CHAOS IS A LADDER: A NEW UNDERSTANDING OF CONTRASTIVE LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Recently, contrastive learning has risen to be a promising approach for large-scale self-supervised learning. However, theoretical understanding of how it works is still unclear. In this paper, we propose a new guarantee on the downstream performance without resort to the conditional independence assumption that is widely adopted in previous work but hardly holds in practice. Our new theory hinges on the insight that different samples from the same class could be bridged together with aggressive data augmentations, thus simply aligning the positive samples (augmented views of the same sample) could make contrastive learning cluster intra-class samples together. We also show that our theory aligns well with existing contrastive methods on both synthetic and real-world datasets. Our work suggests an alternative understanding of contrastive learning: the role of aligning positive samples is more like a surrogate task than an ultimate goal, and it is the overlapping augmented views (i.e., the chaos) that create a ladder for contrastive learning to gradually learn class-separated representations.

1 INTRODUCTION

Contrastive Learning (CL) emerges to be a promising paradigm for learning data representations without labeled data (Oord et al., 2018; Hjelm et al., 2019). Recently, it has achieved impressive results and gradually closed the gap between supervised and unsupervised learning, hopefully leading to a new era that resolves the hunger for labeled data in the deep learning field (He et al., 2020; Chen et al., 2020b). However, despite its intriguing empirical success, a theoretical understanding of how contrastive learning actually works in practice is still under-explored.

The general methodology of contrastive learning is quite simple, that is to maximize the similarity between augmented views of the same image (*a.k.a.* positive samples), and minimize the similarity between that of two random images (*a.k.a.* negative samples). Intuitively, it is *an instance discrimination task* (differing each image from others) instead of *a classification task* (clustering images from the same class together and differing with other classes). Nevertheless, as shown in Figure 1(a), CL representations are also class-separated. Therefore, understanding how the pretraining (CL) task and the downstream (classification) task interact plays a central role in both theoretical understandings and practical designings of contrastive methods.

Previously, Saunshi et al. (2019) and Lee et al. (2020) have tried to establish guarantees on the classification performance for self-supervised representations. However, their analysis relies heavily on the assumption that the two positive samples, as augmented views of the same image, are (nearly) conditionally independent on class y. However, this is hardly practical as the augmented views are still strongly input-dependent (see Figure 1(b)). In fact, if the conditional independence is satisfied, the unsupervised task will become as informative as the supervised task, making this discussion almost unnecessary. This motivates us to find more practical and weaker assumptions to understand how contrastive learning actually works (even without conditional independence). To achieve this, we need to reexamine the contrastive learning process. Previously, Wang & Isola (2020) show that CL objective involves two goals: alignment (for positive samples) and uniformity (for negative samples). Nevertheless, we show that there exist bad cases where the features could still have poor performance even with perfect alignment and uniformity. Thus, contradictory to the common belief of contrastive learning as learning invariance, we note that invariance alone is inadequate to learning useful representations for downstream tasks.

In this paper, we provide a novel understanding of contrastive learning that requires only practical and minimal assumptions, while also guarantee class-separated representations. Our core insight

(a) Contrastive learning learns clustered features.

(b) Intra-class samples are more alike via some augmented views.

Figure 1: (a) t-SNE visualization of representations before and after contrastive learning with SimCLR (Chen et al., 2020a) on CIFAR-10. Each point denotes a sample and its color denotes its class. (b) Applying aggressive data augmentations as in SimCLR (Chen et al., 2020a) to four images from ImageNet (two are cars and two are pens). The 1st column shows the raw (center-cropped) images and the 2-5th colums show the augmented ones.

hinges on the observation that contrastive learning usually adopts much more aggressive data augmentations than that in supervised learning (He et al., 2020; Chen et al., 2020a). As shown in Figure 1(b), we notice that aggressive random cropping of two images can generate views that are very much alike that we could even hardly tell them apart, *e.g.,* the wheels of two different cars. In other words, there will be connections between different intra-class images through aggressively augmented views of them. Thus, the alignment of positive samples will also cluster all the intra-class samples together, and lead to class-separated representations. From our perspective, the role of data augmentation is to create a certain degree of "chaos" between intra-class samples, and the role of contrastive loss is to "climb the ladder of chaos", *i.e.,* the process that we gradually cluster intra-class samples by aligning positive samples.

Following this intuition, we develop a new theory for understanding the effectiveness of contrastive learning. Specifically, we derive the upper and lower bounds for its downstream performance and show how the two will asymptotically converge with our assumptions on the support overlapping effect of data augmentations. Driven by this analysis, we further discuss how varying augmentation will affect the performance of contrastive learning from both synthetic and real-world datasets, and show that the results align well with our theory. In summary,

- We characterize the failure of the previous analysis of contrastive learning, and develop a new understanding through the support overlapping effect of data augmentations. Compared to existing theories on contrastive learning, ours can provide guidance to the practical designing of contrastive methods and evaluation metrics.
- We establish general guarantees (both upper and lower bounds) for the downstream performance without assumptions on conditional independence. And we further show how the two bounds could asymptotically converge under our less restrictive assumptions.
- We provide a quantitative discussion on the effect of augmentation strength, which verifies our theory from both theoretical and empirical aspects. Motivated by our theory, we further propose a new unsupervised evaluation metric for contrastive learning named ARC and show that it aligns well with downstream performance on real-world datasets.

2 Related Work

Contrastive Learning in Practice. Contrastive self-supervised learning originates from a mutual information perspective of representation learning (Oord et al., 2018; Hjelm et al., 2019), and soon becomes a general learning paradigm that contrasts between positive and negative pairs (He et al., 2020; Chen et al., 2020a). It is rapidly closing the performance gap between unsupervised and supervised learning on large-scale dataset like ImageNet (Chen et al., 2021), and outperforms supervised learning when combined with a few (*e.g.,* 10%) labels (Chen et al., 2020b). Several recent works show that similar performance could be achieved without negative samples by adopting certain training techniques (Grill et al., 2020; Chen & He, 2020).

Understanding Contrastive Learning Objectives. Both the original InfoNCE loss (Oord et al., 2018) and its InfoMax variants (Hjelm et al., 2019; Poole et al., 2019) are designed as variational estimates of the mutual information between inputs and representations, but these estimators are shown to have poor bias-variance trade-offs (Song & Ermon, 2020). Instead, Wang & Isola (2020) simply understand contrastive learning through the two terms in the InfoNCE loss: alignment of positive samples and uniformity of negative samples. However, as we show later, this perspective is also insufficient to explain the effectiveness of contrastive learning, and we should take the interplay between augmentation and alignment into consideration.

Understanding Downstream Generalization. Saunshi et al. (2019) propose the first theoretical guarantees by bridging the contrastive and classification objectives. Lee et al. (2020) further link the reconstruction-based objective to the downstream objective. However, both Saunshi et al. (2019) and Lee et al. (2020) rely on the unrealistic assumption that the positive samples are (nearly) conditionally independent. Besides, there are several other perspectives for analyzing contrastive learning, *e.g.,* information-theoretical perspective (Tsai et al., 2021) and spectral graph theory perspective (HaoChen et al., 2021). Nevertheless, their assumptions are usually technical and the analysis does not involve the actual learning process, and thus can hardly provide guides for designing practical contrastive methods. In comparison, our theory is directly built on a new understanding of the contrastive learning process, and we verify the feasibility of each assumption with empirical evidence.

3 Limitations of Previous Understandings

3.1 Problem Setup

We begin by introducing the basic notations and common practice of contrastive learning in the image classification task. In general, it has two stages, unsupervised pretraining, and supervised finetuning. In the first stage, with N unlabeled samples $\mathcal{D}_u = \{x_i\}_{i=1}^N$, we pretrain an encoder mapping from the d-dimensonal input space to a unit hypersphere $f \in \mathcal{F} : \mathbb{R}^d \to \mathbb{S}^{m-1}$ in the m-dimensional space. In the second stage, we evaluate the learned representations z with the labeled data $\mathcal{D}_l = \{(x_i, y_i)\}$ where labels $y_i \in \{1, \dots, K\}$. Specifically, we fix the encoder and learn a linear classification head $g : \mathcal{R}^m \to \mathcal{R}^K$ with the representations, *i.e.,* $\tilde{\mathcal{D}}_l = \{(z, y) | z = f(x) \in \mathcal{R}^m\}$.

Contrastive Pretraining. Taking a training example $x \in \mathcal{D}_u$, we draw its positive sample $x^+ = t(x)$ by applying a random data augmentation $t \sim \mathcal{T}$, and draw M randomly augmented samples $\{x_i^-\}_{i=1}^M$ from $\mathcal{D}_u$ as its negative samples. Then, we can learn the encoder f with the widely used InfoNCE loss (Oord et al., 2018)

$$\begin{aligned}\mathcal{L}_{\text{NCE}}(f) =& \mathbb{E}_{p(x,x^+)} \mathbb{E}_{\{p(x_i^-)\}} \left[-\log \frac{\exp(f(x)^\top f(x^+))}{\sum_{i=1}^M \exp(f(x)^\top f(x_i^-))} \right] \\ =& \mathbb{E}_{p(x)} \left[-\mathbb{E}_{p(x^+|x)} f(x)^\top f(x^+) + \mathbb{E}_{\{p(x_i^-)\}} \log \sum_{i=1}^M \exp(f(x)^\top f(x_i^-)) \right].\end{aligned} \tag{1}$$

Let $p(x)$ be the data distribution, $p(x, x^+)$ be the joint distribution of positive pairs, and we simply assume $p(x, x^+) = p(x^+, x)$ and $p(x) = \int p(x, x^+) dx^+, \forall\, x \in \mathbb{R}^d$ following Wang & Isola (2020).

Linear Evaluation. To evaluate the learned representations by contrastive learning, we usually adopt the Cross Entropy (CE) loss (Chen et al., 2020a) for a labeled pair $(x, y) \in \mathcal{D}_l$

$$\mathcal{L}_{\text{CE}}(f, g) = \mathbb{E}_{p(x,y)} \left[-\log \frac{\exp\left(f(x)^\top w_y\right)}{\sum_{i=1}^K \exp\left(f(x)^\top w_i\right)} \right], \tag{2}$$

with a linear classifier $g(z) = Wz$ where $W = [w_1, w_2, \dots, w_K]$.

3.2 Existing Theoretical Assumptions and Their Limitations

As discussed above, there are some previous understandings on how contrastive learning yields good performance, and they mainly differ by their theoretical assumptions.

First, Wang & Isola (2020) interpret the first and second terms of the InfoNCE loss (Eq. 1) as they are aiming at the following two properties: 1) alignment (the 1st term): positive samples x, x^+ has

similar features, *i.e.,* $f(x) \approx f(x^+)$; 2) uniformity (the 2nd term): features are roughly uniformly distributed in the unit hypersphere $\mathbb{S}^{m-1}$. In particular, they show that InfoNCE can be minimized with 1) perfect alignment and 2) perfect uniformity. However, as we illustrate in Figure 2, the features could still have very poor downstream performance in the finite sample scenario.

Proposition 3.1 (Class-uniform features also minimize the InfoNCE loss)**.** *For N training examples of K classes, consider the case when features $\{f(x_i)\}_{i=1}^N$ are randomly distributed in $\mathbb{S}^{m-1}$ with maximal uniformity while also satisfying $\forall x_i, x_i^+ \sim p(x, x^+), f(x_i) = f(x_i^+)$. Because we have perfect alignment and perfect uniformity, the InfoNCE loss achieves its minimum. However, the downstream classification accuracy is at most $1/K + \varepsilon$ and ε is nearly zero when N is large enough.*

This indicates that the instance discrimination task (alignment + uniformity) alone cannot guarantee to learn class-discriminative features as desired in the downsteam classification. Instead, Saunshi et al. (2019) and Lee et al. (2020) both establish the relationship between pretraining and classification objectives and provide guarantees for the downstream performance. In fact, the two works both assume the conditional independence of the two positive samples, *i.e.,* $p(x, x^+|y) = p(x|y)p(x^+|y)$. However, this assumption is too strong as it is hardly practical. As shown in Figure 1(b), augmented views from the same class are not actually independent as views from the same sample are more alike than that from other samples.

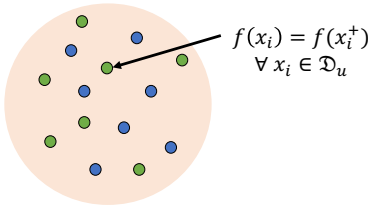


Figure 2: An illustration of the case when contrastive learning fails to learn class-separated features even if the features are uniformly distributed and positive samples are perfect aligned. Each color denotes a class.

4 A New Theory for Contrastive Learning

The analysis above motivates us to find a *minimal* and *practical* assumption: 1) it is enough to guarantee good performance on downstream tasks; 2) it is less restrictive than the i.i.d. assumptions as in Saunshi et al. (2019) and Lee et al. (2020).

4.1 Gap Between Contrastive Learning and Downstream Classification

We start with an assumption on the label consistency between positive samples, that is, any pair of positive samples (x, x^+) should belong to the same class.

Assumption 4.1 (Label Consistency)**.** *$\forall\, x, x^+ \sim p(x, x^+)$, we have $p(y|x) = p(y|x^+)$.*

This is a natural and minimal assumption that is likely to hold in practice. As shown in Figure 1(b), the widely adopted augmentations in contrastive learning (Chen et al., 2020a) like images cropping, color distortion, and horizontal flipping will hardly alter the belonging image classes.

With this minimal assumption, we can characterize the generalization gap between unsupervised and supervised learning risks. We first introduce the mean CE loss, $L_{\mathrm{CE}}^{\mu}(f) = \mathbb{E}_{p(x,y)}\left[-\log \frac{\exp\left(f(x)^\top \mu_y\right)}{\sum_{i=1}^K \exp(f(x)^\top \mu_i)}\right]$, where we use the classwise *mean representation* $\mu_k = \mathbb{E}_{p(x|y=k)}[f(x)]$ as the weight w_k of the classifier g. It is easy to see that the mean CE loss upper bounds the CE loss, *i.e.,* $L_{\mathrm{CE}}^{\mu}(f) \geq \min_g \mathcal{L}_{\mathrm{CE}}(f, g)$ and Saunshi et al. (2019) showed that the mean classifier could achieve comparable performance to learned weights. Then, we have the following upper and lower bounds on the downstream risk (measured by mean CE loss).

Theorem 4.2 (Guarantees for general encoders)**.** *If Assumption 4.1 holds, then, for any $f \in \mathcal{F}$, its downstream classification risk $\mathcal{L}_{\mathrm{CE}}^{\mu}(f)$ can be upper and lower bounded by the contrastive learning risk $\mathcal{L}_{\mathrm{NCE}}(f)$*

$$\begin{aligned} &\mathcal{L}_{\mathrm{NCE}}(f) - \sqrt{\mathrm{Var}(f(x) \mid y)} - \frac{1}{2} \cdot \sum_{j=1}^{m} \sqrt{\mathrm{Var}(f_j(x) \mid y)} - \mathcal{O}\left(M^{-1/2}\right) \\ &\leq \mathcal{L}_{\mathrm{CE}}^{\mu}(f) + \log(M/K) \leq \mathcal{L}_{\mathrm{NCE}}(f) + \sqrt{\mathrm{Var}(f(x)|y)} + \mathcal{O}\left(M^{-1/2}\right), \end{aligned} \tag{3}$$

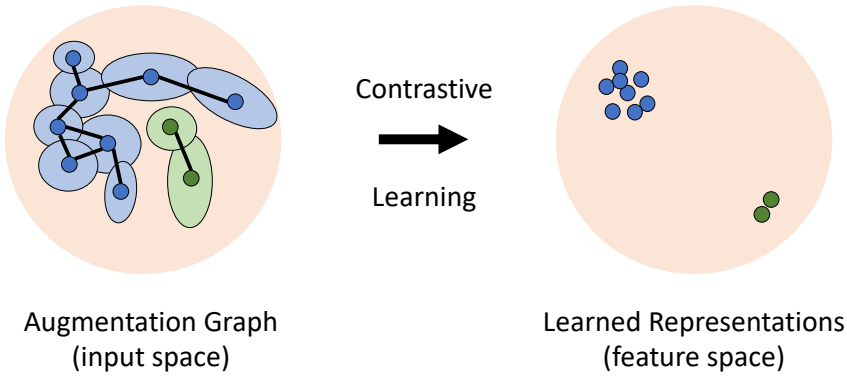


Figure 3: Left: an illustration of a augmentation graph satisfying inter-class connectivity. Each dot denotes a sample $x \in \mathcal{D}_u$ and its color denotes its class. The lighter disks denote the support of the positive samples $p(x^+|x)$. We draw a solid edge for each $\mathcal{T}$-connected pair. The graph is classwise connected. Right: contrastive learning with the left augmentation graph will learn to cluster intra-class samples together (alignment) while pushing inter-class samples away (uniformity).

where $\log(M/K)$ *is a constant* [1]*,* $\mathcal{O}\left(M^{-1/2}\right)$ *denotes the order of the approximation error by using* M *negative samples, and* $\mathrm{Var}(f(x)|y) = \mathbb{E}_{p(y)}\left[\mathbb{E}_{p(x|y)}\|f(x) - \mathbb{E}_{p(x|y)}f(x)\|^2\right]$ *denotes the conditional variance.*

Remark. Our generalization bounds extend the analysis in Saunshi et al. (2019) in several aspects: 1) we do not require the assumption on conditional independence; 2) we directly analyze the widely adopted InfoNCE loss (for contrastive learning) and CE loss (for supervised finetuning), while Saunshi et al. (2019) are restricted to hinge and logistic objectives that have worse performance in practice (Chen et al., 2020a); 3) the class collision terms[2] that are incompressible in Saunshi et al. (2019) now disappear in our bounds by adopting the InfoNCE loss, which also explains why InfoNCE performs better in practice; and 4) the bounds in Saunshi et al. (2019) will become looser with more negative samples, which is contradictory to the common practice (Chen et al., 2020a). While for our bounds, increasing M indeed leads to a lower approximation error and helps close the gap.

Notably, the theorem above shows that Assumption 4.1 alone is still insufficient to guarantee good downstream performance. As there are two intra-variance terms in the upper and lower bounds, $\mathrm{Var}(f(x)|y)$ and $\mathrm{Var}(f_j(x)|y)$. The contrastive learning might still have inferior performance with intra-class variance as in Proposition 3.1. In the next part, we will present a new understanding of how contrastive learning could control this variance term.

4.2 Closing the Gap with Intra-class Connectivity

The theorem above motivates us to study how contrastive learning could effectively control its intra-class variance and learn class-separated features. Here, we propose a new understanding of this clustering ability through a dissection of the augmented views. In particular, we notice that although samples are different from each other, applying aggressive augmentations like that in SimCLR (Chen et al., 2020a) can largely make them more alike. For example, in Figure 1(b), two different cars become very similar when they are both cropped to the wheels. Then, with contrastive learning, the two cars will have closer representations as they share a common view of the wheels. In other words, two different intra-class samples could be aligned together if they have overlapping augmented views. If all intra-class samples could be bridged by data augmentations, we can successfully cluster the whole class together. Below, we formalize the intuition above with the language of graphs.

Notations. A graph $\mathcal{G}$ is represented by a tuple $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where $\mathcal{V} = (v_1, v_2, \dots, v_N)$ is a set of vertices and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ is a set of edges. A path is a sequence of edges that joins a sequence of vertices, *e.g.,* $v_{i_1} - v_{i_2} - \cdots - v_{i_k}$. We say that two vertices v and u are connected if $\mathcal{G}$ contains a path from v to u. A graph is said to be connected if every pair of vertices in the graph is connected. Two graphs are said to be disjoint if any pair of inter-graph vertices are not connected.

[1] $\log(M/K)$ could be absorbed in the loss functions by replacing sum with mean in InfoNCE and CE.

[2] Error terms introduced because there are samples from the same class as x in the negative samples $\{x_i^-\}$.

To begin with, we define the concept of $\mathcal{T}$-connectivity of sample pairs, which describes whether two samples could be connected via the support overlapping of their augmented views.

Definition 4.3 ($\mathcal{T}$-connectivity)**.** *Given a collection of augmentations $\mathcal{T} = \{t \mid t : \mathbb{R}^d \to \mathbb{R}^d\}$, we say that two different images $x_i, x_j \in \mathbb{R}^d$ are $\mathcal{T}$-*connected *if they have overlapping views: $\mathrm{supp}(p(x_i^+|x_i)) \bigcap \mathrm{supp}(p(x_j^+|x_j)) \neq \varnothing$, or equivalently, $\exists\, t_i, t_j \in \mathcal{T}$ such that $t_i(x_i) = t_j(x_j)$.*

Then, we can define an augmentation graph of all training samples in terms of their $\mathcal{T}$-connectivity.

Definition 4.4 (Augmentation Graph)**.** *Given a set of N samples $\mathcal{D} = \{x_i\}_{i=1}^N$ and an augmentation set $\mathcal{T} = \{t \mid t : \mathbb{R}^d \to \mathbb{R}^d\}$, we can define an augmentation graph $\mathcal{G}(\mathcal{D}, \mathcal{T}) = (\mathcal{V}, \mathcal{E})$ as*

- *we take the N natural samples as the vertices of the graph, i.e., $\mathcal{V} = \{x_i\}_{i=1}^N$;*
- *there exists an edge e_{ij} between two vertices x_i and x_j if they are $\mathcal{T}$-connected.*

Based on these concepts, we introduce the following assumption that with a proper choice of data augmentations, all intra-class samples could form a connected graph, as depicted in Figure 3.

Assumption 4.5 (Intra-class Connectivity)**.** *Given a training set $\mathcal{D}_u$, there exists an appropriate augmentation set $\mathcal{T}$ such that the augmentation graph $\mathcal{G}(\mathcal{D}_u, \mathcal{T})$ is class-wise connected, i.e., $\forall\, k \in \{1, \dots, K\}$, the subgraph $\mathcal{G}_k$ (graph $\mathcal{G}$ restricted to vertices in class k) is connected.*

Comparing to Saunshi et al. (2019) and Lee et al. (2020) that require (nearly) conditional independence $p(x, x^+|y) = p(x|y)p(x^+|y)$, while ours only requires the connectivity of intra-class samples as in Figure 1(b), and does not need them to be conditionally independent.

To make this analysis simpler, we need another assumption on the alignment term that by minimizing the InfoNCE loss we can align the positive samples perfectly (empirical evidence in Appendix C).

Assumption 4.6 (Perfect Alignment)**.** *At the minimizer $f^\star$ of the InfoNCE loss, we can achieve perfect alignment, i.e., $\forall\, x, x^+ \sim p(x, x^+), f^\star(x) = f^\star(x^+)$.*

Proposition 4.7. *Under Assumptions 4.5 & 4.6, by minimizing the InfoNCE loss we can conclude that the conditional variance term vanishes, i.e.,*

$$\mathrm{Var}(f(x) \mid y) = \mathrm{Var}(f_j(x) \mid y) = 0,\ \forall j \in \{1, 2, \dots, m\}. \tag{4}$$

Intuitively, for samples in each class k, if the corresponding subgraph $\mathcal{G}_k$ is connected, there exists a path connecting every intra-class pairs (x_i, x_j), as shown in Figure 3. Consequently, aligning the positive pairs will also align all samples on the path, and eventually align x_i and x_j. In this way, all intra-class samples can be clustered together and the intra-class variance shrinks to zero (under Assumption 4.6). Besides, because proper data augmentation will not cause inter-class support overlap (Assumption 4.1), inter-class samples can be well separated with the uniformity term. As a result, we can attain alignment of intra-class samples while maximizing the uniformity of inter-class samples. According to Theorem 4.2, we will have an asymptotically closed generalization gap (with more negative samples $M \to \infty$) for the encoder that minimizes the contrastive loss.

Theorem 4.8 (Guarantees for the optimal encoder)**.** *If Assumption 4.1, 4.5 & 4.6 hold and f is L-smooth, then, for the minimizer $f^\star = \arg\min \mathcal{L}_{\mathrm{NCE}}(f)$, its classification risk can be upper and lower bounded by its contrastive risk as*

$$\mathcal{L}_{\mathrm{NCE}}(f^\star) - \mathcal{O}\left(M^{-1/2}\right) \leq \quad \mathcal{L}_{\mathrm{CE}}^{\mu}(f^\star) + \log(M/K) \leq \mathcal{L}_{\mathrm{NCE}}(f^\star) + \mathcal{O}\left(M^{-1/2}\right). \tag{5}$$

We note that different to previous bounds that hold for any $f \in \mathcal{F}$ as in Theorem 4.2, our results here only stand for the minimizer of the contrastive loss $f^\star$. This indicates that the InfoNCE loss alone cannot simply guarantee good downstream performance, and the learning dynamics matters for the contrastive learning to learn useful features.

4.3 Discussions

Our analysis above suggests a new understanding of the role of data augmentations in contrastive learning. Conventionally, the success of contrastive learning is usually contributed to learning invariance *w.r.t.* various data augmentations by matching positive examples. However, as shown in

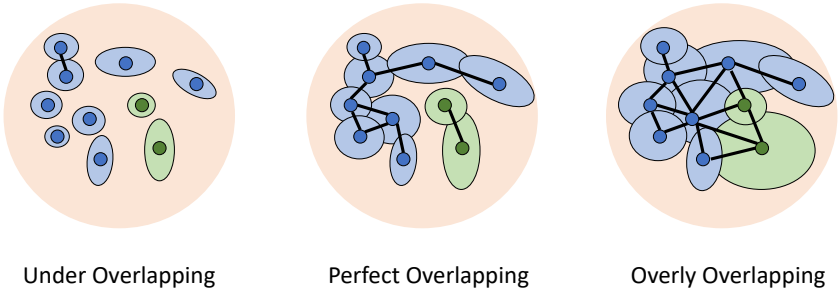


Figure 4: Augmentations graph with increasing augmentation strength (left to right). Each dot denotes a sample $x \in \mathcal{D}_u$ and its color denotes its class. The lighter disks denote the support of the positive samples $p(x^+|x)$. We draw a solid edge for each $\mathcal{T}$-connected pair.

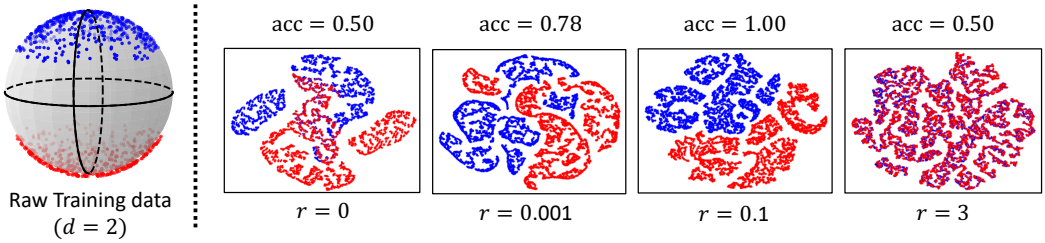


Figure 5: t-SNE visualization of features learned with different augmentation strength r on the random augmentation graph experiment. Each dot denotes a sample and its color denotes its class.

Proposition 3.1, matching positive pairs is theoretically inadequate to learn useful features. Indeed, assume that we have an ideal encoder that possesses invariance *a priori* as desired, like invariance to translation (CNNs), rotation (Cheng et al., 2016), and scaling (Xu et al., 2014), do we obtain class-discriminative features simply by random initialization? NO, since these low-level properties are independent of the high-level class information that we want to learn. Thus, the reason why contrastive learning works cannot simply be contributed to the invariance learning principle.

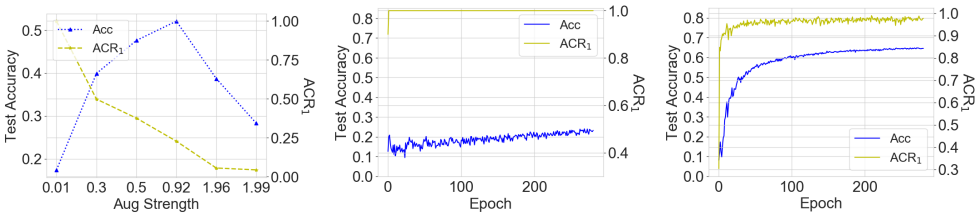
We instead believe that the role of data augmentation is to create a certain degree of "chaos" between different intra-class samples (Figure 1(b)) such that they become more alike (or formally, $\mathcal{T}$-connected). In this way, the chaos serves as a "ladder" for bridging intra-class samples together when labels are absent, and the mission of the contrastive loss is to "climb this ladder", that is, aligning intra-class samples by aligning the overlapping positive samples, as shown in Figure 3. Therefore, from our perspective, instance discrimination by contrastive learning is actually a surrogate for the classification task, and the surrogate could complete its misson when the ladder of chaos is complete (or formally, when intra-class connectivity holds).

5 On the Influence of Augmentation Strength

We have shown that with appropriate augmentations, we can derive guarantees on downstream performance. However, in practice, as illustrated in Figure 4, there could be cases where augmentations are either too weak (intra-class features cannot be clustered together as in Figure 2) or too strong (inter-class features will also collapse to the same point) and lead to sub-optimal results. In this section, we provide a quantitative analysis of how different strengths of data augmentation will affect the final performance both theoretically and empirically.

5.1 Characterization on Random Augmentation Graph

In practice, there are various data augmentation types that are hard to be described precisely. For the ease of analysis, we consider a simple case where for each class k, there are N samples uniformly distributed around the cluster center c_k on a hypersphere $\mathbb{S}^d$. We then augment each sample x_i with random samples in a hyper-disk of radius r on the hypersphere.



(a) ACR *v.s.* aug-strength r. (b) ACR while training (r=0.01). (c) ACR while training (r=0.92).

Figure 6: (a) Average Confusion Rate (ACR) and downstream accuracy *v.s.* different augmentation strength (before training). (b,c): ACR and downstream accuracy while training.

In Appendix B, we provide theoretical analysis on how different augmentation strength (measured by r) will affect the connectivity of the augmentation as a function of the number of samples N, the position of the cluster centers c_k and input dimensions d. Here, we show our simulation results by applying contrastive learning to the problem above. From Figure 5, we can see that when $r = 0$ (no augmentation), the features are mixed together and hardly (linearly) separable, which corresponds to the under-overlapping case in Figure 4. As we increase r from 0 to 0.1, the features become more and more discriminative. And when r is too large ($r = 3$), the inter-class features become mixed and inseparable again (overly overlapping). Overall, our theoretical and synthetic discussions verify our theory that intra-class support overlapping with a proper amount of data augmentation is crucial for contrastive learning to work well.

5.2 New Surrogate Metrics for the Support Overlapping

From our theory and analysis above, we see that the support overlapping between intra-class samples indeed matters from contrastive learning to generalize better. Inspired by this, we propose the Confusion Ratio metric as a measure of the degree of support overlapping. Specifically, for an unlabeled dataset $\mathcal{D}_u$ with N samples, we randomly augment each raw sample $x_i \in \mathcal{D}_u$ for C times, and get an augmented set $\widetilde{\mathcal{D}}_u = \{x_{ij}, i \in [N], j \in [C]\}$. Then, for each $x_{ip} \in \widetilde{\mathcal{D}}_u$ that is an augmented view of $x_i \in \mathcal{D}_u$, denoting its k-nearest neighbors in $\widetilde{\mathcal{D}}_u$ in the feature space of f as $\mathcal{N}_k(x_{ip}, f)$ and other augmented views from the same image as $\mathcal{C}(x_{ip}) = \{x_{ij}, j \neq p\}$, we can define its Confusion Ratio (CR) as the ratio of augmented views from the same raw sample in its k-nearest neighbors,

$$\mathrm{CR}_k(x_{ij}, f) = \frac{\#[\mathcal{N}_k(x_{ip}, f) \bigcap \mathcal{C}(x_{ip})]}{\#\mathcal{N}_k(x_{ip}, f)}. \tag{6}$$

We also define its average as Average Confusion Ratio (ACR):

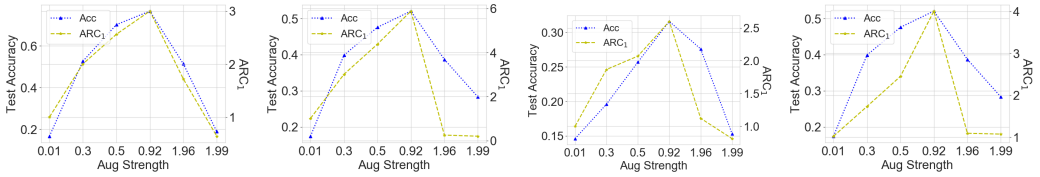
$$\mathrm{ACR}_k(f) = \mathbb{E}_{x_{ij} \sim \widetilde{\mathcal{D}}_u} \mathrm{CR}_k(x_{ij,f}). \tag{7}$$

When support overlapping happens, the nearest neighbors could be augmented views from a different sample, leading to a lower ACR. Thus, ACR measures the degree of support overlapping, and a lower ACR indicates a higher degree of support overlapping. Here we take $k = 1$ by default.

As shown in Figure 6(a), ACR (support overlapping) indeed decreases with the strength of data augmentations, and only a moderate ACR achieves the best accuracy, which is consistent with our theory discussed above. Besides, we also plot the change of ACR along the training process in Figure 6(b) & 6(c). We can notice that for small augmentations, the initial ACR is large, and it rapidly grows to the optimum and seldom changes while training, which leads to poor test accuracy. Instead, with proper augmentations, the initial ACR is lower, and it gradually increases to nearly one and obtains good accuracy. This is also consistent with our theory that we need a certain amount of support overlapping for contrative learning to work well. At the beginning, this will lead to a lower ACR, but as training continues, better alignment (higher ACR) will help bring up the test accuracy.

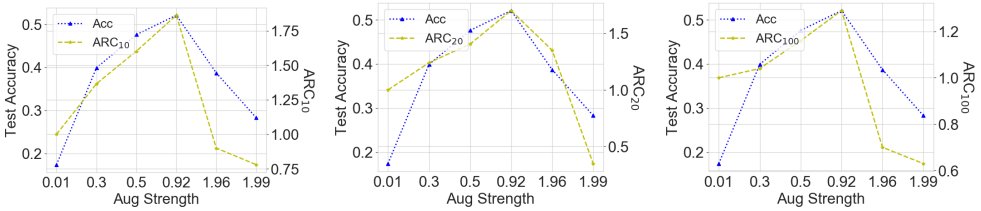
Average Relative Confusion (ARC). In the discussion above, we notice that although ACR itself does not indicate the test accuracy, the relative change of ACR before and after training can be used as such an indicator. A large change of ACR means a large change of support overlapping, which indicates that the contrastive loss can actually cluster intra-class samples together through overlapped views. Based on this observation, we propose Average Relative Confusion (ARC) as

$$\mathrm{ARC}_k = \frac{\mathrm{ACR}_k(f_{\text{final}})}{\mathrm{ACR}_k(f_{\text{init}})}, \tag{8}$$



(a) CIFAR-10 (SimCLR) (b) CIFAR-10 (BYOL) (c) CIFAR-100 (SimCLR) (d) STL-10 (SimCLR)

Figure 7: Average Relative Confusion (ARC) and downstream accuracy *v.s.* different augmentation strength on different datasets (CIFAR-10, CIFAR-100, and STL-10) with different contrastive learning methods: SimCLR (Chen et al., 2020a) and BYOL (Grill et al., 2020).



(a) $k = 10$ (b) $k = 20$ (c) $k = 100$

Figure 8: Average Relative Confusion (ARC) and downstream accuracy *v.s.* different augmentation strength on CIFAR-10 (SimCLR) with different number of nearest neighbors k.

a ratio of ACR with the initialized model f_{init} and the final model f_{final} after contrastive learning. A higher ARC indicates that the contrastive learning process faces a harder task (support overlapping) at the beginning (lower initial ACR), while successfully clustering intra-class samples with good alignment of positive samples at the end (higher final ACR). Therefore, a higher ARC should correspond to higher downstream accuracy.

Here, to measure the augmentation strength in real-world datasets, following the common practice (Chen et al., 2020a), we adopt the $\mathrm{RandomResizedCrop}$ operator with scale range $[a, b]$ for data augmentation, and we define its strength of augmentation as $r = (1 - b) + (1 - a)$. As shown in Figure 7 & 8, as augmentations become stronger, ARC scores align well with the change of downstream accuracy across 1) different datasets, 2) different contrastive methods, and 3) different choices of k. This justifies our understanding of contrastive learning through support overlapping.

Meanwhile, as the calculation of ARC only involves unsupervised data, it could serve as a good surrogate metric for evaluating contrastive learning without using labeled data. Compared to previous evaluation methods like linear classification (Eq. 2), our ARC metric is more preferable as 1) it is theoretically motivated; 2) it does not need labeled data; 3) it does not need to learn additional modules like linear classifiers or rotation tasks (Reed et al., 2021). More experimental details can be found in Appendix D.

6 Conclusion

In this paper, we have proposed a new understanding of contrastive learning through a revisiting of the role of data augmentations. In particular, we notice the aggressive data augmentation applied in contrastive learning can significantly increase the support overlapping between intra-class samples, and as a result, by aligning positive samples, we can also cluster inter-class samples together. Based on this insight, we develop a new theory that could guarantee good downstream performance without relying on conditional independence and obtain asymptotically closed gaps. With this perspective, we also characterize how different augmentation strength affects downstream performance with both random graphs and real-world datasets. Last but not least, we also develop a new surrogate metric for evaluating contrastive learning without labels and show that it aligns well with downstream performance. Overall, we believe that we pave a new way for understanding contrastive learning with insights on the designing of contrastive methods and evaluation metrics.

REFERENCES

Ivan Budimir, Sever S Dragomir, and Josep Pecaric. Further reverse results for jensen's discrete inequality and applications in information theory. *RGMIA research report collection*, 3(1), 2000.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *ICML*, 2020a.

Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. Big self-supervised models are strong semi-supervised learners. *arXiv preprint arXiv:2006.10029*, 2020b.

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. *arXiv preprint arXiv:2011.10566*, 2020.

Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. *arXiv preprint arXiv:2104.02057*, 2021.

Gong Cheng, Peicheng Zhou, and Junwei Han. RIFD-CNN: Rotation-invariant and fisher discriminative convolutional neural networks for object detection. In *CVPR*, 2016.

Jean-Bastien Grill, Florian Strub, Florent Altché, C. Tallec, Pierre H. Richemond, Elena Buchatskaya, C. Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, B. Piot, K. Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning. *NeurIPS*, 2020.

Jeff Z HaoChen, Colin Wei, Adrien Gaidon, and Tengyu Ma. Provable guarantees for self-supervised deep learning with spectral contrastive loss. *NeurIPS*, 2021.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. *CVPR*, 2020.

R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. *ICLR*, 2019.

Jason D Lee, Qi Lei, Nikunj Saunshi, and Jiacheng Zhuo. Predicting what you already know helps: Provable self-supervised learning. *arXiv preprint arXiv:2008.01064*, 2020.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Mathew D. Penrose. A Strong Law for the Longest Edge of the Minimal Spanning Tree. *The Annals of Probability*, 27(1):246 – 260, 1999.

Allon G Percus and Olivier C Martin. Scaling universalities of kth-nearest neighbor distances on closed manifolds. *Advances in Applied Mathematics*, 21(3):424–436, 1998. ISSN 0196-8858.

Ben Poole, Sherjil Ozair, Aaron Van Den Oord, Alex Alemi, and George Tucker. On variational bounds of mutual information. In *ICML*, 2019.

Colorado J Reed, Sean Metzger, Aravind Srinivas, Trevor Darrell, and Kurt Keutzer. Selfaugment: Automatic augmentation policies for self-supervised learning. In *CVPR*, 2021.

Nikunj Saunshi, Orestis Plevrakis, Sanjeev Arora, Mikhail Khodak, and Hrishikesh Khandeparkar. A theoretical analysis of contrastive unsupervised representation learning. In *ICML*, 2019.

Jiaming Song and Stefano Ermon. Understanding the limitations of variational mutual information estimators. In *ICLR*, 2020.

Yao-Hung Hubert Tsai, Yue Wu, Ruslan Salakhutdinov, and Louis-Philippe Morency. Self-supervised learning from a multi-view perspective. In *ICLR*, 2021.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere, 2020.

Yichong Xu, Tianjun Xiao, Jiaxing Zhang, Kuiyuan Yang, and Zheng Zhang. Scale-invariant convolutional neural networks. *arXiv preprint arXiv:1411.6369*, 2014.

A OMITTED PROOFS

A.1 PROOF OF PROPOSITION 3.1

Proposition A.1 (Class-uniform features also minimize the InfoNCE loss)**.** *For N training examples of K classes, consider the case when features $\{f(x_i)\}_{i=1}^N$ are randomly distributed in $\mathbb{S}^{m-1}$ with maximal uniformity while also satisfying $\forall x_i, x_i^+ \sim p(x, x^+), f(x_i) = f(x^i)$. Because we have perfect alignment and perfect uniformity, the InfoNCE loss achieves its minimum. However, the downstream classification accuracy is at most $1/K + \varepsilon$ and ε is nearly zero when N is large enough.*

Proof. Because the features $\{f(x_i)\}_{i=1}^N$ are uniformly distributed in $\mathbb{S}^{m-1}$, according to the law of large number, for any measurable set $\mathcal{U} \in \mathbb{S}^{m-1}$, when N is large enough, there will be almost equal size of features from each class in $\mathcal{U}$. Consequently, any classifier g that classifies $\mathcal{U}$ to class k will only have $1/K$ accuracy asymptotically. □

A.2 PROOF OF THEOREM 4.2

We will prove the upper and lower bounds separately as follows.

A.2.1 THE UPPER BOUND

We first provide the upper bound of the approximation error of the following Monte Carlo estimate.

Lemma A.2. *For $\mathrm{LSE} := \log \mathbb{E}_{p(z)} \exp(f(x)^\top g(z))$, we denote its (biased) Monte Carlo estimate with M random samples $z_i \sim p(z), i = 1, \ldots, M$ as $\widehat{\mathrm{LSE}}_M = \log \frac{1}{M} \sum_{i=1}^M \exp(f(x)^\top g(z_i))$. Then the approximation error $A(M)$ can be upper bounded in expectation as*

$$A(M) := \mathbb{E}_{p(x,z_i)} |\widehat{\mathrm{LSE}}(M) - \mathrm{LSE}| \leq \mathcal{O}(M^{-1/2}). \tag{9}$$

We can see that the approximation error converges to zero in the order of $1/M^{-1/2}$.

Proof. First, we have

$$\begin{aligned}
&\mathbb{E}_{p(x,z_i)} \left[\log \frac{1}{M} \sum_{i=1}^M \exp(f(x)^\top g(z_i)) - \log \mathbb{E}_{p(z_i)} \exp(f(x)^\top g(z_i)) \right] \\
\leq & e \mathbb{E}_{p(x,z_i)} \left[\frac{1}{M} \sum_{i=1}^M \exp(f(x)^\top g(z_i)) - \mathbb{E}_{p(z_i)} \exp(f(x)^\top g(z_i)) \right] = \mathcal{O}(M^{-1/2}),
\end{aligned}$$

where the first inequality follows the Intermediate Value Theorem and e (the natural number) is the upper bound of the absolute derivative of log between two points when $|f(x)^\top g(z_i)| \leq 1$. And the second inequality follows the Berry-Esseen Theorem given the bounded support of $\exp(f(x)^\top g(z_i))$ as following: for i.i.d random variables Y_i with bounded support $\mathrm{supp}(Y) \subset [-\alpha, \alpha]$, zero mean and bounded variance $\sigma_Y^2 < \alpha^2$, we have:

$$\begin{aligned}
&\mathbb{E}\left[\left| \frac{1}{M} \sum_{i=1}^M Y_i \right| \right] = \frac{\sigma_y}{\sqrt{M}} \mathbb{E}\left[\left| \frac{1}{\sqrt{M}\sigma_y} \sum_{i=1}^M Y_i \right| \right] \\
= & \frac{\sigma_Y}{\sqrt{M}} \int_0^{\frac{\alpha\sqrt{M}}{\sigma_Y}} \mathbb{P}\left[\left| \frac{1}{\sqrt{M}\sigma_Y} \sum_{i=1}^M Y_i \right| > x \right] \mathrm{dx} \\
\leq & \frac{\sigma_Y}{\sqrt{M}} \int_0^{\frac{\alpha\sqrt{M}}{\sigma_Y}} \mathbb{P}[|\mathcal{N}(0,1)| > x] + \frac{C_\alpha}{\sqrt{M}} \mathrm{dx} \\
\leq & \frac{\sigma_Y}{\sqrt{M}} \left(\frac{\alpha C_\alpha}{\sigma_Y} + \int_0^\infty \mathbb{P}[|\mathcal{N}(0,1)| > x] \mathrm{dx} \right) \\
\leq & \frac{C_\alpha}{\sqrt{M}} + \frac{\alpha}{\sqrt{M}} \mathbb{E}[|\mathcal{N}(0,1)|] = \mathcal{O}(M^{-1/2})
\end{aligned}$$

where the constant C_α only depends on α. Here, we set $Y_i = \exp(f(x)^\top g(z_i)) - \mathbb{E}_{p(z_i)} \exp(f(x)^\top g(z_i))$. As $|f(x)^\top g(z_i)| \leq 1$, $|Y_i| \leq 2e$. Y_i has zero mean and bounded variance $(2e)^2$. □

Theorem A.3. *For each $f \in \mathcal{F}$, the mean CE loss can be upper bounded by the InfoNCE loss:*

$$\mathcal{L}_{\mathrm{CE}}^{\mu}(x, y; f) \leq \mathcal{L}_{\mathrm{NCE}}(x; f) - \log(M/K) + \sqrt{\mathrm{Var}(f(x) \mid y)} + A(M), \tag{10}$$

where $\mathrm{Var}(f(x)|y) = \mathbb{E}_{p(y)} \left[\mathbb{E}_{p(x|y)} \|f(x) - \mathbb{E}_{p(x|y)} f(x)\|^2\right]$ *denotes the conditional variance.*

Proof. Denote $p(x, x^+, y)$ as the joint distribution of the positive pairs x, x^+ and the label y. Denote the M independently negative smaples as $\{x_i^-\}_{i=1}^M$. According to Assumption 4.1, x^+ and x here has the same label y. Denote μ_y as the center of features of class y, $y = 1, \ldots, K$. Then we have the following lower bounds of the InfoNCE loss,

$$\begin{aligned}
&\mathcal{L}_{\mathrm{NCE}}(f) = -\mathbb{E}_{p(x,x^+)} f(x)^\top f(x^+) + \mathbb{E}_{p(x)} \mathbb{E}_{p(x_i^-)} \sum_{i=1}^M \exp(f(x)^\top f(x_i^-)) \\
&= -\mathbb{E}_{p(x,x^+)} f(x)^\top f(x^+) + \mathbb{E}_{p(x)} \mathbb{E}_{p(x_i^-)} \log \frac{1}{M} \sum_{i=1}^M \exp(f(x)^\top f(x_i^-)) + \log M \\
&\overset{(1)}{\geq} -\mathbb{E}_{p(x,x^+)} f(x)^\top f(x^+) + \mathbb{E}_{p(x)} \log \frac{1}{M} \mathbb{E}_{p(x_i^-)} \sum_{i=1}^M \exp(f(x)^\top f(x_i^-)) - A(M) + \log M \\
&= -\mathbb{E}_{p(x,x^+)} f(x)^\top f(x^+) + \mathbb{E}_{p(x)} \log \mathbb{E}_{p(x^-)} \exp(f(x)^\top f(x^-)) - A(M) + \log M \\
&= -\mathbb{E}_{p(x,x^+,y)} f(x)^\top f(x^+) + \mathbb{E}_{p(x)} \log \mathbb{E}_{p(y^-)} \mathbb{E}_{p(x^-|y^-)} \exp(f(x)^\top f(x^-)) - A(M) + \log M \\
&\overset{(2)}{\geq} -\mathbb{E}_{p(x,x^+,y)} f(x)^\top f(x^+) + \mathbb{E}_{p(x)} \log \mathbb{E}_{p(y^-)} \exp(\mathbb{E}_{p(x^-|y^-)} \left[f(x)^\top f(x^-)\right]) - A(M) + \log M \\
&= -\mathbb{E}_{p(x,x^+,y)} f(x)^\top (\mu_y + f(x^+) - \mu_y) + \mathbb{E}_{p(x)} \log \mathbb{E}_{p(y^-)} \exp(\mathbb{E}_{p(x^-|y^-)} \left[f(x)^\top f(x^-)\right]) - A(M) + \log M \\
&= -\mathbb{E}_{p(x,x^+,y)} [f(x)^\top \mu_y + f(x)^\top (f(x^+) - \mu_y)] + \mathbb{E}_{p(x)} \log \mathbb{E}_{p(y^-)} \exp(f(x)^\top \mu_{y^-}) - A(M) + \log M \\
&\overset{(3)}{\geq} -\mathbb{E}_{p(x,x^+,y)} \left[f(x)^\top \mu_y + \|(f(x^+) - \mu_y)\|\right] + \mathbb{E}_{p(x)} \log \mathbb{E}_{p(y^-)} \exp(f(x)^\top \mu_{y^-}) - A(M) + \log M \\
&\overset{(4)}{\geq} -\mathbb{E}_{p(x,y)} f(x)^\top \mu_y - \sqrt{\mathbb{E}_{p(x,y)} \|f(x) - \mu_y)\|^2} + \mathbb{E}_{p(x)} \log \mathbb{E}_{p(y^-)} \exp(f(x)^\top \mu_{y^-}) - A(M) + \log M \\
&= -\mathbb{E}_{p(x,y)} f(x)^\top \mu_y - \sqrt{\mathrm{Var}(f(x) \mid y)} + \mathbb{E}_{p(x)} \log \frac{1}{K} \sum_{k=1}^K \exp(f(x)^\top \mu_k) - A(M) + \log M \\
&= \mathbb{E}_{p(x,y)} \big[-f(x)^\top \mu_y + \log \sum_{k=1}^K \exp(f(x)^\top \mu_k) \big] - \sqrt{\mathrm{Var}(f(x) \mid y)} - A(M) + \log(M/K) \\
&= \mathcal{L}_{\mathrm{CE}}^{\mu}(f) - \sqrt{\mathrm{Var}(f(x) \mid y)} - A(M) + \log(M/K),
\end{aligned}$$

which is equivalent to our desired results. In the proof above, (1) follows Lemma A.2; (2) follows the Jensen's inequality for the convex function $\exp(\cdot)$; (3) follows from the fact that because $f(x) \in \mathbb{S}^{m-1}$, we have

$$f(x)^\top (f(x^+) - \mu_y) \leq \left(\frac{f(x^+) - \mu_y}{\|f(x^+) - \mu_y\|} \right)^\top (f(x^+) - \mu_y) = \|f(x^+) - \mu_y\|; \tag{11}$$

and (4) follows the Cauchy–Schwarz inequality and the fact that because $p(x, x^+) = p(x^+, x)$ holds, x, x^+ have the same marginal distribution. □

A.2.2 The Lower Bound

In this part, we further show a lower bound on the downstream performance.

Lemma A.4 (Budimir et al. (2000) Corollary 3.5 (restated)). *Let $g : \mathbb{R}^m \to \mathbb{R}$ be a differentiable convex mapping and $z \in \mathbb{R}^n$. Suppose that g is L- smooth with the constant $L > 0$, i.e., $\forall x, y \in \mathbb{R}^m, \|\nabla g(x) - \nabla g(y)\| \leq L\|x - y\|$. Then we have*

$$0 \leq \mathbb{E}_{p(z)} g(z) - g\left(\mathbb{E}_{p(z)} z\right) \leq L\left[\mathbb{E}_{p(z)}\|z\|^2 - \|\mathbb{E}_{p(z)} z\|^2\right] = L\sum_{j=1}^{n} \mathrm{Var}(z^{(j)}), \tag{12}$$

where $x^{(j)}$ denotes the j-th dimension of x.

With the lemma above, we can derive the lower bound of the downstream performance.

Theorem A.5. *For any $f \in \mathcal{F}$, we have*

$$L_{\mathrm{CE}}^{\mu}(f) \geq \mathcal{L}_{\mathrm{NCE}}(x; f) - \sqrt{\mathrm{Var}(f(x) \mid y)} - \frac{1}{2}\sum_{j=1}^{m} \mathrm{Var}(f_j(x^-) \mid y) - A(M) - \log\frac{M}{K}, \tag{13}$$

where $\mathrm{Var}(u(x)|y) = \mathbb{E}_{p(y)}\left[\mathbb{E}_{p(x|y)}\|u(x) - \mathbb{E}_{p(x|y)} u(x)\|^2\right]$ denotes the conditional variance.

Proof. Similar to the proof of Theorem A.3, we have

$$\begin{aligned}
&L_{\mathrm{CE}}^{\mu}(f) = -\mathbb{E}_{p(x,y)} f(x)^\top \mu_y + \mathbb{E}_{p(x)} \log \sum_{i=1}^{K} \exp(f(x)^\top \mu_i) \\
=& -\mathbb{E}_{p(x,y)} f(x)^\top \mu_y + \mathbb{E}_{p(x)} \log \frac{1}{K}\sum_{i=1}^{K} \exp(f(x)^\top \mu_i) + \log K \\
=& -\mathbb{E}_{p(x,y)} f(x)^\top \mu_y + \mathbb{E}_{p(x)} \log \mathbb{E}_{p(y_i^-)} \exp(f(x)^\top \mu_{y_i}) + \log K \\
\overset{(1)}{\geq}& -\mathbb{E}_{p(x,y)}[f(x)^\top f(x^+) + f(x)^\top(\mu_y - f(x^+))] + \mathbb{E}_{p(x)}\mathbb{E}_{p(y_i^-)} \log \frac{1}{M}\sum_{i=1}^{M} \exp(f(x)^\top \mu_{y_i}) - A(M) + \log K \\
\overset{(2)}{\geq}& -\mathbb{E}_{p(x,x^+)} f(x)^\top f(x^+) - \mathbb{E}_{p(x,y)}\|f(x)^\top - \mu_y\| \\
&\qquad + \mathbb{E}_{p(x)}\mathbb{E}_{p(y_i^-)} \log \frac{1}{M}\sum_{i=1}^{M} \exp(\mathbb{E}_{p(x_i^-|y_i^-)} f(x)^\top f(x_i^-)) - A(M) + \log K \\
\overset{(3)}{\geq}& -\mathbb{E}_{p(x,x^+)} f(x)^\top f(x^+) - \sqrt{\mathrm{Var}(f(x) \mid y)} \\
&+ \mathbb{E}_{p(x)}\mathbb{E}_{p(y_i^-)}\mathbb{E}_{p(x_i^-|y)}\left[\log \frac{1}{M}\sum_{i=1}^{M} \exp(f(x)^\top f(x^-))\right] - \frac{1}{2}\sum_{j=1}^{m} \mathrm{Var}(f_j(x^-) \mid y) - A(M) + \log K \\
=& -\mathbb{E}_{p(x,x^+)}\left[f(x)^\top f(x^+) + \mathbb{E}_{p(x_i^-)} \log \sum_{i=1}^{M} \exp(f(x)^\top f(x^-))\right] \\
&\qquad - \sqrt{\mathrm{Var}(f(x) \mid y)} - \frac{1}{2}\sum_{j=1}^{m} \mathrm{Var}(f_j(x^-) \mid y) - A(M) + \log K - \log M \\
=& \mathcal{L}_{\mathrm{NCE}}(x; f) - \sqrt{\mathrm{Var}(f(x) \mid y)} - \frac{1}{2}\sum_{j=1}^{m} \mathrm{Var}(f_j(x^-) \mid y) - A(M) - \log\frac{M}{K},
\end{aligned}$$

which is our desired result. In the proof, (1) we adopt a Monte Carlo estimate with M samples from $p(y)$ and bound the approximation error with Lemma A.2; (2) follows the same deduction in Theorem A.3; (3) the first term is derived following the Cauchy–Schwarz inequality for the alignment term. As for the second term, we first show that the convex function $\mathrm{logsumexp}$ is L-smooth as a function of $f(x_j^-)$ in our scenario. Because $\|f(X)\| \leq 1$, we have $\forall f(x_{j_1}), f(x_{j_2}) \in \mathbb{R}^m$, the

following bound on the difference of their gradients holds

$$
\begin{aligned}
&\left\| \frac{\partial \log[\exp(f(x)^\top f(x_{j_1}^-) + \sum_{i\neq j} \exp(f(x)^\top f(x_i^-)))]}{\partial f(x_{j_1}^-)} - \frac{\partial \log[\exp(f(x)^\top f(x_{j_2}^-) + \sum_{i\neq j} \exp(f(x)^\top f(x_i^-)))]}{\partial f(x_{j_2}^-)} \right\| \\
=&\left\| \left(\frac{\exp(f(x)^\top f(x_{j_1}^-))}{\exp(f(x)^\top f(x_{j_1}^-) + \sum_{i\neq j} \exp(f(x_i^-)))} - \frac{\exp(f(x)^\top f(x_{j_2}^-))}{\exp(f(x)^\top f(x_{j_2}^-) + \sum_{i\neq j} \exp(f(x)^\top f(x_i^-)))} \right) f(x) \right\| \\
\leq&\left| \frac{(\sum_{i\neq j} \exp(f(x)^\top f(x_i^-)) \exp(f(x_{j_1}^-)) - \sum_{i\neq j} \exp(f(x)^\top f(x_i^-)) \exp(f(x)^\top f(x_{j_2}^-))}{(\exp(f(x)^\top f(x_{j_1}^-)) + \sum_{i\neq j} \exp(f(x)^\top f(x_i^-)))(\exp(f(x)^\top f(x_{j_2}^-)) + \sum_{i\neq j} \exp(f(x)^\top f(x_i^-)))} \right| \cdot \|f(x)\| \\
\leq&\|f(x)\| \leq \frac{1}{2} \left\| f(x_{j_1}^-) - f(x_{j_2}^-) \right\|
\end{aligned}
$$

So here the $\mathrm{logsumexp}$ is L-smooth for $L = \frac{1}{2}$. Then, we can apply the reversed Jensen's inequality in Lemma A.4.

□

A.3 Proof of Proposition 4.7

Proposition A.6. *Under Assumptions 4.5, & 4.6, by minimizing the InfoNCE loss, we can conclude that the conditional variance term vanishes, i.e.,*

$$\mathrm{Var}(f(x) \mid y) = \mathrm{Var}(f_j(x) \mid y) = 0, \ \forall j \in \{1, 2, \ldots, m\}. \tag{14}$$

Proof. Consider any $\mathcal{T}$-connected sample x_i, x_j. Accoding to the definition of $\mathcal{T}$-connectivity, there exist $t_i, t_j \in \mathcal{T}$ such that $t_i(x) = t_j(x)$. When perfect alignment holds as in Assumption 4.6, we will have $f(x_i) = f(t_i(x_i))$ and $f(x_j) = f(t_j(x_j))$. Combining with $t_i(x_i) = t_j(x_j)$, we have $f(x_i) = f(x_j)$. That is, any $\mathcal{T}$-connected pair has the same representation. Then, in the augmentation subgraph $\mathcal{G}_k$ that is connected according to Assumption 4.5, there exists a path for any pair of samples $\hat{x}_i, \hat{x}_j \in \mathcal{G}_k$ where any two adjacent samples are $\mathcal{T}$-connected. As a result, $\hat{x}_i$ and $\hat{x}_j$ will also have the same representation by applying $\mathcal{T}$-connectivity recursively. At last, all samples in $\mathcal{G}_k$ will have the same representation and the intra-class variance terms vanish. □

A.4 Proof of Theorem 4.8

Theorem A.7 (Guarantees for the optimal encoder)**.** *If Assumption 4.1, 4.5 & 4.6 hold and f is L-smooth, then, for the minimizer $f^\star = \arg\min \mathcal{L}_{\mathrm{NCE}}(f)$, its classification risk can be upper and lower bounded by its contrastive risk as*

$$\mathcal{L}_{\mathrm{NCE}}(f^\star) - \mathcal{O}\left(M^{-1/2}\right) \leq \quad \mathcal{L}_{\mathrm{CE}}^{\mu}(f^\star) + \log(M/K) \leq \mathcal{L}_{\mathrm{NCE}}(f^\star) + \mathcal{O}\left(M^{-1/2}\right). \tag{15}$$

Proof. A direct combination of Theorem 4.2 and 4.7 will give us the above two-sided bounds. □

B Theoretical Characterization of Augmentation Strength

Following the setting in Section 5.1, we can take the radius r as a notation of augmentation strength, and analyze its effect on the connectivity of the corresponding augmentation graph.

Theorem B.1. *For N random samples taken from a class, while gradually increasing the augmentation strength r, we have the following results.*

(a) ***Under-overlapping.*** *When $0 \leq r \leq r_1 = \frac{[(d/2)!]^{\frac{1}{d}}}{\sqrt{\pi}} (\frac{1}{d})! (\frac{S}{N-1})^{\frac{1}{d}} [1 - \frac{1/d + 1/d^2}{2(N-1)} + O(\frac{1}{(N-1)^2})]$, where r_1 is the minimal distance between N samples, all samples (vertices) in the augmentation graph will be isolated. As a result, the learned features could be totally random as in Proposition 3.1. Instead, if $r \geq r_1$, there are at least two intra-class samples are $\mathcal{T}$-connected and enjoy the same representation.*

(b) ***Perfect Overlapping.*** *When* $r \geq r_2 = \frac{[(d/2)!]^{\frac{1}{d}}}{\sqrt{\pi}} \frac{(N-2+1/d)!}{(N-2)!} (\frac{S}{N-1})^{\frac{1}{d}} [1 - \frac{1/d+1/d^2}{2(N-1)} + O(\frac{1}{(N-1)^2})]$*, where* r_2 *is the maximal distance between* N *samples, all samples in the augmentation graph will be* $\mathcal{T}$*-connected. As a result, the classwise connectivity in Assumption 4.5 will be guaranteed.*

(c) ***Over-overlapping.*** *When* $0 \leq r < r_3 = \frac{1}{2} \min_{i,j} \|c_i - c_j\| - 1$*, where* r_3 *is the (asymptotic) minimal distance between samples from different classes, the label consistency is guaranteed. Otherwise, when the augmentation is too large, e.g.,* $r > r_3$*, there will be inter-class support overlap and Assumption 4.1 not longer holds.*

In the theorem above, we show that the proper augmentation strength is a function of the number of samples N and the input dimensions d. In particular, for each x, as N increases, there will be more natural examples and we only need a smaller r to obtain an overlapping sample. Instead, as d increases, due to the curse of dimensionality, there will be less samples within the same distance, thus it requires a larger r.

Nevertheless, we actually only need the augmentation sub-graph G_k to be connected, instead of being fully connected as in Theorem B.1 (b). While the connectivity is hard to analyze in the finite sample scenario ($N < \infty$), we have the following asymptotic property as $N \to \infty$.

Theorem B.2. *For* N *uniformly distributed samples defined in* $\mathbb{R}^d$ *as above, we denote the minimal augmentation strength needed for connectivity as a function of* N*:* $c_N = \inf\{r > 0 : G_k^{(r)}$ *is connected*$\}$*, and the minimal augmentation strength needed for avoiding isolated points as a function of* N*:* $d_N = \inf\{r > 0 :$ *every vertex at least has a neighbour*$\}$*.* V_u *is the volume of unit hyperball. Then we have the following asymptotic result:*

$$\forall\, d \geq 2,\ \lim_{N\to\infty} \left(c_N^d \frac{N^2}{\log N} \right) = \lim_{N\to\infty} \left(d_N^d \frac{N^2}{\log N} \right) = 2\frac{(1-1/d)S}{V_u}. \tag{16}$$

From the theorem we can see that c_N^d decreases in the order of $\Theta\left(\sqrt[d]{\frac{\log N}{N^2}}\right)$ as $N \to \infty$. First, this result is aligned with the empirical finding that self-supervised learning can benefit more from large scale dataset (Chen et al., 2020b). Second, it also indicate a curse of dimensionality that the required augmentation strength is exponentially large.

B.1 Proof of Theorem B.1

Proof. From definition and notation in section 4. We can construct an augmentation Graph $\mathcal{G}(\mathcal{D}, \mathcal{T})$ given N random samples. We define D_k as the distance from a random point to its k-th nearest neighbour. Percus & Martin (1998) discuss D_k in random grpah and give the estimation of that:

$$D_k \approx \frac{[(d/2)!]^{\frac{1}{d}}}{\sqrt{\pi}} \frac{(k-1+1/d)!}{(k-1)!} (\frac{S}{N-1})^{\frac{1}{d}} [1 - \frac{1/d+1/d^2}{2(N-1)} + O(\frac{1}{(N-1)^2})] \tag{17}$$

where d is the dimension of hypersphere and N is the number of random points. When $r < D_1$ there is no edge in the graph. So the class is separated. When $r > D_{N-1}$, any pair of vertexes have an edge between them,so the graph is full connected. □

B.2 Proof of Theorem B.2

Proof. Denote

$$c_N = \inf\{r_i > 0 : G_N(V, E, r_i) \text{is connected}\}. \tag{18}$$

With Theorem 1.1 from Penrose (1999) and features are uniformly distributed in the surface of unit hypersphere, V_u denotes to the volume of unit hypershpere

$$\lim_{N\to\infty} (c_N^d \frac{N^2}{\log N}) = 2\frac{(1-\frac{1}{d})S}{V_u}, d \geq 2 \tag{19}$$

$\exists N_0$ when $N > N_0$, and augmentation strength is larger than $(\frac{2(d-1)S\log N}{2N^2 V_u d})^{\frac{1}{d}} + \epsilon_1$, the graph is connected,i.e the class is overlapped.

Then we want specify the case the class will be depart begin with some concepts in graph theory. The largest nearest-neighbor link: For a give edge distance x and for each i= 1,...,n,let

$$\deg\ U_{N,j} = \sum_{1 \le j \ne k \le N} 1_{\{} \|U_j - U_k\| \le r_i\} \tag{20}$$

to be the degree of the vertex U_j in the random graph $G_N(V, E, r_i)$, and let

$$\delta_N(r_i) = \min\{\deg\ U_{m,1}(x), ..., \deg\ U_{N,N}(x)\} \tag{21}$$

be the minimum vertex degree.Define the largest nearest-neighbor link, the smallest edge distance for which each vertex has at least one neighbor

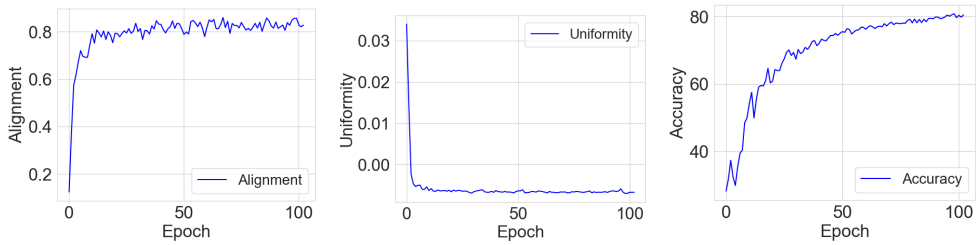
$$d_N = \inf\{r_i : \delta_N(r_i) \ge 1\} \tag{22}$$

With Theorem 1.2 from Penrose (1999),

$$\limsup_{N\to\infty}(d_N^d \frac{N^2}{\log N}) = 2\frac{(1-\frac{1}{d})S}{V_u}, d \ge 2 \tag{23}$$

$\exists N_0'$ when $N > N_0'$, and augmentation strength is less than $(\frac{2(d-1)S\log N}{2N^2 V_u d})^{\frac{1}{d}} - \epsilon_2$, there will be at least 1 isolated point which is not connected to any other point,i.e the class is departed.
Thus $\exists N_1 = \max(N_0, N_0')$,when $N > N_1$, if augmentation strength is larger than $(\frac{2(d-1)S\log N}{2N^2 V_u d})^{\frac{1}{d}} + \epsilon_1$, the graph is connected, if augmentation strength is less than $(\frac{2(d-1)S\log N}{2N^2 V_u d})^{\frac{1}{d}} - \epsilon_2$, there will be at least 1 isolated point. □

C Alignment in the Training Process



(a) Alignment (b) Uniformity (c) Accuracy

Figure 9: Feature alignment, uniformity, and test accuracy along the training process on CIFAR-10 with SimCLR.

Here, we show the empirical learning process with the InfoNCE loss on Figure 9. Comparing the alignment and uniformity terms, we can see that the alignment gradually improves while test accuracy improves. In comparison, the uniformity term becomes very small at the beginning and is relatively stable while training. This indicates that the alignment term is the main driving force in the contrastive learning process and the uniformity term is mainly used to avoid feature collapse.

D Additional Experimental Details

D.1 Simulation on Random Augmentation Graph

Following our setting in Section 5.1, we consider a binary classification task with InfoNCE loss. We generate data from two uniform distribution on a unit ball $\mathbb{S}^2$ in the 3-dimensional space. One center is $(0, 0, 1)$ and another is $(0, 0, -1)$. The area of both parts are 1. We take 5000 sample as train set and 1000 samples as test set. For the encoder class , we use a single-hidden-layer neural network with softmax activation, and we use InfoNCE loss to optimize it.

D.2 Experiments on Real-world Datasets

To better understand and verify our theorem, we conduct experiments on real-world datasets, including CIFAR-10, CIFAR-100 and STL-10. We use SimCLR (Chen et al., 2020a) and BYOL Grill et al. (2020) as our training framework and use ResNet18 as our network. For CIFAR-10 and CIFAR-100, we adopt $C = 10$ augmentations for each image, and search neural neighbors in the entire augmented dataset. For STL-10, we adopt $C = 6$ due to its relatively large size.