Towards the Generalization of Contrastive Self-Supervised Learning

Weiran Huang[1*] Mingyang Yi[2,3*] Xuyang Zhao[4*]

[1]Huawei Noah's Ark Lab
[2]University of Chinese Academy of Sciences
[3]Academy of Mathematics and Systems Science, Chinese Academy of Sciences
[4]Peking University

Abstract

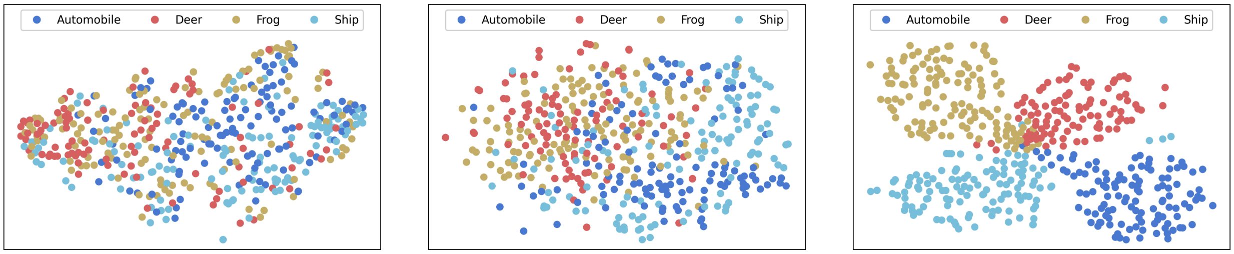
Recently, self-supervised learning has attracted great attention, since it only requires unlabeled data for training. Contrastive learning is one popular method for self-supervised learning and has achieved promising empirical performance. However, the theoretical understanding of its generalization ability is still limited. To this end, we define a kind of (σ, δ)-measure to mathematically quantify the data augmentation, and then provide an upper bound of the downstream classification error based on the measure. We show that the generalization ability of contrastive self-supervised learning depends on three key factors: *alignment* of positive samples, *divergence* of class centers, and *concentration* of augmented data. The first two factors can be optimized by contrastive algorithms, while the third one is priorly determined by pre-defined data augmentation. With the above theoretical findings, we further study two canonical contrastive losses, InfoNCE and cross-correlation loss, and prove that both of them are indeed able to satisfy the first two factors. Moreover, we empirically verify the third factor by conducting various experiments on the real-world dataset, and show that our theoretical inferences on the relationship between the data augmentation and the generalization of contrastive self-supervised learning agree with the empirical observations.

1 Introduction

Contrastive Self-Supervised Learning (SSL) has attracted great attention for its fantastic data efficiency and generalization ability in both computer vision (He et al., 2020; Chen et al., 2020a,b; Grill et al., 2020; Chen and He, 2021; Zbontar et al., 2021) and natural language processing (Fang et al., 2020; Wu et al., 2020; Giorgi et al., 2020; Gao et al., 2021; Yan et al., 2021). It learns the representation through a large number of unlabeled data and artificially defined self-supervision signals (i.e., regarding the augmented views of a data sample as positive samples). The model is updated by encouraging the embeddings of positive samples close to each other. To overcome the feature collapse issue, different losses (e.g., InfoNCE (Chen et al., 2020a; He et al., 2020) and cross-correlation (Zbontar et al., 2021)) and training strategies (e.g., stop gradient (Grill et al., 2020; Chen and He, 2021)) are proposed.

In spite of the empirical success of contrastive SSL in terms of their generalization ability on downstream tasks, the theoretical understanding is still limited. Arora et al. (2019) propose a theoretical framework to show the provable downstream performance of contrastive SSL based on the InfoNCE loss. However, their results rely on the assumption that positive samples are

[*]Equal contribution (alphabetical ordering). Correspondence to Weiran Huang <weiran.huang@outlook.com>.



(a) Original **(b)** Only color distortion **(c)** Multiple transformations

Figure 1: SimCLR's Embedding spaces with different richness of data augmentations on CIFAR-10

drawn from the same latent class, instead of the augmented views of a data point as in practice. Wang and Isola (2020) propose alignment and uniformity to explain the downstream performance, but they are empirical indicators without theoretical generalization guarantees of contrastive SSL. Furthermore, both of the above works *avoid characterizing the important role of data augmentation*, which is the key to the success of contrastive SSL, since the only human knowledge is injected via data augmentation.

Besides the limitations of existing contrastive SSL theories, there are also some interesting empirical observations that have not been unraveled theoretically yet. For example, why does the richer data augmentation lead to the more clustered structure (Figure 1) as well as the better downstream performance (observed in (Chen et al., 2020a))? Why is aligning positive samples (augmented from the "same data point") able to gather the samples from the "same latent class" into a cluster (Figure 1c)? More interestingly, decorrelating components of representation like Barlow Twins (Zbontar et al., 2021) does not directly optimize the geometry of embedding space, but it can still result in the clustered structure. Why?

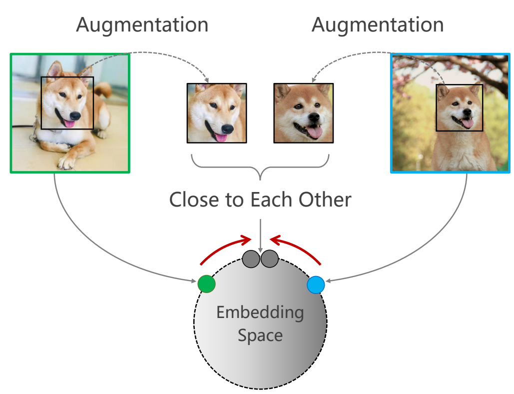


Figure 2: Mechanism of Clustering

In this paper, we focus on exploring the generalization ability of contrastive SSL *provably*, which can also provide insights for understanding the above interesting observations. We start with understanding the role of data augmentation in contrastive SSL. Intuitively, samples from the same latent class are likely to have similar augmented views, which are mapped to the close locations in the embedding space. Since the augmented views of each sample are encouraged to be clustered in the embedding space by contrastive learning, different samples from the same latent class tend to be pulled closer. As an example, let's consider two images of dogs with different backgrounds (Figure 2). If we augment them with operation "crop", we may get two similar views (dog heads), whose representations (gray points in the embedding space) are close. As the augmented views of each dog image are enforced to be close in the embedding space due to the objective of contrastive learning, the representations of two dog images (green and blue points) will be pulled closer to their dog head views. Thus, aligning positive samples can obtain the clustered embedding space. Following the above intuition, we define the *augmented distance* between two samples as the minimum distance between their augmented views, and further introduce the (σ, δ)-augmentation to measure the concentration of augmented data, i.e., for each latent class, the proportion of samples located in a ball with diameter δ (w.r.t. the augmented distance) is larger than σ.

With the mathematical description of data augmentation settled, we then prove the upper bound of downstream classification error in Section 3. It reveals that the generalization of contrastive SSL depends on three key factors. The first one is the *alignment* of positive samples, which is the common objective that contrastive algorithms aim to optimize. The second one is the *divergence* of class centers, which prevents the collapse of representation. We remark that "divergence" characterizes the sufficient condition of generalization more precisely than the "uniformity" indicator of (Wang and Isola, 2020) (see discussion in Section 3). The third factor is the *concentration* of augmented data, i.e., the model trained using the augmented data with sharper concentration exhibits a better generalization ability. Notice that the first two factors, alignment and divergence, can be optimized by SSL algorithms, while the third factor is priorly decided by the pre-determined data augmentation. Therefore, data augmentation plays a role as important as algorithms in contrastive SSL.

To verify the above three factors, we rigorously prove that not only InfoNCE but also cross-correlation loss (which does not directly optimize the geometry of embedding space) can satisfy the first two factors (Section 4). For the third factor, we conduct various experiments on the real-world dataset to verify that the model trained using the augmented data with sharper concentration in terms of the proposed augmented distance indeed has better performance on downstream tasks (Section 5).

In summary, our contributions include: 1) proposing a novel (σ, δ)-measure to quantify the data augmentation; 2) proposing a theoretical framework for contrastive SSL, which indicates that alignment, divergence, and concentration are key factors of generalization; 3) provably verifying that not only InfoNCE but also cross-correlation loss satisfy the first two factors; and 4) empirically verifying that sharper concentration w.r.t. the proposed augmented distance leads to better downstream performance.

Related Work

Algorithms of Contrastive SSL. Early works such as MoCo (He et al., 2020) and SimCLR (Chen et al., 2020a), use the InfoNCE loss to pull the positive samples close while enforcing them away from the negative samples in the embedding space. These methods require large batch sizes (Chen et al., 2020a), memory banks (He et al., 2020), or carefully designed negative sampling strategies (Hu et al., 2021). To obviate these, some recent works get rid of negative samples and prevent representation collapse by cross-correlation loss (Zbontar et al., 2021; Bardes et al., 2021) or training strategies (Grill et al., 2020; Chen and He, 2021). In this paper, we mainly study the effectiveness of InfoNCE and cross-correlation loss, and do not enter the discussion of training strategies.

Theoretical Understandings of Contrastive SSL. Most theoretical analysis is based on the InfoNCE loss, and lack of understanding of recently proposed cross-correlation loss (Zbontar et al., 2021). Early works understand the InfoNCE loss based on maximizing the mutual information (MI) between positive samples (Oord et al., 2018; Bachman et al., 2019; Hjelm et al., 2018; Tian et al., 2019, 2020). However, Tschannen et al. (2019) find that optimizing tighter bounds of MI does not imply better representations. Thus, MI may not fully explain the success of InfoNCE. Besides, Arora et al. (2019) directly analyze the generalization of InfoNCE loss based on the assumption that positive samples are drawn from the same latent classes, which is different from practical contrastive algorithms. Furthermore, HaoChen et al. (2021) study contrastive SSL from a matrix decomposition perspective, but it is only applicable to their spectral contrastive loss. The behavior of InfoNCE loss is also studied from the perspective of alignment and uniformity, under the condition of infinite negative samples (Wang and Isola, 2020), sparse coding model (Wen and Li, 2021), or impractical "expansion" assumption (Wei et al., 2020).

2 Problem Formulation

Given a number of unlabeled training data i.i.d. drawn from an unknown distribution, each sample belongs to one of K latent classes $C_1, C_2, \ldots, C_K$. Based on a transformation set A, the set of potential positive samples generated from a data point $\boldsymbol{x}$ is denoted as $A(\boldsymbol{x})$. We assume that $\boldsymbol{x} \in A(\boldsymbol{x})$ for any $\boldsymbol{x}$, and samples from different classes never transfer to the same augmented sample, i.e., $\cap_{k=1}^{K} A(C_k) = \varnothing$. Notation $\|\cdot\|$ stands for ℓ_2-norm or Frobenius norm for vectors and matrices.

Contrastive SSL aims to learn an encoder f, such that positive samples are closely aligned. In order to make the samples from different latent classes far away from each other, a class of methods (e.g., SimCLR (Chen et al., 2020a)) use the InfoNCE loss to push away negative pairs, formulated as

$$\mathcal{L}_{\text{InfoNCE}} = -\mathop{\mathbb{E}}_{\boldsymbol{x},\boldsymbol{x}'} \mathop{\mathbb{E}}_{\substack{\boldsymbol{x}_1,\boldsymbol{x}_2 \in A(\boldsymbol{x}) \\ \boldsymbol{x}^- \in A(\boldsymbol{x}')}} \log \frac{e^{f(\boldsymbol{x}_1)^\top f(\boldsymbol{x}_2)}}{e^{f(\boldsymbol{x}_1)^\top f(\boldsymbol{x}_2)} + e^{f(\boldsymbol{x}_1)^\top f(\boldsymbol{x}^-)}},$$

where $\boldsymbol{x}, \boldsymbol{x}'$ are two random data points. Some other methods (e.g., Barlow Twins (Zbontar et al., 2021)) use the cross-correlation loss to decorrelate the components of representation, formulated as

$$\mathcal{L}_{\text{Cross-Corr}} = \sum_{i=1}^{d} (1 - F_{ii})^2 + \lambda \sum_{i=1}^{d} \sum_{i \neq j} F_{ij}^2,$$

where $F_{ij} = \mathbb{E}_{\boldsymbol{x}} \mathbb{E}_{\boldsymbol{x}_1,\boldsymbol{x}_2 \in A(\boldsymbol{x})}[f_i(\boldsymbol{x}_1) f_j(\boldsymbol{x}_2)]$, d is the dimension of encoder f, and encoder f is normalized as $\mathbb{E}_{\boldsymbol{x}} \mathbb{E}_{\boldsymbol{x}' \in A(\boldsymbol{x})}[f_i(\boldsymbol{x}')^2] = 1$ for each dimension i.

The standard evaluation of contrastive SSL is to train a linear classifier over the self-supervised learned representation using labeled data and regard its performance as the indicator. To simplify the analysis, we instead consider a non-parametric classifier – nearest neighbor (NN) classifier:

$$G_f(\boldsymbol{x}) = \arg\min_{k \in [K]} \|f(\boldsymbol{x}) - \mu_k\|,$$

where $\mu_k := \mathbb{E}_{\boldsymbol{x} \in C_k} \mathbb{E}_{\boldsymbol{x}' \in A(\boldsymbol{x})}[f(\boldsymbol{x}')]$ is the center of class C_k. In fact, the NN classifier is a special case of linear classifier, since it can be reformulated as $G_f(\boldsymbol{x}) = \arg\max_{k \in [K]} (W f(\boldsymbol{x}) + b)_k$, where the k-th row of W is μ_k and $b_k = -\frac{1}{2}\|\mu_k\|^2$. Therefore, the directly learned linear classifier used in practice should have better performance than the NN classifier. In this paper, we use classification error rate to quantify the performance of G_f, formulated as

$$\text{Err}(G_f) = \sum_{k=1}^{K} \mathbb{P}[G_f(\boldsymbol{x}) \neq k, \forall \boldsymbol{x} \in C_k].$$

In the sequel, we study why contrastive SSL is able to achieve a small error rate $\text{Err}(G_f)$.

3 Generalization Guarantee of Contrastive SSL

Based on the NN classifier, if the samples are well clustered by latent classes in the embedding space, the error rate $\text{Err}(G_f)$ should be small. Thus, one expects to have a small intra-class distance $\mathbb{E}_{\boldsymbol{x}_1,\boldsymbol{x}_2 \in C_k} \|f(\boldsymbol{x}_1) - f(\boldsymbol{x}_2)\|^2$ for a contrastive learned encoder f. However, contrastive algorithms only control the alignment of positive samples $\mathbb{E}_{\boldsymbol{x}_1,\boldsymbol{x}_2\in A(\boldsymbol{x})}\|f(\boldsymbol{x}_1)-f(\boldsymbol{x}_2)\|^2$. To bridge the gap between these two distances, we will show that data augmentation plays an important role.

Motivated by Figure 2 introduced in Section 1, for a given transformation set A, we define the *augmented distance* between two samples as the minimum distance between their augmented views:

$$d_A(\boldsymbol{x}_1,\boldsymbol{x}_2)=\min_{\boldsymbol{x}_1'\in A(\boldsymbol{x}_1),\boldsymbol{x}_2'\in A(\boldsymbol{x}_2)}\left\|\boldsymbol{x}_1'-\boldsymbol{x}_2'\right\|. \tag{1}$$

For the dog images in Figure 2, although they are quite different at the pixel level, they contain similar semantic meanings. Meanwhile, they have a small augmented distance. Thus, the semantic distance can be partially characterized by the proposed augmented distance. Based on the augmented distance, we now introduce the (σ,δ)-augmentation to measure the concentration of augmented data.

Definition 1 (**(σ,δ)-Augmentation**)**.** The data augmentation set A is called a (σ,δ)-augmentation, if for each class C_k, there exists a subset $C_k^0\subseteq C_k$ (called a main part of C_k), such that both $\mathbb{P}[\boldsymbol{x}\in C_k^0]\geq\sigma\,\mathbb{P}[\boldsymbol{x}\in C_k]$ where $\sigma\in(0,1]$ and $\sup_{\boldsymbol{x}_1,\boldsymbol{x}_2\in C_k^0}d_A(\boldsymbol{x}_1,\boldsymbol{x}_2)\leq\delta$ hold.

In other words, the main part samples locate in a ball with diameter δ (w.r.t. the augmented distance) and its proportion is larger than σ. Larger σ and smaller δ indicate the sharper concentration of augmented data. One can verify that for any $A'\supseteq A$ with richer augmentations, we have $d_{A'}(\boldsymbol{x}_1,\boldsymbol{x}_2)\leq d_A(\boldsymbol{x}_1,\boldsymbol{x}_2)$ for any $\boldsymbol{x}_1,\boldsymbol{x}_2$. Therefore, richer data augmentations lead to sharper concentration as δ gets smaller. With Definition 1, our analysis will focus on the samples in the main parts with good alignment, i.e., $(C_1^0\cup\cdots\cup C_K^0)\cap S_\varepsilon$, where $S_\varepsilon:=\{\boldsymbol{x}\in\cup_{k=1}^K C_k\colon\forall\boldsymbol{x}_1,\boldsymbol{x}_2\in A(\boldsymbol{x}),\|f(\boldsymbol{x}_1)-f(\boldsymbol{x}_2)\|\leq\varepsilon\}$ is the set of samples with ε-close representations among augmented data. Furthermore, we let $R_\varepsilon:=\mathbb{P}\left[\overline{S_\varepsilon}\right]$, which should be small for an encoder with good alignment.

Lemma 3.1. *For a (σ,δ)-augmentation with main part C_k^0 of each class C_k, if all samples belonging to $(C_1^0\cup\cdots\cup C_K^0)\cap S_\varepsilon$ can be correctly classified by a classifier G, then its classification error rate $\mathrm{Err}(G)$ is upper bounded by $(1-\sigma)+R_\varepsilon$.*

The above lemma presents a simple sufficient condition to guarantee the generalization ability on downstream tasks. Based on it, we will further explore 1) when samples in $(C_1^0\cup\cdots\cup C_K^0)\cap S_\varepsilon$ can be all correctly classified by the NN classifier, and 2) the upper bound of R_ε. We will tackle the first one below, and leave the second one in Section 3.1.

For simplicity, we assume that encoder f is normalized by $\|f\|=r$, and it is L-Lipschitz continuity, i.e., for any $\boldsymbol{x}_1,\boldsymbol{x}_2$, $\|f(\boldsymbol{x}_1)-f(\boldsymbol{x}_2)\|\leq L\|\boldsymbol{x}_1-\boldsymbol{x}_2\|$. We let $p_k:=\mathbb{P}[\boldsymbol{x}\in C_k]$ for any $k\in[K]$.

Lemma 3.2. *Given a (σ,δ)-augmentation used in contrastive SSL, for any $\ell\in[K]$, if $\mu_\ell^\top\mu_k<r^2\left(1-\rho_\ell(\sigma,\delta,\varepsilon)-\sqrt{2\rho_\ell(\sigma,\delta,\varepsilon)}-\frac{\Delta_\mu}{2}\right)$ holds for all $k\neq\ell$, then every sample $\boldsymbol{x}\in C_\ell^0\cap S_\varepsilon$ can be correctly classified by the NN classifier G_f, where $\rho_\ell(\sigma,\delta,\varepsilon)=2(1-\sigma)+\frac{R_\varepsilon}{p_\ell}+\sigma\left(\frac{L\delta}{r}+\frac{2\varepsilon}{r}\right)$ and $\Delta_\mu=1-\min_{k\in[K]}\|\mu_k\|^2/r^2$.*

With Lemma 3.1 and 3.2, we can directly obtain the generalization guarantee of contrastive SSL:

Theorem 1. *Given a (σ,δ)-augmentation used in contrastive SSL, if*

$$\mu_\ell^\top\mu_k<r^2\left(1-\rho_{max}(\sigma,\delta,\varepsilon)-\sqrt{2\rho_{max}(\sigma,\delta,\varepsilon)}-\frac{\Delta_\mu}{2}\right) \tag{2}$$

holds for any pair of (ℓ, k) *with* $\ell \neq k$*, then the downstream error rate of NN classifier* G_f

$$\mathrm{Err}(G_f) \leq (1-\sigma) + R_\varepsilon, \tag{3}$$

where $\rho_{max}(\sigma,\delta,\varepsilon) = 2(1-\sigma) + \frac{R_\varepsilon}{\min_\ell p_\ell} + \sigma\left(\frac{L\delta}{r} + \frac{2\varepsilon}{r}\right)$ *and* $\Delta_\mu = 1 - \min_{k\in[K]} \|\mu_k\|^2/r^2$.

To better understand the above theorem, let us consider a simple case that any two samples from the latent same class at least own a same augmented view ($\sigma = 1, \delta = 0$), and the positive samples are perfectly aligned after contrastive learning ($\varepsilon = 0, R_\varepsilon = 0$). In this case, the samples from the same latent class are embedded to a single point on the hypersphere, and thus arbitrarily small positive angle $\frac{\langle \mu_\ell, \mu_k \rangle}{\|\mu_\ell\|\cdot\|\mu_k\|} < 1$ is enough to distinguish them by the NN classifier. In fact, one can quickly verify that $\rho_{max}(\sigma,\delta,\varepsilon) = \Delta_\mu = 0$ holds in the above case. According to Theorem 1, if $\mu_\ell^\top \mu_k/r^2 < 1 - \rho_{max}(\sigma,\delta,\varepsilon) - \sqrt{2\rho_{max}(\sigma,\delta,\varepsilon)} - \frac{\Delta_\mu}{2} = 1$, then $\mathrm{Err}(G_f) = 0$, i.e., NN classifier correctly recognize every sample. The condition suggested by Theorem 1 matches the intuition.

Theorem 1 implies three key factors to the success of contrastive SSL. The first one is the *alignment* of positive samples, which is the common objective that contrastive algorithms aim to optimize. The good alignment enables the small R_ε, which directly decreases the upper bound of error rate (3). The second factor is the *divergence* of class centers, i.e., the distance between class centers should be large enough (small $\mu_\ell^\top \mu_k$). The divergence condition is related to the alignment (R_ε) and data augmentation (σ, δ). Better alignment and sharper concentration indicate smaller $\rho_{max}(\sigma,\delta,\varepsilon)$ in (2). Thus, a good alignment property or concentrated augmented data can loosen the divergence condition. The third factor is the *concentration* of augmented data. When δ is given, sharper concentration implies larger σ, which directly affects the upper bound of error rate (3). For example, richer data augmentations can make the augmented data more concentrated (see the paragraph below Definition 1), and thus the better downstream performance. Theorem 1 provides a theoretical framework to analyze the generalization of contrastive SSL. We remark that the first two factors can be optimized by contrastive algorithms, and we will provably verify this argument via two concrete examples in Section 4. In contrast, the third factor is priorly decided by the pre-defined data augmentation and is unrelated to algorithms. We will empirically verify that the concentration of augmented data affects the downstream performance in Section 5.

Compared with the alignment and uniformity proposed by Wang and Isola (2020), both of the works have the same meaning of "alignment" since it is the objective that contrastive algorithms aim to optimize, but our "divergence" is fundamentally different from their "uniformity". Uniformity requires "all data" uniformly distributed on the embedding hypersphere, while our divergence characterizes the cosine distance between "class centers". We do not require the divergence to be as large as better, instead, the divergence condition can be loosened by better alignment and concentration properties. More importantly, alignment and uniformity are empirical predictors for downstream performance, while our alignment and divergence have explicit theoretical guarantees (Theorem 1) for the generalization of contrastive SSL. Moreover, (Wang and Isola, 2020) does not consider the crucial effect of data augmentation on the generalization. In fact, with bad concentration (e.g., using identity transform as data augmentation), perfect alignment and perfect uniformity still can not imply clustered embedding space.

3.1 Upper Bound of R_ε

We now upper bound the term R_ε in the error rate (3) via $\mathbb{E}_{\boldsymbol{x}}\,\mathbb{E}_{\boldsymbol{x}_1,\boldsymbol{x}_2\in A(\boldsymbol{x})}\,\|f(\boldsymbol{x}_1) - f(\boldsymbol{x}_2)\|^2$, which is the common alignment objective of contrastive losses (4).

We separate the augmentation set A as discrete transformations $\{A_\gamma(\cdot)\colon \gamma \in [m]\}$ and continuous transformations $\{A_\theta(\cdot)\colon \theta \in [0,1]^n\}$. For example, random cropping or flipping can be categorized into the discrete transformation, while the others like random color distortion or Gaussian blur can be regarded as the continuous transformation parameterized by the augmentation strength θ. Without loss of generality, we assume that for any given $\boldsymbol{x}$, its augmented data are uniformly random sampled, i.e., $\mathbb{P}[\boldsymbol{x}' = A_\gamma(\boldsymbol{x})] = \frac{1}{2m}$ and $\mathbb{P}[\boldsymbol{x}' \in \{A_\theta(\boldsymbol{x}) \colon \theta \in \Theta\}] = \frac{\mathrm{vol}(\Theta)}{2}$ for any $\Theta \subseteq [0,1]^n$, where $\mathrm{vol}(\Theta)$ denotes the volume of Θ. For the continuous transformation, we further assume that the transformation is M-Lipschitz continuous w.r.t. θ, i.e., $\|A_{\theta_1}(\boldsymbol{x}) - A_{\theta_2}(\boldsymbol{x})\| \le M\|\theta_1 - \theta_2\|$ for any $\boldsymbol{x}, \theta_1, \theta_2$. Based on the above setting, we have the following guarantee of R_ε.

Theorem 2. *If encoder f is L-Lipschitz continuous, then*

$$R_\varepsilon^2 \le \eta(\varepsilon)^2 \cdot \mathop{\mathbb{E}}_{\boldsymbol{x}} \mathop{\mathbb{E}}_{\boldsymbol{x}_1, \boldsymbol{x}_2 \in A(\boldsymbol{x})} \|f(\boldsymbol{x}_1) - f(\boldsymbol{x}_2)\|^2 ,$$

where $\eta(\varepsilon) = \inf_{h \in \left(0, \frac{\varepsilon}{2\sqrt{n}LM}\right)} \frac{4\max\{1, m^2h^{2n}\}}{h^{2n}(\varepsilon - 2\sqrt{n}LMh)} = \mathcal{O}\left(\frac{1}{\varepsilon}\right)$.

The above theorem confirms that with good alignment, R_ε is guaranteed to be small.

4 Contrastive Losses Meet Alignment and Divergence

The objective of most contrastive algorithms can be formulated as

$$\mathcal{L}(f) = \mathcal{L}_{\text{pos}}(f) + \mathcal{L}_{\text{reg}}(f). \tag{4}$$

The first term $\mathcal{L}_{\text{pos}}(f)$ measures the alignment of positive samples, and usually takes the form of $\mathbb{E}_{\boldsymbol{x}} \mathbb{E}_{\boldsymbol{x}_1, \boldsymbol{x}_2 \in A(\boldsymbol{x})} \|f(\boldsymbol{x}_1) - f(\boldsymbol{x}_2)\|^2$. The second term $\mathcal{L}_{\text{reg}}(f)$ is the regularizer for preventing the collapse of representation. The main difference among algorithms lies in the choice of $\mathcal{L}_{\text{reg}}(f)$, i.e., an effective regularizer should make the angles between class centers large enough according to (2). In this section, we study two canonical contrastive objectives, InfoNCE and cross-correlation loss, to see how they meet the alignment and divergence properties required by Theorem 1. We highlight that the cross-correlation loss which decorrelates the components of representation, does NOT directly optimize the geometry of embedding space, but it still can result in clustered embedding space.

4.1 Warmup: InfoNCE Loss

InfoNCE loss is widely adopted by contrastive algorithms such as SimCLR (Chen et al., 2020a) and MoCo (He et al., 2020):

$$\mathcal{L}_{\text{InfoNCE}} = -\mathop{\mathbb{E}}_{\boldsymbol{x}, \boldsymbol{x}'} \mathop{\mathbb{E}}_{\substack{\boldsymbol{x}_1, \boldsymbol{x}_2 \in A(\boldsymbol{x}) \\ \boldsymbol{x}^- \in A(\boldsymbol{x}')}} \log \frac{e^{f(\boldsymbol{x}_1)^\top f(\boldsymbol{x}_2)}}{e^{f(\boldsymbol{x}_1)^\top f(\boldsymbol{x}_2)} + e^{f(\boldsymbol{x}_1)^\top f(\boldsymbol{x}^-)}}.$$

It can be divided into the two parts as the form of (4):

$$\begin{aligned} \mathcal{L}_{\text{InfoNCE}} &= \mathop{\mathbb{E}}_{\boldsymbol{x}, \boldsymbol{x}'} \mathop{\mathbb{E}}_{\substack{\boldsymbol{x}_1, \boldsymbol{x}_2 \in A(\boldsymbol{x}) \\ \boldsymbol{x}^- \in A(\boldsymbol{x}')}} \left[-f(\boldsymbol{x}_1)^\top f(\boldsymbol{x}_2) + \log\left(e^{f(\boldsymbol{x}_1)^\top f(\boldsymbol{x}_2)} + e^{f(\boldsymbol{x}_1)^\top f(\boldsymbol{x}^-)} \right) \right] \\ &= \underbrace{\frac{1}{2} \mathop{\mathbb{E}}_{\boldsymbol{x}} \mathop{\mathbb{E}}_{\boldsymbol{x}_1, \boldsymbol{x}_2 \in A(\boldsymbol{x})} [\|f(\boldsymbol{x}_1) - f(\boldsymbol{x}_2)\|^2] - 1}_{=:\mathcal{L}_1(f)} + \underbrace{\mathop{\mathbb{E}}_{\boldsymbol{x}, \boldsymbol{x}'} \mathop{\mathbb{E}}_{\substack{\boldsymbol{x}_1, \boldsymbol{x}_2 \in A(\boldsymbol{x}) \\ \boldsymbol{x}^- \in A(\boldsymbol{x}')}} \left[\log\left(e^{f(\boldsymbol{x}_1)^\top f(\boldsymbol{x}_2)} + e^{f(\boldsymbol{x}_1)^\top f(\boldsymbol{x}^-)} \right) \right]}_{=:\mathcal{L}_2(f)}, \end{aligned} \tag{5}$$

where the second equality uses the normalization condition of $\|f\| = 1$. Regardless of the constant factors, we can see that $\mathcal{L}_1(f)$ is exactly the term $\mathcal{L}_{\text{pos}}(f)$ introduced in (4), which imposes an upper bound to R_ε by Theorem 2. Thus, the requirement of alignment is naturally full-filled. Next, we show that $\mathcal{L}_2(f)$ can satisfy the requirement of divergence.

Theorem 3. *Assume that encoder f with norm 1 is L-Lipschitz continuous. If the augmented data is (σ, δ)-augmented, then for any $\varepsilon > 0$ and $k \neq \ell$,*

$$\mu_k^\top \mu_\ell \leq \log \left(\exp \left\{ \frac{\mathcal{L}_2(f) + \tau(\varepsilon, \sigma, \delta)}{p_k p_\ell} \right\} - \exp(1 - \varepsilon) \right) \lesssim \mathcal{L}_2(f),$$

where $\tau(\varepsilon, \sigma, \delta)$ is related to the mean of intra-class variance in the embedding space.

The specific formulation of $\tau(\varepsilon, \sigma, \delta)$ is deferred to the appendix. Here we remark that $\tau(\varepsilon, \sigma, \delta)$ depends on both R_ε and augmentation parameters (σ, δ): better alignment (hence less R_ε) and sharper concentration of augmented data imply smaller $\tau(\varepsilon, \sigma, \delta)$. Therefore, with the good alignment property, the divergence $\mu_k^\top \mu_\ell$ is mainly controlled by $\mathcal{L}_2(f)$. In summary, minimizing the InfoNCE loss leads to both small $\mathcal{L}_1(f)$ and $\mathcal{L}_2(f)$, which guarantee good alignment and divergence, respectively. Thus, according to Theorem 1 and Theorem 2, the generalization ability of encoder f on the downstream task is implied, i.e.,

$$\mathrm{Err}(G_f) \leq (1 - \sigma) + \eta(\varepsilon) \sqrt{2 + 2\mathcal{L}_1(f)},$$

when the upper bound of $\mu_k^\top \mu_\ell$ in Theorem 3 is smaller than the threshold in Theorem 1.

It is worth mentioning that the form of InfoNCE is critical to meeting the requirement of divergence, which is found when we prove Theorem 3. For example, let us consider the contrastive loss (5) formulated in a linear form[1] instead of `LogExp` such that

$$\mathcal{L}'(f) = \mathop{\mathbb{E}}_{\boldsymbol{x}, \boldsymbol{x}'} \mathop{\mathbb{E}}_{\substack{\boldsymbol{x}_1, \boldsymbol{x}_2 \in A(\boldsymbol{x}) \\ \boldsymbol{x}^- \in A(\boldsymbol{x}')}} \left[-f(\boldsymbol{x}_1)^\top f(\boldsymbol{x}_2) + \lambda f(\boldsymbol{x}_1)^\top f(\boldsymbol{x}^-) \right] = \mathcal{L}_1(f) + \lambda \mathcal{L}_2'(f),$$

where $\mathcal{L}_2'(f)$ is the regularizer weighted by some $\lambda > 0$. Due to the independence between $\boldsymbol{x}$ and $\boldsymbol{x}'$, we have $\mathcal{L}_2'(f) = \| \mathbb{E}_{\boldsymbol{x}} \mathbb{E}_{\boldsymbol{x}_1 \in A(\boldsymbol{x})} [f(\boldsymbol{x}_1)] \|^2$. Therefore, minimizing $\mathcal{L}_2'(f)$ only leads to the representations with zero mean. Unfortunately, the objective of zero mean with $\|f\| = 1$ can not obviate the dimensional collapse (Hua et al., 2021) of the model. For example, the encoder f can map the input data from multi classes into two points in the opposite directions on the hypersphere. This justifies the observation in (Wang and Liu, 2021): the uniformity of encoder on the embedded hypersphere becomes worse when the temperature of the loss increases, where the loss degenerates to $\mathcal{L}'(f)$ with infinite temperature.

4.2 Cross-Correlation Loss

Cross-correlation loss is a recently proposed loss in Barlow Twins (Zbontar et al., 2021). In contrast to InfoNCE loss, it trains the model via decorrelating the components of representation instead of directly optimizing the geometry of embedding space. In this section, we deeply explore the cross-correlation loss, and show that it secretly optimizes the alignment and divergence factors required by Theorem 1.

The cross-correlation loss can be formulated as

$$\mathcal{L}_{\text{Cross-Corr}} = \sum_{i=1}^{d} \left(1 - \mathop{\mathbb{E}}_{\boldsymbol{x}} \mathop{\mathbb{E}}_{\boldsymbol{x}_1, \boldsymbol{x}_2 \in A(\boldsymbol{x})} [f_i(\boldsymbol{x}_1) f_i(\boldsymbol{x}_2)] \right)^2 + \lambda \sum_{i \neq j} \left(\mathop{\mathbb{E}}_{\boldsymbol{x}} \mathop{\mathbb{E}}_{\boldsymbol{x}_1, \boldsymbol{x}_2 \in A(\boldsymbol{x})} [f_i(\boldsymbol{x}_1) f_j(\boldsymbol{x}_2)] \right)^2,$$

with the normalization of $\mathbb{E}_{\boldsymbol{x}} \mathbb{E}_{\boldsymbol{x}_1 \in A(\boldsymbol{x})} [f_i(\boldsymbol{x}_1)] = 0$ and $\mathbb{E}_{\boldsymbol{x}} \mathbb{E}_{\boldsymbol{x}_1 \in A(\boldsymbol{x})} [f_i(\boldsymbol{x}_1)^2] = 1$ for each $i \in [d]$, where d is the output dimension of encoder f. Positive coefficient λ balances the importance between diagonal and non-diagonal elements of cross-correlation matrix. When $\lambda = 1$, the above

[1] It is also called *simple contrastive loss* in some literature.

loss is exactly the difference between the cross-correlation matrix and identity matrix. Similar to Section 4.1, we first decompose the loss into two terms, by defining

$$\mathcal{L}_1(f) := \sum_{i=1}^{d} \left(1 - \mathop{\mathbb{E}}_{\boldsymbol{x}} \mathop{\mathbb{E}}_{\boldsymbol{x}_1, \boldsymbol{x}_2 \in A(\boldsymbol{x})} [f_i(\boldsymbol{x}_1) f_i(\boldsymbol{x}_2)]\right)^2 \text{ and } \mathcal{L}_2(f) := \left\| \mathop{\mathbb{E}}_{\boldsymbol{x}} \mathop{\mathbb{E}}_{\boldsymbol{x}_1, \boldsymbol{x}_2 \in A(\boldsymbol{x})} [f(\boldsymbol{x}_1) f(\boldsymbol{x}_2)^\top] - I_d \right\|^2.$$

In this way, the cross-correlation loss becomes

$$\mathcal{L}_{\text{Cross-Corr}} = (1 - \lambda)\, \mathcal{L}_1(f) + \lambda\, \mathcal{L}_2(f).$$

Then, we will show that $\mathcal{L}_1(f)$ and $\mathcal{L}_2(f)$ control the alignment and divergence, respectively.

Lemma 4.1. *For a given encoder f, the alignment $\mathcal{L}_{\text{pos}}(f)$ in (4) is upper bounded via $\mathcal{L}_1(f)$, i.e.,*

$$\mathcal{L}_{\text{pos}}(f) = \mathop{\mathbb{E}}_{\boldsymbol{x}} \mathop{\mathbb{E}}_{\boldsymbol{x}_1, \boldsymbol{x}_2 \in A(\boldsymbol{x})} \|f(\boldsymbol{x}_1) - f(\boldsymbol{x}_2)\|^2 \le 2\sqrt{d \mathcal{L}_1(f)},$$

where d is the output dimension of encoder f.

The above lemma connects $\mathcal{L}_{\text{pos}}(f)$ with $\mathcal{L}_1(f)$, indicating that the diagonal elements of the cross-correlation matrix determine the alignment of positive samples. Next, we will connect the divergence $\mu_k^\top \mu_\ell$ with the second loss term $\mathcal{L}_2(f)$. It is challenging because $\mathcal{L}_2(f)$ is designed for reducing the redundancy between encoder's output units, not for optimizing the geometry of embedding space.

When a good alignment is satisfied, one can expect that $f(\boldsymbol{x}_1) \approx f(\boldsymbol{x}_2)$ for any $\boldsymbol{x}_1, \boldsymbol{x}_2 \in A(\boldsymbol{x})$ for any $\boldsymbol{x}$. Thus, $\mathcal{L}_2(f)$ approximately equals to $\| \mathbb{E}_{\boldsymbol{x}} \mathbb{E}_{\boldsymbol{x}_1 \in A(\boldsymbol{x})} [f(\boldsymbol{x}_1) f(\boldsymbol{x}_1)^\top] - I_d\|^2$ by replacing $f(\boldsymbol{x}_2)$ with $f(\boldsymbol{x}_1)$. On the other hand, we prove in the appendix that most samples are embedded to the surrounding area of its class center, i.e., small $\mathbb{E}_{\boldsymbol{x} \in C_k} \mathbb{E}_{\boldsymbol{x}_1 \in A(\boldsymbol{x})} \|f(\boldsymbol{x}_1) - \mu_k\|^2$. Therefore, we can further approximate $\mathbb{E}_{\boldsymbol{x}} \mathbb{E}_{\boldsymbol{x}_1 \in A(\boldsymbol{x})} [f(\boldsymbol{x}_1) f(\boldsymbol{x}_1)^\top]$ by the weighted class center $\sum_{k=1}^{K} p_k \mu_k \mu_k^\top$. Based on the above idea, we conclude that the second loss term $\mathcal{L}_2(f)$ approximately captures the difference between $\sum_{k=1}^{K} p_k \mu_k \mu_k^\top$ and the identity matrix, namely,

$$\mathbb{E}\left[f(\boldsymbol{x}_1) f(\boldsymbol{x}_2)^\top\right] \approx \mathbb{E}\left[f(\boldsymbol{x}_1) f(\boldsymbol{x}_1)^\top\right] \approx \sum_{k=1}^{K} p_k \mu_k \mu_k^\top \;\Rightarrow\; \mathcal{L}_2(f) \approx \left\| \sum_{k=1}^{K} p_k \mu_k \mu_k^\top - I_d \right\|^2.$$

Notice that the form of $\mu_k \mu_k^\top$ is only one step away from our goal $\mu_k^\top \mu_\ell$. In fact, we can further prove that $\| \sum_{k=1}^{K} p_k \mu_k \mu_k^\top - I_d\|^2 \ge p_k p_\ell (\mu_k^\top \mu_\ell)^2 + d - K$, where $d - K$ appears due to the rank mismatch between $\sum_{k=1}^{K} p_k \mu_k \mu_k^\top$ and I_d. In this way, we can derive the following theorem.

Theorem 4. *Assume that encoder f with norm $\sqrt{d}$ is L-Lipschitz continuous. If the augmented data used in Barlow Twins is (σ, δ)-augmented, then for any $k \ne \ell$, we have*

$$\mu_k^\top \mu_\ell \le \sqrt{\frac{2}{p_k p_\ell} \left(\mathcal{L}_2(f) + \tau'(\varepsilon, \sigma, \delta) - \frac{d - K}{2}\right)} \lesssim \mathcal{L}_2(f)^{1/2},$$

where $\tau'(\varepsilon, \sigma, \delta)$ is the upper bound of $\| \mathbb{E}_{\boldsymbol{x}} \mathbb{E}_{\boldsymbol{x}_1, \boldsymbol{x}_2 \in A(\boldsymbol{x})} [f(\boldsymbol{x}_1) f(\boldsymbol{x}_2)^\top] - \sum_{k=1}^{K} p_k \mu_k \mu_k^\top\|^2$.

The specific formulation of $\tau'(\varepsilon, \sigma, \delta)$ is deferred to the appendix. Here we remark that $\tau'(\varepsilon, \sigma, \delta)$ depends on both R_ε and augmentation parameters (σ, δ): better alignment (hence the less R_ε) and sharper concentration of augmented data imply smaller $\tau'(\varepsilon, \sigma, \delta)$. Therefore, with the good alignment property, the divergence $\mu_k^\top \mu_\ell$ is mainly controlled by $\mathcal{L}_2(f)$.

In summary, decorrelating the components of representation leads to the small $\mathcal{L}_1(f)$ as well as the $\mathcal{L}_2(f)$. With Lemma 4.1 and Theorem 4, we can conclude that the good alignment and divergence are guaranteed. Thus, according to Theorem 1 and Theorem 2, the generalization ability of f on the downstream task is implied, i.e.,

$$\mathrm{Err}(G_f) \le (1-\sigma) + \sqrt{2}\,\eta(\varepsilon)\,d^{\frac{1}{4}}\mathcal{L}_1(f)^{\frac{1}{4}},$$

when the upper bound of $\mu_k^\top \mu_\ell$ in Theorem 4 is smaller than the threshold in Theorem 1.

5 Experimental Verification of Augmented Data Concentration

We have provably verified that alignment and divergence can be achieved via widely used contrastive losses in the last section. We now move to the third one – concentration of augmented data. In this section, we empirically verify the argument made in Section 3 that *sharper concentration* of augmented data w.r.t. the augmented distance implies *better generalization* of contrastive SSL.

To construct the augmented data with different concentration levels, we design three groups of experiments: 1) since richer data augmentations imply sharper concentration (see the paragraph below Definition 1), we conduct experiments with an increasing number of transformations; 2) when the transform types are fixed, stronger transformations exhibit sharper concentration, and thus we conduct experiments with different strength levels of transformations; 3) to more directly verify our (σ, δ)-augmentation modeling, we fix the number of transformations to two and conduct experiments with different transformation pairs to observe the correlation between the downstream performance and concentration of augmented data.

Basic Setup. Our experiments are conducted on CIFAR-10 (Krizhevsky, 2009), which is a colorful image dataset with 50000 training samples and 10000 test samples from 10 categories. We consider 5 kinds of transformations: (a) random cropping; (b) random Gaussian blur; (c) color dropping (i.e., randomly converting images to grayscale); (d) color distortion; (e) random horizontal flipping. We use ResNet-18 (He et al., 2016) as the encoder, and the other settings such as projection head all use the original settings of algorithms. To evaluate the quality of the encoder, we follow the KNN evaluation protocol (Wu et al., 2018). We train each model via various algorithms (Chen et al., 2020a; Zbontar et al., 2021; He et al., 2020; Chen and He, 2021) with batch size of 512 and 800 epochs.

Different Richness of Augmentations. We compose all 5 kinds of transformations together, and successively drop one of the composed transformations from (e) to (b) to conduct a total of 5 experiments. Accordingly, the concentration of augmented data goes from sharp to flat.

Different Strength Levels of Transformations. We fix (a) and (d) as transformations, and vary the strength of (d) in $\{1, \frac{1}{2}, \frac{1}{4}, \frac{1}{8}\}$, where the concentration of augmented data goes from sharp to flat.

Different Transformation Pairs. We apply 5 kinds of transforms in pairs to conduct a total of 10 experiments. We observe the correlation between $\mathrm{Err}(G_f)$ and $(1-\sigma)$ for different δ, suggested by Theorem 1. Each model is trained by SimCLR with batch size of 512 and 200 epochs.

Compute σ. When δ is given, for each class C_k, we construct an auxiliary graph G_k whose nodes correspond to the samples of C_k and edge $(\boldsymbol{x}_1, \boldsymbol{x}_2)$ exists if $d_A(\boldsymbol{x}_1, \boldsymbol{x}_2) \le \delta$. According to Definition 1, we can compute the main part of C_k by finding the maximum clique of graph G_k. Then σ can be estimated by $\min_{k\in[K]} |\textsc{MaxClique}(G_k)|/|C_k|$. We solve $\textsc{MaxClique}$ via its dual problem – vertex cover, and adopt the Approx-Vertex-Cover (Papadimitriou and Steiglitz, 1998) to compute the solution.

Table 1: Performance with different data augmentations.

Transformations					Accuracy	
(a)	(b)	(c)	(d)	(e)	SimCLR (Chen et al., 2020a)	Barlow Twins (Zbontar et al., 2021)
✓	✓	✓	✓	✓	$\mathbf{89.92 \pm 0.05}$	$\mathbf{83.93 \pm 0.57}$
✓	✓	✓	✓	×	88.41 ± 0.11	83.37 ± 0.43
✓	✓	✓	×	×	83.62 ± 0.19	73.70 ± 0.99
✓	✓	×	×	×	62.91 ± 0.25	49.56 ± 0.11
✓	×	×	×	×	62.37 ± 0.09	48.54 ± 0.29

Color Distortion Strength	SimCLR (Chen et al., 2020a)	Barlow Twins (Zbontar et al., 2021)
1	$\mathbf{82.64 \pm 0.57}$	$\mathbf{78.79 \pm 0.54}$
1/2	78.49 ± 0.09	72.76 ± 1.50
1/4	76.25 ± 0.16	68.30 ± 0.15
1/8	73.60 ± 0.11	61.13 ± 2.81

Main Results. In the first table of Table 1, the downstream performance monotonously increases with the number of operations, under different algorithms. This verifies that richer transformations lead to sharper concentration and thus better generalization. We also observe that color dropping (c) and distortion (d) have a great impact on the performance of all algorithms. This is because they enable the augmented data to vary in a very wide range, which makes the augmented distance (1) largely decrease. Thus, the concentration of augmented data becomes sharper (smaller δ) and the generalization gets better. As an intuitive example, if the right dog image in Figure 2 is replaced by a Husky image, with random cropping, one can only get two dog heads with similar shapes but different colors, which still have a large augmented distance. Instead, if color distortion is further applied, they will be similar both in shape and color. Therefore, concentration will get sharper and the performance will be better. In fact, fixing cropping and color distortion as transformations, we indeed observe in the second table of Table 1 that the downstream performance increases with stronger color distortions.

More directly, Figure 3 shows that $\mathrm{Err}(G_f)$ is highly correlated to the computed value of $(1-\sigma)$: when δ is fixed, smaller $(1-\sigma)$ indicating sharper concentration, results in small $\mathrm{Err}(G_f)$. If we fix one transform as (a), both $\mathrm{Err}(G_f)$ and $(1-\sigma)$ have the same order that $(a,d) < (a,c) < (a,e) < (a,b)$. Furthermore, both of them indicate that the combination of (a) and (d) is the most effective pair. In addition, we observe that the choice of δ is not sensitive to the variation tendency of $(1-\sigma)$ but affects their absolute values. This matches our Theorem 1 that, when the divergence condition is satisfied, a larger δ implies a larger σ, leading to a tighter upper bound of (3).

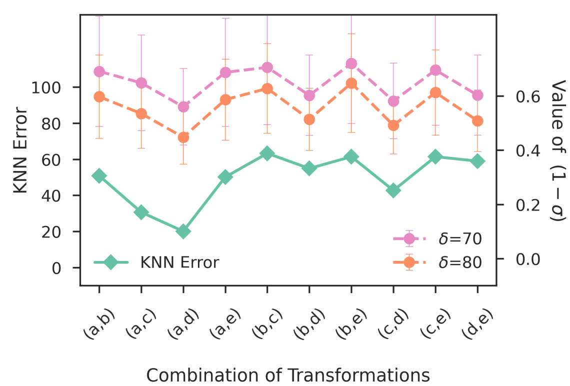


Figure 3: The correlation between observed $\mathrm{Err}(G_f)$ and computed value of $(1-\sigma)$.

6 Future Work

In this paper, we theoretically study the generalization guarantees of contrastive SSL. One future work can be relaxing the assumption that $\cap_{k=1}^{K} A(C_k) = \varnothing$, corresponding to the situation that the data augmentation is too aggressive. Another possible direction is to refine the solution by taking the sample size into account, to justify the phenomenon that batch size is sensitive to the performance of SimCLR but not sensitive for Barlow Twins. A more challenging direction can be studying the out-of-distribution generalization of contrastive SSL based on our framework, where the upstream data and downstream data have different distributions.

References

Sanjeev Arora, Hrishikesh Khandeparkar, Mikhail Khodak, Orestis Plevrakis, and Nikunj Saunshi. A theoretical analysis of contrastive unsupervised representation learning. *arXiv preprint arXiv:1902.09229*, 2019.

Philip Bachman, R Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. In *Advances in Neural Information Processing Systems*, pages 15535–15545, 2019.

Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *arXiv preprint arXiv:2105.04906*, 2021.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020a.

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15750–15758, 2021.

Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020b.

Hongchao Fang, Sicheng Wang, Meng Zhou, Jiayuan Ding, and Pengtao Xie. Cert: Contrastive self-supervised learning for language understanding. *arXiv preprint arXiv:2005.12766*, 2020.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*, 2021.

John M Giorgi, Osvald Nitski, Gary D Bader, and Bo Wang. Declutr: Deep contrastive learning for unsupervised textual representations. *arXiv preprint arXiv:2006.03659*, 2020.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint arXiv:2006.07733*, 2020.

Jeff Z HaoChen, Colin Wei, Adrien Gaidon, and Tengyu Ma. Provable guarantees for self-supervised deep learning with spectral contrastive loss. *Advances in Neural Information Processing Systems*, 34, 2021.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 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. *arXiv preprint arXiv:1808.06670*, 2018.

Qianjiang Hu, Xiao Wang, Wei Hu, and Guo-Jun Qi. Adco: Adversarial contrast for efficient learning of unsupervised representations from self-trained negative adversaries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1074–1083, 2021.

Tianyu Hua, Wenxiao Wang, Zihui Xue, Sucheng Ren, Yue Wang, and Hang Zhao. On feature decorrelation in self-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9598–9608, 2021.

Alex Krizhevsky. Learning multiple layers of features from tiny images. *University of Toronto*, 2009.

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

Christos H Papadimitriou and Kenneth Steiglitz. *Combinatorial optimization: algorithms and complexity*. Courier Corporation, 1998.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. *arXiv preprint arXiv:1906.05849*, 2019.

Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning? *arXiv preprint arXiv:2005.10243*, 2020.

Michael Tschannen, Josip Djolonga, Paul K Rubenstein, Sylvain Gelly, and Mario Lucic. On mutual information maximization for representation learning. *arXiv preprint arXiv:1907.13625*, 2019.

Feng Wang and Huaping Liu. Understanding the behaviour of contrastive loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2495–2504, 2021.

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

Colin Wei, Kendrick Shen, Yining Chen, and Tengyu Ma. Theoretical analysis of self-training with deep networks on unlabeled data. *arXiv preprint arXiv:2010.03622*, 2020.

Zixin Wen and Yuanzhi Li. Toward understanding the feature learning process of self-supervised contrastive learning. *arXiv preprint arXiv:2105.15134*, 2021.

Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3733–3742, 2018.

Zhuofeng Wu, Sinong Wang, Jiatao Gu, Madian Khabsa, Fei Sun, and Hao Ma. Clear: Contrastive learning for sentence representation. *arXiv preprint arXiv:2012.15466*, 2020.

Yuanmeng Yan, Rumei Li, Sirui Wang, Fuzheng Zhang, Wei Wu, and Weiran Xu. Consert: A contrastive framework for self-supervised sentence representation transfer. *arXiv preprint arXiv:2105.11741*, 2021.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. *arXiv preprint arXiv:2103.03230*, 2021.

Appendix

We list all the proofs of lemmas and theorems in the appendix, which are organized by sections.

A Proofs for Section 3

Lemma 3.1. *For a (σ, δ)-augmentation with main part C_k^0 of each class C_k, if all samples belonging to $(C_1^0 \cup \cdots \cup C_K^0) \cap S_\varepsilon$ can be correctly classified by a classifier G, then its classification error rate $\mathrm{Err}(G)$ is upper bounded by $(1-\sigma) + R_\varepsilon$.*

Proof. Since every sample $\boldsymbol{x} \in (C_1^0 \cup \cdots \cup C_K^0) \cap S_\varepsilon$ can be correctly classified by G, then the classification error rate

$$
\begin{aligned}
\mathrm{Err}(G) &= \sum_{k=1}^{K} \mathbb{P}[G(\boldsymbol{x}) \neq k, \forall \boldsymbol{x} \in C_k] \\
&\leq \mathbb{P}\left[\overline{(C_1^0 \cup \cdots \cup C_K^0) \cap S_\varepsilon}\right] \\
&= \mathbb{P}\left[\overline{C_1^0 \cup \cdots \cup C_K^0} \cup \overline{S_\varepsilon}\right] \\
&\leq (1-\sigma) + \mathbb{P}\left[\overline{S_\varepsilon}\right] \\
&= (1-\sigma) + R_\varepsilon.
\end{aligned}
$$

This finishes the proof. □

Lemma 3.2. *Given a (σ, δ)-augmentation used in contrastive SSL, for any $\ell \in [K]$, if $\mu_\ell^\top \mu_k < r^2\left(1 - \rho_\ell(\sigma,\delta,\varepsilon) - \sqrt{2\rho_\ell(\sigma,\delta,\varepsilon)} - \frac{\Delta_\mu}{2}\right)$ holds for all $k \neq \ell$, then every sample $\boldsymbol{x} \in C_\ell^0 \cap S_\varepsilon$ can be correctly classified by the NN classifier G_f, where $\rho_\ell(\sigma,\delta,\varepsilon) = 2(1-\sigma) + \frac{R_\varepsilon}{p_\ell} + \sigma\left(\frac{L\delta}{r} + \frac{2\varepsilon}{r}\right)$ and $\Delta_\mu = 1 - \min_{k\in[K]} \|\mu_k\|^2/r^2$.*

Proof. Without loss of generality, we consider $\ell = 1$. To show that every sample $\boldsymbol{x}_0 \in C_1^0 \cap S_\varepsilon$ can be correctly classified by G_f, we need to prove that for all $k \neq 1$, $\|f(\boldsymbol{x}_0) - \mu_1\| < \|f(\boldsymbol{x}_0) - \mu_k\|$. It is equivalent to prove that

$$
f(\boldsymbol{x}_0)^\top \mu_1 - f(\boldsymbol{x}_0)^\top \mu_k - \left(\frac{1}{2}\|\mu_1\|^2 - \frac{1}{2}\|\mu_k\|^2\right) > 0. \tag{6}
$$

Let $\tilde{f}(\boldsymbol{x}) := \mathbb{E}_{\boldsymbol{x}'\in A(\boldsymbol{x})}[f(\boldsymbol{x}')]$. Then $\|\tilde{f}(\boldsymbol{x})\| = \|\mathbb{E}_{\boldsymbol{x}'\in A(\boldsymbol{x})}[f(\boldsymbol{x}')]\| \leq \mathbb{E}_{\boldsymbol{x}'\in A(\boldsymbol{x})}[\|f(\boldsymbol{x}')\|] = r$.

On the one hand,

$$
\begin{aligned}
f(\boldsymbol{x}_0)^\top \mu_1 &= \frac{1}{p_1} f(\boldsymbol{x}_0)^\top \mathop{\mathbb{E}}_{\boldsymbol{x}}[\tilde{f}(\boldsymbol{x})\mathbb{I}(\boldsymbol{x}\in C_1)] \\
&= \frac{1}{p_1} f(\boldsymbol{x}_0)^\top \mathop{\mathbb{E}}_{\boldsymbol{x}}[\tilde{f}(\boldsymbol{x})\mathbb{I}(\boldsymbol{x}\in C_1 \cap C_1^0 \cap S_\varepsilon)] + \frac{1}{p_1} f(\boldsymbol{x}_0)^\top \mathop{\mathbb{E}}_{\boldsymbol{x}}\left[\tilde{f}(\boldsymbol{x})\mathbb{I}(\boldsymbol{x}\in C_1 \cap \overline{C_1^0 \cap S_\varepsilon})\right] \\
&= \frac{\mathbb{P}[C_1^0 \cap S_\varepsilon]}{p_1} f(\boldsymbol{x}_0)^\top \mathop{\mathbb{E}}_{\boldsymbol{x}\in C_1^0\cap S_\varepsilon}[\tilde{f}(\boldsymbol{x})] + \frac{1}{p_1}\mathop{\mathbb{E}}_{\boldsymbol{x}}\left[f(\boldsymbol{x}_0)^\top \tilde{f}(\boldsymbol{x}) \cdot \mathbb{I}(\boldsymbol{x}\in C_1 \setminus C_1^0 \cap S_\varepsilon)\right] \\
&\geq \frac{\mathbb{P}[C_1^0 \cap S_\varepsilon]}{p_1} f(\boldsymbol{x}_0)^\top \mathop{\mathbb{E}}_{\boldsymbol{x}\in C_1^0\cap S_\varepsilon}[\tilde{f}(\boldsymbol{x})] - \frac{r^2}{p_1}\mathbb{P}[C_1 \setminus C_1^0 \cap S_\varepsilon],
\end{aligned} \tag{7}
$$

where $\mathbb{I}(\cdot)$ is the indicator function. Note that

$$\mathbb{P}[C_1 \setminus C_1^0 \cap S_\varepsilon] = \mathbb{P}[(C_1 \setminus C_1^0) \cup (C_1^0 \cap \overline{S_\varepsilon})] \leq (1-\sigma)p_1 + R_\varepsilon, \tag{8}$$

and

$$\mathbb{P}[C_1^0 \cap S_\varepsilon] = \mathbb{P}[C_1] - \mathbb{P}[C_1 \setminus C_1^0 \cap S_\varepsilon] \geq p_1 - ((1-\sigma)p_1 + R_\varepsilon) = \sigma p_1 - R_\varepsilon. \tag{9}$$

Plugging to (7), we have

$$\begin{aligned} f(\boldsymbol{x}_0)^\top \mu_1 &\geq \frac{\mathbb{P}[C_1^0 \cap S_\varepsilon]}{p_1} f(\boldsymbol{x}_0)^\top \mathop{\mathbb{E}}_{\boldsymbol{x} \in C_1^0 \cap S_\varepsilon} [\tilde{f}(\boldsymbol{x})] - \frac{r^2}{p_1} \mathbb{P}[C_1 \setminus C_1^0 \cap S_\varepsilon] \\ &\geq \left(\sigma - \frac{R_\varepsilon}{p_1}\right) f(\boldsymbol{x}_0)^\top \mathop{\mathbb{E}}_{\boldsymbol{x} \in C_1^0 \cap S_\varepsilon} [\tilde{f}(\boldsymbol{x})] - r^2 \left(1 - \sigma + \frac{R_\varepsilon}{p_1}\right). \end{aligned} \tag{10}$$

Notice that $\boldsymbol{x}_0 \in C_1^0 \cap S_\varepsilon$. For any $\boldsymbol{x} \in C_1^0 \cap S_\varepsilon$, we have $d_A(\boldsymbol{x}_0, \boldsymbol{x}) \leq \delta$. Let $(\boldsymbol{x}_0^*, \boldsymbol{x}^*) = \arg\min_{\boldsymbol{x}_0' \in A(\boldsymbol{x}_0), \boldsymbol{x}' \in A(\boldsymbol{x})} \|\boldsymbol{x}_0' - \boldsymbol{x}'\|$. We have $\|\boldsymbol{x}_0^* - \boldsymbol{x}^*\| \leq \delta$. Since f is L-Lipschitz continuous, we have $\|f(\boldsymbol{x}_0^*) - f(\boldsymbol{x}^*)\| \leq L \cdot \|\boldsymbol{x}_0^* - \boldsymbol{x}^*\| \leq L\delta$. Since $\boldsymbol{x} \in S_\varepsilon$, for any $\boldsymbol{x}' \in A(\boldsymbol{x})$, $\|f(\boldsymbol{x}') - f(\boldsymbol{x}^*)\| \leq \varepsilon$. Similarly, since $\boldsymbol{x}_0 \in S_\varepsilon$ and $\boldsymbol{x}_0, \boldsymbol{x}_0^* \in A(\boldsymbol{x}_0)$, we have $\|f(\boldsymbol{x}_0) - f(\boldsymbol{x}_0^*)\| \leq \varepsilon$.

The first term of (10) can be bounded by

$$\begin{aligned} & f(\boldsymbol{x}_0)^\top \mathop{\mathbb{E}}_{\boldsymbol{x} \in C_1^0 \cap S_\varepsilon} [\tilde{f}(\boldsymbol{x})] \\ &= \mathop{\mathbb{E}}_{\boldsymbol{x} \in C_1^0 \cap S_\varepsilon} \mathop{\mathbb{E}}_{\boldsymbol{x}' \in A(\boldsymbol{x})} [f(\boldsymbol{x}_0)^\top f(\boldsymbol{x}')] \\ &= \mathop{\mathbb{E}}_{\boldsymbol{x} \in C_1^0 \cap S_\varepsilon} \mathop{\mathbb{E}}_{\boldsymbol{x}' \in A(\boldsymbol{x})} [f(\boldsymbol{x}_0)^\top (f(\boldsymbol{x}') - f(\boldsymbol{x}_0) + f(\boldsymbol{x}_0))] \\ &= r^2 + \mathop{\mathbb{E}}_{\boldsymbol{x} \in C_1^0 \cap S_\varepsilon} \mathop{\mathbb{E}}_{\boldsymbol{x}' \in A(\boldsymbol{x})} [f(\boldsymbol{x}_0)^\top (f(\boldsymbol{x}') - f(\boldsymbol{x}_0))] \\ &= r^2 + \mathop{\mathbb{E}}_{\boldsymbol{x} \in C_1^0 \cap S_\varepsilon} \mathop{\mathbb{E}}_{\boldsymbol{x}' \in A(\boldsymbol{x})} [f(\boldsymbol{x}_0)^\top (\underbrace{f(\boldsymbol{x}') - f(\boldsymbol{x}^*)}_{\|\cdot\| \leq \varepsilon} + \underbrace{f(\boldsymbol{x}^*) - f(\boldsymbol{x}_0^*)}_{\|\cdot\| \leq L\delta} + \underbrace{f(\boldsymbol{x}_0^*) - f(\boldsymbol{x}_0)}_{\|\cdot\| \leq \varepsilon})] \\ &\geq r^2 - [r\varepsilon + rL\delta + r\varepsilon] \\ &= r^2 - r(L\delta + 2\varepsilon). \end{aligned}$$

Therefore, (10) turns to

$$\begin{aligned} f(\boldsymbol{x}_0)^\top \mu_1 &\geq \left(\sigma - \frac{R_\varepsilon}{p_1}\right) f(\boldsymbol{x}_0)^\top \mathop{\mathbb{E}}_{\boldsymbol{x} \in C_1^0 \cap S_\varepsilon} [\tilde{f}(\boldsymbol{x})] - r^2 \left(1 - \sigma + \frac{R_\varepsilon}{p_1}\right) \\ &\geq \left(\sigma - \frac{R_\varepsilon}{p_1}\right) (r^2 - r(L\delta + 2\varepsilon)) - r^2 \left(1 - \sigma + \frac{R_\varepsilon}{p_1}\right) \\ &= r^2 \left((2\sigma - 1) - \frac{R_\varepsilon}{p_1} - \left(\sigma - \frac{R_\varepsilon}{p_1}\right)\left(\frac{L\delta}{r} + \frac{2\varepsilon}{r}\right)\right) \\ &= r^2 \left(1 - 2(1-\sigma) - \frac{R_\varepsilon}{p_1} - \left(\sigma - \frac{R_\varepsilon}{p_1}\right)\left(\frac{L\delta}{r} + \frac{2\varepsilon}{r}\right)\right) \\ &= r^2(1 - \rho_1(\sigma, \delta, \varepsilon)). \end{aligned} \tag{11}$$

On the other hand,

$$\begin{aligned}
f(\boldsymbol{x}_0)^\top \mu_k &= (f(\boldsymbol{x}_0)-\mu_1)^\top \mu_k + \mu_1^\top \mu_k \\
&\le \|f(\boldsymbol{x}_0)-\mu_1\| \cdot \|\mu_k\| + \mu_1^\top \mu_k \\
&\le r\sqrt{\|f(\boldsymbol{x}_0)\|^2 - 2f(\boldsymbol{x}_0)^\top \mu_1 + \|\mu_1\|^2} + \mu_1^\top \mu_k \\
&\le r\sqrt{2r^2 - 2f(\boldsymbol{x}_0)^\top \mu_1} + \mu_1^\top \mu_k \\
&\le \sqrt{2\rho_1(\sigma,\delta,\varepsilon)} r^2 + \mu_1^\top \mu_k.
\end{aligned} \tag{12}$$

Note that $\Delta_\mu = 1 - \min_k \|\mu_k\|^2/r^2$, the LHS of (6) is

$$\begin{aligned}
&f(\boldsymbol{x}_0)^\top \mu_1 - f(\boldsymbol{x}_0)^\top \mu_k - \left(\frac{1}{2}\|\mu_1\|^2 - \frac{1}{2}\|\mu_k\|^2\right) \\
&\ge f(\boldsymbol{x}_0)^\top \mu_1 - f(\boldsymbol{x}_0)^\top \mu_k - \frac{1}{2} r^2 \Delta_\mu \\
&\ge r^2(1-\rho_1(\sigma,\delta,\varepsilon)) - \sqrt{2\rho_1(\sigma,\delta,\varepsilon)} r^2 - \mu_1^\top \mu_k - \frac{1}{2} r^2 \Delta_\mu \\
&= r^2\left(1-\rho_1(\sigma,\delta,\varepsilon) - \sqrt{2\rho_1(\sigma,\delta,\varepsilon)} - \frac{1}{2}\Delta_\mu\right) - \mu_1^\top \mu_k > 0,
\end{aligned}$$

where the second iequality is due to (11) and (12). This finishes the proof. □

Theorem 1. *Given a (σ,δ)-augmentation used in contrastive SSL, if*

$$\mu_\ell^\top \mu_k < r^2\left(1-\rho_{max}(\sigma,\delta,\varepsilon) - \sqrt{2\rho_{max}(\sigma,\delta,\varepsilon)} - \frac{\Delta_\mu}{2}\right) \tag{2}$$

holds for any pair of (ℓ,k) with $\ell \neq k$, then the downstream error rate of NN classifier G_f

$$\mathrm{Err}(G_f) \le (1-\sigma) + R_\varepsilon, \tag{3}$$

where $\rho_{max}(\sigma,\delta,\varepsilon) = 2(1-\sigma) + \frac{R_\varepsilon}{\min_\ell p_\ell} + \sigma\left(\frac{L\delta}{r} + \frac{2\varepsilon}{r}\right)$ and $\Delta_\mu = 1 - \min_{k\in[K]} \|\mu_k\|^2/r^2$.

Proof. Since the augmentation A is (σ,δ)-augmented, there exists a main part C_k^0 for each class C_k such that $\mathbb{P}[C_k^0] \ge \sigma p_k$ and $\sup_{\boldsymbol{x}_1,\boldsymbol{x}_2\in C_k^0} d_A(\boldsymbol{x}_1,\boldsymbol{x}_2) \le \delta$. Since for any $\ell \neq k$, we have $\mu_\ell^\top \mu_k < r^2\left(1-\rho_{max}(\sigma,\delta,\varepsilon) - \sqrt{2\rho_{max}(\sigma,\delta,\varepsilon)} - \frac{\Delta_\mu}{2}\right) \le r^2\left(1-\rho_\ell(\sigma,\delta,\varepsilon) - \sqrt{2\rho_\ell(\sigma,\delta,\varepsilon)} - \frac{\Delta_\mu}{2}\right)$. According to Lemma 3.2, every sample $\boldsymbol{x} \in C_\ell^0 \cap S_\varepsilon$ can be correctly classified by G_f. Therefore, every sample $\boldsymbol{x} \in (C_1^0 \cap \cdots \cap C_K^0) \cap S_\varepsilon$ can be correctly classified by G_f. According to Lemma 3.1, the error rate $\mathrm{Err}(G_f) \le 1-\sigma+R_\varepsilon$. □

Theorem 2. *If encoder f is L-Lipschitz continuous, then*

$$R_\varepsilon^2 \le \eta(\varepsilon)^2 \cdot \mathop{\mathbb{E}}_{\boldsymbol{x}} \mathop{\mathbb{E}}_{\boldsymbol{x}_1,\boldsymbol{x}_2\in A(\boldsymbol{x})} \|f(\boldsymbol{x}_1) - f(\boldsymbol{x}_2)\|^2,$$

where $\eta(\varepsilon) = \inf_{h\in\left(0,\frac{\varepsilon}{2\sqrt{n}LM}\right)} \frac{4\max\{1,m^2h^{2n}\}}{h^{2n}(\varepsilon-2\sqrt{n}LMh)} = \mathcal{O}\left(\frac{1}{\varepsilon}\right)$.

Proof. The parameter space $[0,1]^n$ of θ can be separated to cubes $\Theta_1, \dots, \Theta_{m'}$ where $m' = 1/h^n$ and each cube's edge length is $h \in (0, \frac{\varepsilon}{2\sqrt{n}LM})$. Then for any given $\boldsymbol{x}$, we have

$$\begin{aligned}
\mathop{\mathbb{E}}_{\boldsymbol{x}_1,\boldsymbol{x}_2 \in A(\boldsymbol{x})} \|f(\boldsymbol{x}_1) - f(\boldsymbol{x}_2)\| &= \underbrace{\frac{1}{4m^2}\sum_{\gamma=1}^{m}\sum_{\beta=1}^{m} \|f(A_\gamma(\boldsymbol{x})) - f(A_\beta(\boldsymbol{x}))\|}_{\Lambda_1} \\
&\quad + \underbrace{\frac{1}{2mm'}\sum_{\gamma=1}^{m}\sum_{j=1}^{m'} \int_{\Theta_j} \frac{1}{h^n} \|f(A_\gamma(\boldsymbol{x})) - f(A_\theta(\boldsymbol{x}))\| \, d\theta}_{\Lambda_2} \\
&\quad + \underbrace{\frac{1}{4m'^2}\sum_{i=1}^{m'}\sum_{j=1}^{m'} \int_{\Theta_i}\int_{\Theta_j} \frac{1}{h^{2n}} \|f(A_{\theta_1}(\boldsymbol{x})) - f(A_{\theta_2}(\boldsymbol{x}))\| \, d\theta_2 \, d\theta_1}_{\Lambda_3}.
\end{aligned}$$

By Cauchy-Schwarz inequality,

$$\forall\theta, \|f(A_\gamma(\boldsymbol{x})) - f(A_{\theta'}(\boldsymbol{x}))\| \le \|f(A_\gamma(\boldsymbol{x})) - f(A_\theta(\boldsymbol{x}))\| + \|f(A_\theta(\boldsymbol{x})) - f(A_{\theta'}(\boldsymbol{x}))\|.$$

Then for any given θ,

$$\begin{aligned}
\sup_{\theta'} \|f(A_\gamma(\boldsymbol{x})) - f(A_{\theta'}(\boldsymbol{x}))\| &\le \|f(A_\gamma(\boldsymbol{x})) - f(A_\theta(\boldsymbol{x}))\| + \sup_{\theta} \|f(A_\theta(\boldsymbol{x})) - f(A_{\theta'}(\boldsymbol{x}))\| \\
&\le \|f(A_\gamma(\boldsymbol{x})) - f(A_\theta(\boldsymbol{x}))\| + \sup_{\theta_1,\theta_2} \|f(A_{\theta_1}(\boldsymbol{x})) - f(A_{\theta_2}(\boldsymbol{x}))\|.
\end{aligned}$$

Therefore, for any $\gamma \in [m], j \in [m']$, we have

$$\begin{aligned}
&\sup_{\theta' \in \Theta_j} \|f(A_\gamma(\boldsymbol{x})) - f(A_{\theta'}(\boldsymbol{x}))\| \\
&= \int_{\Theta_j} \frac{1}{h^n} \sup_{\theta' \in \Theta_j} \|f(A_\gamma(\boldsymbol{x})) - f(A_{\theta'}(\boldsymbol{x}))\| \, d\theta \\
&\le \int_{\Theta_j} \frac{1}{h^n} \|f(A_\gamma(\boldsymbol{x})) - f(A_\theta(\boldsymbol{x}))\| \, d\theta + \sup_{\theta_1,\theta_2 \in \Theta_j} \|f(A_{\theta_1}(\boldsymbol{x})) - f(A_{\theta_2}(\boldsymbol{x}))\| \\
&\le \int_{\Theta_j} \frac{1}{h^n} \|f(A_\gamma(\boldsymbol{x})) - f(A_\theta(\boldsymbol{x}))\| \, d\theta + L \sup_{\theta_1,\theta_2 \in \Theta_j} \|A_{\theta_1}(\boldsymbol{x}) - A_{\theta_2}(\boldsymbol{x})\| \\
&\le \int_{\Theta_j} \frac{1}{h^n} \|f(A_\gamma(\boldsymbol{x})) - f(A_\theta(\boldsymbol{x}))\| \, d\theta + LM \sup_{\theta_1,\theta_2 \in \Theta_j} \|\theta_1 - \theta_2\| \\
&= \int_{\Theta_j} \frac{1}{h^n} \|f(A_\gamma(\boldsymbol{x})) - f(A_\theta(\boldsymbol{x}))\| \, d\theta + LM\sqrt{n}h \\
&= \int_{\Theta_j} \frac{1}{h^n} \|f(A_\gamma(\boldsymbol{x})) - f(A_\theta(\boldsymbol{x}))\| \, d\theta + \sqrt{n}LMh.
\end{aligned}$$

Similarly, we can obtain

$$\begin{aligned}
&\sup_{\theta\in\Theta_i,\theta'\in\Theta_j} \|f(A_\theta(\boldsymbol{x}))-f(A_{\theta'}(\boldsymbol{x}))\| \\
&=\int_{\Theta_i}\int_{\Theta_j}\frac{1}{h^{2n}}\sup_{\theta\in\Theta_i,\theta'\in\Theta_j}\|f(A_\theta(\boldsymbol{x}))-f(A_{\theta'}(\boldsymbol{x}))\|\,d\theta_2\,d\theta_1 \\
&\le\int_{\Theta_i}\int_{\Theta_j}\frac{1}{h^{2n}}\|f(A_{\theta_1}(\boldsymbol{x}))-f(A_{\theta_2}(\boldsymbol{x}))\|\,d\theta_2\,d\theta_1 \\
&+\int_{\Theta_i}\int_{\Theta_j}\frac{1}{h^{2n}}\sup_{\theta\in\Theta_i}\|f(A_\theta(\boldsymbol{x}))-f(A_{\theta_1}(\boldsymbol{x}))\|\,d\theta_2\,d\theta_1 \\
&+\int_{\Theta_i}\int_{\Theta_j}\frac{1}{h^{2n}}\sup_{\theta'\in\Theta_j}\|f(A_{\theta_2}(\boldsymbol{x}))-f(A_{\theta'}(\boldsymbol{x}))\|\,d\theta_2\,d\theta_1 \\
&\le\int_{\Theta_i}\int_{\Theta_j}\frac{1}{h^{2n}}\|f(A_{\theta_1}(\boldsymbol{x}))-f(A_{\theta_2}(\boldsymbol{x}))\|\,d\theta_2\,d\theta_1 \\
&+\sup_{\theta,\theta'\in\Theta_i}\|f(A_\theta(\boldsymbol{x}))-f(A_{\theta'}(\boldsymbol{x}))\|+\sup_{\theta,\theta'\in\Theta_j}\|f(A_\theta(\boldsymbol{x}))-f(A_{\theta'}(\boldsymbol{x}))\| \\
&\le\int_{\Theta_i}\int_{\Theta_j}\frac{1}{h^{2n}}\|f(A_{\theta_1}(\boldsymbol{x}))-f(A_{\theta_2}(\boldsymbol{x}))\|\,d\theta_2\,d\theta_1+2\sqrt{n}LMh.
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\sup_{\boldsymbol{x}_1,\boldsymbol{x}_2\in A(\boldsymbol{x})}\|f(\boldsymbol{x}_1)-f(\boldsymbol{x}_2)\| \\
&=\max\left\{\begin{matrix}\sup_{\gamma,\beta\in[m]}\|f(A_\gamma(\boldsymbol{x}))-f(A_\beta(\boldsymbol{x}))\| \\ \sup_{\gamma\in[m],j\in[m']}\sup_{\theta'\in\Theta_j}\|f(A_\gamma(\boldsymbol{x}))-f(A_{\theta'}(\boldsymbol{x}))\| \\ \sup_{i,j\in[m']}\sup_{\theta\in\Theta_i,\theta'\in\Theta_j}\|f(A_\theta(\boldsymbol{x}))-f(A_{\theta'}(\boldsymbol{x}))\|\end{matrix}\right\} \\
&\le\max\left\{\begin{matrix}\sup_{\gamma,\beta\in[m]}\|f(A_\gamma(\boldsymbol{x}))-f(A_\beta(\boldsymbol{x}))\| \\ \sup_{\gamma\in[m],j\in[m']}\int_{\Theta_j}\frac{1}{h^n}\|f(A_\gamma(\boldsymbol{x}))-f(A_\theta(\boldsymbol{x}))\|\,d\theta+\sqrt{n}LMh \\ \sup_{i,j\in[m']}\int_{\Theta_i}\int_{\Theta_j}\frac{1}{h^{2n}}\|f(A_{\theta_1}(\boldsymbol{x}))-f(A_{\theta_2}(\boldsymbol{x}))\|\,d\theta_2\,d\theta_1+2\sqrt{n}LMh\end{matrix}\right\} \\
&\le\max\left\{\begin{matrix}\sum_{\gamma=1}^m\sum_{\beta=1}^m\|f(A_\gamma(\boldsymbol{x}))-f(A_\beta(\boldsymbol{x}))\| \\ \sum_{\gamma=1}^m\sum_{j=1}^{m'}\int_{\Theta_j}\frac{1}{h^n}\|f(A_\gamma(\boldsymbol{x}))-f(A_\theta(\boldsymbol{x}))\|\,d\theta+\sqrt{n}LMh \\ \sum_{i=1}^{m'}\sum_{j=1}^{m'}\int_{\Theta_i}\int_{\Theta_j}\frac{1}{h^{2n}}\|f(A_{\theta_1}(\boldsymbol{x}))-f(A_{\theta_2}(\boldsymbol{x}))\|\,d\theta_2\,d\theta_1+2\sqrt{n}LMh\end{matrix}\right\} \\
&\le\max\left\{\begin{matrix}4m^2\Lambda_1 \\ 2mm'\Lambda_2+\sqrt{n}LMh \\ 4m'^2\Lambda_3+2\sqrt{n}LMh\end{matrix}\right\} \\
&\le\max\left\{\begin{matrix}4m^2\Lambda_1 \\ 2mm'\Lambda_2 \\ 4m'^2\Lambda_3\end{matrix}\right\}+2\sqrt{n}LMh \\
&\le\max\{4m^2,2mm',4m'^2\}(\Lambda_1+\Lambda_2+\Lambda_3)+2\sqrt{n}LMh \\
&=\max\{4m^2,2mm',4m'^2\}\mathop{\mathbb{E}}_{\boldsymbol{x}_1,\boldsymbol{x}_2\in A(\boldsymbol{x})}\|f(\boldsymbol{x}_1)-f(\boldsymbol{x}_2)\|+2\sqrt{n}LMh.
\end{aligned}$$

Thus, the following set S is a subset of S_ε:

$$S = \left\{ \boldsymbol{x} \colon \mathop{\mathbb{E}}_{\boldsymbol{x}_1, \boldsymbol{x}_2 \in A(\boldsymbol{x})} \| f(\boldsymbol{x}_1) - f(\boldsymbol{x}_2) \| \leq \frac{\varepsilon - 2\sqrt{n} L M h}{\max\{4m^2, 2mm', 4m'^2\}} \right\} \subseteq S_\varepsilon.$$

Then by Markov's inequality, we have

$$\begin{aligned} R_\varepsilon = \mathbb{P}\left[\overline{S_\varepsilon}\right] &\leq \mathbb{P}\left[\overline{S}\right] \\ &\leq \frac{\mathbb{E}_{\boldsymbol{x}} \mathbb{E}_{\boldsymbol{x}_1, \boldsymbol{x}_2 \in A(\boldsymbol{x})} \| f(\boldsymbol{x}_1) - f(\boldsymbol{x}_2) \|}{\frac{\varepsilon - 2\sqrt{n} L M h}{\max\{4m^2, 2mm', 4m'^2\}}} \\ &= \frac{\max\{4, 2mh^n, 4m^2 h^{2n}\}}{h^{2n}(\varepsilon - 2\sqrt{n} L M h)} \mathop{\mathbb{E}}_{\boldsymbol{x}} \mathop{\mathbb{E}}_{\boldsymbol{x}_1, \boldsymbol{x}_2 \in A(\boldsymbol{x})} \| f(\boldsymbol{x}_1) - f(\boldsymbol{x}_2) \| \\ &= \frac{4 \max\{1, m^2 h^{2n}\}}{h^{2n}(\varepsilon - 2\sqrt{n} L M h)} \mathop{\mathbb{E}}_{\boldsymbol{x}} \mathop{\mathbb{E}}_{\boldsymbol{x}_1, \boldsymbol{x}_2 \in A(\boldsymbol{x})} \| f(\boldsymbol{x}_1) - f(\boldsymbol{x}_2) \|. \end{aligned}$$

The above inequality holds for all $h \in (0, \frac{\varepsilon}{2\sqrt{n}LM})$, thus

$$R_\varepsilon \leq \inf_{0 < h < \frac{\varepsilon}{2\sqrt{n}LM}} \frac{4 \max\{1, m^2 h^{2n}\}}{h^{2n}(\varepsilon - 2\sqrt{n} L M h)} \mathop{\mathbb{E}}_{\boldsymbol{x}} \mathop{\mathbb{E}}_{\boldsymbol{x}_1, \boldsymbol{x}_2 \in A(\boldsymbol{x})} \| f(\boldsymbol{x}_1) - f(\boldsymbol{x}_2) \| = \eta(\varepsilon) \cdot \mathop{\mathbb{E}}_{\boldsymbol{x}} \mathop{\mathbb{E}}_{\boldsymbol{x}_1, \boldsymbol{x}_2 \in A(\boldsymbol{x})} \| f(\boldsymbol{x}_1) - f(\boldsymbol{x}_2) \|.$$

Therefore, we have

$$R_\varepsilon^2 \leq \eta(\varepsilon)^2 \cdot (\mathop{\mathbb{E}}_{\boldsymbol{x}} \mathop{\mathbb{E}}_{\boldsymbol{x}_1, \boldsymbol{x}_2 \in A(\boldsymbol{x})} \| f(\boldsymbol{x}_1) - f(\boldsymbol{x}_2) \|)^2 \leq \eta(\varepsilon)^2 \cdot \mathop{\mathbb{E}}_{\boldsymbol{x}} \mathop{\mathbb{E}}_{\boldsymbol{x}_1, \boldsymbol{x}_2 \in A(\boldsymbol{x})} \| f(\boldsymbol{x}_1) - f(\boldsymbol{x}_2) \|^2.$$

This finishes the proof. □

B Proofs for Section 4

Before providing our proofs, we give the following useful lemma, which upper bounds the first and second order moment of intra-difference within each class C_k via ε and R_ε.

Lemma B.1. *Suppose that $\|f(\boldsymbol{x})\| = r$ for every $\boldsymbol{x}$. For each $k \in [K]$,*

$$\mathop{\mathbb{E}}_{\boldsymbol{x} \in C_k} \mathop{\mathbb{E}}_{\boldsymbol{x}_1 \in A(\boldsymbol{x})} \| f(\boldsymbol{x}_1) - \mu_k \| \leq 4r \left(1 - \sigma \left(1 - \frac{\varepsilon}{2r} - \frac{L\delta}{4r} \right) + \frac{R_\varepsilon}{p_k} \right),$$

and

$$\mathop{\mathbb{E}}_{\boldsymbol{x} \in C_k} \mathop{\mathbb{E}}_{\boldsymbol{x}_1 \in A(\boldsymbol{x})} \| f(\boldsymbol{x}_1) - \mu_k \|^2 \leq (2\varepsilon + L\delta)^2 + 4r \left(1 - \sigma + \frac{R_\varepsilon}{p_k} \right) [r + 2\varepsilon + L\delta] + 4r^2 \left(1 - \sigma + \frac{R_\varepsilon}{p_k} \right)^2.$$

Proof. For each $k \in [K]$,

$$\begin{aligned} &\mathop{\mathbb{E}}_{\boldsymbol{x} \in C_k} \mathop{\mathbb{E}}_{\boldsymbol{x}_1 \in A(\boldsymbol{x})} \| f(\boldsymbol{x}_1) - \mu_k \| \\ &= \frac{1}{p_k} \mathop{\mathbb{E}}_{\boldsymbol{x}} \mathop{\mathbb{E}}_{\boldsymbol{x}_1 \in A(\boldsymbol{x})} [\mathbb{I}(\boldsymbol{x} \in C_k) \| f(\boldsymbol{x}_1) - \mu_k \|] \\ &= \frac{1}{p_k} \mathop{\mathbb{E}}_{\boldsymbol{x}} \mathop{\mathbb{E}}_{\boldsymbol{x}_1 \in A(\boldsymbol{x})} \left[\mathbb{I}(\boldsymbol{x} \in C_k^0 \cap S_\varepsilon) \| f(\boldsymbol{x}_1) - \mu_k \|\right] + \frac{1}{p_k} \mathop{\mathbb{E}}_{\boldsymbol{x}} \mathop{\mathbb{E}}_{\boldsymbol{x}_1 \in A(\boldsymbol{x})} \left[\mathbb{I}(\boldsymbol{x} \in C_k \setminus C_k^0 \cap S_\varepsilon) \| f(\boldsymbol{x}_1) - \mu_k \|\right] \\ &\leq \frac{1}{p_k} \mathop{\mathbb{E}}_{\boldsymbol{x}} \mathop{\mathbb{E}}_{\boldsymbol{x}_1 \in A(\boldsymbol{x})} \left[\mathbb{I}(\boldsymbol{x} \in C_k^0 \cap S_\varepsilon) \| f(\boldsymbol{x}_1) - \mu_k \|\right] + \frac{2r \mathbb{P}[C_k \setminus C_k^0 \cap S_\varepsilon]}{p_k} \end{aligned}$$

$$\leq \frac{1}{p_k} \mathop{\mathbb{E}}_{\boldsymbol{x}} \mathop{\mathbb{E}}_{\boldsymbol{x}_1 \in A(\boldsymbol{x})} \left[\mathbb{I}(\boldsymbol{x} \in C_k^0 \cap S_\varepsilon) \| f(\boldsymbol{x}_1) - \mu_k \| \right] + 2r \left(1 - \sigma + \frac{R_\varepsilon}{p_k} \right) \quad \text{(using (8))}$$

$$\leq \frac{\mathbb{P}[C_k^0 \cap S_\varepsilon]}{p_k} \mathop{\mathbb{E}}_{\boldsymbol{x} \in C_k^0 \cap S_\varepsilon} \mathop{\mathbb{E}}_{\boldsymbol{x}_1 \in A(\boldsymbol{x})} \| f(\boldsymbol{x}_1) - \mu_k \| + 2r \left(1 - \sigma + \frac{R_\varepsilon}{p_k} \right)$$

$$\leq \sigma \mathop{\mathbb{E}}_{\boldsymbol{x} \in C_k^0 \cap S_\varepsilon} \mathop{\mathbb{E}}_{\boldsymbol{x}_1 \in A(\boldsymbol{x})} \| f(\boldsymbol{x}_1) - \mu_k \| + 2r \left(1 - \sigma + \frac{R_\varepsilon}{p_k} \right) \tag{13}$$

where

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\boldsymbol{x} \in C_k^0 \cap S_\varepsilon} \mathop{\mathbb{E}}_{\boldsymbol{x}_1 \in A(\boldsymbol{x})} \| f(\boldsymbol{x}_1) - \mu_k \| \\
&= \mathop{\mathbb{E}}_{\boldsymbol{x} \in C_k^0 \cap S_\varepsilon} \mathop{\mathbb{E}}_{\boldsymbol{x}_1 \in A(\boldsymbol{x})} \| f(\boldsymbol{x}_1) - \mathop{\mathbb{E}}_{\boldsymbol{x}' \in C_k} \mathop{\mathbb{E}}_{\boldsymbol{x}_2 \in A(\boldsymbol{x}')} f(\boldsymbol{x}_2) \| \\
&= \mathop{\mathbb{E}}_{\boldsymbol{x} \in C_k^0 \cap S_\varepsilon} \mathop{\mathbb{E}}_{\boldsymbol{x}_1 \in A(\boldsymbol{x})} \left\| f(\boldsymbol{x}_1) - \frac{\mathbb{P}[C_k^0 \cap S_\varepsilon]}{p_k} \mathop{\mathbb{E}}_{\boldsymbol{x}' \in C_k^0 \cap S_\varepsilon} \mathop{\mathbb{E}}_{\boldsymbol{x}_2 \in A(\boldsymbol{x}')} f(\boldsymbol{x}_2) - \frac{\mathbb{P}[C_k \setminus C_k^0 \cap S_\varepsilon]}{p_k} \mathop{\mathbb{E}}_{\boldsymbol{x}' \in C_k \setminus C_k^0 \cap S_\varepsilon} \mathop{\mathbb{E}}_{\boldsymbol{x}_2 \in A(\boldsymbol{x}')} f(\boldsymbol{x}_2) \right\| \\
&= \mathop{\mathbb{E}}_{\boldsymbol{x} \in C_k^0 \cap S_\varepsilon} \mathop{\mathbb{E}}_{\boldsymbol{x}_1 \in A(\boldsymbol{x})} \left\| \frac{\mathbb{P}[C_k^0 \cap S_\varepsilon]}{p_k} \left(f(\boldsymbol{x}_1) - \mathop{\mathbb{E}}_{\boldsymbol{x}' \in C_k^0 \cap S_\varepsilon} \mathop{\mathbb{E}}_{\boldsymbol{x}_2 \in A(\boldsymbol{x}')} f(\boldsymbol{x}_2) \right) \right. \\
&\quad \left. + \frac{\mathbb{P}[C_k \setminus C_k^0 \cap S_\varepsilon]}{p_k} \left(f(\boldsymbol{x}_1) - \mathop{\mathbb{E}}_{\boldsymbol{x}' \in C_k \setminus C_k^0 \cap S_\varepsilon} \mathop{\mathbb{E}}_{\boldsymbol{x}_2 \in A(\boldsymbol{x}')} f(\boldsymbol{x}_2) \right) \right\| \\
&\leq \frac{\mathbb{P}[C_k^0 \cap S_\varepsilon]}{p_k} \mathop{\mathbb{E}}_{\boldsymbol{x} \in C_k^0 \cap S_\varepsilon} \mathop{\mathbb{E}}_{\boldsymbol{x}_1 \in A(\boldsymbol{x})} \left\| f(\boldsymbol{x}_1) - \mathop{\mathbb{E}}_{\boldsymbol{x}' \in C_k^0 \cap S_\varepsilon} \mathop{\mathbb{E}}_{\boldsymbol{x}_2 \in A(\boldsymbol{x}')} f(\boldsymbol{x}_2) \right\| + \frac{\mathbb{P}[C_k \setminus C_k^0 \cap S_\varepsilon]}{p_k} \cdot 2r \\
&\leq \sigma \sup_{\boldsymbol{x}, \boldsymbol{x}' \in C_k^0 \cap S_\varepsilon} \sup_{\substack{\boldsymbol{x}_1 \in A(\boldsymbol{x}) \\ \boldsymbol{x}_2 \in A(\boldsymbol{x}')}} \| f(\boldsymbol{x}_1) - f(\boldsymbol{x}_2) \| + 2r \left(1 - \sigma + \frac{R_\varepsilon}{p_k} \right).
\end{aligned} \tag{14}$$

For any $\boldsymbol{x}, \boldsymbol{x}' \in C_1^0 \cap S_\varepsilon$, we have $d_A(\boldsymbol{x}, \boldsymbol{x}') \leq \delta$. Let $(\boldsymbol{x}_1^*, \boldsymbol{x}_2^*) = \arg\min_{\boldsymbol{x}_1 \in A(\boldsymbol{x}), \boldsymbol{x}_2 \in A(\boldsymbol{x}')} \|\boldsymbol{x}_1 - \boldsymbol{x}_2\|$. We have $\|\boldsymbol{x}_1^* - \boldsymbol{x}_2^*\| \leq \delta$. Since f is L-Lipschitz continuous, we have $\|f(\boldsymbol{x}_1^*) - f(\boldsymbol{x}_2^*)\| \leq L \cdot \|\boldsymbol{x}_1^* - \boldsymbol{x}_2^*\| \leq L\delta$. Since $\boldsymbol{x} \in S_\varepsilon$, for any $\boldsymbol{x}_1 \in A(\boldsymbol{x})$, $\|f(\boldsymbol{x}_1) - f(\boldsymbol{x}_1^*)\| \leq \varepsilon$. Similarly, since $\boldsymbol{x}' \in S_\varepsilon$, for any $\boldsymbol{x}_2 \in A(\boldsymbol{x}')$, we have $\|f(\boldsymbol{x}_2) - f(\boldsymbol{x}_2^*)\| \leq \varepsilon$. Therefore, for any $\boldsymbol{x}, \boldsymbol{x}' \in C_1^0 \cap S_\varepsilon$ and $\boldsymbol{x}_1 \in A(\boldsymbol{x}), \boldsymbol{x}_2 \in A(\boldsymbol{x}')$,

$$\|f(\boldsymbol{x}_1) - f(\boldsymbol{x}_2)\| \leq \|f(\boldsymbol{x}_1) - f(\boldsymbol{x}_1^*)\| + \|f(\boldsymbol{x}_1^*) - f(\boldsymbol{x}_2^*)\| + \|f(\boldsymbol{x}_2^*) - f(\boldsymbol{x}_2)\| \leq 2\varepsilon + L\delta.$$

Plugging into (13) and (14), we obtain

$$\begin{aligned}
\mathop{\mathbb{E}}_{\boldsymbol{x} \in C_k} \mathop{\mathbb{E}}_{\boldsymbol{x}_1 \in A(\boldsymbol{x})} \| f(\boldsymbol{x}_1) - \mu_k \| &\leq \sigma(2\varepsilon + L\delta) + 4r \left(1 - \sigma + \frac{R_\varepsilon}{p_k} \right) \\
&= 4r \left(1 - \sigma \left(1 - \frac{\varepsilon}{2r} - \frac{L\delta}{4r} \right) + \frac{R_\varepsilon}{p_k} \right).
\end{aligned}$$

Similar to (13) and (14), we have

$$\mathop{\mathbb{E}}_{\boldsymbol{x} \in C_k} \mathop{\mathbb{E}}_{\boldsymbol{x}_1 \in A(\boldsymbol{x})} \| f(\boldsymbol{x}_1) - \mu_k \|^2 \leq \sigma \mathop{\mathbb{E}}_{\boldsymbol{x} \in C_k^0 \cap S_\varepsilon} \mathop{\mathbb{E}}_{\boldsymbol{x}_1 \in A(\boldsymbol{x})} \| f(\boldsymbol{x}_1) - \mu_k \|^2 + 4r^2 \left(1 - \sigma + \frac{R_\varepsilon}{p_k} \right),$$

and

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\boldsymbol{x} \in C_k^0 \cap S_\varepsilon} \mathop{\mathbb{E}}_{\boldsymbol{x}_1 \in A(\boldsymbol{x})} \| f(\boldsymbol{x}_1) - \mu_k \|^2 \\
&= \mathop{\mathbb{E}}_{\boldsymbol{x} \in C_k^0 \cap S_\varepsilon} \mathop{\mathbb{E}}_{\boldsymbol{x}_1 \in A(\boldsymbol{x})} \left\| \frac{\mathbb{P}[C_k^0 \cap S_\varepsilon]}{p_k} \left(f(\boldsymbol{x}_1) - \mathop{\mathbb{E}}_{\boldsymbol{x}' \in C_k^0 \cap S_\varepsilon} \mathop{\mathbb{E}}_{\boldsymbol{x}_2 \in A(\boldsymbol{x}')} f(\boldsymbol{x}_2) \right) \right.
\end{aligned}$$

$$
\begin{aligned}
&\quad + \frac{\mathbb{P}[C_k \setminus C_k^0 \cap S_\varepsilon]}{p_k} \left(f(\boldsymbol{x}_1) - \mathop{\mathbb{E}}_{\boldsymbol{x}' \in C_k \setminus C_k^0 \cap S_\varepsilon} \mathop{\mathbb{E}}_{\boldsymbol{x}_2 \in A(\boldsymbol{x}')} f(\boldsymbol{x}_2) \right) \Bigg\|^2 \\
&\le \mathop{\mathbb{E}}_{\boldsymbol{x} \in C_k^0 \cap S_\varepsilon} \mathop{\mathbb{E}}_{\boldsymbol{x}_1 \in A(\boldsymbol{x})} \left[\sigma \left\| f(\boldsymbol{x}_1) - \mathop{\mathbb{E}}_{\boldsymbol{x}' \in C_k^0 \cap S_\varepsilon} \mathop{\mathbb{E}}_{\boldsymbol{x}_2 \in A(\boldsymbol{x}')} f(\boldsymbol{x}_2) \right\| + 2r \left(1 - \sigma + \frac{R_\varepsilon}{p_k} \right) \right]^2 \\
&\le \left[\sigma(2\varepsilon + L\delta) + 2r \left(1 - \sigma + \frac{R_\varepsilon}{p_k} \right) \right]^2.
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{\boldsymbol{x} \in C_k} \mathop{\mathbb{E}}_{\boldsymbol{x}_1 \in A(\boldsymbol{x})} \| f(\boldsymbol{x}_1) - \mu_k \|^2 \\
&\le \sigma^3 (2\varepsilon + L\delta)^2 + 4r \left(1 - \sigma + \frac{R_\varepsilon}{p_k} \right) [r + \sigma^2 (2\varepsilon + L\delta)] + 4\sigma r^2 \left(1 - \sigma + \frac{R_\varepsilon}{p_k} \right)^2 \\
&\le (2\varepsilon + L\delta)^2 + 4r \left(1 - \sigma + \frac{R_\varepsilon}{p_k} \right) [r + 2\varepsilon + L\delta] + 4r^2 \left(1 - \sigma + \frac{R_\varepsilon}{p_k} \right)^2.
\end{aligned}
$$

This finishes the proof. $\square$

Now we are ready to give our proofs of theorems.

B.1 InfoNCE Loss

Theorem 3. *Assume that encoder f with norm 1 is L-Lipschitz continuous. If the augmented data is (σ, δ)-augmented, then for any $\varepsilon > 0$ and $k \neq \ell$,*

$$
\mu_k^\top \mu_\ell \le \log \left(\exp \left\{ \frac{\mathcal{L}_2(f) + \tau(\varepsilon, \sigma, \delta)}{p_k p_\ell} \right\} - \exp(1 - \varepsilon) \right) \lesssim \mathcal{L}_2(f),
$$

where $\tau(\varepsilon, \sigma, \delta)$ is related to the mean of intra-class variance in the embedding space.

Proof. Given $\boldsymbol{x} \in S_\varepsilon$, for any $\boldsymbol{x}_1, \boldsymbol{x}_2 \in A(\boldsymbol{x})$, we have

$$
\begin{aligned}
\log \left(e^{f(\boldsymbol{x}_1)^\top f(\boldsymbol{x}_2)} + e^{f(\boldsymbol{x}_1)^\top f(\boldsymbol{x}^-)} \right) &= \log \left(e^{f(\boldsymbol{x}_1)^\top f(\boldsymbol{x}_1)} e^{f(\boldsymbol{x}_1)^\top (f(\boldsymbol{x}_2) - f(\boldsymbol{x}_1))} + e^{f(\boldsymbol{x}_1)^\top f(\boldsymbol{x}^-)} \right) \\
&\ge \log \left(e^{\|f(\boldsymbol{x}_1)\|^2} e^{-\|f(\boldsymbol{x}_1)\| \cdot \varepsilon} + e^{f(\boldsymbol{x}_1)^\top f(\boldsymbol{x}^-)} \right) \\
&= \log \left(e^{1 - \varepsilon} + e^{f(\boldsymbol{x}_1)^\top f(\boldsymbol{x}^-)} \right).
\end{aligned}
$$

Therefore, we have

$$
\begin{aligned}
&\mathcal{L}_2(f) = \mathop{\mathbb{E}}_{\boldsymbol{x}, \boldsymbol{x}'} \mathop{\mathbb{E}}_{\substack{\boldsymbol{x}_1, \boldsymbol{x}_2 \in A(\boldsymbol{x}) \\ \boldsymbol{x}^- \in A(\boldsymbol{x}')}} \left[\log \left(e^{f(\boldsymbol{x}_1)^\top f(\boldsymbol{x}_2)} + e^{f(\boldsymbol{x}_1)^\top f(\boldsymbol{x}^-)} \right) \right] \\
&= \mathop{\mathbb{E}}_{\boldsymbol{x}, \boldsymbol{x}'} \mathop{\mathbb{E}}_{\substack{\boldsymbol{x}_1, \boldsymbol{x}_2 \in A(\boldsymbol{x}) \\ \boldsymbol{x}^- \in A(\boldsymbol{x}')}} \left[(\mathbb{I}(\boldsymbol{x} \in S_\varepsilon) + \mathbb{I}(\boldsymbol{x} \in \bar{S}_\varepsilon)) \log \left(e^{f(\boldsymbol{x}_1)^\top f(\boldsymbol{x}_2)} + e^{f(\boldsymbol{x}_1)^\top f(\boldsymbol{x}^-)} \right) \right] \\
&\ge \sum_{k=1}^K \sum_{\ell=1}^K \mathop{\mathbb{E}}_{\boldsymbol{x}, \boldsymbol{x}'} \left[\mathbb{I}(\boldsymbol{x} \in S_\varepsilon \cap C_k) \mathbb{I}(\boldsymbol{x}' \in C_\ell) \mathop{\mathbb{E}}_{\substack{\boldsymbol{x}_1 \in A(\boldsymbol{x}) \\ \boldsymbol{x}^- \in A(\boldsymbol{x}')}} \log \left(e^{1 - \varepsilon} + e^{f(\boldsymbol{x}_1)^\top f(\boldsymbol{x}^-)} \right) \right] + \mathop{\mathbb{E}}_{\boldsymbol{x}} \left[\mathbb{I}(\boldsymbol{x} \in \bar{S}_\varepsilon) \log \left(e^{-1} + e^{-1} \right) \right]
\end{aligned}
$$

$$
\begin{aligned}
&= \left(\sum_{k=1}^{K} \sum_{\ell=1}^{K} \mathop{\mathbb{E}}_{\boldsymbol{x},\boldsymbol{x}'} \left[\mathbb{I}(\boldsymbol{x} \in C_k) \mathbb{I}(\boldsymbol{x}' \in C_\ell) \log \left(e^{1-\varepsilon} + e^{\mu_k^\top \mu_\ell} \right) \right] + \Delta_1 \right) - (1 - \log 2) R_\varepsilon \\
&= \sum_{k=1}^{K} \sum_{\ell=1}^{K} \left[p_k p_l \log \left(e^{1-\varepsilon} + e^{\mu_k^\top \mu_\ell} \right) \right] - (1 - \log 2) R_\varepsilon + \Delta_1 \\
&\geq p_k p_\ell \log \left(e^{1-\varepsilon} + e^{\mu_k^\top \mu_\ell} \right) + \sum_{k=1}^{K} p_k^2 \log \left(e^{1-\varepsilon} + e^{\|\mu_k\|^2} \right) - (1 - \log 2) R_\varepsilon + \Delta_1 \\
&\geq p_k p_\ell \log \left(e^{1-\varepsilon} + e^{\mu_k^\top \mu_\ell} \right) + (1-\varepsilon) \sum_{k=1}^{K} p_k^2 - (1 - \log 2) R_\varepsilon + \Delta_1 \\
&\geq p_k p_\ell \log \left(e^{1-\varepsilon} + e^{\mu_k^\top \mu_\ell} \right) + \frac{1-\varepsilon}{K} - (1 - \log 2) R_\varepsilon + \Delta_1, \qquad (15)
\end{aligned}
$$

where Δ_1 is defined as

$$
\begin{aligned}
\Delta_1 :=& \sum_{k=1}^{K} \sum_{\ell=1}^{K} \mathop{\mathbb{E}}_{\boldsymbol{x},\boldsymbol{x}'} \left[\mathbb{I}(\boldsymbol{x} \in S_\varepsilon \cap C_k) \mathbb{I}(\boldsymbol{x}' \in C_\ell) \mathop{\mathbb{E}}_{\substack{\boldsymbol{x}_1 \in A(\boldsymbol{x}) \\ \boldsymbol{x}^- \in A(\boldsymbol{x}')}} \log \left(e^{1-\varepsilon} + e^{f(\boldsymbol{x}_1)^\top f(\boldsymbol{x}^-)} \right) \right] \\
&- \sum_{k=1}^{K} \sum_{\ell=1}^{K} \mathop{\mathbb{E}}_{\boldsymbol{x},\boldsymbol{x}'} \left[\mathbb{I}(\boldsymbol{x} \in C_k) \mathbb{I}(\boldsymbol{x}' \in C_\ell) \log \left(e^{1-\varepsilon} + e^{\mu_k^\top \mu_\ell} \right) \right] \\
=& - \sum_{k=1}^{K} \sum_{\ell=1}^{K} \mathop{\mathbb{E}}_{\boldsymbol{x},\boldsymbol{x}'} \left[\left(\mathbb{I}(\boldsymbol{x} \in C_k) - \mathbb{I}(\boldsymbol{x} \in S_\varepsilon \cap C_k) \right) \mathbb{I}(\boldsymbol{x}' \in C_\ell) \mathop{\mathbb{E}}_{\substack{\boldsymbol{x}_1 \in A(\boldsymbol{x}) \\ \boldsymbol{x}^- \in A(\boldsymbol{x}')}} \left[\log \left(e^{1-\varepsilon} + e^{f(\boldsymbol{x}_1)^\top f(\boldsymbol{x}^-)} \right) \right] \right] \\
&+ \sum_{k=1}^{K} \sum_{\ell=1}^{K} \mathop{\mathbb{E}}_{\boldsymbol{x},\boldsymbol{x}'} \left[\mathbb{I}(\boldsymbol{x} \in C_k) \mathbb{I}(\boldsymbol{x}' \in C_\ell) \mathop{\mathbb{E}}_{\substack{\boldsymbol{x}_1 \in A(\boldsymbol{x}) \\ \boldsymbol{x}^- \in A(\boldsymbol{x}')}} \left[\log \left(e^{1-\varepsilon} + e^{f(\boldsymbol{x}_1)^\top f(\boldsymbol{x}^-)} \right) - \log \left(e^{1-\varepsilon} + e^{\mu_k^\top \mu_\ell} \right) \right] \right].
\end{aligned}
$$

Then,

$$
\begin{aligned}
&|\Delta_1| \\
&\leq \log(2e) \sum_{k=1}^{K} \sum_{\ell=1}^{K} \mathop{\mathbb{E}}_{\boldsymbol{x},\boldsymbol{x}'} \left[\left(\mathbb{I}(\boldsymbol{x} \in C_k) - \mathbb{I}(\boldsymbol{x} \in S_\varepsilon \cap C_k) \right) \mathbb{I}(\boldsymbol{x}' \in C_\ell) \right] \\
&\quad + \sum_{k=1}^{K} \sum_{\ell=1}^{K} \mathop{\mathbb{E}}_{\boldsymbol{x},\boldsymbol{x}'} \left[\mathbb{I}(\boldsymbol{x} \in C_k) \mathbb{I}(\boldsymbol{x}' \in C_\ell) \mathop{\mathbb{E}}_{\substack{\boldsymbol{x}_1 \in A(\boldsymbol{x}) \\ \boldsymbol{x}^- \in A(\boldsymbol{x}')}} \left[\log \left(e^{1-\varepsilon} + e^{f(\boldsymbol{x}_1)^\top f(\boldsymbol{x}^-)} \right) - \log \left(e^{1-\varepsilon} + e^{\mu_k^\top \mu_\ell} \right) \right] \right] \\
&\leq (1 + \log 2) R_\varepsilon + \sum_{k=1}^{K} \sum_{\ell=1}^{K} \mathop{\mathbb{E}}_{\boldsymbol{x},\boldsymbol{x}'} \left[\mathbb{I}(\boldsymbol{x} \in C_k) \mathbb{I}(\boldsymbol{x}' \in C_\ell) \mathop{\mathbb{E}}_{\substack{\boldsymbol{x}_1 \in A(\boldsymbol{x}) \\ \boldsymbol{x}^- \in A(\boldsymbol{x}')}} \left[\log \left(e^{1-\varepsilon} + e^{f(\boldsymbol{x}_1)^\top f(\boldsymbol{x}^-)} \right) - \log \left(e^{1-\varepsilon} + e^{\mu_k^\top \mu_\ell} \right) \right] \right] \\
&\leq (1 + \log 2) R_\varepsilon + \sum_{k=1}^{K} \sum_{\ell=1}^{K} \mathop{\mathbb{E}}_{\boldsymbol{x},\boldsymbol{x}'} \left[\mathbb{I}(\boldsymbol{x} \in C_k) \mathbb{I}(\boldsymbol{x}' \in C_\ell) \mathop{\mathbb{E}}_{\substack{\boldsymbol{x}_1 \in A(\boldsymbol{x}) \\ \boldsymbol{x}^- \in A(\boldsymbol{x}')}} \left[\frac{e^\xi}{e^{1-\varepsilon} + e^\xi} \left| f(\boldsymbol{x}_1)^\top f(\boldsymbol{x}^-) - \mu_k^\top \mu_\ell \right| \right] \right] \\
&\hfill \text{(mean value theorem, } \xi \in [-1, 1])
\end{aligned}
$$

$$
\begin{aligned}
&\le (1+\log 2)R_\varepsilon + \sum_{k=1}^{K}\sum_{\ell=1}^{K} \mathop{\mathbb{E}}_{\boldsymbol{x},\boldsymbol{x}'}\left[\mathbb{I}(\boldsymbol{x}\in C_k)\mathbb{I}(\boldsymbol{x}'\in C_\ell) \mathop{\mathbb{E}}_{\substack{\boldsymbol{x}_1\in A(\boldsymbol{x})\\ \boldsymbol{x}^-\in A(\boldsymbol{x}')}} \left| f(\boldsymbol{x}_1)^\top f(\boldsymbol{x}^-) - \mu_k^\top \mu_\ell \right|\right] \\
&\le (1+\log 2)R_\varepsilon + \sum_{k=1}^{K}\sum_{\ell=1}^{K} \mathop{\mathbb{E}}_{\boldsymbol{x},\boldsymbol{x}'}\Bigg[\mathbb{I}(\boldsymbol{x}\in C_k)\mathbb{I}(\boldsymbol{x}'\in C_\ell) \mathop{\mathbb{E}}_{\substack{\boldsymbol{x}_1\in A(\boldsymbol{x})\\ \boldsymbol{x}^-\in A(\boldsymbol{x}')}} \Big[\left|(f(\boldsymbol{x}_1)-\mu_k)^\top (f(\boldsymbol{x}^-)-\mu_\ell)\right| \\
&\quad + \|f(\boldsymbol{x}_1)-\mu_k\|\cdot\|\mu_\ell\| + \|\mu_k\|\cdot\|f(\boldsymbol{x}^-)-\mu_\ell\|\Big]\Bigg] \\
&\le (1+\log 2)R_\varepsilon + \sum_{k=1}^{K}\sum_{\ell=1}^{K} \mathop{\mathbb{E}}_{\boldsymbol{x},\boldsymbol{x}'}\left[\mathbb{I}(\boldsymbol{x}\in C_k)\mathbb{I}(\boldsymbol{x}'\in C_\ell) \mathop{\mathbb{E}}_{\substack{\boldsymbol{x}_1\in A(\boldsymbol{x})\\ \boldsymbol{x}^-\in A(\boldsymbol{x}')}} \left|(f(\boldsymbol{x}_1)-\mu_k)^\top (f(\boldsymbol{x}^-)-\mu_\ell)\right|\right] \\
&\quad + 2\sum_{k=1}^{K} \mathop{\mathbb{E}}_{\boldsymbol{x}}\left[\mathbb{I}(\boldsymbol{x}\in C_k) \mathop{\mathbb{E}}_{\boldsymbol{x}_1\in A(\boldsymbol{x})} \|f(\boldsymbol{x}_1)-\mu_k\|\right] \\
&\le (1+\log 2)R_\varepsilon + \left[\sum_{k=1}^{K} p_k \mathop{\mathbb{E}}_{\boldsymbol{x}\in C_k}\mathop{\mathbb{E}}_{\boldsymbol{x}_1\in A(\boldsymbol{x})} \|f(\boldsymbol{x}_1)-\mu_k\|\right]^2 + 2\sum_{k=1}^{K} p_k \mathop{\mathbb{E}}_{\boldsymbol{x}\in C_k}\mathop{\mathbb{E}}_{\boldsymbol{x}_1\in A(\boldsymbol{x})} \|f(\boldsymbol{x}_1)-\mu_k\| \\
&\le (1+\log 2)R_\varepsilon + \left[\sum_{k=1}^{K} p_k\cdot 4\left(1-\sigma\left(1-\frac{\varepsilon}{2}-\frac{L\delta}{4}\right)+\frac{R_\varepsilon}{p_k}\right)\right]^2 + 2\sum_{k=1}^{K} p_k\cdot 4\left(1-\sigma\left(1-\frac{\varepsilon}{2}-\frac{L\delta}{4}\right)+\frac{R_\varepsilon}{p_k}\right) \\
&\hspace{30em}\text{(Lemma B.1)} \\
&\le (1+\log 2)R_\varepsilon + 16\left(1-\sigma\left(1-\frac{1}{2}\varepsilon-\frac{1}{4}L\delta\right)+KR_\varepsilon\right)^2 + 8\left(1-\sigma\left(1-\frac{1}{2}\varepsilon-\frac{1}{4}L\delta\right)+KR_\varepsilon\right),
\end{aligned}
$$

which is a small. Then (15) turns to

$$
\begin{aligned}
&p_k p_\ell \log\left(e^{1-\varepsilon}+e^{\mu_k^\top\mu_\ell}\right) \\
&\le \mathcal{L}_2(f) - \frac{1-\varepsilon}{K} + (1-\log 2)R_\varepsilon + |\Delta_1| \\
&\le \mathcal{L}_2(f) + 16\left(1-\sigma\left(1-\frac{1}{2}\varepsilon-\frac{1}{4}L\delta\right)+KR_\varepsilon\right)^2 + 8\left(1-\sigma\left(1-\frac{1}{2}\varepsilon-\frac{1}{4}L\delta\right)+KR_\varepsilon\right) + \frac{\varepsilon-1}{K} + 2R_\varepsilon.
\end{aligned}
$$

Let

$$
\tau(\varepsilon,\sigma,\delta) := 16\left(1-\sigma\left(1-\frac{1}{2}\varepsilon-\frac{1}{4}L\delta\right)+KR_\varepsilon\right)^2 + 8\left(1-\sigma\left(1-\frac{1}{2}\varepsilon-\frac{1}{4}L\delta\right)+KR_\varepsilon\right) + \frac{\varepsilon-1}{K} + 2R_\varepsilon,
$$

and we obtain

$$
\mu_k^\top\mu_\ell \le \log\left(\exp\left\{\frac{\mathcal{L}_2(f)+\tau(\varepsilon,\sigma,\delta)}{p_k p_\ell}\right\} - \exp(1-\varepsilon)\right).
$$

This finishes the proof. □

B.2 Cross-Correlation Loss

Lemma 4.1. *For a given encoder f, the alignment $\mathcal{L}_{\text{pos}}(f)$ in (4) is upper bounded via $\mathcal{L}_1(f)$, i.e.,*

$$
\mathcal{L}_{\text{pos}}(f) = \mathop{\mathbb{E}}_{\boldsymbol{x}}\mathop{\mathbb{E}}_{\boldsymbol{x}_1,\boldsymbol{x}_2\in A(\boldsymbol{x})} \|f(\boldsymbol{x}_1)-f(\boldsymbol{x}_2)\|^2 \le 2\sqrt{d\mathcal{L}_1(f)},
$$

where d is the output dimension of encoder f.

Proof. Since $\mathbb{E}_{\boldsymbol{x}} \mathbb{E}_{\boldsymbol{x}_1 \in A(\boldsymbol{x})} f_i(\boldsymbol{x}_1)^2 = 1$, for each coordinate component i, we have

$$\begin{aligned}
1 - \mathbb{E}_{\boldsymbol{x}} \mathbb{E}_{\boldsymbol{x}_1,\boldsymbol{x}_2 \in A(\boldsymbol{x})} [f_i(\boldsymbol{x}_1) f_i(\boldsymbol{x}_2)] &= \frac{1}{2} \mathbb{E}_{\boldsymbol{x}} \mathbb{E}_{\boldsymbol{x}_1,\boldsymbol{x}_2 \in A(\boldsymbol{x})} \left[f_i(\boldsymbol{x}_1)^2 + f_i(\boldsymbol{x}_2)^2\right] - \mathbb{E}_{\boldsymbol{x}} \mathbb{E}_{\boldsymbol{x}_1,\boldsymbol{x}_2 \in A(\boldsymbol{x})} [f_i(\boldsymbol{x}_1) f_i(\boldsymbol{x}_2)] \\
&= \frac{1}{2} \mathbb{E}_{\boldsymbol{x}} \mathbb{E}_{\boldsymbol{x}_1,\boldsymbol{x}_2 \in A(\boldsymbol{x})} [f_i(\boldsymbol{x}_1) - f_i(\boldsymbol{x}_2)]^2 .
\end{aligned}$$

Then

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{x}} \mathbb{E}_{\boldsymbol{x}_1,\boldsymbol{x}_2 \in A(\boldsymbol{x})} \|f(\boldsymbol{x}_1) - f(\boldsymbol{x}_2)\|^2 &= \sum_{i=1}^d \mathbb{E}_{\boldsymbol{x}} \mathbb{E}_{\boldsymbol{x}_1,\boldsymbol{x}_2 \in A(\boldsymbol{x})} [f_i(\boldsymbol{x}_1) - f_i(\boldsymbol{x}_2)]^2 \\
&= 2 \sum_{i=1}^d \left(1 - \mathbb{E}_{\boldsymbol{x}} \mathbb{E}_{\boldsymbol{x}_1,\boldsymbol{x}_2 \in A(\boldsymbol{x})} [f_i(\boldsymbol{x}_1) f_i(\boldsymbol{x}_2)]\right) \\
&\leq 2 \left(d \sum_{i=1}^d \left(1 - \mathbb{E}_{\boldsymbol{x}} \mathbb{E}_{\boldsymbol{x}_1,\boldsymbol{x}_2 \in A(\boldsymbol{x})} (f_i(\boldsymbol{x}_1) f_i(\boldsymbol{x}_2)]\right)^2 \right)^{\frac{1}{2}} \\
&= 2 d^{\frac{1}{2}} \mathcal{L}_1(f)^{\frac{1}{2}},
\end{aligned}$$

where the inequality holds due to Cauchy inequality. □

Lemma B.2. *Assume that encoder f with norm $\sqrt{d}$ is L-Lipschitz continuous. If the augmented data used in Barlow Twins is (σ, δ)-augmented, then for any $\varepsilon > 0$,*

$$\left\| \mathbb{E}_{\boldsymbol{x}} \mathbb{E}_{\boldsymbol{x}_1,\boldsymbol{x}_2 \in A(\boldsymbol{x})} \left[f(\boldsymbol{x}_1) f(\boldsymbol{x}_2)^\top\right] - \sum_{k=1}^K p_k \mu_k \mu_k^\top \right\|^2 \leq \tau'(\varepsilon, \sigma, \delta),$$

where

$$\tau'(\varepsilon, \sigma, \delta) := 4d \left(1 - \sigma + \frac{2\varepsilon + L\delta}{2\sqrt{d}}\right)^2 + 4(1-\sigma)d + 8dKR_\varepsilon \left(\frac{3}{2} - \sigma + \frac{2\varepsilon + L\delta}{2\sqrt{d}}\right) + 4dR_\varepsilon^2 \left(\sum_{k=1}^K \frac{1}{p_k}\right) + \sqrt{2} d^{\frac{3}{4}} \mathcal{L}_1(f)^{\frac{1}{4}}.$$

Proof. We first decompose the LHS as

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{x}} \mathbb{E}_{\boldsymbol{x}_1,\boldsymbol{x}_2 \in A(\boldsymbol{x})} \left[f(\boldsymbol{x}_1) f(\boldsymbol{x}_2)^\top\right] - \sum_{k=1}^K p_k \mu_k \mu_k^\top \\
&= \sum_{k=1}^K p_k \mathbb{E}_{\boldsymbol{x} \in C_k} \mathbb{E}_{\boldsymbol{x}_1,\boldsymbol{x}_2 \in A(\boldsymbol{x})} \left[f(\boldsymbol{x}_1) f(\boldsymbol{x}_2)^\top\right] - \sum_{k=1}^K p_k \mu_k \mu_k^\top \\
&= \sum_{k=1}^K p_k \mathbb{E}_{\boldsymbol{x} \in C_k} \mathbb{E}_{\boldsymbol{x}_1 \in A(\boldsymbol{x})} \left[f(\boldsymbol{x}_1) f(\boldsymbol{x}_1)^\top\right] - \sum_{k=1}^K p_k \mu_k \mu_k^\top + \sum_{k=1}^K p_k \mathbb{E}_{\boldsymbol{x} \in C_k} \mathbb{E}_{\boldsymbol{x}_1,\boldsymbol{x}_2 \in A(\boldsymbol{x})} \left[f(\boldsymbol{x}_1)(f(\boldsymbol{x}_2)^\top - f(\boldsymbol{x}_1)^\top)\right] \\
&= \sum_{k=1}^K p_k \mathbb{E}_{\boldsymbol{x} \in C_k} \mathbb{E}_{\boldsymbol{x}_1 \in A(\boldsymbol{x})} \left[(f(\boldsymbol{x}_1) - \mu_k)(f(\boldsymbol{x}_1) - \mu_k)^\top\right] + \mathbb{E}_{\boldsymbol{x}} \mathbb{E}_{\boldsymbol{x}_1,\boldsymbol{x}_2 \in A(\boldsymbol{x})} \left[f(\boldsymbol{x}_1)(f(\boldsymbol{x}_2)^\top - f(\boldsymbol{x}_1)^\top)\right] .
\end{aligned}$$

Then its norm is

$$
\begin{aligned}
&\left\|\mathop{\mathbb{E}}_{\boldsymbol{x}} \mathop{\mathbb{E}}_{\boldsymbol{x}_1,\boldsymbol{x}_2\in A(\boldsymbol{x})}\left[f(\boldsymbol{x}_1)f(\boldsymbol{x}_2)^\top\right] - \sum_{k=1}^K p_k\mu_k\mu_k^\top\right\| \\
&\leq \sum_{k=1}^K p_k \mathop{\mathbb{E}}_{\boldsymbol{x}\in C_k} \mathop{\mathbb{E}}_{\boldsymbol{x}_1\in A(\boldsymbol{x})}\left[\left\|(f(\boldsymbol{x}_1)-\mu_k)(f(\boldsymbol{x}_1)-\mu_k)^\top\right\|\right] + \mathop{\mathbb{E}}_{\boldsymbol{x}} \mathop{\mathbb{E}}_{\boldsymbol{x}_1,\boldsymbol{x}_2\in A(\boldsymbol{x})}\left[\left\|f(\boldsymbol{x}_1)(f(\boldsymbol{x}_2)^\top - f(\boldsymbol{x}_1)^\top)\right\|\right] \\
&\leq \sum_{k=1}^K p_k \mathop{\mathbb{E}}_{\boldsymbol{x}\in C_k} \mathop{\mathbb{E}}_{\boldsymbol{x}_1\in A(\boldsymbol{x})}\left[\|f(\boldsymbol{x}_1)-\mu_k\|^2\right] + \mathop{\mathbb{E}}_{\boldsymbol{x}} \mathop{\mathbb{E}}_{\boldsymbol{x}_1,\boldsymbol{x}_2\in A(\boldsymbol{x})}\left[\|f(\boldsymbol{x}_1)\|\,\|f(\boldsymbol{x}_2)-f(\boldsymbol{x}_1)\|\right] \\
&\leq \sum_{k=1}^K p_k \mathop{\mathbb{E}}_{\boldsymbol{x}\in C_k} \mathop{\mathbb{E}}_{\boldsymbol{x}_1\in A(\boldsymbol{x})}\left[\|f(\boldsymbol{x}_1)-\mu_k\|^2\right] + \left[\mathop{\mathbb{E}}_{\boldsymbol{x}} \mathop{\mathbb{E}}_{\boldsymbol{x}_1\in A(\boldsymbol{x})}\|f(\boldsymbol{x}_1)\|^2\right]^{\frac{1}{2}}\left[\mathop{\mathbb{E}}_{\boldsymbol{x}} \mathop{\mathbb{E}}_{\boldsymbol{x}_1,\boldsymbol{x}_2\in A(\boldsymbol{x})}\|f(\boldsymbol{x}_2)-f(\boldsymbol{x}_1)\|^2\right]^{\frac{1}{2}} \\
&\hfill \text{(Cauchy–Schwarz inequality)} \\
&\leq \sum_{k=1}^K p_k \mathop{\mathbb{E}}_{\boldsymbol{x}\in C_k} \mathop{\mathbb{E}}_{\boldsymbol{x}_1\in A(\boldsymbol{x})}\left[\|f(\boldsymbol{x}_1)-\mu_k\|^2\right] + \sqrt{d}\left(2\sqrt{d\mathcal{L}_1(f)}\right)^{\frac{1}{2}} \qquad \text{(Lemma 4.1)} \\
&\leq (2\varepsilon+L\delta)^2 + 4r(1-\sigma)(r+2\varepsilon+L\delta) + 4r^2(1-\sigma)^2 + KR_\varepsilon[4r(r+2\varepsilon+L\delta)+8r^2(1-\sigma)] + 4r^2R_\varepsilon^2\left(\sum_{k=1}^K\frac{1}{p_k}\right) \\
&\quad + \sqrt{2}d^{\frac{3}{4}}\mathcal{L}_1(f)^{\frac{1}{4}} \qquad \text{(Lemma B.1)} \\
&= 4r^2\left(1-\sigma+\frac{2\varepsilon+L\delta}{2r}\right)^2 + 4(1-\sigma)r^2 + 8r^2KR_\varepsilon\left(\frac{3}{2}-\sigma+\frac{2\varepsilon+L\delta}{2r}\right) + 4r^2R_\varepsilon^2\left(\sum_{k=1}^K\frac{1}{p_k}\right) + \sqrt{2}d^{\frac{3}{4}}\mathcal{L}_1(f)^{\frac{1}{4}} \\
&= 4d\left(1-\sigma+\frac{2\varepsilon+L\delta}{2\sqrt{d}}\right)^2 + 4(1-\sigma)d + 8dKR_\varepsilon\left(\frac{3}{2}-\sigma+\frac{2\varepsilon+L\delta}{2\sqrt{d}}\right) + 4dR_\varepsilon^2\left(\sum_{k=1}^K\frac{1}{p_k}\right) + \sqrt{2}d^{\frac{3}{4}}\mathcal{L}_1(f)^{\frac{1}{4}} \\
&= \tau'(\varepsilon,\sigma,\delta).
\end{aligned}
$$

This finishes the proof. □

Theorem 4. *Assume that encoder f with norm $\sqrt{d}$ is L-Lipschitz continuous. If the augmented data used in Barlow Twins is (σ,δ)-augmented, then for any $k\neq\ell$, we have*

$$
\mu_k^\top\mu_\ell \leq \sqrt{\frac{2}{p_kp_\ell}\left(\mathcal{L}_2(f)+\tau'(\varepsilon,\sigma,\delta)-\frac{d-K}{2}\right)} \lesssim \mathcal{L}_2(f)^{1/2},
$$

where $\tau'(\varepsilon,\sigma,\delta)$ is the upper bound of $\|\mathbb{E}_{\boldsymbol{x}}\mathbb{E}_{\boldsymbol{x}_1,\boldsymbol{x}_2\in A(\boldsymbol{x})}[f(\boldsymbol{x}_1)f(\boldsymbol{x}_2)^\top] - \sum_{k=1}^K p_k\mu_k\mu_k^\top\|^2$.

Proof. Let $U = (\sqrt{p_1}\mu_1,\ldots,\sqrt{p_K}\mu_K)\in\mathbb{R}^{d\times K}$.

$$
\begin{aligned}
\left\|\sum_{k=1}^K p_k\mu_k\mu_k^\top - I_d\right\|^2 &= \left\|UU^\top - I_d\right\|^2 \\
&= \mathrm{Tr}(UU^\top UU^\top - 2UU^\top + I_d) \qquad \text{(due to } \|A\|^2 = \mathrm{Tr}(A^\top A)\text{)} \\
&= \mathrm{Tr}(U^\top UU^\top U - 2U^\top U + I_K) + d - K \\
&= \left\|U^\top U - I_K\right\|^2 + d - K
\end{aligned}
$$

$$= \sum_{k=1}^{K}\sum_{\ell=1}^{K}(\sqrt{p_k p_\ell}\mu_k^\top \mu_\ell - \delta_{k\ell})^2 + d - K$$
$$\geq p_k p_\ell (\mu_k^\top \mu_\ell)^2 + d - K.$$

where δ_{kl} is the Dirichlet function.

Therefore,

$$\begin{aligned}
&\left(\mu_k^\top \mu_l\right)^2 \\
&\leq \frac{\left\|\sum_{k=1}^{K} p_k \mu_k \mu_k^\top - I_d\right\|^2 - (d-K)}{p_k p_\ell} \\
&= \frac{\left\|\mathbb{E}_{\boldsymbol{x}}\,\mathbb{E}_{\boldsymbol{x}_1,\boldsymbol{x}_2 \in A(\boldsymbol{x})}[f(\boldsymbol{x}_1)f(\boldsymbol{x}_2)^\top] - I_d + \sum_{k=1}^{K} p_k \mu_k \mu_k^\top - \mathbb{E}_{\boldsymbol{x}}\,\mathbb{E}_{\boldsymbol{x}_1,\boldsymbol{x}_2 \in A(\boldsymbol{x})}[f(\boldsymbol{x}_1)f(\boldsymbol{x}_2)^\top]\right\|^2 - (d-K)}{p_k p_\ell} \\
&\leq \frac{2\left\|\mathbb{E}_{\boldsymbol{x}}\,\mathbb{E}_{\boldsymbol{x}_1,\boldsymbol{x}_2 \in A(\boldsymbol{x})}[f(\boldsymbol{x}_1)f(\boldsymbol{x}_2)^\top] - I_d\right\|^2 + 2\left\|\sum_{k=1}^{K} p_k \mu_k \mu_k^\top - \mathbb{E}_{\boldsymbol{x}}\,\mathbb{E}_{\boldsymbol{x}_1,\boldsymbol{x}_2 \in A(\boldsymbol{x})}[f(\boldsymbol{x}_1)f(\boldsymbol{x}_2)^\top]\right\|^2 - (d-K)}{p_k p_\ell} \\
&\leq \frac{2\mathcal{L}_2(f) + 2\tau'(\varepsilon,\sigma,\delta) - (d-K)}{p_k p_\ell} \qquad \text{(Lemma B.2)} \\
&= \frac{2}{p_k p_\ell}\left(\mathcal{L}_2(f) + \tau'(\varepsilon,\sigma,\delta) - \frac{d-K}{2}\right).
\end{aligned}$$

This finishes the proof. □