML*, 2022.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. In *ICML*, 2020.

Chen, X. and He, K. Exploring simple siamese representation learning. In *CVPR*, 2021.

Cheng, H., Zhu, Z., Sun, X., and Liu, Y. Demystifying how self-supervised features improve training from noisy labels. *arXiv preprint arXiv:2110.09022*, 2021.

Chuang, C.-Y., Hjelm, R. D., Wang, X., Vineet, V., Joshi, N., Torralba, A., Jegelka, S., and Song, Y. Robust contrastive learning against noisy views. In *CVPR*, 2022.

Fulton, W. Eigenvalues, invariant factors, highest weights, and schubert calculus. *Bulletin of the American Mathematical Society*, 37(3):209–249, 2000.

Ghosh, A. and Lan, A. Contrastive learning improves model robustness under label noise. In *CVPR*, 2021.

HaoChen, J. Z., Wei, C., Gaidon, A., and Ma, T. Provable guarantees for self-supervised deep learning with spectral contrastive loss. In *NeurIPS*, 2021.

He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020.

Hu, T., Liu, Z., Zhou, F., Wang, W., and Huang, W. Your contrastive learning is secretly doing stochastic neighbor embedding. In *ICLR*, 2023.

Huang, W., Yi, M., Zhao, X., and Jiang, Z. Towards the generalization of contrastive self-supervised learning. In *ICLR*, 2023.

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 *ICCV*, pp. 8845–8855, 2021.

Johnson, D. D., Hanchi, A. E., and Maddison, C. J. Contrastive learning can find an optimal basis for approximately view-invariant functions. In *ICLR*, 2023.

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

Kim, B., Choo, J., Kwon, Y.-D., Joe, S., Min, S., and Gwon, Y. Selfmatch: Combining contrastive self-supervision and consistency for semi-supervised learning. *arXiv preprint arXiv:2101.06480*, 2021a.

Kim, Y., Yim, J., Yun, J., and Kim, J. Nlnl: Negative learning for noisy labels. In *ICCV*, 2019.

Kim, Y., Yun, J., Shon, H., and Kim, J. Joint negative and positive learning for noisy labels. In *CVPR*, 2021b.

Ko, C.-Y., Mohapatra, J., Liu, S., Chen, P.-Y., Daniel, L., and Weng, L. Revisiting contrastive learning through the lens of neighborhood component analysis: an integrated framework. In *ICML*, 2022.

Lee, D., Kim, S., Kim, I., Cheon, Y., Cho, M., and Han, W.-S. Contrastive regularization for semi-supervised learning. In *CVPR*, 2022.

Li, S., Xia, X., Ge, S., and Liu, T. Selective-supervised contrastive learning with noisy labels. In *CVPR*, 2022.

Navaneet, K., Abbasi Koohpayegani, S., Tejankar, A., Pourahmadi, K., Subramanya, A., and Pirsiavash, H. Constrained mean shift using distant yet related neighbors for representation learning. In *ECCV 2022*, 2022.

Nozawa, K. and Sato, I. Understanding negative samples in instance discriminative self-supervised representation learning. In *NeurIPS*, 2021.

Ortego, D., Arazo, E., Albert, P., O'Connor, N. E., and McGuinness, K. Multi-objective interpolation training for robustness to label noise. In *CVPR*, 2021.

Shen, K., Jones, R. M., Kumar, A., Xie, S. M., HaoChen, J. Z., Ma, T., and Liang, P. Connect, not collapse: Explaining contrastive learning for unsupervised domain adaptation. In *ICML*, 2022.

Wang, Y., Geng, Z., Jiang, F., Li, C., Wang, Y., Yang, J., and Lin, Z. Residual relaxation for multi-view representation learning. In *NeurIPS*, 2021a.

Wang, Y., Zhang, Q., Wang, Y., Yang, J., and Lin, Z. Chaos is a ladder: A new theoretical understanding of contrastive learning via augmentation overlap. In *ICLR*, 2021b.

Wang, Y., Zhang, Q., Du, T., Yang, J., Lin, Z., and Wang, Y. A message passing perspective on learning dynamics of contrastive learning. In *ICLR*, 2023.

Xue, Y., Whitecross, K., and Mirzasoleiman, B. Investigating why contrastive learning benefits robustness against label noise. In *ICML*, 2022.

Yan, J., Luo, L., Xu, C., Deng, C., and Huang, H. Noise is also useful: Negative correlation-steered latent contrastive learning. In *CVPR*, 2022.

Yang, F., Wu, K., Zhang, S., Jiang, G., Liu, Y., Zheng, F., Zhang, W., Wang, C., and Zeng, L. Class-aware contrastive semi-supervised learning. In *CVPR*, 2022.

Yao, Y., Sun, Z., Zhang, C., Shen, F., Wu, Q., Zhang, J., and Tang, Z. Jo-src: A contrastive approach for combating noisy labels. In *CVPR*, 2021.

Zhang, M., Yuan, C., Yao, J., and Huang, W. Learning with noisily-labeled class-imbalanced data. *arXiv preprint arXiv:2211.10955*, 2022a.

Zhang, Y., Zhang, X., Li, J., Qiu, R., Xu, H., and Tian, Q. Semi-supervised contrastive learning with similarity co-calibration. *IEEE Transactions on Multimedia*, 2022b.

Zimmermann, R. S., Sharma, Y., Schneider, S., Bethge, M., and Brendel, W. Contrastive learning inverts the data generating process. In *ICML*, 2021.

Appendix

The appendix consists of the proofs for lemmas and theorems (Section A) and additional experimental results (Section B).

A. Proofs

A.1. Proof of Lemma 4.1

Proof. Under Assumption 3.1, we have

$$
\begin{aligned}
(\boldsymbol{T}^2)_{i,j} &= \begin{cases} (1-\gamma)^2 + \gamma^2/(r-1), & i = j \\ 2\gamma(1-\gamma)/(r-1) + (r-2)\gamma^2/(r-1)^2, & i \neq j \end{cases} \\
&= \begin{cases} (1-\gamma)^2 + \gamma^2/(r-1), & i = j \\ \dfrac{\gamma}{r-1}\Big(2 - \dfrac{r}{r-1}\gamma\Big), & i \neq j. \end{cases}
\end{aligned} \tag{27}
$$

That is, we have

$$
\begin{aligned}
\boldsymbol{T}^2 &= \Big[(1-\gamma)^2 + \gamma^2/(r-1) - \frac{\gamma}{r-1}\Big(2 - \frac{r}{r-1}\gamma\Big)\Big]\boldsymbol{I}_{r\times r} \\
&\quad + \frac{\gamma}{r-1}\Big(2 - \frac{r}{r-1}\gamma\Big)\vec{1}_r\vec{1}_r^\top \\
&= \Big(1 - \frac{r}{r-1}\gamma\Big)^2\boldsymbol{I}_{r\times r} + \frac{\gamma}{r-1}\Big(2 - \frac{r}{r-1}\gamma\Big)\vec{1}_r\vec{1}_r^\top \\
&:= \alpha\boldsymbol{I}_{r\times r} + \beta\vec{1}_r\vec{1}_r^\top.
\end{aligned} \tag{28}
$$

Given $\gamma \in [0, 1)$, we have

$$
\begin{aligned}
\boldsymbol{A}_L^\star = \boldsymbol{Y}_L\boldsymbol{T}^2\boldsymbol{Y}_L^\top &= \boldsymbol{Y}_L\big(\alpha\boldsymbol{I}_{r\times r} + \beta\vec{1}_r\vec{1}_r^\top\big)\boldsymbol{Y}_L^\top \\
&= \alpha\boldsymbol{Y}_L\boldsymbol{Y}_L^\top + \beta\boldsymbol{Y}_L\vec{1}_r\vec{1}_r^\top\boldsymbol{Y}_L^\top \\
&= \alpha\boldsymbol{A}_L + \beta\vec{1}_{n_L}\vec{1}_{n_L}^\top,
\end{aligned} \tag{29}
$$

where the last equality holds because $\sum_j \eta_j(x_i) = 1$ for $i \in [n]$. □

and the normalized augmentation graph is

$$
\bar{\boldsymbol{A}}^\star = \tilde{\boldsymbol{D}}^{-1/2}\boldsymbol{A}^\star\tilde{\boldsymbol{D}}^{-1/2}, \tag{30}
$$

where

$$
\tilde{\boldsymbol{D}} = \begin{bmatrix} \tilde{\boldsymbol{D}}_L & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}_{n_U\times n_U} \end{bmatrix}, \tag{31}
$$

$$
\tilde{\boldsymbol{D}}_L = \mathrm{diag}(d_i), \tag{32}
$$

and

$$
\begin{aligned}
d_i &= \sum_{j\in[n_L]} \boldsymbol{A}_{i,j}^\star = \alpha\sum_{j\in[n_L]}\sum_{\ell\in[r]}\eta_\ell(x_i)\eta_\ell(x_j) + n_L\beta \\
&= \alpha\sum_{\ell\in[r]}\eta_\ell(x_i)\sum_{j\in[n_L]}\eta_\ell(x_j) + n_L\beta = \alpha\sum_{\ell\in[r]}\eta_\ell(x_i)n_\ell + n_L\beta
\end{aligned} \tag{33}
$$

Specifically, when the labeled data is class-balanced, i.e. $n_1 = \ldots = n_r = n_L/r$. Then we have

$$
d_i = \frac{n_L}{r}\alpha\sum_{\ell\in[r]}\eta_\ell(x_i) + n_L\beta = \frac{n_L}{r}\alpha + n_L\beta = \frac{n_L}{r}, \tag{34}
$$

and thus

$$\bar{\boldsymbol{A}}^\star = \begin{bmatrix} \alpha \frac{r}{n_L} \boldsymbol{A}_L + \beta \frac{r}{n_L} \vec{1}_{n_L} \vec{1}_{n_L}^\top & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}_{n_U \times n_U} \end{bmatrix}. \tag{35}$$

A.2. Proof of Proposition 5.1

Proof. We first prove that $v_1 = \frac{1}{\sqrt{n_L}} \vec{1}_{n_L}$ is an eigenvector of $\bar{\boldsymbol{A}}_L := \frac{r}{n_L} \boldsymbol{A}_L$ with eigenvalue $\mu_1 = 1$. To be specific,

$$\begin{aligned} \bar{\boldsymbol{A}}_L \cdot \frac{1}{\sqrt{n_L}} \vec{1}_{n_L} &= \frac{1}{\sqrt{n_L}} \cdot \frac{r}{n_L} \boldsymbol{A}_L \cdot \vec{1}_{n_L} \\ &= \frac{1}{\sqrt{n_L}} \cdot \frac{r}{n_L} \boldsymbol{Y}_L \boldsymbol{Y}_L^\top \vec{1}_{n_L} \\ &= \frac{1}{\sqrt{n_L}} \cdot \frac{r}{n_L} \boldsymbol{Y}_L \frac{n_L}{r} \vec{1}_r \\ &= \frac{1}{\sqrt{n_L}} \vec{1}_{n_L}, \end{aligned} \tag{36}$$

where the second last equality is due to class balance, i.e. $\sum_{i \in [n_L]} \eta_j(x_i) = n_L / r$ for $j \in [r]$, and the last equality holds because $\sum_{j \in [r]} \eta_j(x_i) = 1$ for $i \in [n_L]$.

Therefore, we can rewrite $\bar{\boldsymbol{A}}_L$ as

$$\bar{\boldsymbol{A}}_L = \left[\tfrac{1}{\sqrt{n_L}} \vec{1}_{n_L}, v_2, \ldots, v_{n_L} \right] \begin{bmatrix} 1 & 0 & \ldots & 0 \\ 0 & \mu_2 & \ldots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \ldots & \mu_{n_L} \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{n_L}} \vec{1}_{n_L}^\top \\ v_2^\top \\ \vdots \\ v_{n_L}^\top \end{bmatrix}. \tag{37}$$

Note that $\frac{1}{n_L} \vec{1}_{n_L} \vec{1}_{n_L}^\top$ can be decomposed as

$$\begin{aligned} \frac{1}{n_L} \vec{1}_{n_L} \vec{1}_{n_L}^\top &= \left(\frac{1}{\sqrt{n_L}} \vec{1}_{n_L} \right) \left(\frac{1}{\sqrt{n_L}} \vec{1}_{n_L} \right)^\top \\ &= \left[\tfrac{1}{\sqrt{n_L}} \vec{1}_{n_L}, v_2, \ldots, v_{n_L} \right] \begin{bmatrix} 1 & 0 & \ldots & 0 \\ 0 & 0 & \ldots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \ldots & 0 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{n_L}} \vec{1}_{n_L}^\top \\ v_2^\top \\ \vdots \\ v_{n_L}^\top \end{bmatrix}. \end{aligned} \tag{38}$$

Then we have

$$\begin{aligned} \bar{\boldsymbol{A}}_L^\star &:= \alpha \frac{r}{n_L} \boldsymbol{A}_L + r\beta \frac{1}{n_L} \vec{1}_{n_L} \vec{1}_{n_L}^\top \\ &= \left[\tfrac{1}{\sqrt{n_L}} \vec{1}_{n_L}, v_2, \ldots, v_{n_L} \right] \begin{bmatrix} \alpha & 0 & \ldots & 0 \\ 0 & \alpha\mu_2 & \ldots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \ldots & \alpha\mu_{n_L} \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{n_L}} \vec{1}_{n_L}^\top \\ v_2^\top \\ \vdots \\ v_{n_L}^\top \end{bmatrix} \\ &\quad \cdot \left[\tfrac{1}{\sqrt{n_L}} \vec{1}_{n_L}, v_2, \ldots, v_{n_L} \right] \begin{bmatrix} r\beta & 0 & \ldots & 0 \\ 0 & 0 & \ldots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \ldots & 0 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{n_L}} \vec{1}_{n_L}^\top \\ v_2^\top \\ \vdots \\ v_{n_L}^\top \end{bmatrix} \\ &= \left[\tfrac{1}{\sqrt{n_L}} \vec{1}_{n_L}, v_2, \ldots, v_{n_L} \right] \begin{bmatrix} \alpha + r\beta & 0 & \ldots & 0 \\ 0 & \alpha\mu_2 & \ldots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \ldots & \alpha\mu_{n_L} \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{n_L}} \vec{1}_{n_L}^\top \\ v_2^\top \\ \vdots \\ v_{n_L}^\top \end{bmatrix}. \end{aligned} \tag{39}$$

Since $\alpha + r\beta = 1$, the eigenvalues of $\bar{\boldsymbol{A}}_L^\star$ are $1, \alpha\mu_2, \ldots, \alpha\mu_{n_L}$. Thus the eigenvalues of

$$\bar{\boldsymbol{A}}^\star = \begin{bmatrix} \bar{\boldsymbol{A}}_L^\star & \mathbf{0} \\ \mathbf{0} & \boldsymbol{I}_{n_U \times n_U} \end{bmatrix} \tag{40}$$

are

$$\tilde{\mu}_1 = \ldots = \tilde{\mu}_{n_U+1} = 1, \tag{41}$$
$$\tilde{\mu}_j = \alpha\mu_j, \text{ for } j = n_U + 2, \ldots, n. \tag{42}$$

□

A.3. Proof of Proposition 5.2

Proof. By equation 13 in Fulton (2000), for two real symmetric n by n matrix $(1-\theta)\bar{\boldsymbol{A}}_0$ and $\theta\bar{\boldsymbol{A}}^\star$, the $k+1$-th largest eigenvalue of $\boldsymbol{A}_{\theta,\lambda,n_L} := (1-\theta)\bar{\boldsymbol{A}}_0 + \theta\bar{\boldsymbol{A}}^\star$ can take any value in the interval

$$\max_{i+j=n+k+1} (1-\theta)\nu_i + \theta\tilde{\mu}_j \le \lambda_{k+1} \le \min_{i+j=k+2} (1-\theta)\nu_i + \theta\tilde{\mu}_j. \tag{43}$$

By Proposition 5.1, we have

$$\tilde{\mu}_j = \begin{cases} 1, & j = 1, \ldots, n_U + 1; \\ \alpha\mu_j, & j = n_U + 2, \ldots, n_U + r; \\ 0, & j = n_U + r + 1, \ldots, n. \end{cases} \tag{44}$$

Therefore, we have

$$\begin{aligned} &\max_{i+j=n+k+1} (1-\theta)\nu_i + \theta\tilde{\mu}_j \\ &= \max_{1\le i\le n+k+1} (1-\theta)\nu_i + \theta\tilde{\mu}_{n+k+1-i} \\ &= \max \begin{cases} \theta + (1-\theta)\nu_i, & i = n_L + k, \ldots, n; \\ \theta\alpha\mu_{n+k+1-i} + (1-\theta)\nu_i, & i = n_L + k - r + 1, \ldots, n_L + k - 1; \\ (1-\theta)\nu_i, & i = k+1, \ldots, n_L + k - r \end{cases} \\ &= \max \begin{cases} \theta + (1-\theta)\nu_{n_L+k} \\ \theta\alpha\mu_{n+k+1-i} + (1-\theta)\nu_i, & i = n_L + k - r + 1, \ldots, n_L + k - 1; \\ (1-\theta)\nu_{k+1}, \end{cases} \end{aligned} \tag{45}$$

where the last equality holds because $\{\nu_i\}_{i\in[n]}$ is ranked in descending order. Then when $k \le n_U$,

$$\lambda_{k+1} \ge \max\left\{\theta + (1-\theta)\nu_{n_L+k}, \max_{i=n_L+k-r+1,\ldots,n_L+k-1} \{\theta\alpha\mu_{n+k+1-i} + (1-\theta)\nu_i\}, (1-\theta)\nu_{k+1}\right\}, \tag{46}$$

when $n_U < k < n_U + r$,

$$\lambda_{k+1} \ge \max\left\{(1-\theta)\nu_{k+1}, \max_{i=n_L+k-r+1,\ldots,n_L+k-1} \{\theta\alpha\mu_{n+k+1-i} + (1-\theta)\nu_i\}\right\}, \tag{47}$$

and when $k \ge n_U + r$,

$$\lambda_{k+1} \ge (1-\theta)\nu_{k+1}. \tag{48}$$

On the other hand, we have

$$\begin{aligned} &\min_{i+j=k+2} (1-\theta)\nu_i + \theta\tilde{\mu}_j \\ &= \min_{1\le i\le k+2} (1-\theta)\nu_i + \theta\tilde{\mu}_{k+2-i} \end{aligned}$$

$$
\begin{aligned}
&= \min \begin{cases} \theta + (1-\theta)\nu_i, & i = k+1-n_U, \dots, k+1; \\ \theta\alpha\mu_{k+2-i} + (1-\theta)\nu_i, & i = k+2-r-n_U, \dots, k-n_U; \\ (1-\theta)\nu_i, & i = k+2-n, \dots, k+1-r-n_U. \end{cases} \\
&= \min \begin{cases} \theta + (1-\theta)\nu_{k+1}; \\ \theta\alpha\mu_{k+2-i} + (1-\theta)\nu_i, & i = k+2-r-n_U, \dots, k-n_U; \\ (1-\theta)\nu_{k+1-r-n_U}. \end{cases}
\end{aligned} \tag{49}
$$

Then when $k \le n_U$, there holds

$$
\lambda_{k+1} \le \theta + (1-\theta)\nu_{k+1}, \tag{50}
$$

when $n_U < k < n_U + r$, there holds

$$
\lambda_{k+1} \le \min\Big\{\theta + (1-\theta)\nu_{k+1}, \min_{i=k+2-r-n_U,\dots,k-n_U} \{\theta\alpha\mu_{k+2-i} + (1-\theta)\nu_i\}\Big\}, \tag{51}
$$

and when $k \ge n_U + r$, there holds

$$
\lambda_{k+1} \le \min\Big\{\theta + (1-\theta)\nu_{k+1}, \min_{i=k+2-r-n_U,\dots,k-n_U} \{\theta\alpha\mu_{k+2-i} + (1-\theta)\nu_i\}, (1-\theta)\nu_{k+1-r-n_U}\Big\}. \tag{52}
$$

□

A.4. Proof of Theorem 5.4

To prove Theorem 5.4, we first prove Proposition A.1.

Proposition A.1. *For arbitrary $\boldsymbol{Y}$, assume that the labeled data is class-balanced, i.e. $\sum_{i\in[n_L]} \eta_j(x_i) = n_L/r$ for $j \in [r]$. Denote $\nu_1, \dots, \nu_n$ as the eigenvalues of $\bar{\boldsymbol{A}}_0$ (in descending order). Denote $\mathcal{E} := \mathrm{P}_{\bar{x}\sim\mathcal{P}_{\bar{\mathcal{X}}}, x\sim\mathcal{A}(\cdot|\bar{x})}\big(g_{f^*_{\mathrm{pop}},B^*}(x) \ne y(\bar{x})\big)$ as the linear evaluation error, where $B^* \in \mathbb{R}^{r\times k}$ with norm $\|B^*\|_F \le 1/\lambda_k$. Assume there exists $\rho > 0$, such that $w_i/w_j < \rho$, for $i, j \in [n]$. Then under the deterministic scenario and Assumptions 3.1 and 5.3, for $k \le n_U$, there holds*

$$
\mathcal{E} \le \frac{2\big[2\delta_u + [\alpha(1+\rho)\delta_s - 2\delta_u + (1-\alpha)]\theta\big]}{1-\theta-(1-\theta)\nu_{k+1}} + 8\delta_u, \tag{53}
$$

for $n_U + 1 \le k \le n_U + r - 1$, there holds

$$
\mathcal{E} \le \frac{2\big[2\delta_u + [\alpha(1+\rho)\delta_s - 2\delta_u + (1-\alpha)]\theta\big]}{1-\min\{\theta+(1-\theta)\nu_{k+1}, \theta\alpha + (1-\theta)\nu_{k-n_U}\}} + 8\delta_u, \tag{54}
$$

and for $k \ge n_U + r$, there holds

$$
\mathcal{E} \le \frac{2\big[2\delta_u + [\alpha(1+\rho)\delta_s - 2\delta_u + (1-\alpha)]\theta\big]}{1-\lambda(\nu;\theta,\alpha)} + 8\delta_u, \tag{55}
$$

where

$$
\begin{aligned}
\lambda(\boldsymbol{\nu};\theta,\alpha) = \min\{&\theta + (1-\theta)\nu_{k+1}, \theta\alpha + (1-\theta)\nu_{k-n_U}, \\ &(1-\theta)\nu_{k+1-r-n_U}\}
\end{aligned} \tag{56}
$$

and $\alpha := \big(1 - \frac{r}{r-1}\gamma\big)^2$.

Proof of Proposition A.1. By Lemma B.3 of HaoChen et al. (2021), for any labeling function $\vec{y} : \mathcal{X} \to [r]$, there exists a linear probe $B^* \in \mathbb{R}^{r\times k}$ with norm $\|B^*\|_F \le 1/\lambda_k$ such that

$$
\mathrm{P}_{\bar{x}\sim\mathcal{P}_{\bar{\mathcal{X}}}, x\sim\mathcal{A}(\cdot|\bar{x})}\Big(g_{f^*_{\mathrm{pop}},B^*}(x) \ne y(\bar{x})\Big) \le \frac{2\phi^{\hat{y}}}{1-\lambda_{k+1}} + 8\delta_u, \tag{57}
$$

where according to the definition of $\boldsymbol{A}_{\theta,\lambda,n_L}$,

$$\begin{aligned}\phi^{\hat{y}} &= \sum_{i,j\in[n]} \big[(1-\theta)\frac{w_{ij}}{\sqrt{w_i w_j}} + \theta \bar{\boldsymbol{A}}^{\star}_{i,j}\big]\mathbf{1}[\hat{y}(x_i)\neq\hat{y}(x_j)] \\ &= \frac{1}{\sqrt{w_i w_j}}\big[(1-\theta)\sum_{i,j\in[n]} w_{i,j}\mathbf{1}[\hat{y}(x_i)\neq\hat{y}(x_j)] + \theta\sum_{i,j\in[n]}\sqrt{w_i w_j}\bar{\boldsymbol{A}}^{\star}_{i,j}\mathbf{1}[\hat{y}(x_i)\neq\hat{y}(x_j)]\big]. \end{aligned} \tag{58}$$

We investigate the RHS of (58) respectively. The first term is

$$\begin{aligned}&\sum_{i,j\in[n]} w_{i,j}\mathbf{1}[\hat{y}(x_i)\neq\hat{y}(x_j)] \\ &= \sum_{i,j\in[n]} \mathbb{E}_{\bar{x}\sim\mathcal{P}_{\bar{\mathcal{X}}}}\mathcal{A}(x_i|\bar{x})\mathcal{A}(x_j|\bar{x})\mathbf{1}[\hat{y}(x_i)\neq\hat{y}(x_j)] \\ &\leq \sum_{i,j\in[n]} \mathbb{E}_{\bar{x}\sim\mathcal{P}_{\bar{\mathcal{X}}}}\mathcal{A}(x_i|\bar{x})\mathcal{A}(x_j|\bar{x})\big(\mathbf{1}[\hat{y}(x_i)\neq\hat{y}(\bar{x})]+\mathbf{1}[\hat{y}(x_j)\neq\hat{y}(\bar{x})]\big) \\ &= 2\sum_{i\in[n]} \mathbb{E}_{\bar{x}\sim\mathcal{P}_{\bar{\mathcal{X}}}}\mathcal{A}(x_i|\bar{x})\mathbf{1}[\hat{y}(x_i)\neq\hat{y}(\bar{x})] \\ &\leq 2\delta_u. \end{aligned} \tag{59}$$

The second term is

$$\begin{aligned}&\sum_{i,j\in[n]}\sqrt{w_i w_j}\bar{\boldsymbol{A}}^{\star}_{i,j}\mathbf{1}[\hat{y}(x_i)\neq\hat{y}(x_j)] \\ &= \sum_{i,j\in[n_L]}\sqrt{w_i w_j}(\bar{\boldsymbol{A}}^{\star}_{L})_{i,j}\mathbf{1}[\hat{y}(x_i)\neq\hat{y}(x_j)] + \sum_{i>n_L}\sqrt{w_i w_j}\mathbf{1}[\hat{y}(x_i)\neq\hat{y}(x_i)] \\ &\quad + 2\sum_{i\leq n_L, j>n_L}\sqrt{w_i w_j}\bar{\boldsymbol{A}}^{\star}_{i,j}\mathbf{1}[\hat{y}(x_i)\neq\hat{y}(x_i)]. \end{aligned} \tag{60}$$

According to the definition of $\bar{\boldsymbol{A}}^{\star}$, the last two terms are equal to 0. Then by Lemma 4.1, the second term on the RHS of (58) becomes

$$\begin{aligned}&\sum_{i,j\in[n]}\sqrt{w_i w_j}\bar{\boldsymbol{A}}^{\star}_{i,j}\mathbf{1}[\hat{y}(x_i)\neq\hat{y}(x_j)] \\ &= \sum_{i,j\in[n_L]}\sqrt{w_i w_j}\Big(\alpha\bar{\boldsymbol{A}}_{i,j}+\beta\frac{r}{n_L}\Big)\mathbf{1}[\hat{y}(x_i)\neq\hat{y}(x_j)] \\ &\leq \alpha\frac{r}{n_L}\sum_{i,j\in[n_L]}\sum_{\ell\in[r]}\sqrt{w_i w_j}\eta_\ell(x_i)\eta_\ell(x_j)\Big(\mathbf{1}[\hat{y}(x_i)\neq\ell]+\mathbf{1}[\hat{y}(x_j)\neq\ell]\Big) \\ &\quad + (1-\alpha)\frac{1}{n_L}\sum_{i,j\in[n_L]}\sqrt{w_i w_j} \\ &\leq \alpha\frac{r}{n_L}\sum_{i,j\in[n_L]}\sum_{\ell\in[r]}\frac{1}{2}(w_i+w_j)\eta_\ell(x_i)\eta_\ell(x_j)\Big(\mathbf{1}[\hat{y}(x_i)\neq\ell]+\mathbf{1}[\hat{y}(x_j)\neq\ell]\Big) \\ &\quad + (1-\alpha)\frac{1}{n_L}\sum_{i,j\in[n_L]}\frac{1}{2}(w_i+w_j) \\ &= \frac{1}{2}\alpha\frac{r}{n_L}\sum_{\ell\in[r]}\sum_{i,j\in[n_L]} w_i\eta_\ell(x_i)\eta_\ell(x_j)\mathbf{1}[\hat{y}(x_i)\neq\ell] \\ &+ \frac{1}{2}\alpha\frac{r}{n_L}\sum_{\ell\in[r]}\sum_{i,j\in[n_L]} w_i\eta_\ell(x_i)\eta_\ell(x_j)\mathbf{1}[\hat{y}(x_j)\neq\ell] \end{aligned}$$

$$
\begin{aligned}
&+\frac{1}{2}\alpha\frac{r}{n_L}\sum_{\ell\in[r]}\sum_{i,j\in[n_L]} w_j\eta_\ell(x_i)\eta_\ell(x_j)\mathbf{1}[\hat{y}(x_i)\neq\ell] \\
&+\frac{1}{2}\alpha\frac{r}{n_L}\sum_{\ell\in[r]}\sum_{i,j\in[n_L]} w_j\eta_\ell(x_i)\eta_\ell(x_j)\mathbf{1}[\hat{y}(x_j)\neq\ell]+(1-\alpha) \\
&\leq\frac{1}{2}\alpha\frac{r}{n_L}\sum_{\ell\in[r]}\sum_{i,j\in[n_L]} \rho w_j\eta_\ell(x_i)\eta_\ell(x_j)\mathbf{1}[\hat{y}(x_i)\neq\ell] \\
&+\frac{1}{2}\alpha\frac{r}{n_L}\sum_{\ell\in[r]}\sum_{i,j\in[n_L]} w_i\eta_\ell(x_i)\eta_\ell(x_j)\mathbf{1}[\hat{y}(x_j)\neq\ell] \\
&+\frac{1}{2}\alpha\frac{r}{n_L}\sum_{\ell\in[r]}\sum_{i,j\in[n_L]} w_j\eta_\ell(x_i)\eta_\ell(x_j)\mathbf{1}[\hat{y}(x_i)\neq\ell] \\
&+\frac{1}{2}\alpha\frac{r}{n_L}\sum_{\ell\in[r]}\sum_{i,j\in[n_L]} \rho w_i\eta_\ell(x_i)\eta_\ell(x_j)\mathbf{1}[\hat{y}(x_j)\neq\ell]+(1-\alpha) \\
&=\frac{1}{2}\alpha\frac{r}{n_L}\sum_{\ell\in[r]}\sum_{i\in[n_L]} \rho\pi_\ell\eta_\ell(x_i)\mathbf{1}[\hat{y}(x_i)\neq\ell] \\
&+\frac{1}{2}\alpha\frac{r}{n_L}\sum_{\ell\in[r]}\sum_{j\in[n_L]} \pi_\ell\eta_\ell(x_j)\mathbf{1}[\hat{y}(x_j)\neq\ell] \\
&+\frac{1}{2}\alpha\frac{r}{n_L}\sum_{\ell\in[r]}\sum_{i\in[n_L]} \pi_\ell\eta_\ell(x_i)\mathbf{1}[\hat{y}(x_i)\neq\ell] \\
&+\frac{1}{2}\alpha\frac{r}{n_L}\sum_{\ell\in[r]}\sum_{j\in[n_L]} \rho\pi_\ell\eta_\ell(x_j)\mathbf{1}[\hat{y}(x_j)\neq\ell]+(1-\alpha) \\
&=\alpha(1+\rho)\frac{1}{n_L}\sum_{\ell\in[r]}\sum_{i\in[n_L]} \eta_\ell(x_i)\mathbf{1}[\hat{y}(x_i)\neq\ell]+(1-\alpha) \\
&\leq\alpha(1+\rho)\delta_s+(1-\alpha),
\end{aligned}
\tag{61}
$$

where we denote $\pi_\ell = \mathrm{P}(Y=\ell)$ and by class-balance, $\pi_\ell = \frac{1}{r}$. Then combining (58), (59) and (61), we have

$$
\phi^{\hat{y}} \leq 2(1-\theta)\delta_u + \theta\alpha(1+\rho)\delta_s + \theta(1-\alpha) = 2\delta_u + [\alpha(1+\rho)\delta_s - 2\delta_u + (1-\alpha)]\theta. \tag{62}
$$

Therefore, by (57), we have

$$
\mathcal{E} := \mathrm{P}_{\bar{x}\sim\mathcal{P}_{\bar{X}},x\sim\mathcal{A}(\cdot|\bar{x})}\Big(g_{f^*_{\mathrm{pop}},B^*}(x)\neq y(\bar{x})\Big) \leq \frac{2\big[2\delta_u + [\alpha(1+\rho)\delta_s - 2\delta_u + (1-\alpha)]\theta\big]}{1-\lambda_{k+1}} + 8\delta_u, \tag{63}
$$

Combined with Proposition 5.2, we have for $k \leq n_U$, there holds

$$
\mathcal{E} \leq \frac{2\big[2\delta_u + [\alpha(1+\rho)\delta_s - 2\delta_u + (1-\alpha)]\theta\big]}{1-\theta-(1-\theta)\nu_{k+1}} + 8\delta_u, \tag{64}
$$

for $n_U + 1 \leq k \leq n_U + r - 1$, there holds

$$
\mathcal{E} \leq \frac{2\big[2\delta_u + [\alpha(1+\rho)\delta_s - 2\delta_u + (1-\alpha)]\theta\big]}{1-\min\{\theta+(1-\theta)\nu_{k+1}, \theta\alpha+(1-\theta)\nu_{k-n_U}\}} + 8\delta_u, \tag{65}
$$

and for $k \geq n_U + r$, there holds

$$
\mathcal{E} \leq \frac{2\big[2\delta_u + [\alpha(1+\rho)\delta_s - 2\delta_u + (1-\alpha)]\theta\big]}{1-\min\{\theta+(1-\theta)\nu_{k+1}, \theta\alpha+(1-\theta)\nu_{k-n_U}, (1-\theta)\nu_{k+1-r-n_U}\}} + 8\delta_u. \tag{66}
$$

□

Proof. By taking $\gamma = 0$, i.e. $\alpha = 1$, and $k \leq n_U$ in Proposition A.1, we reach the bound in (20).

Besides, by Proposition 5.2, we have

$$\begin{aligned}\|B^*\|_F &\leq 1/\lambda_k \\ &\leq 1/\max\left\{\theta + (1-\theta)\nu_{n_L+k-1}, (1-\theta)\nu_k, \theta\alpha + (1-\theta)\nu_{n_L+k-r}\right\} \\ &\leq 1/\max\left\{\theta + (1-\theta)\nu_{n_L+k-1}, (1-\theta)\nu_k, \theta + (1-\theta)\nu_{n_L+k-r}\right\} \\ &\leq 1/\max\left\{(1-\theta)\nu_k, \theta + (1-\theta)\nu_{n_L+k-r}\right\}.\end{aligned}$$

□

A.5. Proof of Theorem 5.5

Proof. By taking $n_U = 0$ and $k > r$ in Proposition A.1, we reach the bound in (21).

Besides, by Proposition 5.2, we have

$$\|B^*\|_F \leq 1/\lambda_k \leq 1/(1-\theta)\nu_k.$$

□

A.6. Proof of Theorem 5.6

To prove Theorem 5.6, we first derive the following generalization bound for spectral contrastive pretraining under label noise inspired by the proofs of Theorem 4.1 in HaoChen et al. (2021). We note that the form seems to be the same as that of Theorem 4.1 in HaoChen et al. (2021), but the underlying distribution of positive samples is different since we include the label similarity into consideration.

Proposition A.2. *For some $\kappa > 0$, assume $\|f\|_\infty \leq \kappa$ for all $f \in \mathcal{F}$ and $x \in \mathcal{X}$. Let $f^*_{pop} \in \mathcal{F}$ be a minimizer of the population loss $\mathcal{L}(f)$. Given a random dataset of size n, let $\hat{f}_{emp} \in \mathcal{F}$ be a minimizer of empirical loss $\widehat{\mathcal{L}}_n(f)$. Then, with probability at least $1-\varepsilon$ over the randomness of data, we have*

$$\mathcal{L}(\hat{f}_{emp}) \leq \mathcal{L}(f^*_{pop}) + c_1 \cdot \widehat{R}_{n/2}(\mathcal{F}) + c_2 \cdot \left(\sqrt{\frac{\log 2/\delta}{n}} + \varepsilon\right), \tag{67}$$

where constants $c_1 \lesssim k^2\kappa^2 + k\kappa$ and $c_2 \lesssim k\kappa^2 + k^2\kappa^4$.

Proof. Inspired by the proofs of Theorem D.8 in HaoChen et al. (2021), for $\hat{f} \in \arg\min_{f:\mathcal{X}\to\mathbb{R}^k} \widehat{\mathcal{L}}_n(f)$ such that $\mathcal{L}(\hat{f}) \leq \min_{f:\mathcal{X}\to\mathbb{R}^k} \mathcal{L}(f) + \epsilon_0$, there holds

$$\widehat{\mathcal{E}} \leq \min_{1\leq k'\leq k}\left(\frac{2\phi^{\hat{y}}}{1-\lambda_{k'+1}} + \frac{4k'\epsilon_0}{\left(\lambda_{k'} - \lambda_{k+1}\right)^2}\right) + 8\delta_u. \tag{68}$$

According to Proposition 5.2, we have for $n_U = 0$ and $k \geq r$,

$$(1-\theta)\nu_{k+1} \leq \lambda_{k+1} \leq \min\{\theta + (1-\theta)\nu_{k+1}, \theta\alpha + (1-\theta)\nu_k, (1-\theta)\nu_{k+1-r}\}. \tag{69}$$

By $k' \leq k$, there holds $\lambda_{k'} > \lambda_{k+1}$, and thus we have

$$\lambda_{k'} - \lambda_{k+1} \geq (1-\theta)\nu_{k'} - \min\{\theta + (1-\theta)\nu_{k+1}, \theta\alpha + (1-\theta)\nu_k, (1-\theta)\nu_{k+1-r}\} \tag{70}$$

Here we without loss of generality assume $1 \leq k' \leq k+1-r$ to ensure that (70) ≥ 0. Then pluging (62), (69), and (70) into (68), we finish the proof. □

B. Additional Experiments

B.1. Algorithms for Joint Training

Joint Training for Semi-supervised Learning. For semi-supervised contrastive learning, both labeled and unlabeled data are inputs of L_{SimCLR}, that is, the unsupervised contrastive loss L_{SimCLR} is calculated based on all samples (denoted as $\{x_i\}_{i=1}^n$). The supervised contrastive loss L_{SupCon} is calculated based only on the labeled samples (without loss of generality, we denote the labeled samples as $\{(x_i, y_i)\}_{i=1}^{n_L}$). We show the procedures for semi-supervised contrastive learning in Algorithm 1.

Algorithm 1 Joint Training for Semi-supervised Learning

Input: Labeled data $\{(x_i, y_i)\}_{i=1}^{n_L}$; unlabeled data $\{x_i\}_{i=n_L+1}^{n}$; parameter θ.
Initialize: Encoder f.
repeat
  Compute SupCon loss on the labeled data without using labels $\{(x_i, y_i)\}_{i=1}^{n_L}$;
  Compute SimCLR loss on all data $\{x_i\}_{i=1}^{n}$;
  Update encoder f according to the joint training loss $(1-\theta)\frac{1}{n_L}\sum_{i=1}^{n_L}\mathcal{L}_{\text{SupCon}}(f(x_i), y_i)+\theta\frac{1}{n}\sum_{i=1}^{n}\mathcal{L}_{\text{SimCLR}}(f(x_i))$;
until Converge.
Output: Encoder f.

Note that in Section 4.2, it seems that the unlabeled data is used in computing the SupCon loss, but this is not true. We include the unlabeled samples in the label similarity graph only for the convinience of mathematical formulations. In fact, as we view the unlabeled samples as having unique class labels, this formulation does not affect the selection of positive pairs and therefore does not affect the calculation of the SupCon loss.

Joint Training for Noisy Label Learning. For joint training contrastive learning with label noise, the L_{SupCon} loss is calculated based on the noisy-labeled samples $(x_i, \tilde{y}_i)_{i=1}^n$, and the L_{SimCLR} loss is calculated based on all samples without using their labels. We show the procedures for joint training contrastive learning under label noise in Algorithm 2.

Algorithm 2 Joint Training for Noisy Label Learning

Input: Noisy labeled data $\{(x_i, \tilde{y}_i)\}_{i=1}^{n}$; parameter θ.
Initialize: Encoder f.
repeat
  Compute SupCon loss on the (noisy) labeled data $\{(x_i, \tilde{y}_i)\}_{i=1}^{n}$;
  Compute SimCLR loss on all data without using labels $\{x_i\}_{i=1}^{n}$;
  Update encoder f according to the joint training loss $(1-\theta)\frac{1}{n}\sum_{i=1}^{n}\mathcal{L}_{\text{SupCon}}(f(x_i), \tilde{y}_i)+\theta\frac{1}{n}\sum_{i=1}^{n}\mathcal{L}_{\text{SimCLR}}(f(x_i))$;
until Converge.
Output: Encoder f.

B.2. Algorithmic Comparisons with Similar Methods

Islam et al. (2021) and Chen et al. (2022) adopt similar methods as ours, but focus on different aspects. Specifically, Islam et al. (2021) and Chen et al. (2022) empirically investigate the transferability of the joint training of SimCLR ad SupCon, whereas ours focuses on theoretically analyzing the performance of their joint training under linear probing. We discuss the connections and differences as follows.

Islam et al. (2021) finds that the combination of SimCLR and SupCon significantly improves transfer learning performance over Cross-Entropy. Chen et al. (2022) investigates the problems of coarse-to-fine transfer learning by adding a weighted class-conditional InfoNCE loss and a class-conditional autoencoder to SupCon. Similar to our results, according to Tables 2 and 3 of Islam et al. (2021) and Table 3 of Chen et al. (2022), the transfer learning performance of the combination of SupCon and InfoNCE (or class-conditional InfoNCE) is comparable to but has no significant improvement over the winner of SupCon and SimCLR. In addition, the robustness investigated in these two papers is conceptually different from ours. The robustness in Islam et al. (2021) is to *image corruptions* and the robustness in Chen et al. (2022) measures *how well an algorithm can recover hidden subgroups in an unsupervised setting*, whereas our manuscript investigates the robustness to *label corruptions* (label noise).

B.3. Transfer Learning Performances

We conduct additional experiments to evaluate the representation quality with the transfer learning performance on downstream tasks including detection, segmentation, and fine-tuning. We evaluate transfer learning performance of models pretrained on TinyImageNet-200 with SimCLR, SupCon, and Mix. All models are pretrained under the settings of Section 6.1. For Mix, we set $\theta = 0.2$. For SupCon and Mix, the models are pretrained under noise rate $\gamma \in \{20\%, 40\%, 60\%, 80\%\}$. We list the results as follows. The best results are marked in **bold** and the second best marked in underline.

Detection. Object detection is fine-tuned on PASCAL VOC 07+12 dataset. The detector is Faster R-CNN. All models are fine-tuned for 12 epochs. We evaluate the models by the default VOC metric of AP50.

Table 2: Performance comparisons of object detection.

γ	20%	40%	60%	80%
SimCLR	65.3	65.3	65.3	65.3
SupCon	**68.5**	**66.9**	**66.2**	**67.0**
Mix	66.9	65.5	63.8	63.2

Segmentation. Segmentation is fine-tuned on Pascal VOC 12 dataset. The model is DeeplabV3. All models are fine-tuned for 20,000 iterations. We evaluate the models by mean IoU.

Table 3: Performance comparisons of segmentation.

γ	20%	40%	60%	80%
SimCLR	36.61	36.61	36.61	36.61
SupCon	**49.07**	**47.58**	**46.57**	**42.47**
Mix	40.33	38.14	33.66	34.33

Fine-tuning. We finetune the pretrained models on the labeled CIFAR-10 data. The weights of the linear classifier used to fine-tune the encoder network are initialized to zero. On CIFAR-10, models are fine-tuned for 90 epochs. All results are reported on the standard CIFAR-10 test set and are summarized in the following table.

Table 4: Performance comparisons of fine-tuning.

γ	20%	40%	60%	80%
SimCLR	68.01	**68.01**	**68.01**	**68.01**
SupCon	**71.35**	67.46	61.62	49.03
Mix	70.01	66.89	63.86	55.38

As shown in the Tables 2, 3, and 4, the performance of joint training (Mix) is no better than the winner of SimCLR and InfoNCE across all noise rates, which is similar to the conclusion on linear probing discussed in Section 6.4.