On the Surrogate Gap between Contrastive and Supervised Losses

Han Bao [* 1 2]   **Yoshihiro Nagano** [* 1 2]   **Kento Nozawa** [* 1 2]

Abstract

Contrastive representation learning encourages data representation to make semantically similar pairs closer than randomly drawn negative samples, which has been successful in various domains such as vision, language, and graphs. Recent theoretical studies have attempted to explain the benefit of the large negative sample size by upper-bounding the downstream classification loss with the contrastive loss. However, the previous surrogate bounds have two drawbacks: they are only legitimate for a limited range of negative sample sizes and prohibitively large even within that range. Due to these drawbacks, there still does not exist a consensus on *how negative sample size theoretically correlates with downstream classification performance*. Following the simplified setting where positive pairs are drawn from the true distribution (not generated by data augmentation; as supposed in previous studies), this study establishes surrogate upper and lower bounds for the downstream classification loss for all negative sample sizes that best explain the empirical observations on the negative sample size in the earlier studies. Our bounds suggest that the contrastive loss can be viewed as a surrogate objective of the downstream loss and larger negative sample sizes improve downstream classification because the surrogate gap between contrastive and supervised losses decays. We verify that our theory is consistent with experiments on synthetic, vision, and language datasets.

1. Introduction

The contrastive loss (Chopra et al., 2005) is one of the popular loss functions in metric learning (Kulis, 2012) and representation learning (Bengio et al., 2013). The contrastive

---
[*]Equal contribution [1]The University of Tokyo, Tokyo, Japan [2]RIKEN AIP, Tokyo, Japan. Correspondence to: Han Bao (currently with Kyoto University) <bao@i.kyoto-u.ac.jp>.
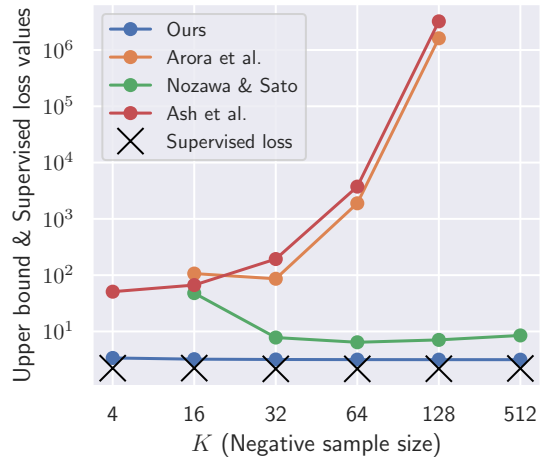
Figure 1: Empirical comparison of our upper bound and the existing bounds of the classification loss with CIFAR-10 ($C = 10$), showing that *we improve the classification loss bound by exponentially large margins*. Arora et al.'s and Nozawa & Sato's bounds are valid only at $K + 1 \geq C$. Note that Arora et al.'s and Ash et al.'s bounds become infinity at $K = 512$, where K and C are the numbers of negative samples and classes, respectively. As can be seen, our bound is the closest approximator of the true mean supervised losses. The detailed setup is in Section 5.2.

loss forces data representation of semantically similar pairs closer in some metric space than multiple random samples, called negative samples. Many state-of-the-art representation learning algorithms use a type of contrastive losses in natural language processing (Mikolov et al., 2013; Logeswaran & Lee, 2018), vision (Chopra et al., 2005; He et al., 2019; Chen et al., 2020), and graph (Lirong et al., 2021) domains. A simple model built on top of the learned representation can achieve almost the same accuracy as supervised learning does.

Recent empirical studies observed that downstream classification performance could be improved with a sufficiently large negative sample size (denoted by K), compared with the number of classes (denoted by C) (He et al., 2020; Chen et al., 2020). To better understand the underlying mechanism of this large K benefit, several studies attempted to derive surrogate upper bounds of the downstream classification loss by the contrastive loss. Arora et al. (2019) success-

fully established the first upper bound, which exponentially deteriorates with larger K because the labels of negative samples frequently collide with the positive sample (called *label collision*), contradicting the larger-K benefit. By contrast, Nozawa & Sato (2021) argued that supervised classes could be *covered* by negative samples with higher probability as K becomes larger (called *label coverage*), supported by their bound. While their claim agrees with the large-K benefit, their bound holds only when $K > C + 1$, and hence does not explain the empirical observation that contrastive learning works to some extent even with small K (Chen et al., 2021; Tomasev et al., 2022). Furthermore, Ash et al. (2022) advocated the existence of *collision-coverage trade-off* so that their upper bound has an *optimal K*. Why have we yet to reach a consensus? We observe that the lack of a consensus is due to the existing upper bounds having the following drawbacks: some bounds are only valid within a limited range of K, and prohibitively large even within that range. Figure 1 shows the comparison of the existing bounds and the classification loss. Hence, we ask the following research question: *How does negative sample size K affect the downstream classification performance?*

In this study, we derive a surrogate gap bound of the downstream classification loss that is applicable to any K and shrinks with larger K. In particular, we derive not only the upper (Theorem 1) but also the *lower* bound (Theorem 2) of the downstream loss, and show the tightness in K. As the gap between upper and lower bounds shrinks in $O(K^{-1})$, the contrastive loss can be viewed as a surrogate objective of the downstream classification loss, and the downstream performance is improved by larger K. This is consistent with the empirical observations that larger K improves the downstream performance whereas contrastive learning can work to some extent even with small K (Section 3.3). In addition, our bounds resolve the controversy among the existing bounds so that there is no collision-coverage trade-off in K and the large-K benefit is witnessed (Section 4). Finally, we empirically verify our theory by experiments (Section 5) on a synthetic dataset, CIFAR-10/100 (Krizhevsky, 2009) datasets, and Wiki-3029 dataset (Arora et al., 2019). Note that we assume that positive pairs are drawn from the true underlying distribution instead of generated by data augmentation for simplicity, as supposed in the previous studies.

2. Formulation of Contrastive Learning

First, this section briefly summarizes the problem setup and formulation of contrastive unsupervised representation learning (CURL).[1]

---

[1] We refer to our problem setting as contrastive *unsupervised* representation learning by following Arora et al. (2019) while CURL is provided with the contrastive supervision.

Notation. The C-dimensional vector whose elements are all ones is denoted by $\mathbf{1}_C := \begin{bmatrix} 1 & 1 & \dots & 1 \end{bmatrix}^\top$. When it is clear from context, the subscript is abbreviated. For a vector $\mathbf{a} \in \mathbb{R}^p$, $a_{(i)}$ denotes the i-th largest element of $\mathbf{a}$, namely, $a_{(1)} \geq a_{(2)} \geq \cdots \geq a_{(p)}$. Likewise, $a_{(-i)}$ denotes the i-th smallest element of the vector $\mathbf{a}$. The indicator function is denoted by $\mathbb{1}_{\{A\}}$ for a predicate A. Let $\triangle^C := \{\mathbf{p} \in [0,1]^C \mid \mathbf{p}^\top \mathbf{1} = 1\}$ be the C-dimensional probability simplex. For $\mathbf{p} \in \triangle^C$, the Shannon entropy is denoted by $\mathbb{H}(\mathbf{p}) := -\sum_{c \in [C]} p_c \ln p_c$.

Supervised classification. One of the goals in machine learning is supervised classification, while we consider the setup where the label supervision is unavailable. Here, we first formulate C-class classification problem for $C \in \mathbb{N}$. Let $\mathcal{X}$ be d-dimensional feature space and $\mathcal{Y} := [C]$ be the supervised label set. In the supervised setup, we are interested in the following risk quantity, the *supervised loss*, for a multi-class classifier $\mathbf{g} : \mathcal{X} \to \mathbb{R}^C$:

$$R_{\text{supv}}(\mathbf{g}) := \mathop{\mathbb{E}}_{\mathbf{x}, y \sim \mathbb{P}} \left[-\ln \frac{e^{g_y(\mathbf{x})}}{\sum_{c \in \mathcal{Y}} e^{g_c(\mathbf{x})}} \right], \qquad (1)$$

which is specialized for the softmax cross-entropy loss. The expectation is taken over the unknown underlying joint distribution $\mathbb{P}$. Test prediction is given by $\text{argmax}_{y \in \mathcal{Y}} g_y(\cdot)$.

Contrastive unsupervised representation learning. In the CURL framework (Arora et al., 2019), we target to learn meaningful data representation by training a similarity model to make the representation of *positive* pairs more similar than randomly drawn K *negative* samples. The class-conditional distribution is denoted by $\mathcal{D}_c := \mathbb{P}(X \mid Y = c)$ for each $c \in \mathcal{Y}$ and the class-prior distribution by $\boldsymbol{\pi} := [\mathbb{P}(Y = c)]_{c \in \mathcal{Y}} \in \triangle^C$. The data generating process is described as follows: (i) draw positive/negative classes: $c^+, \{c_k^-\}_{k=1}^K \sim \boldsymbol{\pi}^{K+1}$ (ii) draw an anchor sample $\mathbf{x} \sim \mathcal{D}_{c^+}$ (iii) draw a positive sample $\mathbf{x}^+ \sim \mathcal{D}_{c^+}$ (iv) draw K negative samples $\mathbf{x}_k^- \sim \mathcal{D}_{c_k^-}$ (for each $k \in [K]$).

In CURL, the representation is learned through minimization of the following *contrastive loss* $R_{\text{cont}}(\mathbf{f})$

$$\mathop{\mathbb{E}}_{\substack{c^+, \{c_k^-\} \\ \mathbf{x}, \mathbf{x}^+, \{\mathbf{x}_k^-\}}} \left[-\ln \frac{e^{\mathbf{f}(\mathbf{x})^\top \mathbf{f}(\mathbf{x}^+)}}{e^{\mathbf{f}(\mathbf{x})^\top \mathbf{f}(\mathbf{x}^+)} + \sum_{k \in [K]} e^{\mathbf{f}(\mathbf{x})^\top \mathbf{f}(\mathbf{x}_k^-)}} \right]. \quad (2)$$

Evaluation of representations. Now we specify our model of classifiers to evaluate learned representations. A multi-class classifier $\mathbf{g} : \mathcal{X} \to \mathbb{R}^C$ consists of learned representation $\mathbf{f} : \mathcal{X} \to \mathbb{R}^h$ (frozen) and linear parameters $\mathbf{W} \in \mathbb{R}^{C \times h}$ as $\mathbf{g}(\cdot) := \mathbf{W}\mathbf{f}(\cdot)$, where $h \in \mathbb{N}$ denotes the dimensionality of the representation given in advance.

For the sake of evaluation, a specific linear classifier called *mean classifier* is introduced. Given representation $\mathbf{f}$, the

mean classifier $\mathbf{W}^\mu$ is defined as $\mathbf{W}^\mu := [\boldsymbol{\mu}_1 \cdots \boldsymbol{\mu}_C]^\top$, where $\boldsymbol{\mu}_c := \mathbb{E}_{\mathbf{x} \sim \mathcal{D}_c}[\mathbf{f}(\mathbf{x})]$. This will later be used for evaluating the representation $\mathbf{f}$ combined with the supervised loss, which is denoted by $R_{\mu\text{-supv}}(\mathbf{f}) := R_{\text{supv}}(\mathbf{W}^\mu \mathbf{f})$. We call it the *mean supervised loss*. If the mean supervised loss is successfully bounded from above, we end up a bound on the supervised loss through $\inf_{\mathbf{W} \in \mathbb{R}^{C \times h}} R_{\text{supv}}(\mathbf{W}\mathbf{f}) \leq R_{\mu\text{-supv}}(\mathbf{f})$. For this reason, an upper bound on $R_{\mu\text{-supv}}$ is an intermediate milestone that we seek in this paper.

3. Surrogate Bounds for Contrastive Learning

In this section, our main theoretical results are provided. We aim at showing that the contrastive loss $R_{\text{cont}}(\mathbf{f})$ serves as a good estimator of the mean supervised loss $R_{\mu\text{-supv}}(\mathbf{f})$ for any $\mathbf{f}$. We show this by establishing upper and lower bounds of $R_{\mu\text{-supv}}(\mathbf{f})$ by $R_{\text{cont}}(\mathbf{f})$. Eventually, the minimization of $R_{\text{cont}}(\mathbf{f})$ may lead to a good minimizer of $R_{\mu\text{-supv}}(\mathbf{f})$. All proofs are provided in Appendix B.

3.1. Assumptions

Before proceeding with the main results, we explicitly state assumptions used throughout this paper and discuss their validity.

Conditional independence. In Section 2, we assumed that anchor and positive samples are conditionally independent: $\mathbf{x} \perp\!\!\!\perp \mathbf{x}^+ \mid c^+$, whereas data augmentation (DA) is commonly combined with contrastive learning, and the assumption no longer holds. While a concurrent work (Wang et al., 2022) attempted to mitigate this assumption, we work with this assumption to concentrate on the theoretical relationship between K and the downstream performance. With this assumption, it is possible to compare our result with the previous bounds in a relatively fair manner since the previous studies assumed the same assumption (Arora et al., 2019; Nozawa & Sato, 2021; Ash et al., 2022). The conditional independence assumption has been used in metric learning (Bellet et al., 2012) and weakly-supervised learning (Bao et al., 2018). In Appendix C, we discuss how to relax the conditional independence assumption.

Existence of supervised classes. In unsupervised representation learning, the latent classes $[C]$ and the downstream supervised classes $\mathcal{Y}$ are often distinguished. To draw the connection between learned representation and downstream classification, we must suppose the relationship between $[C]$ and $\mathcal{Y}$. We assume $\mathcal{Y} = [C]$ for ease of exposition in the main part. This assumption can be relaxed to some extent (similarly to Arora et al. (2019)), which will be discussed in Appendix C.

Bounded feature representation. The size of the representation $\|\mathbf{f}(\mathbf{x})\|_2$ is assumed to be bounded. This assumption is reasonable from the experimental perspective since it is common to normalize representation to employ the cosine similarity as the similarity metric. Several works reported that the normalized embeddings improve the performance (Chen et al., 2020; Wang & Isola, 2020). The existing theoretical work (Arora et al., 2019) also assumes the bounded feature extractor. Unlike the existing analyses (reviewed in Section 4), we take advantage of this assumption to derive the sharp bounds.

Loss function. We focus on the cross-entropy-type contrastive loss (2) because this is the most commonly used loss function (Mikolov et al., 2013; Logeswaran & Lee, 2018; Chen et al., 2020) and its connection to mutual information has been actively discussed (Tian et al., 2020; Tschannen et al., 2020), while other contrastive loss functions have been proposed in a few recent studies (Li et al., 2021; HaoChen et al., 2021; Chuang et al., 2022).

3.2. Main Results

Below, we investigate the *surrogate gap* $R_{\mu\text{-supv}}(\mathbf{f}) - R_{\text{cont}}(\mathbf{f})$ for a fixed representation $\mathbf{f}$. If the surrogate gap is bounded sufficiently small, the contrastive loss $R_{\text{cont}}(\mathbf{f})$ can be regarded a good surrogate objective for $R_{\mu\text{-supv}}(\mathbf{f})$.

First, we show a sharp upper bound of the mean supervised loss. Unlike the existing surrogate bounds of CURL, the upper bound obtained here has a constant coefficient in the contrastive loss and is applicable for all C and K (see discussions in Section 4).

Theorem 1. *For all $\mathbf{f}$ such that $\|\mathbf{f}(\mathbf{x})\|_2 \leq L$ ($\forall \mathbf{x} \in \mathcal{X}$), the following inequality holds.*

$$R_{\mu\text{-supv}}(\mathbf{f}) \leq R_{\text{cont}}(\mathbf{f}) + \Delta_{\text{U}}, \qquad (3)$$

where $\Delta_{\text{U}} := \ln\{\pi_{(1)} K^{-1} C^2 \cosh^2(L^2)\}$.

Next, the lower bound of the mean supervised loss is provided. While the existing theoretical analyses often provided upper bounds with a huge coefficient in the contrastive loss, our lower bound provided below has the same constant coefficient and intercept (Δ_{U} and Δ_{L}) rate as our upper bound, ensuring the tightness of our analysis.

Theorem 2. *For all $\mathbf{f}$ such that $\|\mathbf{f}(\mathbf{x})\|_2 \leq L$ ($\forall \mathbf{x} \in \mathcal{X}$), the following inequality holds.*

$$R_{\mu\text{-supv}}(\mathbf{f}) \geq R_{\text{cont}}(\mathbf{f}) + \Delta_{\text{L}}, \qquad (4)$$

where $\Delta_{\text{L}} := \mathbb{H}(\boldsymbol{\pi}) + \ln \frac{K}{(K+1)^2} - 2 \ln \cosh(L^2)$.

Our proofs leverage that the contrastive loss and mean supervised loss share the similar log-sum-exp functional form
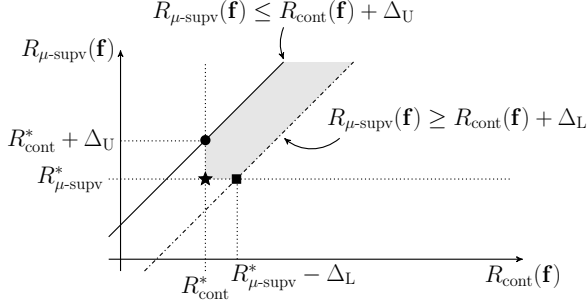
Figure 2: The surrogate bounds and feasible region. The point ★ , $(R_{\text{cont}}^*, R_{\mu\text{-supv}}^*)$, is the optimal point in the feasible region. The points ● and ■ are mentioned in the texts.

and directly apply the Jensen's inequality. This is in contrast to the existing works including Arora et al. (2019), which approximate the mean supervised loss with the contrastive loss by taking the expectation over latent classes, leading to an exponentially large coefficient.

As we see in Section 3.3, Δ_{U} and Δ_{L} are the same order in K under the uniform class prior assumption. By applying either the high-probability bound (Arora et al., 2019) or PAC-Bayesian analysis (Nozawa et al., 2020), Theorem 1 (Theorem 2 as well) can be naturally extended to the form $R_{\mu\text{-supv}}(\widehat{\mathbf{f}}) \leq R_{\text{cont}}(\mathbf{f}) + \Delta_{\text{U}} + \chi$ with a complexity term χ, where $\widehat{\mathbf{f}}$ is the empirical minimizer of the contrastive loss. Since this is a routine and does not affect the surrogate gap, we omit the high-probability bounds.

3.3. Discussion

Subsequently, we discuss implications of our main results on the relationship between the mean supervised loss and K. For the sake of simplicity, we assume $\pi_c = 1/C$ for all $c \in [C]$ (the uniform class prior) in this section.

Gap between upper and lower bounds. Both of our upper (Theorem 1) and lower (Theorem 2) bounds draw the linear relationship between the mean supervised loss $R_{\mu\text{-supv}}$ and the contrastive loss R_{cont}, with the additional intercept terms Δ_{U} and Δ_{L}. Under the uniform class prior assumption, the intercepts are in the same order:

$$\Delta_{\text{U}} = \ln\left(C/K\right) + 2\ln\cosh(L^2) = O\left(\ln\left(1/K\right)\right),$$
$$\Delta_{\text{L}} = \ln\left(CK/(K+1)^2\right) - 2\ln\cosh(L^2) = O\left(\ln\left(1/K\right)\right),$$

and the gap between two bounds $\Delta_{\text{U}} - \Delta_{\text{L}}$ is[2]

$$4\ln\cosh(L^2) + 2\ln\left(1 + 1/K\right) = O\left(K^{-1}\right), \quad (5)$$

meaning that the gap shrinks to $4\ln\cosh(L^2)$ as K increases. Hence, our bounds have the tight intercepts, and

---

[2]The approximation $\ln(1 + z) \approx z$ is used (for $0 < z \ll 1$).

the larger K is beneficial for CURL from the viewpoint of the surrogate gap of the mean supervised loss.

Surrogate bounds and feasible region. Next, we consider the $(R_{\text{cont}}, R_{\mu\text{-supv}})$-plot, in which a point indicates $(R_{\text{cont}}(\mathbf{f}), R_{\mu\text{-supv}}(\mathbf{f}))$ for some $\mathbf{f}$ (see Figure 2). Here, let us focus on the feasible region in the $(R_{\text{cont}}, R_{\mu\text{-supv}})$-plot by assuming $\|\mathbf{f}\|_2 \leq L$ for any $\mathbf{f}$ (same as Theorems 1 and 2). Then, the mean supervised loss and contrastive loss are essentially lower-bounded by the constants[3]

$$R_{\mu\text{-supv}}^* := \ln\{1 + (C-1)e^{-2L^2}\}, \quad (6)$$

$$R_{\text{cont}}^* := \sum_{m=0}^{K} r_{K,C,m} \ln\{1 + m + (K-m)e^{-2L^2}\}, \quad (7)$$

respectively, where $r_{K,C,m} := \binom{K}{m}\left(\frac{1}{C}\right)^m\left(1 - \frac{1}{C}\right)^{K-m}$. Hence, the feasible region is

$$R_{\mu\text{-supv}}(\mathbf{f}) \leq R_{\text{cont}}(\mathbf{f}) + \Delta_{\text{U}}, \quad (8a)$$
$$R_{\mu\text{-supv}}(\mathbf{f}) \geq R_{\text{cont}}(\mathbf{f}) + \Delta_{\text{L}}, \quad (8b)$$
$$R_{\mu\text{-supv}}(\mathbf{f}) \geq R_{\mu\text{-supv}}^*, \quad (8c)$$
$$R_{\text{cont}}(\mathbf{f}) \geq R_{\text{cont}}^*, \quad (8d)$$

as illustrated in Figure 2. The first two bounds (8a) and (8b) restrict the mean supervised loss by the contrastive loss. We specifically refer to these bounds as *surrogate bounds*. The remaining two bounds (8c) and (8d) represent the achievable limits for each loss separately. One of the important questions is how the smallest possible value of $R_{\mu\text{-supv}}$ in the feasible region (8) changes as K and C change. In other words, we are interested in whether the optimal point $(R_{\text{cont}}^*, R_{\mu\text{-supv}}^*)$ (★ in Figure 2) is always achievable regardless of the values of K and C. To investigate it, we check whether the optimal point (★) crosses the surrogate gaps (● or ■) under the following two conditions.

- The feasible region at $R_{\text{cont}}(\mathbf{f}) = R_{\text{cont}}^*$ (Figure 3a): We plot the value $R_{\text{cont}}^* + \Delta_{\text{U}}$ (solid line; the $R_{\mu\text{-supv}}$-value of the point ● in Figure 2) and the minimum possible $R_{\mu\text{-supv}}$ ($R_{\mu\text{-supv}}^*$; dotted line) numerically. These two curves do not cross for all K, which means $R_{\mu\text{-supv}}(\mathbf{f}) = R_{\mu\text{-supv}}^*$ is attainable no matter the values K and C. In addition, the bound becomes sharper as K increases, but the gap between the upper bound and $R_{\mu\text{-supv}}^*$ does remain even at the limit $K \nearrow \infty$.

- The feasible region at $R_{\mu\text{-supv}}(\mathbf{f}) = R_{\mu\text{-supv}}^*$ (Figure 3b): When $R_{\mu\text{-supv}}(\mathbf{f}) = R_{\mu\text{-supv}}^*$, the contrastive loss $R_{\text{cont}}(\mathbf{f})$ is upper-bounded by $R_{\mu\text{-supv}}^* - \Delta_{\text{L}}$ (the R_{cont}-value of the point ■ in Figure 2). The curve

---

[3]The derivations of $R_{\mu\text{-supv}}^*$ and R_{cont}^* are detailed in Appendix D.

(a) The feasible region at $R_{\text{cont}}(\mathbf{f}) = R_{\text{cont}}^*$.

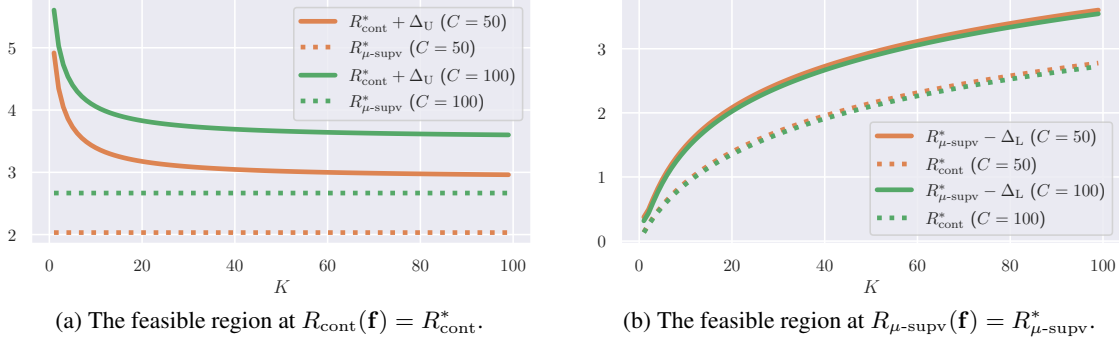(b) The feasible region at $R_{\mu\text{-supv}}(\mathbf{f}) = R_{\mu\text{-supv}}^*$.

Figure 3: Visualization of the smallest possible value of $R_{\mu\text{-supv}}$ in the feasible region (8) for different K and C. The dotted lines show the essential lower bounds which come from each loss separately. The solid lines show the surrogate upper bounds in which $R_{\mu\text{-supv}}$ and R_{cont} restrict each other.

of this value does not cross R_{cont}^*, which tells us that the lower bound does not exclude the optimal point $(R_{\text{cont}}^*, R_{\mu\text{-supv}}^*)$ from the feasible region (8) at any K. Note that the gap between R_{cont}^* and $R_{\mu\text{-supv}}^* - \Delta_{\text{L}}$ gradually increases, meaning that it becomes much easier to attain $R_{\mu\text{-supv}}^*$ as K increases.

Hence, the optimal point ★ stays in the feasible region (8) no matter the value K. From this viewpoint, *smaller K is not necessarily disadvantageous because the optimal point ★ remains in the feasible region.* Note again that the estimation of $R_{\mu\text{-supv}}$ may become harder with the smaller K because of the gap $\Delta_{\text{U}} - \Delta_{\text{L}} = O(K^{-1})$, even if the optimal solution is unaffected by K.

Summary. We draw a connection between the mean supervised loss and the negative sample size K by the following claim: the gap between the contrastive loss and mean supervised loss shrinks with larger K but the optimal mean supervised loss can nevertheless be achieved with small K.

4. Comparison with Existing Work

This section first discusses the detailed difference between our main results and the existing theoretical results on CURL. Then, we briefly review the other related literatures.

Surrogate bounds comparison. Here, we compare our results with the existing works by Arora et al. (2019), Nozawa & Sato (2021), and Ash et al. (2022). We assume the uniform class prior for comparison. We introduce a notation $\text{Col} := \sum_{k \in [K]} \mathbb{1}_{\{c^+ = c_k^-\}}$. Let v_K be the probability that sampled K negative classes contains all classes $c \in [C]$.

$$v_K := \sum_{n=1}^{K} \sum_{m=0}^{C-1} \binom{C-1}{m} (-1)^m \left(1 - \frac{m+1}{C}\right)^{n-1}. \quad (9)$$
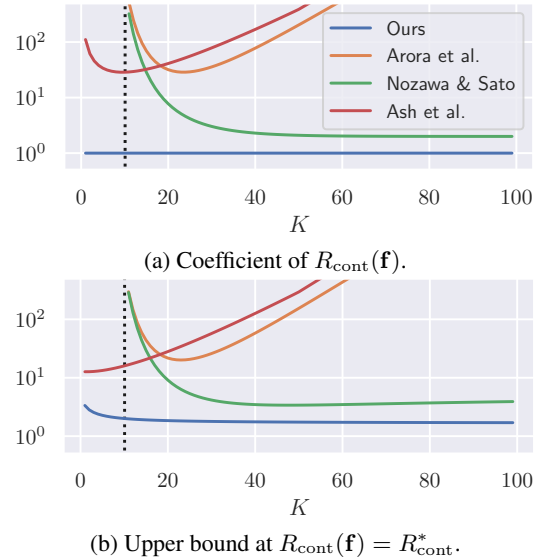

(a) Coefficient of $R_{\text{cont}}(\mathbf{f})$.

(b) Upper bound at $R_{\text{cont}}(\mathbf{f}) = R_{\text{cont}}^*$.

Figure 4: Theoretical comparison of surrogate bounds ($C = 10$), log-scaled. Arora et al.'s and Nozawa & Sato's bounds are valid only at $K + 1 \geq C$ (the dotted vertical lines).

The value v_K is often referred to as the *coupon collector's probability*. Let τ_K be the probability that at least one of the negative classes c_k^- is the same as the positive class c^+. Under the uniform class prior, $\tau_K = 1 - (1 - 1/C)^K$. The surrogate bounds are summarized in Table 1.[4]

We discuss the applicability and the dependence of existing and our bounds with respect to K. As summarized in Table 1, the coefficients of $R_{\text{cont}}(\mathbf{f})$ for existing bounds strongly depend on C and K through the *coverage* (v_K)

---

[4]More precisely, Arora et al. (2019) bound the averaged supervised loss over a part of the latent classes rather than $R_{\mu\text{-supv}}$. Thus we can obtain a slightly better upper bound than Arora et al.'s bound shown in Table 1. Nevertheless, the scale of the upper bound is dominated by the coefficient $(1 - \tau_K)v_{K+1}$.

Table 1: Surrogate bounds of the existing works. H_n denotes the n-th harmonic number. Remark that Arora et al.'s and Nozawa & Sato's bounds are valid only $K + 1 \geq C$. The detailed derivations are discussed in Appendix E.

	UPPER BOUND	REFERENCE
$R_{\mu\text{-supv}}(\mathbf{f}) \leq$	$\frac{1}{(1-\tau_K)v_{K+1}}\left\{R_{\text{cont}}(\mathbf{f}) - \mathbb{E}\ln(\text{Col}+1)\right\}$	Arora et al. (2019)
	$\frac{1}{v_{K+1}}\left\{2R_{\text{cont}}(\mathbf{f}) - \mathbb{E}\ln(\text{Col}+1)\right\}$	Nozawa & Sato (2021)
	$\frac{2}{1-\tau_K}\left\lceil\frac{2(C-1)H_{C-1}}{K}\right\rceil\left\{R_{\text{cont}}(\mathbf{f}) - \mathbb{E}\ln(\text{Col}+1)\right\}$	Ash et al. (2022)

and the *collision* (τ_K) probability whereas our bounds have the constant coefficient. As Arora et al.'s and Nozawa & Sato's bounds depend on the *coverage* probability v_K in the denominator of the coefficients, the coefficients diverge in the range of $K + 1 < C$ (where the negative sample size is insufficient to cover the entire $[C]$). In addition, the coefficients of the Arora et al.'s and Ash et al.'s bounds increase exponentially with increasing K due to the *collision* probability τ_K in the denominator, which are not consistent with the experimental facts. Compared to these bounds, our upper bound has the coefficient independent of C and K.

We numerically demonstrated the abovementioned dependencies on K in Figure 4. As we can see in Figure 4a, the coefficients of $R_{\text{cont}}(\mathbf{f})$ of Arora et al.'s and Ash et al.'s bounds have unique minima, Nozawa & Sato's coefficient has monotonically decreasing nature, and our coefficient is constant. On the other hand, the tendencies of the bound values at $R_{\text{cont}}(\mathbf{f}) = R^*_{\text{cont}}$, namely, the best possible mean supervised loss in terms of the upper bounds (Figure 4b) are slightly different from the coefficient: Ash et al.'s bound is monotonically increasing, Arora et al.'s and Nozawa & Sato's bounds have a unique minimum, and ours is monotonically decreasing. Among the compared bounds, only ours is legitimate for all K and moderately decreases with K, which agrees well with the experimental fact observed as well in Figure 1; the details are stated in Section 5.2.[5] Such a moderate dependence on K is due to the mechanism that the contrastive loss behaves as a surrogate objective.

4.1. Related literatures.

Wang & Isola (2020) showed that the contrastive loss asymptotically favors data representation uniformly distributed over the unit sphere yet aligning across semantically similar samples. Li et al. (2021) proposed an alternative loss function to the contrastive loss based on a kernel metric, following the similar idea to Wang & Isola (2020). Tosh et al. (2021) showed that a (linear) mean classifier learned in CURL can approximate the (potentially nonlinear) Bayes

classifier well.

More recently, Awasthi et al. (2022) showed the existence of representations that well separate each class and optimize the contrastive loss regardless of K, under the assumptions of the conditional independence and the disjoint class-conditional distributions. Their claim aligns well with our claim that larger K is not harmful. While our analysis focuses on the relationship between the contrastive and (mean) supervised losses, they investigated the property of representations that minimize the contrastive loss.

While our work does not handle DA, several works analyzed the effect of DA on the performance. Wen & Li (2021) showed that DA is necessary to recover sparse signals under a specific assumption on the model architecture. HaoChen et al. (2021) introduced a notion of the augmentation graph, representing how likely the nearby samples are generated via DA and showed that a type of contrastive loss could be viewed as a low-rank approximation of the adjacency matrix of the augmentation graph. von Kügelgen et al. (2021) proposed a loss function that enables the model to identify invariant factors across DA.

We mention a few works analyzing the other types of self-supervised learning; Garg & Liang (2020) analyzed masked self-supervised learning, Wei et al. (2021) analyzed the input consistency loss for unsupervised learning, and Saunshi et al. (2021) analyzed auto-regressive language models. Grill et al. (2020); Chen & He (2021) proposed self-supervised learning without negative samples.

Lastly, multi-sample estimators (van den Oord et al., 2018; Poole et al., 2019; Song & Ermon, 2020) popularly used in mutual information estimation are substantially related to the contrastive loss. We defer its discussion to Appendix F.

Remark. A concurrent work (Wang et al., 2022) recently established the surrogate bound that has similar order in K with ours without conditional independence assumption. We stress that our results were obtained independently of theirs. In addition, the purpose of our research is to clarify the mechanism of how K affects the downstream performance, which is different from their motivation to discuss the validity of assumptions in contrastive learning. In Ap-

---

[5]Note that Nozawa & Sato (2021)'s bound also implies larger K is better. Still, our argument on how contrastive learning works differs from theirs. See Appendix E for the further discussions.

pendix C, we discuss how our surrogate bounds hold without the conditional independence assumption.

5. Experiments

We verified our theoretical findings with experiments on synthetic (Section 5.1), vision, and language datasets (Section 5.2). The details of the setup are in Appendix G. The experimental codes to reproduce all figures in the paper are available at `https://github.com/nzw0301/gap-contrastive-and-supervised-losses`.

5.1. Small-scale Experiments on Synthetic Dataset

Dataset and learning setups. We create a synthetic dataset `circle`, which is a 2D dataset created as follows: for each class $c \in [C]$ ($C = 10$), $1\,000$ samples are drawn from $\mathsf{Uniform}([-0.5 \ -0.5], [0.5, 0.5])$, normalized, and multiplied by $c+1/2$. The generated samples are nonlinear and require disentanglement to be linearly separable. We treated 60% of the generated samples as a training dataset and the rest of the samples as a test dataset.

As a feature extractor $\mathbf{f}$, we used a multi-layer perceptron (the number of units 2-256-256-256) with the ReLU activation functions following after each hidden layer. During the training, the extracted feature representations are normalized. For negative samples, we sampled $K \in \{1, 4, 16, 64, 256\}$ samples without replacement from $2B - 2$ points included in the same mini-batch to avoid the influence of mini-batch size B, inspired by Ash et al. (2022).[6]

Results. Figure 5 shows a single trajectory in the $(R_{\mathrm{cont}}, R_{\mu\text{-supv}})$-plot and the feasible region (confer Figure 2) for each K. We plotted the trajectories by tracking $(R_{\mathrm{cont}}(\mathbf{f}^{(t)}), R_{\mu\text{-supv}}(\mathbf{f}^{(t)}))$ at each epoch t computed with the test dataset. All trajectories were located in between the upper ($R_{\mathrm{cont}} + \Delta_{\mathrm{U}}$) and lower ($R_{\mathrm{cont}} + \Delta_{\mathrm{L}}$) bounds as a matter of course. Given that the existing surrogate bounds provide the much larger upper bounds (Figure 4), our surrogate bounds provide the finest estimate of the mean supervised loss. In addition, it is remarkable that all trajectories have nearly the same slopes as our surrogate bounds, which constitutes solid evidence that our surrogate bounds capture the learning dynamics well.

In Figure 6, the mean supervised loss and accuracy are compared with the different K. We plotted the standard deviations of the same experiments with eight different random seeds for each K. From these figures, it can be concluded that the contrastive loss performance becomes better with the larger K in the sense that the supervised loss improved

---

[6]Each mini-batch consists of B pairs of positive pairs. The candidates of the negative samples are the $2B - 2$ samples excluding the anchor and its paired point.

and the variance shrank. The variance improvement is theoretically suggested by Figure 5 as well; the larger K is, the smaller the gap between upper and lower bounds becomes.

5.2. Large-scale Experiments on Vision and Language Datasets

We used the same datasets as Arora et al. (2019): CIFAR-100 (Krizhevsky, 2009) and Wiki-3029 (Arora et al., 2019) datasets, along with CIFAR-10 (Krizhevsky, 2009) dataset.

Learning setups. We treated the supervised classes as latent classes as in Arora et al. (2019) and Ash et al. (2022) for creating positive pairs. We used the original supervised classes of CIFAR-10/100 as $[C]$; $C = 10$ and $C = 100$, respectively. We used K in $\{4, 16, 32, 64, 128, 512\}$ and in $\{4, 64, 128, 512\}$ for CIFAR-10/100, respectively. For Wiki-3029, we used $C \in \{500, 1\,000, 2\,000, 3\,029\}$ and $K \in \{8, 64, 256, 1\,024\}$. For each different (C, K), we trained the feature extractor $\mathbf{f}$ on the training dataset. We then evaluated its performance on the test dataset with mean and linear classifiers. We used ResNet-18 (He et al., 2016)-based feature extractor $\mathbf{f}$ for CIFAR-10/100 and the fasttext (Joulin et al., 2017)-based feature extractor for Wiki-3029.

Results. Figure 1 shows the comparison between the estimated upper bounds using Theorem 1 and actual supervised loss on the CIFAR-10 test dataset. We estimated the bounds by substituting the actual $R_{\mathrm{cont}}(\mathbf{f})$ to the equations shown in Theorem 1 and Table 1. Our bound gave the closest bound to the experimental value of the supervised loss. The existing surrogate bounds of Arora et al. (2019) and Ash et al. (2022) were prohibitively large to explain the classification performance. Although Nozawa & Sato's bound was comparable with ours, it was valid only in $K + 1 \geq C$ and tended to diverge near $K + 1 = C$, as shown in Section 4.

We investigated how K affects the test accuracy for different C in Figure 7. The test accuracy improved or was saturated with the larger K for all C on Wiki-3029. In contrast, it was degraded as K increased in mean and linear classifiers on CIFAR-10/100. This behavior could be partly because of the gap between the cross-entropy loss and the supervised accuracy—the theory of CURL, including the existing studies, usually focuses on the cross-entropy loss only. Figure 9 in Appendix G.6 revealed that the supervised loss was not significantly worse with the larger K on CIFAR-10/100.

With the smaller K and large C, we found that long epochs were more effective to improve classification accuracy than increasing the negative sample size K (Figure 8b). While similar results were reported by Chen et al. (2020, Figure 9), it is important to remark that we randomly drew K negative samples from the $2B - 2$ samples in the given mini-batch at each iteration as in Ash et al. (2022)—a different approach
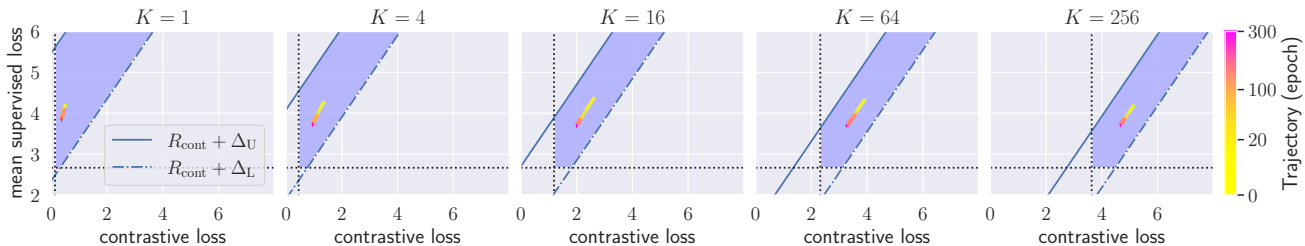
Figure 5: Learning trajectories of the `circle` dataset in the $(R_{\text{cont}}, R_{\mu\text{-supv}})$-plot. The trajectories are plotted with gradient color lines, indicating the epochs.
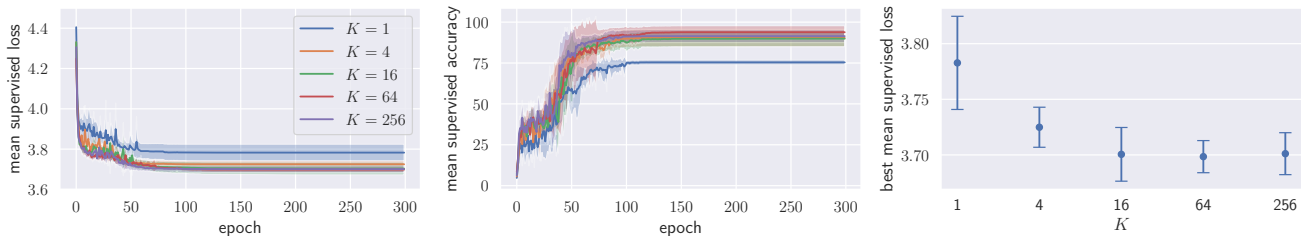

Figure 6: For each K, eight runs on the `circle` dataset are averaged with the standard deviations plotted. **(Left)** the test mean supervised losses at each epoch with the different negative sample sizes K. **(Middle)** the test mean supervised accuracy at each epoch with the different negative sample sizes K. **(Right)** the best test mean supervised loss with the different negative sample sizes K.

was used by Chen et al. (2020) to regard the all samples in the mini-batch except an anchor sample as negative samples. Under our experimental setup, a learner may encounter less diverse samples with the smaller K even if the mini-batch size B is the same, which could make the downstream performance worse—the longer epochs are necessary to mitigate the issue. Since the CIFAR-10 dataset has a smaller C and is simpler than the CIFAR-100, all accuracies were saturated with similar epochs for all K (Figure 8a).

6. Conclusion

We established novel surrogate bounds for contrastive learning. In contrast to existing theories, our bounds are applicable for all negative sample sizes and have a constant coefficient. We verified that our bounds well explained learning dynamics on the synthetic dataset, and the surrogate gap shrinks with large negative samples. For the vision and language datasets, the downstream classification losses were also best explained by our bounds in contrast to existing theories. Our bounds provided a perspective on the effect of negative sample size that the contrastive loss behaves as a surrogate objective of the downstream loss, and its surrogate gap decays with larger negative samples.

References

Arora, S., Khandeparkar, H., Khodak, M., Plevrakis, O., and Saunshi, N. A Theoretical Analysis of Contrastive Unsupervised Representation Learning. In *ICML*, pp. 5628–5637, 2019.

Ash, J. T., Goel, S., Krishnamurthy, A., and Misra, D. Investigating the Role of Negatives in Contrastive Representation Learning. In *AISTATS*, pp. 7187–7209, 2022.

Awasthi, P., Dikkala, N., and Kamath, P. Do More Negative Samples Necessarily Hurt in Contrastive Learning? *arXiv:2205.01789v1 [cs.LG]*, 2022.

Bao, H., Niu, G., and Sugiyama, M. Classification from Pairwise Similarity and Unlabeled Data. In *ICML*, pp. 461–470, 2018.

Bellet, A., Habrard, A., and Sebban, M. Similarity Learning for Provably Accurate Sparse Linear Classification. In *ICML*, pp. 1491–1498, 2012.

Bengio, Y., Courville, A., and Vincent, P. Representation Learning: A Review and New Perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35 (8):1798–1828, 2013.

mean classifier

C	$K=8$	$K=64$	$K=256$	$K=1024$
3029	37.86 (0.19)	40.36 (0.13)	40.85 (0.09)	40.94 (0.13)
2000	42.09 (0.17)	44.06 (0.16)	44.38 (0.09)	44.24 (0.23)
1000	48.07 (0.12)	48.77 (0.23)	48.75 (0.05)	48.66 (0.10)
500	52.66 (0.71)	52.66 (0.69)	53.04 (0.67)	53.72 (0.63)

linear classifier

C	$K=8$	$K=64$	$K=256$	$K=1024$
3029	41.70 (0.14)	43.00 (0.15)	43.26 (0.14)	43.30 (0.21)
2000	45.13 (0.29)	46.33 (0.14)	46.46 (0.15)	46.37 (0.16)
1000	50.86 (0.25)	51.21 (0.33)	51.06 (0.24)	50.94 (0.23)
500	55.41 (0.39)	55.41 (0.43)	55.52 (0.51)	55.69 (0.47)

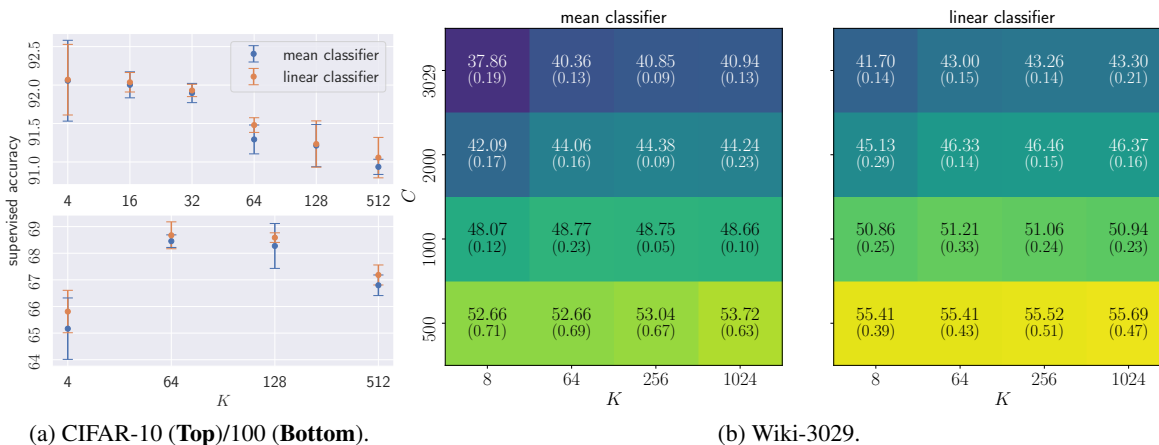(a) CIFAR-10 (**Top**)/100 (**Bottom**). (b) Wiki-3029.

Figure 7: Mean and linear classifier's test accuracy on CIFAR-10/100 and Wiki-3029 when varying the negative samples size K. For Wiki-3029, we also change the number of latent classes C. The error bars in (**a**) and parenthesized number in (**b**) indicate the standard deviation of three runs.
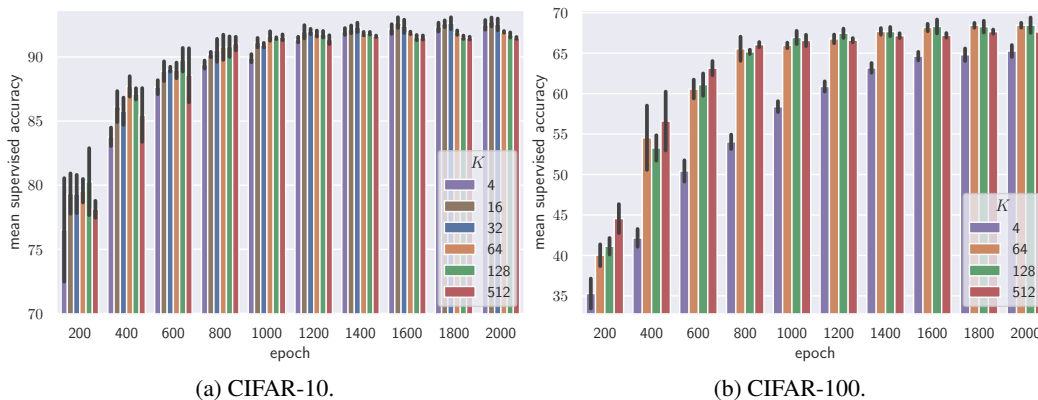

(a) CIFAR-10. (b) CIFAR-100.

Figure 8: Test accuracy of mean classifier at every 200 epochs on CIFAR-10/100. In CIFAR-100, the accuracy of $K=4$ and the others' have a large gap at smaller epochs at epoch 200, but the gap become smaller when epochs increase. The error bars indicate the standard deviation of three runs.

Biewald, L. Experiment Tracking with Weights and Biases, 2020. URL https://www.wandb.com/.

Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., and Joulin, A. Unsupervised Learning of Visual Features by Contrasting Cluster Assignments. In *NeurIPS*, pp. 9912–9924, 2020.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A Simple Framework for Contrastive Learning of Visual Representations. In *ICML*, pp. 1597–1607, 2020.

Chen, T., Luo, C., and Li, L. Intriguing Properties of Contrastive Losses. In *NeurIPS*, pp. 11834–11845, 2021.

Chen, X. and He, K. Exploring Simple Siamese Representation Learning. In *CVPR*, pp. 15750–15758, 2021.

Chopra, S., Hadsell, R., and LeCun, Y. Learning a Similarity Metric Discriminatively, with Application to Face Verification. In *CVPR*, pp. 539–546, 2005.

Chuang, C.-Y., Hjelm, R. D., Wang, X., Vinnet, V., Joshi, N., Torralba, A., Jegelka, S., and Song, Y. Robust Contrastive Learning against Noisy Views. *arXiv:2201.04309v1 [cs.CV]*, 2022.

Gao, S., Ver Steeg, G., and Galstyan, A. Efficient Estimation of Mutual Information for Strongly Dependent Variables. In *AISTATS*, pp. 277–286, 2015.

Garg, S. and Liang, Y. Functional Regularization for Representation Learning: A Unified Theoretical Perspective. In *NeurIPS*, pp. 17187–17199, 2020.

Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P. H., Buchatskaya, E., Doersch, C., Pires, B. A., Guo, Z. D., Azar, M. G., et al. Bootstrap Your Own Latent: A

New Approach to Self-supervised Learning. In *NeurIPS*, pp. 21271–21284, 2020.

Guo, Q., Chen, J., Wang, D., Yang, Y., Deng, X., Carin, L., Li, F., and Tao, C. Tight Mutual Information Estimation With Contrastive Fenchel-Legendre Optimization. *arXiv:2107.01131v1 [stat.ML]*, 2021.

HaoChen, J. Z., Wei, C., Gaidon, A., and Ma, T. Provable Guarantees for Self-Supervised Deep Learning with Spectral Contrastive Loss. In *NeurIPS*, pp. 5000–5011, 2021.

He, K., Zhang, X., Ren, S., and Sun, J. Deep Residual Learning for Image Recognition. In *CVPR*, pp. 770–778, 2016.

He, K., Girshick, R., and Dollár, P. Rethinking ImageNet Pre-training. In *CVPR*, pp. 14918–14927, 2019.

He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. Momentum Contrast for Unsupervised Visual Representation Learning. In *CVPR*, pp. 9726–9735, 2020.

Hunter, J. D. Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering*, 9(3):90–95, 2007. doi: 10.1109/MCSE.2007.55.

Joulin, A., Grave, E., Bojanowski, P., and Mikolov, T. Bag of Tricks for Efficient Text Classification. In *EACL*, volume 2, pp. 427–431, 2017.

Kingma, D. P. and Ba, J. Adam: A Method for Stochastic Optimization. In *ICLR*, 2015.

Krizhevsky, A. Learning Multiple Layers of Features from Tiny Images. Technical report, 2009.

Kulis, B. Metric Learning: A Survey. *Foundations and Trends® in Machine Learning*, 5(4):287–364, 2012.

Li, S., Zhao, Y., Varma, R., Salpekar, O., Noordhuis, P., Li, T., Paszke, A., Smith, J., Vaughan, B., Damania, P., and Chintala, S. PyTorch Distributed: Experiences on Accelerating Data Parallel Training. In *VLDB*, pp. 3005–3018, 2020.

Li, Y., Pogodin, R., Sutherland, D. J., and Gretton, A. Self-Supervised Learning with Kernel Dependence Maximization. In *NeurIPS*, pp. 15543–15556, 2021.

Lirong, W., Lin, H., Gao, Z., Tan, C., and Li, S. Z. Self-supervised on Graphs: Contrastive, Generative, or Predictive. *arXiv:2105.07342v3 [cs.LG]*, 2021.

Logeswaran, L. and Lee, H. An Efficient Framework for Learning Sentence Representations. In *ICLR*, 2018.

Loshchilov, I. and Hutter, F. SGDR: Stochastic Gradient Descent with Warm Restarts. In *ICLR*, 2017.

McAllester, D. and Stratos, K. Formal Limitations on the Measurement of Mutual Information. In *AISTATS*, pp. 875–884, 2020.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. Distributed Representations of Words and Phrases and their Compositionality. In *NeurIPS*, pp. 3111–3119, 2013.

Nozawa, K. and Sato, I. Understanding Negative Samples in Instance Discriminative Self-supervised Representation Learning. In *NeurIPS*, pp. 5784–5797, 2021.

Nozawa, K., Germain, P., and Guedj, B. PAC-Bayesian Contrastive Unsupervised Representation Learning. In *UAI*, pp. 21–30, 2020.

Pardalos, P. M. and Rosen, J. B. Methods for Global Concave Minimization: A Bibliographic Survey. *Siam Review*, 28(3):367–379, 1986.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *NeurIPS*, pp. 8024–8035, 2019.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Édouard Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12(85):2825–2830, 2011.

Pennington, J., Socher, R., and Manning, C. D. GloVe: Global Vectors for Word Representation. In *EMNLP*, pp. 1532–1543, 2014.

Poole, B., Ozair, S., van den Oord, A., Alemi, A., and Tucker, G. On Variational Bounds of Mutual Information. In *ICML*, pp. 5171–5180, 2019.

Reback, J., McKinney, W., jbrockmendel, den Bossche, J. V., Augspurger, T., Cloud, P., gfyoung, Sinhrks, Klein, A., Roeschke, M., Hawkins, S., Tratner, J., She, C., Ayd, W., Petersen, T., Garcia, M., Schendel, J., Hayden, A., MomIsBestFriend, Jancauskas, V., Battiston, P., Seabold, S., chris-b1, h-vetinari, Hoyer, S., Overmeire, W., alimcmaster1, Dong, K., Whelan, C., and Mehyar, M. pandas-dev/pandas: Pandas 1.0.3, March 2020. URL https://doi.org/10.5281/zenodo.3715232.

Saunshi, N., Malladi, S., and Arora, S. A Mathematical Exploration of Why Language Models Help Solve Downstream Tasks. In *ICLR*, 2021.

Song, J. and Ermon, S. Understanding the Limitations of Variational Mutual Information Estimators. In *ICLR*, 2020.

Tange, O. GNU Parallel 20210922 ('Vindelev'), September 2021. URL https://doi.org/10.5281/zenodo.5523272.

Tian, Y., Sun, C., Poole, B., Krishnan, D., Schmid, C., and Isola, P. What Makes for Good Views for Contrastive Learning? In *NeurIPS*, volume 33, pp. 6827–6839, 2020.

Tomasev, N., Bica, I., McWilliams, B., Buesing, L., Pascanu, R., Blundell, C., and Mitrovic, J. Pushing the Limits of Self-supervised ResNets: Can we Outperform Supervised Learning without Labels on ImageNet? *arXiv:2201.05119v1 [cs.CV]*, 2022.

Tosh, C., Krishnamurthy, A., and Hsu, D. Contrastive Learning, Multi-view Redundancy, and Linear Models. In *ALT*, pp. 1179–1206, 2021.

Tschannen, M., Djolonga, J., Rubenstein, P. K., Gelly, S., and Lucic, M. On Mutual Information Maximization for Representation Learning. In *ICLR*, 2020.

van den Oord, A., Li, Y., and Vinyals, O. Representation Learning with Contrastive Predictive Coding. *arXiv:1807.03748v2 [cs.LG]*, 2018.

von Kügelgen, J., Sharma, Y., Gresele, L., Brendel, W., Schölkopf, B., Besserve, M., and Locatello, F. Self-Supervised Learning with Data Augmentations Provably Isolates Content from Style. In *NeurIPS*, pp. 16451–16467, 2021.

Wang, T. and Isola, P. Understanding Contrastive Representation Learning through Alignment and Uniformity on the Hypersphere. In *ICML*, pp. 9929–9939, 2020.

Wang, Y., Zhang, Q., Wang, Y., Jiansheng, Y., and Lin, Z. Chaos is a Ladder: A New Understanding of Contrastive Learning via Augmentation Overlap. In *ICLR*, 2022.

Waskom, M. L. seaborn: Statistical Data Visualization. *Journal of Open Source Software*, 6(60):3021, 2021. doi: 10.21105/joss.03021.

Wei, C., Shen, K., Chen, Y., and Ma, T. Theoretical Analysis of Self-Training with Deep Networks on Unlabeled Data. In *ICLR*, 2021.

Wen, Z. and Li, Y. Toward Understanding the Feature Learning Process of Self-supervised Contrastive Learning. In *ICML*, pp. 11112–11122, 2021.

Wu, Y. and Johnson, J. Rethinking "Batch" in BatchNorm. *arXiv:2105.07576v1 [cs.CV]*, 2021.

Yadan, O. Hydra - A Framework for Elegantly Configuring Complex Applications. GitHub, 2019. URL https://github.com/facebookresearch/hydra.

You, Y., Gitman, I., and Ginsburg, B. Large Batch Training of Convolutional Networks. *arXiv:1708.03888v3 [cs.CV]*, 2017.

«APPENDIX»

ON THE SURROGATE GAP BETWEEN CONTRASTIVE AND SUPERVISED LOSSES

As an additional notation, the d-dimensional ball of radius r associated with the $\mathbb{L}_p$-norm is denoted by $\mathbb{B}_p^d(r) := \{\mathbf{x} \in \mathbb{R}^d \mid \|\mathbf{x}\|_p \leq r\}$. For $\mathbf{z} \in \mathbb{R}^C$, the log-sum-exp function is denoted by $\mathrm{LSE}\,(\mathbf{z}) := \ln(\sum_{c \in [C]} \exp(z_c))$.

A. Useful Lemmas

In this section, a few lemmas are introduced in order to prove the main results.

Lemma 3. *For $\mathbf{z} \in [-L^2, L^2]^N$,*

$$2 \ln N \leq \mathrm{LSE}\,(\mathbf{z}) + \mathrm{LSE}\,(-\mathbf{z}) \leq 2 \ln(N \cosh(L^2)). \tag{10}$$

Proof. Define $H(\mathbf{z}) := \mathrm{LSE}\,(\mathbf{z}) + \mathrm{LSE}\,(-\mathbf{z})$. First, we prove the lower bound of $H(\mathbf{z})$. Since

$$\frac{\partial H}{\partial z_i} = \frac{\exp(z_i)}{\sum_{n \in [N]} \exp(z_n)} - \frac{\exp(-z_i)}{\sum_{n \in [N]} \exp(-z_n)} \tag{11}$$

for all $i \in [N]$, $\mathbf{z} = \mathbf{0}_N$ satisfies the first-order optimality condition of H. By noting that H is convex due to the convexity of the log-sum-exp functions, H is minimized at $\mathbf{z} = \mathbf{0}_N$: $H(\mathbf{z}) \geq H(\mathbf{0}_N) = 2 \ln N$. Note that $\mathbf{z}$ can be any vector in $\mathbb{R}^N$ for this lower bound.

Next, we prove the upper bound of $H(\mathbf{z})$. Observe that finding the maximum of $H(\mathbf{z})$ in $\mathbf{z} \in \mathbb{B}_\infty^N(L^2)$ is equivalent to a concave minimization problem over a convex polytope. It is known that every vertex of the polytope $\mathbf{z} \in \{-L^2, L^2\}^N$ is a local optimum for concave minimization over a convex polytope (Pardalos & Rosen, 1986). Hence, it is sufficient to test the vertices $\mathbf{z} \in \{-L^2, L^2\}^N$ to find the maximum of $H(\mathbf{z})$. Define

$$\tilde{H}(m) := H\Big(\underbrace{L^2, L^2, \ldots, L^2}_{m}, \underbrace{-L^2, -L^2, \ldots, -L^2}_{N-m}\Big) \tag{12}$$

$$= \ln\Big\{m \exp(L^2) + (N - m) \exp(-L^2)\Big\} + \ln\Big\{m \exp(-L^2) + (N - m) \exp(L^2)\Big\}. \tag{13}$$

Note that the maximizer of $H(\mathbf{z})$ in $\mathbf{z} \in \{-L^2, L^2\}^N$ is equivalent to that of $\tilde{H}(m)$ in $m \in [N]$ because $H(\mathbf{z})$ is symmetric in every z_n for $n \in [N]$. We verify the following by simple algebra:

$$\exp \tilde{H}(m) = -\big(\exp(L^2) + \exp(-L^2)\big)\left\{\Big(m - \frac{N}{2}\Big)^2 + \mathrm{Const}\right\}, \tag{14}$$

meaning that $\tilde{H}(m)$ is maximized at $m = \lfloor \frac{N}{2} \rfloor$. Hence, $H(\mathbf{z}) \leq \tilde{H}(\lfloor N/2 \rfloor) \leq \tilde{H}(N/2) = 2 \ln(N \cosh(L^2))$. $\qquad \square$

Lemma 4. *For all $z_0 \in \mathbb{R}$ and $\mathbf{z} \in \mathbb{R}^K$ such that $z_0, z_k \in [-L^2, L^2]$ ($\forall k \in [K]$),*

$$\ln \frac{\exp(-z_0)}{\exp(-z_0) + \sum_{k \in [K]} \exp(-z_k)}$$

$$\geq -\ln \frac{\exp(z_0)}{\exp(z_0) + \sum_{k \in [K]} \exp(z_k)} - 2 \ln \big\{(K + 1) \cosh(L^2)\big\}. \tag{15}$$

Proof. We write $\tilde{\mathbf{z}} := [z_0 \ \mathbf{z}^\top]^\top$. Let $H(\tilde{\mathbf{z}})$ be a function such that

$$H(\tilde{\mathbf{z}}) := -\ln \frac{\exp(z_0)}{\exp(z_0) + \sum_{k \in [K]} \exp(z_k)} - \ln \frac{\exp(-z_0)}{\exp(-z_0) + \sum_{k \in [K]} \exp(-z_k)} \tag{16}$$

$$= \ln \sum_{k=1}^{K+1} \exp(\tilde{z}_k) + \ln \sum_{k=1}^{K+1} \exp(-\tilde{z}_k). \tag{17}$$

Our goal is to find a tight upper bound of $H(\tilde{\mathbf{z}})$ for $\tilde{\mathbf{z}} \in \mathbb{B}_\infty^{K+1}(L^2)$.

Observe that $H(\tilde{\mathbf{z}})$ is the sum of the two log-sum-exp functions hence it is convex in $\tilde{\mathbf{z}}$. In addition, the domain $\mathbb{B}_\infty^{K+1}(L^2)$ is a compact convex polytope. Henceforth, every vertex of the polytope, $\tilde{\mathbf{z}} \in \{-L^2, L^2\}^{K+1}$, is a local maximizer because maximizing $H(\tilde{\mathbf{z}})$ is concave minimization over a convex polytope (Pardalos & Rosen, 1986). Since $H(\tilde{\mathbf{z}})$ is symmetric in every element $\tilde{z}_k$, it is sufficient to test the vertices and see the difference between

$$\underbrace{H\left((\underbrace{L^2, \ldots, L^2}_{\# = j}, -L^2, \ldots, -L^2)\right)}_{:=\tilde{H}(j)} \text{ and } \underbrace{H\left((\underbrace{L^2, \ldots, L^2}_{\# = j+1}, -L^2, \ldots, -L^2)\right)}_{:=\tilde{H}(j+1)} \tag{18}$$

for $j \in \{0, \ldots, K\}$ to seek out the global maximum. For $0 \leq j \leq K$, a simple algebra shows

$$\exp\left(\tilde{H}(j)\right) - \exp\left(\tilde{H}(j+1)\right) = (K - 2j) \underbrace{\left\{2 - \left(\exp(2L^2) + \exp(-2L^2)\right)\right\}}_{\substack{\leq 0 \\ \text{because of AM-GM inequality}}}, \tag{19}$$

from which we can tell that $\tilde{H}(j)$ is maximized at $j = K/2$ when K is even and $j = (K+1)/2$ when K is odd. In addition, it is confirmed that

$$\exp\left(\tilde{H}\left(\frac{K}{2}\right)\right) - \exp\left(\tilde{H}\left(\frac{K+1}{2}\right)\right) = \frac{2 - \left(\exp(2L^2) + \exp(-2L^2)\right)}{4} \tag{20}$$

$$\leq 0, \tag{21}$$

where the AM-GM inequality is invoked at the last line. Eventually, $\tilde{H}((K+1)/2)$ turns out to be a tight upper bound of $H(\tilde{\mathbf{z}})$ for $\tilde{\mathbf{z}} \in \mathbb{B}_\infty^{K+1}(L^2)$. It is elementary to confirm $\tilde{H}((K+1)/2) = 2\ln\left\{(K+1)\cosh(L^2)\right\}$. $\qquad\square$

B. Proofs of Main Results

In this section, we provide proofs for the main results, Theorems 1 and 2.

Theorem 1. *For all* $\mathbf{f}$ *such that* $\|\mathbf{f}(\mathbf{x})\|_2 \leq L$ *(*$\forall \mathbf{x} \in \mathcal{X}$*), the following inequality holds.*

$$R_{\mu\text{-supv}}(\mathbf{f}) \leq R_{\text{cont}}(\mathbf{f}) + \Delta_{\text{U}}, \tag{3}$$

where $\Delta_{\text{U}} := \ln\{\pi_{(1)} K^{-1} C^2 \cosh^2(L^2)\}$.

Proof of Theorem 1. The proof largely relies on the Jensen's inequality. To apply the Jensen's inequality in the *reversed* way, we occasionally transform a convex function into a concave function by applying the log-sum-exp bound in Lemma 3.

$$R_{\text{cont}}(\mathbf{f}) = -\mathop{\mathbb{E}}_{c^+,\{c_k^-\},\mathbf{x},\mathbf{x}^+,\{\mathbf{x}_k^-\}} \ln \frac{\exp(\mathbf{f}(\mathbf{x})^\top \mathbf{f}(\mathbf{x}^+))}{\exp(\mathbf{f}(\mathbf{x})^\top \mathbf{f}(\mathbf{x}^+)) + \sum_{k\in[K]} \exp(\mathbf{f}(\mathbf{x})^\top \mathbf{f}(\mathbf{x}_k^-))} \tag{22}$$

$$= -\mathop{\mathbb{E}}_{c^+,\mathbf{x},\mathbf{x}^+}[\mathbf{f}(\mathbf{x})^\top \mathbf{f}(\mathbf{x}^+)] + \mathop{\mathbb{E}}_{c^+,\{c_k^-\},\mathbf{x},\mathbf{x}^+,\{\mathbf{x}_k^-\}} \ln\left(\exp(\mathbf{f}(\mathbf{x})^\top \mathbf{f}(\mathbf{x}^+)) + \sum_{k\in[K]} \exp(\mathbf{f}(\mathbf{x})^\top \mathbf{f}(\mathbf{x}_k^-))\right) \tag{23}$$

$$\geq -\mathop{\mathbb{E}}_{c^+,\mathbf{x},\mathbf{x}^+}[\mathbf{f}(\mathbf{x})^\top \mathbf{f}(\mathbf{x}^+)] + \mathop{\mathbb{E}}_{c^+,\{c_k^-\},\mathbf{x},\mathbf{x}^+,\{\mathbf{x}_k^-\}} \underbrace{\ln \sum_{k\in[K]} \exp(\mathbf{f}(\mathbf{x})^\top \mathbf{f}(\mathbf{x}_k^-))}_{=\text{LSE}\left(\{\mathbf{f}(\mathbf{x})^\top \mathbf{f}(\mathbf{x}_k^-)\}_{k\in[K]}\right)} \quad \text{(monotonicity of ln)} \tag{24}$$

$$\geq -\mathop{\mathbb{E}}_{c^+,\mathbf{x},\mathbf{x}^+}[\mathbf{f}(\mathbf{x})^\top \mathbf{f}(\mathbf{x}^+)] + \mathop{\mathbb{E}}_{c^+,\{c_k^-\},\mathbf{x}} \underbrace{\ln \sum_{k\in[K]} \exp(\mathbf{f}(\mathbf{x})^\top \boldsymbol{\mu}_{c_k^-})}_{=\text{LSE}\left(\{\mathbf{f}(\mathbf{x})^\top \boldsymbol{\mu}_{c_k^-}\}_{k\in[K]}\right)} \quad \text{(Jensen's inequality)} \tag{25}$$

$$\geq - \mathop{\mathbb{E}}_{c^+,\mathbf{x},\mathbf{x}^+}[\mathbf{f}(\mathbf{x})^\top\mathbf{f}(\mathbf{x}^+)] - \mathop{\mathbb{E}}_{c^+,\{c_k^-\},\mathbf{x}}\underbrace{\ln\sum_{k\in[K]}\exp(-\mathbf{f}(\mathbf{x})^\top\boldsymbol{\mu}_{c_k^-})}_{=\mathrm{LSE}\left(\{-\mathbf{f}(\mathbf{x})^\top\boldsymbol{\mu}_{c_k^-}\}_{k\in[K]}\right)}+2\ln K \qquad \text{(Lemma 3)} \tag{26}$$

$$\geq - \mathop{\mathbb{E}}_{c^+,\mathbf{x},\mathbf{x}^+}[\mathbf{f}(\mathbf{x})^\top\mathbf{f}(\mathbf{x}^+)] - \mathop{\mathbb{E}}_{c^+,\mathbf{x}}\ln\sum_{k\in[K]}\mathop{\mathbb{E}}_{\{c_k^-\}}\left[\exp(-\mathbf{f}(\mathbf{x})^\top\boldsymbol{\mu}_{c_k^-})\right] + 2\ln K \qquad \text{(Jensen's inequality)} \tag{27}$$

$$\overset{(a)}{\geq} - \mathop{\mathbb{E}}_{c^+,\mathbf{x},\mathbf{x}^+}[\mathbf{f}(\mathbf{x})^\top\mathbf{f}(\mathbf{x}^+)] - \mathop{\mathbb{E}}_{c^+,\mathbf{x}}\ln\left(K\sum_{c\in[C]}\pi_{(1)}\exp(-\mathbf{f}(\mathbf{x})^\top\boldsymbol{\mu}_c)\right) + 2\ln K \tag{28}$$

$$= - \mathop{\mathbb{E}}_{c^+,\mathbf{x},\mathbf{x}^+}[\mathbf{f}(\mathbf{x})^\top\mathbf{f}(\mathbf{x}^+)] - \mathop{\mathbb{E}}_{c^+,\mathbf{x}}\underbrace{\ln\sum_{c\in[C]}\exp(-\mathbf{f}(\mathbf{x})^\top\boldsymbol{\mu}_c)}_{=\mathrm{LSE}\left(\{-\mathbf{f}(\mathbf{x})^\top\boldsymbol{\mu}_c\}_{c\in[C]}\right)}+2\ln K - \ln K\pi_{(1)} \tag{29}$$

$$\geq - \mathop{\mathbb{E}}_{c^+,\mathbf{x},\mathbf{x}^+}[\mathbf{f}(\mathbf{x})^\top\mathbf{f}(\mathbf{x}^+)] + \mathop{\mathbb{E}}_{c^+,\mathbf{x}}\underbrace{\ln\sum_{c\in[C]}\exp(\mathbf{f}(\mathbf{x})^\top\boldsymbol{\mu}_c)}_{=\mathrm{LSE}\left(\{\mathbf{f}(\mathbf{x})^\top\boldsymbol{\mu}_c\}_{c\in[C]}\right)}+2\ln(C\cosh(L^2)) - 2\ln K - \ln K\pi_{(1)} \quad \text{(Lemma 3)}$$

$$\tag{30}$$

$$\overset{(*)}{=} - \mathop{\mathbb{E}}_{c^+,\mathbf{x}}[\mathbf{f}(\mathbf{x})^\top\boldsymbol{\mu}_{c^+}] + \mathop{\mathbb{E}}_{c^+,\mathbf{x}}\underbrace{\ln\sum_{c\in[C]}\exp(\mathbf{f}(\mathbf{x})^\top\boldsymbol{\mu}_c)}_{=\mathrm{LSE}\left(\{\mathbf{f}(\mathbf{x})^\top\boldsymbol{\mu}_c\}_{c\in[C]}\right)}-2\ln(C\cosh(L^2)) + 2\ln K - \ln K\pi_{(1)} \tag{31}$$

$$= R_{\mu\text{-supv}}(\mathbf{f}) - 2\ln(C\cosh(L^2)) + 2\ln K - \ln K\pi_{(1)}, \tag{32}$$

where we use $\mathbb{E}_c[A] \leq \sum_{c\in[C]}\pi_{(1)}A$ and the monotonicity of $-\ln$ at (a). Note that the conditional independence is used only at (*). $\qquad\square$

Theorem 2. *For all* $\mathbf{f}$ *such that* $\|\mathbf{f}(\mathbf{x})\|_2 \leq L \ (\forall \mathbf{x} \in \mathcal{X})$, *the following inequality holds.*

$$R_{\mu\text{-supv}}(\mathbf{f}) \geq R_{\mathrm{cont}}(\mathbf{f}) + \Delta_{\mathrm{L}}, \tag{4}$$

where $\Delta_{\mathrm{L}} := \mathbb{H}(\boldsymbol{\pi}) + \ln\frac{K}{(K+1)^2} - 2\ln\cosh(L^2)$.

Proof of Theorem 2. The proof is essentially a consequence of the Fenchel's inequality and the Jensen's inequality. First, by noting that the convex conjugate of the log-sum-exp function is the negative Shannon entropy, the following identity is obtained.

$$R_{\mu\text{-supv}}(\mathbf{f}) = \mathop{\mathbb{E}}_{\mathbf{x},y}\left[-\mathbf{f}(\mathbf{x})^\top\boldsymbol{\mu}_y + \mathrm{LSE}\left(\mathbf{W}^\mu\mathbf{f}(\mathbf{x})\right)\right] \tag{33}$$

$$= \mathop{\mathbb{E}}_{\mathbf{x},y}\left[-\mathbf{f}(\mathbf{x})^\top\boldsymbol{\mu}_y + \sup_{\mathbf{p}\in\triangle^C}\left\{\mathbf{p}^\top(\mathbf{W}^\mu\mathbf{f}(\mathbf{x})) + \mathbb{H}(\mathbf{p})\right\}\right]. \tag{34}$$

If we choose an arbitrary $\mathbf{p} \in \triangle^C$, $R_{\mu\text{-supv}}(\mathbf{f})$ is lower bounded (Fenchel's inequality). Our choice is $\mathbf{p} = \boldsymbol{\pi}$. Recall that K is the number of negative samples. Then,

$$R_{\mu\text{-supv}}(\mathbf{f}) \geq \mathop{\mathbb{E}}_{c^+,\mathbf{x}}\left[-\mathbf{f}(\mathbf{x})^\top\boldsymbol{\mu}_{c^+} + \sum_{c^-\in\mathcal{Y}}\pi_{c^-}\mathbf{f}(\mathbf{x})^\top\boldsymbol{\mu}_{c^-}\right] + \mathbb{H}(\boldsymbol{\pi}) \tag{35}$$

$$\overset{(*)}{=} \mathop{\mathbb{E}}_{c^+}\mathop{\mathbb{E}}_{\mathbf{x},\mathbf{x}^+\sim\mathcal{D}_{c^+}^2}\left[-\mathbf{f}(\mathbf{x})^\top\mathbf{f}(\mathbf{x}^+) + \mathbf{f}(\mathbf{x})^\top\left(\mathop{\mathbb{E}}_{c^-}\mathop{\mathbb{E}}_{\mathbf{x}^-\sim\mathcal{D}_{c^-}}\left[\mathbf{f}(\mathbf{x}^-)\right]\right)\right] + \mathbb{H}(\boldsymbol{\pi}) \tag{36}$$

$$= \mathop{\mathbb{E}}_{c^+}\mathop{\mathbb{E}}_{\mathbf{x},\mathbf{x}^+}\left[-\mathbf{f}(\mathbf{x})^\top\mathbf{f}(\mathbf{x}^+) + \frac{1}{K}\sum_{k\in[K]}\mathop{\mathbb{E}}_{c_k^-}\mathop{\mathbb{E}}_{\mathbf{x}_k^-}[\mathbf{f}(\mathbf{x})^\top\mathbf{f}(\mathbf{x}_k^-)]\right] + \mathbb{H}(\boldsymbol{\pi}) \tag{37}$$

$$= \mathop{\mathbb{E}}_{c^+,\{c_k^-\}_k} \mathop{\mathbb{E}}_{\mathbf{x},\mathbf{x}^+,\{\mathbf{x}_k^-\}_k} \left[-\frac{1}{K} \sum_{k \in [K]} \left(\mathbf{f}(\mathbf{x})^\top \mathbf{f}(\mathbf{x}^+) - \mathbf{f}(\mathbf{x})^\top \mathbf{f}(\mathbf{x}_k^-) \right) \right] + \mathbb{H}(\boldsymbol{\pi}) \tag{38}$$

$$= \mathop{\mathbb{E}}_{c^+,\{c_k^-\}_k} \mathop{\mathbb{E}}_{\mathbf{x},\mathbf{x}^+,\{\mathbf{x}_k^-\}_k} \left[-\frac{1}{K} \sum_{k \in [K]} \ln \exp(\mathbf{f}(\mathbf{x})^\top (\mathbf{f}(\mathbf{x}^+) - \mathbf{f}(\mathbf{x}_k^-))) \right] + \mathbb{H}(\boldsymbol{\pi}). \tag{39}$$

Note that the conditional independence is used at (*). Here, we can proceed with the Jensen's inequality to lower bound the first term: for a non-negative vector $\mathbf{z} \in \mathbb{R}_{\geq 0}^N$, the inequality $-N^{-1} \sum_{i \in [N]} \ln z_i \geq -\ln(N^{-1} \sum_{i \in [N]} z_i)$ holds. If we set $z_k = \exp(\mathbf{f}(\mathbf{x})^\top (\mathbf{f}(\mathbf{x}^+) - \mathbf{f}(\mathbf{x}_k^-)))$ for $k \in [K]$,

$$R_{\mu\text{-supv}}(\mathbf{f}) - \mathbb{H}(\boldsymbol{\pi})$$

$$\geq \mathop{\mathbb{E}}_{\substack{c^+,\{c_k^-\}_k, \\ \mathbf{x},\mathbf{x}^+,\{\mathbf{x}_k^-\}_k}} \left[-\ln \frac{\sum_{k \in [K]} \exp(\mathbf{f}(\mathbf{x})^\top (\mathbf{f}(\mathbf{x}^+) - \mathbf{f}(\mathbf{x}_k^-)))}{K} \right] \tag{40}$$

$$\geq \mathop{\mathbb{E}}_{\substack{c^+,\{c_k^-\}_k, \\ \mathbf{x},\mathbf{x}^+,\{\mathbf{x}_k^-\}_k}} \left[-\ln \frac{\exp(\mathbf{f}(\mathbf{x})^\top (\mathbf{f}(\mathbf{x}^+) - \mathbf{f}(\mathbf{x}^+))) + \sum_{k \in [K]} \exp(\mathbf{f}(\mathbf{x})^\top (\mathbf{f}(\mathbf{x}^+) - \mathbf{f}(\mathbf{x}_k^-)))}{K} \right] \tag{41}$$

$$= \mathop{\mathbb{E}}_{\substack{c^+,\{c_k^-\}_k, \\ \mathbf{x},\mathbf{x}^+,\{\mathbf{x}_k^-\}_k}} \left[\ln \frac{\exp(-\mathbf{f}(\mathbf{x})^\top \mathbf{f}(\mathbf{x}^+))}{\exp(-\mathbf{f}(\mathbf{x})^\top \mathbf{f}(\mathbf{x}^+)) + \sum_{k \in [K]} \exp(-\mathbf{f}(\mathbf{x})^\top \mathbf{f}(\mathbf{x}_k^-))} \right] + \ln K. \tag{42}$$

Finally, by using Lemma 4,

$$R_{\mu\text{-supv}}(\mathbf{f}) - \mathbb{H}(\boldsymbol{\pi})$$

$$\geq \mathop{\mathbb{E}}_{c^+,\{c_k^-\}_k} \mathop{\mathbb{E}}_{\mathbf{x},\mathbf{x}^+,\{\mathbf{x}_k^-\}_k} \left[-\ln \frac{\exp(\mathbf{f}(\mathbf{x})^\top \mathbf{f}(\mathbf{x}^+))}{\exp(\mathbf{f}(\mathbf{x})^\top \mathbf{f}(\mathbf{x}^+)) + \sum_{k \in [K]} \exp(\mathbf{f}(\mathbf{x})^\top \mathbf{f}(\mathbf{x}_k^-))} \right] + \ln K$$
$$- 2\ln\left\{ (K+1)\cosh(L^2) \right\} \tag{43}$$

$$= R_{\text{cont}}(\mathbf{f}) + \ln K - 2\ln(K+1) - 2\ln\cosh(L^2), \tag{44}$$

which concludes the proof. □

C. Discussion on Relaxing Assumptions

In this section, we discuss relaxation of our main results (Theorems 1 and 2) from the following perspectives: the conditional independence assumption $\mathbf{x} \perp\!\!\!\perp \mathbf{x}^+ \mid c^+$, incorporating DA, and the correspondence between the supervised and latent classes $\mathcal{Y} = [C]$. Note that each relaxation is conceptually orthogonal and can be combined together.

C.1. When Conditionally Independent Assumption Is Violated

Theorems 1 and 2 initially rely on the conditional independence, which is used only when we mutually transform the following terms:

$$\mathop{\mathbb{E}}_{c^+,\mathbf{x},\mathbf{x}^+} [\mathbf{f}(\mathbf{x})^\top \mathbf{f}(\mathbf{x}^+)] = \mathop{\mathbb{E}}_{c^+,\mathbf{x}} [\mathbf{f}(\mathbf{x})^\top \boldsymbol{\mu}_{c^+}]. \tag{45}$$

This operation appears only once in each proof of Theorems 1 and 2 at (*). The conditional independence can be removed by the following bound:

$$-2L^2 \leq \mathop{\mathbb{E}}_{c^+,\mathbf{x},\mathbf{x}^+} [\mathbf{f}(\mathbf{x})^\top \mathbf{f}(\mathbf{x}^+)] - \mathop{\mathbb{E}}_{c^+,\mathbf{x}} [\mathbf{f}(\mathbf{x})^\top \boldsymbol{\mu}_{c^+}] = \mathop{\mathbb{E}}_{c^+,\mathbf{x}} [\mathbf{f}(\mathbf{x})^\top (\mathbf{f}(\mathbf{x}^+) - \boldsymbol{\mu}_{c^+})] \leq 2L^2, \tag{46}$$

where the inequalities is due to the Cauchy-Schwarz inequality: $|\mathbf{f}(\mathbf{x})^\top (\mathbf{f}(\mathbf{x}^+) - \boldsymbol{\mu}_{c^+})| \leq \|\mathbf{f}(\mathbf{x})\| \cdot \|\mathbf{f}(\mathbf{x}^+) - \boldsymbol{\mu}_{c^+}\| \leq 2L^2$. An excessive term incurs in the upper/lower bounds by invoking this bound. Wang et al. (2022) used the same idea to remove the conditional independence assumption. Nevertheless, we show Theorems 1 and 2 with the conditional assumption to focus on the influence of the negative sample size K on the surrogate gap.

C.2. When Supervised Class Differs from Latent Class

In unsupervised/self-supervised representation learning, it is often natural not to suppose any relationship between the supervised class set used in a downstream task and the latent classes. For example, unsupervised data in hand may represent concepts such as 'dog', 'cat', 'taxi', 'bus', and 'bird', while one expects to classify 'animal' or 'vehicle' in downstream. In our main results presented so far, we suppose that the supervised class set $\mathcal{Y}$ is the same as the latent classes $[C]$. Hereafter, we consider several cases where $\mathcal{Y} = [C]$ does not hold.

(Case I) $\mathcal{Y}$ is a subset of $[C]$: The upper bound (Theorem 1) can be extended but the lower bound (Theorem 2) cannot for this case. In the proof of Theorem 1 (Appendix B), we showed

$$R_{\text{cont}}(\mathbf{f}) \geq \underbrace{- \mathop{\mathbb{E}}_{c^+, \mathbf{x} \sim \mathcal{D}_{c^+}} \ln \frac{\exp(\mathbf{f}(\mathbf{x})^\top \boldsymbol{\mu}_{c^+})}{\sum_{c \in [C]} \exp(\mathbf{f}(\mathbf{x})^\top \boldsymbol{\mu}_c)}}_{\text{cross-entropy loss defined over class set } [C]} - 2\ln(C \cosh(L^2)) + 2\ln K - \ln K \pi_{(1)}. \tag{47}$$

The cross-entropy loss defined over latent classes $[C]$ can be lower-bounded as follows:

$$\mathop{\mathbb{E}}_{c^+ \sim \mathbb{P}(c^+ = y \in [C]), \mathbf{x} \sim \mathcal{D}_{c^+}} \left[- \ln \frac{\exp(\mathbf{f}(\mathbf{x})^\top \boldsymbol{\mu}_{c^+})}{\sum_{c \in [C]} \exp(\mathbf{f}(\mathbf{x})^\top \boldsymbol{\mu}_c)} \right] \geq \mathop{\mathbb{E}}_{c^+ \sim \mathbb{P}(c^+ = y \in \mathcal{Y}), \mathbf{x} \sim \mathcal{D}_{c^+}} \left[- \ln \frac{\exp(\mathbf{f}(\mathbf{x})^\top \boldsymbol{\mu}_{c^+})}{\sum_{c \in [C]} \exp(\mathbf{f}(\mathbf{x})^\top \boldsymbol{\mu}_c)} \right] \tag{48}$$

$$= - \mathop{\mathbb{E}}_{c^+, \mathbf{x}} [\mathbf{f}(\mathbf{x})^\top \boldsymbol{\mu}_{c^+}] + \mathop{\mathbb{E}}_{c^+, \mathbf{x}} \text{LSE}\left(\{ \mathbf{f}(\mathbf{x})^\top \boldsymbol{\mu}_c \}_{c \in [C]} \right) \tag{49}$$

$$\geq - \mathop{\mathbb{E}}_{c^+, \mathbf{x}} [\mathbf{f}(\mathbf{x})^\top \boldsymbol{\mu}_{c^+}] + \mathop{\mathbb{E}}_{c^+, \mathbf{x}} \text{LSE}\left(\{ \mathbf{f}(\mathbf{x})^\top \boldsymbol{\mu}_c \}_{c \in \mathcal{Y}} \right) \tag{50}$$

$$= \mathop{\mathbb{E}}_{c^+ \sim \mathbb{P}(c^+ = y \in \mathcal{Y}), \mathbf{x} \sim \mathcal{D}_{c^+}} \left[- \ln \frac{\exp(\mathbf{f}(\mathbf{x})^\top \boldsymbol{\mu}_{c^+})}{\sum_{c \in \mathcal{Y}} \exp(\mathbf{f}(\mathbf{x})^\top \boldsymbol{\mu}_c)} \right], \tag{51}$$

where the first inequality is resulted from the non-negativity of the cross-entropy loss and the second inequality uses monotonicity of the log-sum-exp: $\text{LSE}(z_1, \ldots, z_N) \geq \text{LSE}(z_1, \ldots, z_{N-1})$. Hence, the mean supervised loss defined over the supervised class set $\mathcal{Y}$ is bounded by the contrastive loss.

(Case II) $\mathcal{Y}$ is a coarse-grained set of $[C]$: In this case, we consider $\mathcal{Y}$ such as

$$\mathcal{Y} = \left\{ (c_1, \ldots, c_{j_1}), (c_{j_1+1}, \ldots, c_{j_2}), \ldots, (c_{j_{J-1}+1}, \ldots, c_{j_J}) \; \middle| \; \begin{array}{l} c_j \in [C] \text{ for any } j = 1, 2, \ldots, c_{j_J} \\ \text{and } c_j \neq c_{j'} \text{ for any } j, j' = 1, 2, \ldots, c_{j_J} \end{array} \right\}, \tag{52}$$

where $j_J = C$. Intuitively speaking, we split the latent classes $[C]$ into disjoint tuples $(c_j)_{j \in [j_1]}, \ldots, (c_j)_{j \in [j_J - j_{J-1}]}$ and regard each tuple in the disjoint set as a coarse-grained class of the original latent classes. This case aligns with the initial example: 'dog', 'cat', and 'bird' in the latent classes are combined into a single class 'animal', while 'taxi' and 'bus' are combined into 'vehicle'. Both the upper (Theorem 1) and lower (Theorem 2) bounds can be extended for this case. We omit the discussion on the upper bound because this is an immediate result by noting the linearity of the expectation and summation over classes. For the lower bound, after Equation (35) in the proof (Appendix B), we need to replace the expectation and summation over the supervised class $\mathcal{Y}$ with those over the latent class $[C]$, which is immediate as is the case of the upper bound.

D. Essential Bounds of Mean Supervised and Contrastive Losses

This section provides a supplementary explanation of the essential lower bounds of the mean supervised and contrastive losses. The common approaches of CURL applies the normalization on representation, in order to employ the cosine similarity $\frac{\mathbf{f}(\mathbf{x})^\top \mathbf{f}(\mathbf{x}')}{\|\mathbf{f}(\mathbf{x})\|_2 \cdot \|\mathbf{f}(\mathbf{x}')\|_2}$ as the similarity metric. Then, it is reasonable to assume $\|\mathbf{f}(\mathbf{x})\|_2 \leq L$ for all $\mathbf{x}$ with our data representation $\mathbf{f}$. The normalized representation corresponds to the case $L = 1$.

When we introduce the constraint $\|\mathbf{f}(\mathbf{x})\|_2 \leq L$, the mean supervised loss and contrastive loss are restricted as well. As for

the mean supervised loss,

$$R_{\mu\text{-supv}}(\mathbf{f}) \geq \inf_{\|\mathbf{f}'\|_2 \leq L} R_{\mu\text{-supv}}(\mathbf{f}') \tag{53}$$

$$= \inf_{\|\mathbf{f}'\|_2 \leq L} \mathbb{E}\left[\ln\left(1 + \sum_{c \neq y} \exp(\mathbf{f}'(\mathbf{x})^\top(\boldsymbol{\mu}_c - \boldsymbol{\mu}_y))\right)\right] \tag{54}$$

$$= \ln\left(1 + (C-1)\exp(-2L^2)\right) \tag{55}$$

$$(:= R_{\mu\text{-supv}}^*). \tag{56}$$

As for the contrastive loss,

$$R_{\text{cont}}(\mathbf{f}) \geq \inf_{\|\mathbf{f}'\|_2 \leq L} R_{\text{cont}}(\mathbf{f}') \tag{57}$$

$$= \inf_{\|\mathbf{f}'\|_2 \leq L} \mathbb{E}_{c^+, \{c_k^-\}, \mathbf{x}\,\mathbf{x}^+, \{\mathbf{x}_k^-\}} \left[\ln\left(1 + \sum_{k \in [K]} \exp(\mathbf{f}'(\mathbf{x})^\top(\mathbf{f}'(\mathbf{x}_k^-) - \mathbf{f}'(\mathbf{x}^+)))\right)\right] \tag{58}$$

$$\geq \inf_{\|\mathbf{f}'\|_2 \leq L} \mathbb{E}_{c^+, \{c_k^-\}, \mathbf{x}} \left[\ln\left(1 + \sum_{k \in [K]} \exp(\mathbf{f}'(\mathbf{x})^\top(\boldsymbol{\mu}_{c_k^-} - \boldsymbol{\mu}_{c^+}))\right)\right] \tag{59}$$

$$= \sum_{m=0}^{K} \binom{K}{m}\left(\frac{1}{C}\right)^m\left(1 - \frac{1}{C}\right)^{K-m} \ln\{1 + m + (K-m)\exp(-2L^2)\} \tag{60}$$

$$(:= R_{\text{cont}}^*), \tag{61}$$

where the Jensen's inequality is applied in the second inequality.

E. Discussion of Existing Surrogate Bounds

In this section, we describe the existing surrogate bounds in details to make them comparable with our main results. Then, we further discuss the detailed comparison between our theory and existing works. Before the discussion, we need to introduce the *sub-class* loss (of the mean classifier), which is the supervised classification loss over a subset of classes:

$$R_{\text{sub}}(\mathbf{f}, T) := \mathbb{E}_{\mathbf{x},y}\left[-\ln\frac{\exp(\boldsymbol{\mu}_y^\top \mathbf{f}(\mathbf{x}))}{\sum_{c \in T} \exp(\boldsymbol{\mu}_c^\top \mathbf{f}(\mathbf{x}))}\right], \tag{62}$$

where $T \subseteq [C]$ is a subset of classes and y is drawn from the subset of $\boldsymbol{\pi}$ with respect to T.

Arora et al.'s bound. We introduce additional notation that Arora et al. (2019) use. For a subset of classes T,

- $Q \subseteq [C]$ is the set of distinct classes in $c^+, c_1^-, \ldots, c_K^-$
- $I^+ := \{k \in [K] \mid c_k^- = c^+\}$
- $\text{Col} := \sum_{k \in [K]} \mathbb{1}_{\{c^+ = c_k^-\}} = |I^+|$
- $\rho_{\max}(T) := \max_{c \in T} \pi_c$
- $\rho_{\min}^+(T) := \min_{c \in T} \mathbb{P}_{c^+, \{c_k^-\}_k \sim \boldsymbol{\pi}^{K+1}}(c^+ = c \mid Q = T, I^+ = \emptyset)$
- $\tau_K := \mathbb{P}(I^+ \neq \emptyset)$

Arora et al. (2019) prove a finite-sample surrogate bound in Theorem B.1. In its proof, Eq. (26) is a surrogate bound established for a fixed $\mathbf{f}$. For the comparison, we focus on their Eq. (26):

$$(1 - \tau_K) \mathbb{E}_{T \sim \boldsymbol{\pi}^{K+1}}\left[\frac{\rho_{\min}^+(T)}{\rho_{\max}(T)} R_{\text{sub}}(\mathbf{f}, T)\right]$$
$$\leq R_{\text{cont}}(\mathbf{f}) - \tau_K \mathbb{E}_{c^+, \{c_k^-\}_k \sim \boldsymbol{\pi}^{K+1}}\left[\ln(\text{Col} + 1)\big| I^+ \neq \emptyset\right]. \tag{63}$$

We split the expectation term in the left-hand side as follows.

$$\mathbb{E}\left[\frac{\rho_{\min}^+(T)}{\rho_{\max}(T)}R_{\text{sub}}(\mathbf{f},T)\right]$$

$$= \underbrace{\mathbb{P}(T \text{ covers } [C])}_{=v_{K+1}} \cdot \mathbb{E}\left[\frac{\rho_{\min}^+(T)}{\rho_{\max}(T)}R_{\text{sub}}(\mathbf{f},T)\middle| T \text{ covers } [C]\right]$$

$$+ \mathbb{P}(T \text{ does not cover } [C]) \cdot \mathbb{E}\left[\frac{\rho_{\min}^+(T)}{\rho_{\max}(T)}R_{\text{sub}}(\mathbf{f},T)\middle| T \text{ does not cover } [C]\right] \tag{64}$$

$$\geq v_{K+1}\frac{\rho_{\min}^+([C])}{\rho_{\max}([C])}\underbrace{R_{\text{sub}}(\mathbf{f},[C])}_{=R_{\mu\text{-supv}}(\mathbf{f})}. \tag{65}$$

Under the uniform class prior assumption ($\boldsymbol{\pi} = 1/C \cdot \mathbf{1}$), $\rho_{\max}([C]) = 1/C$, and we can pick any class $c_0 \in [C]$ by the symmetry and $\rho_{\min}^+([C]) = \mathbb{P}(c^+ = c_0 \mid Q = [C], I^+ = \emptyset) = 1/C$. In addition,

$$\tau_K \mathbb{E}_{c^+,\{c_k^-\}_k\sim\boldsymbol{\pi}^{K+1}}\left[\ln(\text{Col}+1)\middle|I^+ \neq \emptyset\right] = \mathbb{E}\left[\ln(\text{Col}+1)\right] - (1-\tau_K)\mathbb{E}\left[\ln(\text{Col}+1)\middle|I^+ = \emptyset\right] \tag{66}$$

$$= \mathbb{E}\left[\ln(\text{Col}+1)\right]. \tag{67}$$

As a result, we obtain the following simplified expression in Table 1:

$$R_{\mu\text{-supv}}(\mathbf{f}) \leq \frac{1}{(1-\tau_K)v_{K+1}}\left\{R_{\text{cont}}(\mathbf{f}) - \mathbb{E}[\ln(\text{Col}+1)]\right\}. \tag{68}$$

Nozawa & Sato's bound. The surrogate bound provided by Nozawa & Sato (2021, Theorem 8) involves a factor resulting from DA and self-supervised learning setting. By dropping this (negative) factor, the surrogate bound is

$$R_{\text{cont}}(\mathbf{f}) \geq \frac{1}{2}\left\{v_{K+1}R_{\mu\text{-supv}}(\mathbf{f}) + (1-v_{K+1})\mathbb{E}_{T\sim\boldsymbol{\pi}^{K+1}}[R_{\text{sub}}(\mathbf{f},T)] + \mathbb{E}[\ln(\text{Col}+1)]\right\} \tag{69}$$

$$\geq \frac{1}{2}\left\{v_{K+1}R_{\mu\text{-supv}}(\mathbf{f}) + \mathbb{E}[\ln(\text{Col}+1)]\right\}, \tag{70}$$

resulting in the bound in Table 1. The sub-class loss may be safely dropped because it has the coefficient $1 - v_{K+1}$, which is expected to be exponentially small in K.

Ash et al.'s bound. Ash et al. (2022, Theorem 5) provides the following surrogate bound

$$R_{\mu\text{-supv}}(\mathbf{f}) \leq \frac{2\left\lceil\frac{2(1-\pi_{(-1)})H_{C-1}}{K\pi_{(-1)}}\right\rceil}{(1-\pi_{(1)})K}\left\{R_{\text{cont}}(\mathbf{f}) - \tau_K\mathbb{E}_{c^+,\{c_k^-\}_k\sim\boldsymbol{\pi}^{K+1}}\left[\ln(\text{Col}+1)\middle|I^+ \neq \emptyset\right]\right\}. \tag{71}$$

By substituting $\boldsymbol{\pi} = 1/C \cdot \mathbf{1}$ and $\tau_K\mathbb{E}[\ln(\text{Col}+1) \mid I^+ \neq \emptyset] = \mathbb{E}[\ln(\text{Col}+1)]$, the bound in Table 1 is obtained.

Detailed comparisons. As we stated in Section 4 of the main text, only our bound agrees well with the experimental fact that the larger K is better for *all* K regions:

- Arora et al. (2019): Large K degrades the performance because of the label collision.

- Nozawa & Sato (2021): Large K improves the performance for $K > C$.

- Ash et al. (2022): The optimal K exists by the collision-coverage trade-off.

Even though the claim by Nozawa & Sato (2021) is similar to ours, we discovered a different underlying mechanism to support this idea, which leads to better explainability of empirical facts.

The proof of Nozawa & Sato (2021) is based on the idea of label coverage: The more negative samples we draw (larger K), the more likely the negative samples can cover all class labels. The upper bound based on this idea is only activated when $K > C$ because label coverage is impossible with $K \leq C$. This inability contradicts the real experiments including Chen et al. (2021); Tomasev et al. (2022), which showed that CURL exhibits reasonable performance even with small K.

Our proof leverages the idea that R_{cont} and $R_{\mu\text{-supv}}$ have the similar log-sum-exp functional forms. This similarity casts R_{cont} as a surrogate objective of $R_{\mu\text{-supv}}$ and its surrogate gap is reduced with larger K. Even with small K, the upper bound of $R_{\mu\text{-supv}}$ is loose but not prohibitively large thereby the surrogate bound of $R_{\mu\text{-supv}}$ is still valid. Our theoretical claim reveals that the surrogate gap improves in $O(K^{-1})$ for all K regions, which is in good agreement with the real experiments. Eventually, our theory provides practical feedback such that one may reduce K (even smaller than C) to trade off the downstream performance with the computational cost.

F. Relationship to Mutual Information (MI) Estimation

The contrastive loss R_{cont} we studied in this paper is also known as the InfoNCE loss (van den Oord et al., 2018), which is known to be deeply related to the multi-sample estimation of mutual information (MI) (van den Oord et al., 2018; Poole et al., 2019; Song & Ermon, 2020). Although the multi-sample estimators have high bias and low variance compared to variational estimators in general (Poole et al., 2019; Song & Ermon, 2020; Guo et al., 2021), the quantitative analysis of the bias-variance trade-off of the multi-sample MI estimators has yet to be clearly known. Tian et al. (2020) and Tschannen et al. (2020) experimentally showed that maximizing *tighter* MI bound does not necessarily lead to good representation; there is no guarantee that the model can achieve higher MI by the tighter bound.

Recently, the theoretical limitations of sample-based MI estimation have been analyzed (Gao et al., 2015; McAllester & Stratos, 2020). These studies revealed that a particular type of sample-based estimator of MI (Gao et al., 2015) or its lower bound (McAllester & Stratos, 2020) can be upper bounded by $O(\ln N)$ for the number of samples N. In this section, we discuss the implications of these limitations in the CURL setting.

Given two random variables X and Y, suppose that we have $K + 1$ randomly drawn pairs $\{(x_i, y_i)\}_{i=1}^{K+1}$ from these random variables such that for all (i, j), (x_i, y_j) can be regarded as a positive pair when $i = j$, and otherwise can be regarded as a negative pair. Poole et al. (2019, Equation 10) derived the following lower bound for MI:

$$I(X;Y) \geq I_{\text{NCE}}^{K+1} := \mathbb{E}\left[\frac{1}{K+1} \sum_{i=1}^{K+1} \ln \frac{\exp(s(x_i, y_i))}{\frac{1}{K+1} \sum_{j=1}^{K+1} \exp(s(x_i, y_j))} \right], \tag{72}$$

where $I(X;Y)$ is the MI between X and Y, and $s(x, y)$ is a critic function. This lower bound estimator can be rewritten using R_{cont} as follows:

$$I_{\text{NCE}}^{K+1} := \mathbb{E}\left[\frac{1}{K+1} \sum_{i=1}^{K+1} \ln \frac{\exp(s(x_i, y_i))}{\frac{1}{K+1} \sum_{j=1}^{K+1} \exp(s(x_i, y_j))} \right] \tag{73}$$

$$= \mathbb{E}\left[\frac{1}{K+1} \sum_{i=1}^{K+1} \ln \frac{\exp(s(x_i, y_i))}{\sum_{j=1}^{K+1} \exp(s(x_i, y_j))} \right] + \ln(K+1) \tag{74}$$

$$= \mathbb{E}\left[\frac{1}{K+1} \sum_{i=1}^{K+1} \ln \frac{\exp(s(x_i, y_i))}{\exp(s(x_i, y_i)) + \sum_{j \neq i} \exp(s(x_i, y_j))} \right] + \ln(K+1) \tag{75}$$

$$= \mathbb{E}\left[\ln \frac{\exp(s(x, x^+))}{\exp(s(x, x^+)) + \sum_{j=1}^{K} \exp(s(x, x_j^-))} \right] + \ln(K+1) \tag{76}$$

$$= -R_{\text{cont}}(\mathbf{f}) + \ln(K+1). \tag{77}$$

The first equality is obtained by putting the constant in the denominator outside. The third equality comes by replacing the notation (x_i, y_i) with (x, x^+) under the assumption that all (x_i, y_i) come from the same iid distribution. By setting $s(x, y) := \mathbf{f}(x)^\top \mathbf{f}(y)$, we obtain the last equation.

Here, McAllester & Stratos (2020) gave the following theorem for the sample-based estimator of the lower bound on MI.

Theorem 5 ([McAllester & Stratos (2020)](#) Theorem 1.1, informal). *Let $\widehat{I}^N$ be any mapping from N samples of (X,Y) to $\mathbb{R}$ that satisfies*

$$I(X;Y) \geq \widehat{I}^N(\{(x_i, y_i)\}_{i=1}^N) \tag{78}$$

in high probability, then the following relationship holds in high probability:

$$\widehat{I}^N(\{(x_i, y_i)\}_{i=1}^N) \leq 2\ln N + 5. \tag{79}$$

Since I_{NCE}^{K+1} satisfies the condition for $\widehat{I}^{K+1}$, we now have the following:

$$-R_{\text{cont}}(\mathbf{f}) + \ln(K+1) \leq 2\ln(K+1) + 5 \implies R_{\text{cont}}(\mathbf{f}) \geq -\ln(K+1) - 5. \tag{80}$$

However, the right-hand statement always holds by the construction of R_{cont} for all $\mathbf{f}$ ($\forall \mathbf{f}, R_{\text{cont}}(\mathbf{f}) \geq 0$). In other words, in the case of the CURL setting, [McAllester & Stratos (2020)](#)'s theorem does not restrict R_{cont}, which means that the large K effect investigated in our paper comes from a completely different mechanism from the above theorem. While the theoretical studies on MI aim to guarantee for multi-sample estimation of ground-truth MI, a series of CURL studies, including ours, differ in that we aim to derive a surrogate gap bound between two different losses, namely supervised loss and contrastive loss. Specifically, the above derivation does not address supervised loss in that both the left-hand ($I(X;Y)$ or I_{NCE}^{K+1}) and the right-hand ($R_{\text{cont}}(\mathbf{f})$) quantities represent the amount of information between different views rather than MI between the view and its label. The existing studies on sample-based MI estimation are worthwhile in the sense that these works revealed the $O(\ln N)$ effect on the non-trivial estimators such as k-NN based estimator ([Gao et al., 2015](#)) or *any* kind of lower bound estimator ([McAllester & Stratos, 2020](#)).

G. Experimental details

G.1. Synthetic Dataset

We used Adam ([Kingma & Ba, 2015](#)) optimizer with the weight decay of coefficient 0.01 to all parameters. The mini-batch size was set to $B = 1\,024$ and the number of epochs was 300. The learning rate was set to 0.01 with `ReduceLROnPlateau` scheduler (patience: 10 epochs) provided by PyTorch ([Paszke et al., 2019](#)).

G.2. CIFAR-10/100

We treated 10% training samples as a validation dataset by sampling class uniformly. We used the original test dataset for testing. We used the same data-augmentation as in the CIFAR-10 experiment by [Chen et al. (2020)](#) during contrastive learning and linear supervised training of the linear classifier.

As a feature extractor $\mathbf{f}$, we modified the ResNet-18 ([He et al., 2016](#)) by following the convention of self-supervised representation learning ([Chen et al., 2020](#), B.9); replacement of the first convolutional layer with a smaller one, removal of the first max-pooling layer, and replacement of the final fully-connected layer with a nonlinear projection head whose dimensional is 32.[7]

Since we need to enlarge the negative samples size K that depends on the size of mini-batches, we followed a large mini-batch training setting used in recent self-supervised learning ([Chen et al., 2020](#); [Caron et al., 2020](#)). We used LARC ([You et al., 2017](#)) optimizer wrapping the momentum SGD, whose momentum term was 0.9. We applied weights decay of coefficient 10^{-4} to all parameters except for all bias terms and batch norm's parameters. The base learning rate was initialized at $\text{lr} \times \sqrt{B}$, where $\text{lr} \in \{2, 4, 6\} \times 1/64$ and mini-batch size $B = 1\,024$ inspired by SimCLR's squared learning rate scaling. As a learning rate scheduler for each iteration, we used linear warmup during the first 10 epochs and cosine annealing without restart ([Loshchilov & Hutter, 2017](#)) during the rest epochs. The number of epochs was $2\,000$.

We implemented our experimental code by using PyTorch ([Paszke et al., 2019](#))'s distributed data-parallel training ([Li et al., 2020](#)) on 8 NVIDIA A100 GPUs provided by the internal cluster. Therefore we replaced the all batch normalization layer with `SyncBatchNorm` module provided by PyTorch.[8] To accelerate contrastive learning, we used automatic mixed-precision training provided by PyTorch.

---

[7]Unlike the reported results by [Chen et al. (2021)](#), smaller dimensionality, i.e., 32 gives better downstream accuracy on CIFAR-100 than 64 or 128. This difference might come from the differences in the loss function and positive pair's generation process.

[8]See [Wu & Johnson (2021](#), Sec. 6.2) for more detailed discussion of this replacement for contrastive learning.

G.3. Wiki-3029

Wiki-3029 contains $3\,029$ English Wikipedia article pages. Each page consists of 200 sentences. Since the dataset does not have the explicit train/validation/test splits, we split the dataset into $70\%/10\%/20\%$ train/validation/test datasets, respectively. As a pre-processing, we tokenized the dataset using torchtext's `basic_english` tokenizer. After tokenization, we removed the tokens whose frequency is less than 5 in the training dataset. We did not use DA.

We used fasttext (Joulin et al., 2017)'s based feature extractor.[9] In our preliminary experiments, only using a word embedding layer and average pooling among words perform better than either additional linear or nonlinear projection heads. A similar model to ours is also used in Ash et al. (2022). The dimensionality of the word embedding layer was 256.

We mainly followed the same optimization setting as our CIFAR-10/100 experiments. We note that the mini-batch size $B = 2\,048$; the initial learning rate lr was selected in $\{1, 2, 3, 4\} \times \text{\textonehalf}/40$; no weights decay; the number of epochs was 90; and perform linear warmup during the first 3 epochs. When we decrease C, the number of epochs is multiplied by $3\,000/C$ for simplicity.[10]

G.4. Contrastive Learning

By following the data generation process in contrastive representation learning and existing work (Arora et al., 2019; Ash et al., 2022), we treated the supervised classes $\mathcal{Y}$ as latent classes $[C]$. After obtaining training/validation/test datasets as described above, we carefully constructed positive pairs for contrastive learning *before* training[11] as follows; We treated each sample in the training data as an anchor sample. We drew a different sample from the same latent class of each anchor sample as a positive sample in the training dataset. For negative samples, we drew K negative samples from other samples in the same mini-batch by following the convention of self-supervised representation learning such as SimCLR (Chen et al., 2020). Since Chen et al. (2020) used all other samples as negative samples, the negative samples size and the size of mini-batches depend on each other: $K = 2B - 2$. To relax the effect of the difference of the mini-batch size when we change K, we drew K samples without replacement from $2B - 2$ inspired by Ash et al. (2022). In this sampling, we guaranteed to draw at most one sample from each positive pair because we are concerned about the relation between the number of latent classes and K. We did not use validation and test datasets during contrastive representation learning.

G.5. Mean and Linear Classifiers' Evaluation

For evaluation, we reported the test accuracy values of mean and linear classifiers. For a linear classifier, we used Nesterov's momentum SGD, whose momentum coefficient was 0.9 without weight decay. We set the mini-batch size $B = 256$ and $B = 512$ for CIFAR-10/100 and Wiki-3029, respectively. We used cosine annealing without restart as a learning rate scheduler for each iteration. We set 100 and 30 epochs for CIFAR-10/100 and Wiki-3029 datasets, respectively. For CIFAR-10/100, we set learning rate as 0.03. For Wiki-3029, we searched the learning rate in $\{0.5, 1, 5, 10, 50\} \times \text{\textonehalf}/10^3$. The learning rate was scaled by using squared learning rate scaling. For linear evaluation of CIFAR-10/100, we used PyTorch's distributed data-parallel training. We calculated the test accuracy by using the best combination of the contrastive model and the hyper-parameter of a linear classifier that maximizes the validation accuracy. We repeated contrastive learning and downstream task's evaluation three times with different random seeds and reported the averaged values.

G.6. Details of Figure 1

Before computing the upper bounds and supervised loss, we normalized feature representations $\mathbf{f}(\mathbf{x})$ learned in Appendix G.4 to ensure $L = 1$, which is the upper bound of $\|\mathbf{f}(\mathbf{x})\|_2, \forall \mathbf{x}$. For each random seed and the number of negative samples K, we selected learned feature encoder $\mathbf{f}$ that got the highest validation mean supervised accuracy in different learning rates of the optimizer of the contrastive learning. Then we calculated the test supervised loss value by using the selected contrastive models.

Using the same feature encoder with L_2 normalization, we calculated the contrastive loss on the test dataset. To do so, we created positive pairs by the same procedure on the test dataset as described in Appendix G.4. Negative samples were also

---

[9]Arora et al. (2019) uses GRU-based feature encoder with frozen word embeddings of GloVe (Pennington et al., 2014) trained on commonCrawl.

[10]We found the contrastive learning did not yield good feature representations for a downstream task without this longer training.

[11]We can create the labeled dataset, especially with non-overlapped latent classes, if we draw positive samples at each iteration or epoch during optimization using stochastic gradient descent.
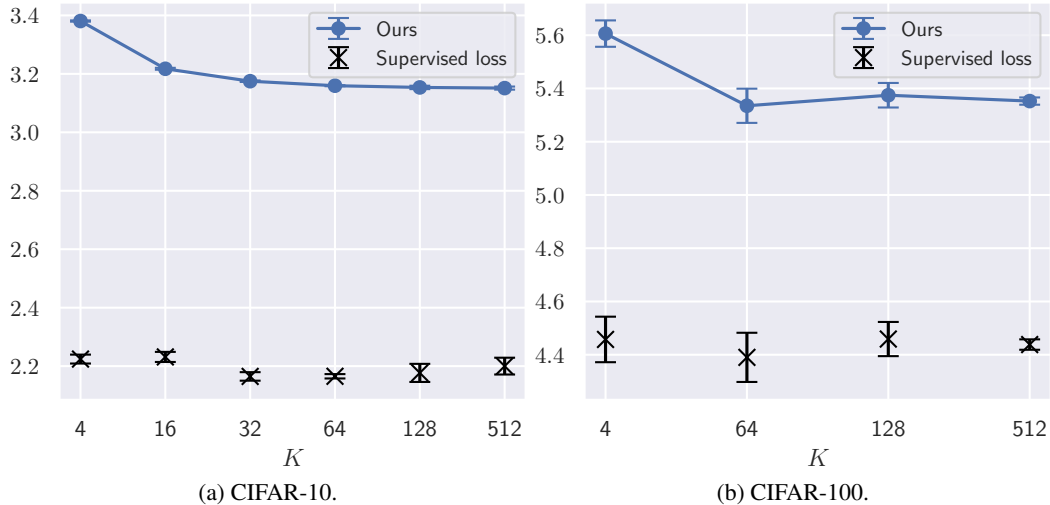
(a) CIFAR-10.   (b) CIFAR-100.

Figure 9: Enlarged Figure 1 for the detailed comparison between the proposed bound and the supervised loss on CIFAR-10/100 datasets. All value is an averaged value among three runs with a different random seed. Error bar indicates the standard deviation.

drawn from the other samples in the mini-batches as the contrastive learning step described in Appendix G.4. To calculate the contrastive loss, we used the same batch size as the contrastive learning step and only one epoch. Since this contrastive loss calculation was stochastic due to the sampling of positive and negative samples, we repeated the contrastive loss calculation 25 times and averaged them to create plot Figure 1. Note that we used the theoretical values of $\tau_K, v_{K+1}, \mathbb{E}\ln(\mathrm{Col}+1)$ that are shown in the existing upper bounds on Table 1 rather than the simulated values.

Figure 9 shows the enlarged version of Figure 1 and the same plot using CIFAR-100. This figure focuses on the detailed comparison between the test datasets' empirical supervised loss values and theoretical bounds. For both CIFAR-10/100 datasets, there were almost no changes in the supervised loss as K varied, and the losses were slightly larger in the region where K was small. These results are consistent with the theoretical estimation of the upper bounds (solid lines).

G.7. Details of Figure 8

During minimization of the contrastive loss to learn **f** in Appendix G.4, we saved the model's weight at every 200 epochs. We reported the test mean supervised accuracy using **f** that maximized validation accuracy among different learning rate values.

G.8. Additionally Used Libraries

In our experiments, we also used scikit-learn (Pedregosa et al., 2011) for train/val/test data splits. We created all plots by using matplotlib (Hunter, 2007) and seaborn (Waskom, 2021) via pandas (Reback et al., 2020) except for Figure 2. We managed our experiments' configuration using hydra (Yadan, 2019) and experimental results using Weights & Biases (Biewald, 2020). For effective parallelized execution of our experimental codes, we use GNU Parallel (Tange, 2021).