Large-Margin Contrastive Learning with Distance Polarization Regularizer

Shuo Chen [1 2] **Gang Niu** [1] **Chen Gong** [2] **Jun Li** [2] **Jian Yang** [2] **Masashi Sugiyama** [1 3]

Abstract

Contrastive learning (CL) pretrains models in a pairwise manner, where given a data point, other data points are all regarded as dissimilar, including some that are *semantically* similar. The issue has been addressed by properly weighting similar and dissimilar pairs as in *positive-unlabeled learning*, so that the objective of CL is *unbiased* and CL is *consistent*. However, in this paper, we argue that this great solution is still not enough: its weighted objective *hides* the issue where the semantically similar pairs are still pushed away; as CL is pretraining, this phenomenon is not our desideratum and might affect downstream tasks. To this end, we propose *large-margin contrastive learning* (LMCL) with *distance polarization regularizer*, motivated by the distribution characteristic of pairwise distances in *metric learning*. In LMCL, we can distinguish between *intra-cluster* and *inter-cluster* pairs, and then only push away inter-cluster pairs, which *solves* the above issue explicitly. Theoretically, we prove a tighter error bound for LMCL; empirically, the superiority of LMCL is demonstrated across multiple domains, *i.e.*, image classification, sentence representation, and reinforcement learning.

1. Introduction

Machine learning without human annotation is a long-standing and important problem. Recently, the unsupervised learning approach has been greatly promoted by *contrastive learning* (CL), which shows encouraging performance compared to fully supervised learning methods (Wu et al., 2018; Saunshi et al., 2019). CL directly learns a generic feature embedding for original data, and the learned embedding can be widely employed in many downstream recognition tasks such as classification (Chen et al., 2020a) and clustering (Zhong et al., 2020). Thereby, CL has become one of the most important unsupervised learning approaches.

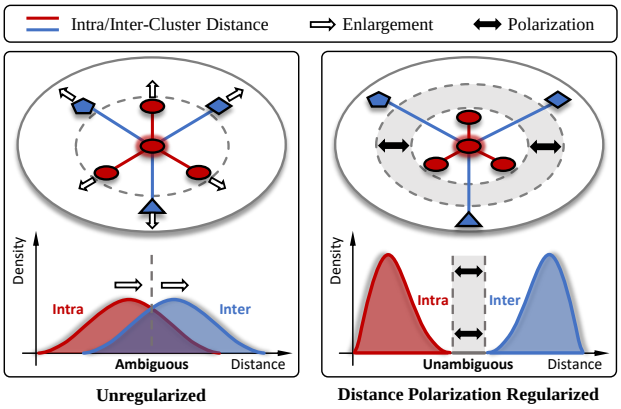

Figure 1. Conceptual illustration of unregularized contrastive learning and distance polarization (DP) regularized contrastive learning. The conventional unregularized model enlarges distances between all pairs of instances and potentially leading to some ambiguous intra/inter-cluster distances. We propose a DP regularized learning algorithm to encourage pairwise distances to be extremely large or small, and thus gaining the unambiguous distance determination with a large margin between intra-cluster and inter-cluster.

As human annotation is not available in an unsupervised learning problem setting, CL algorithms usually consider building the pseudo supervision in their learning objectives (Saunshi et al., 2019; Jing & Tian, 2020). In general, most existing CL frameworks regard any two instances in the training data as a negative pair (including those false-negative pairs consisted of semantically similar instances), and meanwhile construct the positive pair by combining each instance with its perturbation (Wu et al., 2018; Song & Ermon, 2020). Due to the continued success from positive pairs, many recent efforts have increasingly focused on various data augmentation techniques to further enrich training data (Oord et al., 2018; Tian et al., 2020a) and simultaneously preserve semantic contents (Logeswaran & Lee, 2018; Tian et al., 2020b).

While positive pair sampling has drawn much attention, relatively fewer works consider the effectiveness of negative pair in CL (Jing & Tian, 2020). Actually, as most existing

[1]RIKEN Center for Advanced Intelligence Project, Japan; [2]PCA-Lab, School of Computer Science and Engineering, Nanjing University of Science and Technology, China; [3]Graduate School of Frontier Sciences, The University of Tokyo, Japan. Correspondence to: Shuo Chen <shuo.chen.ya@riken.jp>, Jun Li <junli@njust.edu.cn>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

CL methods directly repel all pairs of instances in the training data, the semantically similar instances are undesirably pushed apart. Recent works propose weighting the positive and negative pairs as in *positive-unlabeled learning* (Chen et al., 2020b) to counteract the impact of false-negative pairs (Chuang et al., 2020; Robinson et al., 2020). Nevertheless, the weighted learning objectives still encourage repelling each pair of original instances in the training data (Huynh et al., 2020), so they are not able to faithfully reflect the similarity between two semantically similar instances.

Although the above existing CL algorithms have achieved promising results to some extent, most of their objectives do not explicitly discriminate the semantic similarity of each instance pair, and thus they cannot adequately capture intrinsic features in the training data. To address this issue, we provide analytical results to reveal that when the conventional CL encourages repelling each pair of original instances, the finally learned pairwise distances nearly obey a *unimodal distribution* in the region $(0,\, 1)$. It implies that the conventional CL fails to yield an explicit margin to discriminate the similarities of data pairs (see the left panel of Fig. 1). Therefore, this inspires us to propose *large-margin contrastive learning* (LMCL) with *distance polarization* (DP) regularizer, which clearly separates the similar pairs from dissimilar pairs with a large margin. Such a DP regularizer is motivated by the general goal of *metric learning* (Weinberger et al., 2006), which casts penalty onto all pairwise distances within the margin region, and thereby encouraging polarized distances for similarity determinations (see the *bimodal distribution* in the right panel of Fig. 1). Theoretically, we prove that the proposed DP regularizer effectively tightens the error bound of conventional CL algorithm. Experimentally, our approach consistently improves the state-of-the-art methods on vision, language, and reinforcement learning benchmarks. Our proposed DP regularizer is simple yet generic, which can be easily deployed in many existing CL methods. Our main contributions are summarized below:

- We propose a new distance polarization (DP) regularizer to enhance the generalizability of the conventional CL algorithm by explicitly discriminating the pairwise similarity between two original instances.
- We establish the complete theoretical guarantee for our method to analyze the error bounds of similarity measure and downstream classification.
- We conduct intensive experiments on synthesis datasets and real-world datasets to validate the superiority of our method over the state-of-the-art CL approaches.

2. Background & Related Work

In this section, we first introduce some necessary notations. Then, we briefly review the background of contrastive learning. We also introduce the main concepts of metric learning and regularization technique, which are related to this paper.

Notations. We write matrices and vectors as bold uppercase characters and bold lowercase characters, respectively. We denote the training dataset $\mathcal{X} = \{\boldsymbol{x}_i \in \mathbb{R}^m | i = 1, 2, \ldots, N\}$ where m is the data dimensionality and N is the total number of instances. Operator $\odot$ denotes the element-wise product of two vectors/matrices. Operators $\|\cdot\|_0$ and $\|\cdot\|_1$ denote the vector/matrix ℓ_0-norm and ℓ_1-norm, respectively.

2.1. Contrastive Learning

As an unsupervised / self-supervised learning approach, the basic goal of contrastive learning (CL) algorithm is to learn a generic feature embedding $\boldsymbol{\varphi} : \mathbb{R}^m \mapsto \mathbb{R}^d$, which transforms the data point from m-dimensional sample space to d-dimensional embedding space for extracting intrinsic features. The primitive CL method called *instance discrimination* learns such an embedding by directly repelling each pair of two instances in the training data (Wu et al., 2018). Subsequent works such as *momentum contrastive* (MoCo) encourage using larger negative pair batch size for better learning results (He et al., 2020). Recently, the SimCLR framework further introduces data augmentation to generate positive pairs which incorporate more semantic information into the learning objective (Chen et al., 2020a). In general, the effectiveness of existing CL algorithms relies on two key components: the negative pairs $(\boldsymbol{x},\, \boldsymbol{x}^-)$ sampling from every two original instances in the training data, and the positive pairs $(\boldsymbol{x},\, \boldsymbol{x}^+)$ built by each single instance $\boldsymbol{x}$ and its perturbation $\boldsymbol{x}^+$. When the *noise contrastive estimation* (NCE) loss (Gutmann & Hyvärinen, 2010) is employed to learn a feature embedding $\boldsymbol{\varphi}$ from positive and negative pairs, the general learning objective can be formulated as

$$\mathcal{L}_{\mathrm{NCE}}(\boldsymbol{\varphi}) = \mathbb{E}_{\boldsymbol{x},\, \boldsymbol{x}_j^- \in \mathcal{X}} \left[-\log \frac{\mathrm{e}^{\boldsymbol{\varphi}(\boldsymbol{x})^\top \boldsymbol{\varphi}(\boldsymbol{x}^+)}}{\mathrm{e}^{\boldsymbol{\varphi}(\boldsymbol{x})^\top \boldsymbol{\varphi}(\boldsymbol{x}^+)} + \sum_{j=1}^n \mathrm{e}^{\boldsymbol{\varphi}(\boldsymbol{x})^\top \boldsymbol{\varphi}(\boldsymbol{x}_j^-)}} \right], \quad (1)$$

where $\boldsymbol{x}$ and $\{\boldsymbol{x}_j^-\}_{j=1}^n$ are uniformly sampled from the training data $\mathcal{X}$. Here n is the batch size of negative pairs.

It is worth noting that the conventional NCE loss for contrastive learning is biased, as the semantically similar (*i.e.*, false-negative) data pairs might be pushed apart during the repelling of all negative pairs. To alleviate this issue, the clustering approach (Li et al., 2020) is applied on the learned embedding to gather similar instances, though the reliability of clustering results can be easily influenced by the learned embedding itself. Recent works adopted popular practices in positive-unlabeled (PU) learning (Chen et al., 2020b) to reweight the NCE loss by increasing the importance of positive pairs (Chuang et al., 2020) or allocating different importance for negative pairs (Robinson et al., 2020).

Although few works have been proposed to alleviate the undesirable repelling of semantically similar instances, their learning objectives still cannot clearly discriminate the pairwise similarity between two original instances. In this paper, we address this issue from a different viewpoint, which employs the basic property of metric learning (Chu et al., 2020) to constrain the similarity of negative pairs.

2.2. Metric Learning

As a supervised learning problem, metric learning aims to learn a distance metric to faithfully measure the pairwise similarity between two instances in the sample space (Davis et al., 2007; Chu et al., 2020). For the training data $\mathcal{X} = \{\boldsymbol{x}_i \in \mathbb{R}^m | i = 1, 2, \ldots, N\}$, the class labels $\{y_i \in \{1, 2, \ldots, C\} | i = 1, 2, \ldots, N\}$ are provided for supervision, where C is the number of classes. As the supervisory information is available, the positive pairs and negative pairs in metric learning can be directly built by the semantics labels $\{y_i\}_{i=1}^n$, and thus formulating the well-known $(n+1)$-*tuplet loss* (Sohn, 2016)

$$\mathcal{L}_{\text{TUP}}(\boldsymbol{\varphi}) = \mathbb{E}_{y_i = y_k \neq y_{b_j}} \left[-\log \frac{\mathrm{e}^{\boldsymbol{\varphi}(\boldsymbol{x}_i)^\top \boldsymbol{\varphi}(\boldsymbol{x}_k)}}{\mathrm{e}^{\boldsymbol{\varphi}(\boldsymbol{x}_i)^\top \boldsymbol{\varphi}(\boldsymbol{x}_k)} + \sum_{j=1}^n \mathrm{e}^{\boldsymbol{\varphi}(\boldsymbol{x}_i)^\top \boldsymbol{\varphi}(\boldsymbol{x}_{b_j})}} \right], \quad (2)$$

which encourages to reduce the intra-class distance $\|\boldsymbol{\varphi}(\boldsymbol{x}_i) - \boldsymbol{\varphi}(\boldsymbol{x}_k)\|_2^2$ and enlarge inter-class distance $\|\boldsymbol{\varphi}(\boldsymbol{x}_i) - \boldsymbol{\varphi}(\boldsymbol{x}_{b_j})\|_2^2$ for $i, k, b_j = 1, 2, \ldots, N$, in which $j = 1, 2, \ldots, n$ and $\{b_j\}_{j=1}^n$ is the index set of batch negative points. Similar to Eq. (1), here n is the batch size of negative pairs. Minimizing such a supervised learning objective will lead to a margin between the intra-class and the inter-class distances, and thereby discriminating the pairwise similarity between each two original instances (Yu & Tao, 2019).

Although the above Eq. (2) has a very similar form to Eq. (1), we can find that here Eq. (2) is fully supervised, so its negative pairs are unbiased. In this paper, we convert the basic property of the above metric learning model to a regularizer for constraining the learning objective of the CL algorithm.

2.3. Regularization Technique

Regularization is a generic and effective technique that has been well studied and widely applied in statistics and machine learning (Dong et al., 2014; Scholkopf & Smola, 2018). Generally speaking, a regularization term (*i.e.*, regularizer) usually considers introducing a specific inductive bias into the empirical loss, and thus reducing the hypothesis space complexity and improving the model generalizability (Guo et al., 2017). For example, the well-known ℓ_2-norm regularizer (*i.e.*, weight decay (Krogh & Hertz, 1992) in some deep learning models) restricts the scale of learning parameter so that the learned embedding can successfully capture scale-invariant features (Yang et al., 2011). The ℓ_1-norm regularizer (*i.e.*, sparse regularization) assumes that only a few learning parameters should be activated in practical recognition tasks, and thereby alleviating the impact from over-fitting results (Arpit et al., 2016).

Our proposed method in this paper can also be regarded as a type of regularization technique. Similar to most existing regularizers, our method effectively reduces the hypothesis space complexity by introducing critical priori knowledge, which is acquired from the metric learning algorithm.

3. Methodology

In this section, we first investigate the distribution of pairwise distances learned by the conventional CL algorithm. After that, we propose a new large-margin contrastive learning algorithm by building a distance polarization regularizer. The learning objective and the corresponding optimization algorithm are finally designed with convergence guarantee.

3.1. Understanding The Distance Distribution of CL

As we mentioned before, the key element of CL is the similarity relation between pairwise instances. For a learnable mapping $\boldsymbol{\varphi} : \mathbb{R}^m \mapsto \mathbb{R}^d$, the (squared) Euclidean distance $\|\boldsymbol{\varphi}(\boldsymbol{x}_i) - \boldsymbol{\varphi}(\boldsymbol{x}_j)\|_2^2$ measures the similarity between two original instances $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ from training data $\mathcal{X}$. Since $\boldsymbol{\varphi}(\boldsymbol{x})$ is usually normalized to reduce over-fitting, the pairwise distance in embedding space satisfies $\|\boldsymbol{\varphi}(\boldsymbol{x}_i) - \boldsymbol{\varphi}(\boldsymbol{x}_j)\|_2^2 = 2 - 2\boldsymbol{\varphi}(\boldsymbol{x}_i)^\top \boldsymbol{\varphi}(\boldsymbol{x}_j)$. For simplicity, we further denote the following normalized Euclidean distance

$$\mathcal{D}_{ij}^{\boldsymbol{\varphi}} = (1 - \boldsymbol{\varphi}(\boldsymbol{x}_i)^\top \boldsymbol{\varphi}(\boldsymbol{x}_j))/2, \quad (3)$$

which measures the similarity between instances $\boldsymbol{x}_i, \boldsymbol{x}_j \in \mathcal{X}$ with a real value $\mathcal{D}_{ij}^{\boldsymbol{\varphi}} \in [0, 1]$. Then, from both empirical and theoretical aspects, we investigate the distribution of the distance $\mathcal{D}_{ij}^{\boldsymbol{\varphi}}$ for all $1 \leq i < j \leq N$.

As we know, CL aims to repel each pair of instances away, *i.e.*, enlarging the distance $\mathcal{D}_{ij}^{\boldsymbol{\varphi}}$ to the maximal value 1 for all $1 \leq i < j \leq N$. Now, we conduct simple experiments to investigate the distribution of $\mathcal{D}_{ij}^{\boldsymbol{\varphi}}$ in the range $[0, 1]$ where $\boldsymbol{\varphi}$ is learned by a conventional CL algorithm. Specifically, here we choose the popular method SimCLR (Chen et al., 2020a) as our framework to learn the embedding $\boldsymbol{\varphi}$ on *CIFAR-10* (Krizhevsky et al., 2009) dataset using the Adam optimizer (Reddi et al., 2018). Then we gather all pairwise distances and plot the histogram in Fig. 2(a)).

From Fig. 2(a), we can clearly observe that a significant portion of the distances lie in the range of $[0, 1/2]$. It means that the finally learned embedding cannot equivalently enlarge all pairwise distances to the maximal value 1, although the learning objective of CL algorithm enforces to repel each pair of original instances in the training data.

We further provide theoretical analyses to support the above

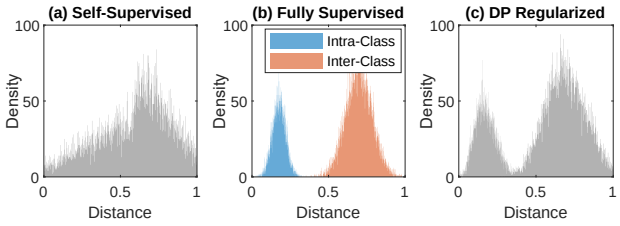

Figure 2. Distance histograms obtained by different methods on *CIFAR-10* dataset, including the conventional self-supervised CL, the fully supervised metric learning, and our proposed DP regularized CL (which is also self-supervised).

empirical observation. To be specific, we assume that the feature embedding $\boldsymbol{\varphi}(\boldsymbol{x}) = (\varphi_1(\boldsymbol{x}), \varphi_2(\boldsymbol{x}), \ldots, \varphi_d(\boldsymbol{x}))^\top$ is learned from a general hypothesis set

$$\mathcal{H} = \{\boldsymbol{\varphi} \,|\, \|\boldsymbol{\varphi}(\boldsymbol{x})\|_2 = 1 \text{ and } \varphi_i(\boldsymbol{x}) \text{ is differentiable for any } i = 1, 2, \ldots, d\}, \quad (4)$$

where $\|\boldsymbol{\varphi}(\boldsymbol{x})\|_2 = 1$ denotes that the embedding result is finally normalized for any data point $\boldsymbol{x} \in \mathbb{R}^m$. Then, we investigate the maximal value of $\mathbb{E}_{1\le i<j\le N}[\mathcal{D}_{ij}^{\boldsymbol{\varphi}}]$ where $\boldsymbol{\varphi}$ is learned from the above hypothesis set $\mathcal{H}$. For sufficiently large sample size, we have that [1]

$$\lim_{N\to\infty} \max_{\boldsymbol{\varphi}\in\mathcal{H}} \mathbb{E}_{1\le i<j\le N}\left[\mathcal{D}_{ij}^{\boldsymbol{\varphi}}\right] \le \lim_{N\to\infty} N/(2N-2) = 1/2, \quad (5)$$

which implies that the mean value of pairwise distances can be maximally enlarged to $1/2$ rather than the ideal value 1. To further investigate the overall distribution of pairwise distances, we provide the following Theorem 1 to reveal the continuity of distance distribution in the range $[0,\ 1]$, even though the intrinsic data distribution is unknown to us.

Theorem 1. *Assume that the optimal feature embedding* $\widehat{\boldsymbol{\varphi}} \in \arg\min_{\boldsymbol{\varphi}\in\mathcal{H}} \mathcal{L}_{\text{NCE}}(\boldsymbol{\varphi})$ *and the corresponding distance value* $\mathcal{D}_{ij}^{\widehat{\boldsymbol{\varphi}}} = (1 - \widehat{\boldsymbol{\varphi}}(\boldsymbol{x}_i)^\top \widehat{\boldsymbol{\varphi}}(\boldsymbol{x}_j))/2$. *Then for any given* $\mu \in [0,\ 1]$ *and* $\epsilon > 0$, *there exists sufficiently large* N *such that* $\min_{1\le i<j\le N} \{|\mathcal{D}_{ij}^{\widehat{\boldsymbol{\varphi}}} - \mu|\} < \epsilon$.

The above Theorem 1 reveals that although conventional CL algorithms repel each pair of original instances, the optimal solution of their learning objectives will still contain many small distance values in $[0, 1/2]$ (*e.g.*, the result in Fig. 2(a)), and all pairwise distances will gradually cover the whole range $[0,\ 1]$ with the increasing of sample size.

According to the above empirical and theoretical analyses, now the good news is that the conventional CL algorithms could adaptively capture the similarity and dissimilarity between pairwise instances during the repelling pairwise instances. The CL algorithms will discard some negative pairs and regard them as semantically similar pairs, even though their learning objective treat each pair of original instances as dissimilar. This can be seen as a new interpretation to understand the effectiveness of existing CL algorithms from the viewpoint of similarity metrics.

However, the bad news is that conventional CL algorithms are still not good enough since they fail to maintain a large margin in the distance space for reliable instance discrimination. As revealed in Theorem 1, the pairwise distances will gradually cover the whole region of $[0, 1]$, which makes it difficult to put the decision plane. To overcome this issue, we propose a distance polarization regularizer to constrain the learning objective of conventional CL algorithm.

3.2. Model Setup

As we revealed, the distance space $[0, 1]$ can be gradually covered by pairwise distances and thus losing a margin region to clearly discriminate the distances of similar and dissimilar pairs. However, when the supervisory information is available, the intra-class and inter-class distances obtained by metric learning algorithms should be clearly discriminated with an explicit margin region (see Fig. 2(b)), so that the metric learning algorithms can adequately capture the intrinsic features.

Actually, most metric learning methods aim to enlarge the inter-class distances and reduce the intra-class distances simultaneously, so they usually yield a margin region between the intra-class and inter-class. It means that the final distances obtained by metric learning methods should be reasonably polarized outside of an intermediate margin region, whatever the class labels are. Therefore, we employ such critical a priori to build a new regularizer which constrains the pairwise distances learned by the CL algorithms.

Distance Polarization (DP) Regularizer. We suppose that the matrix $\boldsymbol{\mathcal{D}}^{\boldsymbol{\varphi}} = [\mathcal{D}_{ij}^{\boldsymbol{\varphi}}] \in \mathbb{R}^{N\times N}$ consisting of pairwise distance $\mathcal{D}_{ij}^{\boldsymbol{\varphi}}$ measures the similarity between instances $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ for $i, j = 1, 2, \ldots, N$. We further assume that the underlying intra-class distances are smaller than δ^+ while the inter-class distances are larger than δ^-, where $0 < \delta^+ < \delta^- < 1$. Then we construct the following distance polarization (DP) regularizer

$$\mathcal{R}_0(\boldsymbol{\varphi}) = \|\min((\boldsymbol{\mathcal{D}}^{\boldsymbol{\varphi}} - \boldsymbol{\Delta}^+) \odot (\boldsymbol{\mathcal{D}}^{\boldsymbol{\varphi}} - \boldsymbol{\Delta}^-), 0)\|_0, \quad (6)$$

where $\boldsymbol{\Delta}^+ = \delta^+ \cdot \mathbf{1}_{N\times N}$ and $\boldsymbol{\Delta}^- = \delta^- \cdot \mathbf{1}_{N\times N}$ are threshold parameters. Here the region $(\delta^+, \delta^-) \subseteq [0, 1]$ can be regarded as the large margin to discriminate the similarity of data pairs. The above ℓ_0-norm (Liu et al., 2010) based regularizer will encourage the sparse distance distribution in the margin region (δ^+, δ^-), because any distance $\mathcal{D}_{ij}^{\boldsymbol{\varphi}}$ fallen into the margin region (δ^+, δ^-) will increase the value of

[1] For detailed calculations, the mean value $\mathbb{E}_{1\le i<j\le N}[\mathcal{D}_{ij}^{\boldsymbol{\varphi}}] = (\sum_{1\le i<j\le N}(1 - \boldsymbol{\varphi}(\boldsymbol{x}_i)^\top \boldsymbol{\varphi}(\boldsymbol{x}_j)))/(N(N-1)) = (\binom{N}{2} + N/2 - \|\sum_{i=1}^N \boldsymbol{\varphi}(\boldsymbol{x}_i)\|_2^2/2)/(N(N-1)) \le (\binom{N}{2} + N/2))/(N(N-1)) = (N(N-1)/2 + N/2)/(N(N-1)) = N/(2N-2)$.

$\mathcal{R}_0(\boldsymbol{\varphi})$ [2]. Thereby, minimizing such a regularizer will encourage all pairwise distances $\{\mathcal{D}^{\boldsymbol{\varphi}}_{ij}\}^N_{i,j=1}$ to distribute in the regions $[0, \delta^+]$ or $[\delta^-, 1]$, and thus adaptively separating each data pair into similar or dissimilar result (see Fig. 2(c)).

Determination of $\boldsymbol{\Delta}^+$ and $\boldsymbol{\Delta}^-$. The above DP regularizer in Eq. (6) involves two critical parameters $\boldsymbol{\Delta}^+$ and $\boldsymbol{\Delta}^-$. Here we demonstrate how to determine these two parameters. We let $\tau = \delta^- - \delta^+ \in (0,\, 1)$, and then we can regard τ as the margin width which can be easily tuned. Thereby, we just need to determine the threshold δ^-. Intuitively, we expect to employ a large δ^- to repel the dissimilar pairs of instances as far as possible, but the pairwise distances cannot be really enlarged to an ideal maximal value 1 as we discussed in Section 3.1. Here we provide the following Theorem 2 to reveal that $\delta^- = 1/2$ is a good choice to yield a margin width $\tau \in (0,\, 1/2)$.

Theorem 2. *For training data $\{\boldsymbol{x}_i\}^N_{i=1}$ with underling class labels $\{y_i\}^N_{i=1}$ and any given $\tau \in (0,\, 1/2)$, there exists a feature embedding $\overline{\boldsymbol{\varphi}} \in \mathcal{H}$ such that*

$$\max_{(i,\,j)\in I^+} \mathcal{D}^{\overline{\boldsymbol{\varphi}}}_{ij} \le 1/2 - \tau < 1/2 \le \min_{(k,\,l)\in I^-} \mathcal{D}^{\overline{\boldsymbol{\varphi}}}_{kl}, \quad (7)$$

where $y_i = 1, 2, \ldots, C$ for $i = 1, 2, \ldots, N$ and $C < d$. Here the bivariate index sets $I^+ = \{(i,\, j)|y_i = y_j,\, i, j = 1, 2, \ldots, N\}$ and $I^- = \{(i,\, j)|y_i \neq y_j,\, i, j = 1, 2, \ldots, N\}$.

With the above Theorem 2, we can easily implement the proposed DP regularizer and deploy it in the learning objective of conventional CL algorithms. Without loss of generality, for most existing CL models equipped with NCE loss $\mathcal{L}_{\text{NCE}}(\boldsymbol{\varphi})$ in Eq. (1), we build the following large-margin contrastive learning (LMCL) model

$$\min_{\boldsymbol{\varphi}\in\mathcal{H}} \mathcal{L}_{\text{NCE}}(\boldsymbol{\varphi}) + \lambda \mathcal{R}_0(\boldsymbol{\varphi}), \quad (8)$$

where the regularization parameter $\lambda > 0$ is tuned by users. As a regularized learning objective, LMCL is simple and generic because here the loss term $\mathcal{L}_{\text{NCE}}(\boldsymbol{\varphi})$ can be implemented by many existing CL algorithms. In the next subsection, we show that Eq. (8) can be easily solved by existing stochastic optimization methods.

3.3. Optimization

Minimizing the objective function in Eq. (8) is a classical ℓ_0-norm optimization problem which is usually non-continuous and non-convex. Fortunately, for the original ℓ_0-norm based regularizer Eq. (6), here we can easily find that $\mathcal{D}^{\boldsymbol{\varphi}}_{ij} - \delta^+ \in (0, 1)$ and $\delta^- - \mathcal{D}^{\boldsymbol{\varphi}}_{ij} \in (0, 1)$ for any $i, j = 1, 2, \ldots, N$, so we have that $\min((\boldsymbol{\mathcal{D}}^{\boldsymbol{\varphi}} - \boldsymbol{\Delta}^+) \odot (\boldsymbol{\mathcal{D}}^{\boldsymbol{\varphi}} - \boldsymbol{\Delta}^-), 0) \in [0,\, 1]^{N\times N}$. As the ℓ_1-norm is a *convex envelope* of ℓ_0-norm in the unit hypercube $[0,\, 1]^{N\times N}$, we can simply convert the ℓ_0-norm based regularizer in Eq. (6) to the ℓ_1-norm based form $\mathcal{R}_1(\boldsymbol{\varphi})$ [3] which is a good approximation to ℓ_0-norm in the unit hypercube. By integrating such a differentiable almost everywhere (*a.e.*) function, we finally have the following learning objective $\mathcal{F}(\boldsymbol{\varphi})$

$$\min_{\boldsymbol{\varphi}\in\mathcal{H}} \{\mathcal{F}(\boldsymbol{\varphi}) = \mathcal{L}_{\text{NCE}}(\boldsymbol{\varphi}) + \lambda \mathcal{R}_1(\boldsymbol{\varphi})\}. \quad (9)$$

For the above objective function, we show that it can be solved by existing stochastic optimization methods. For $n + 1$ (*i.e.*, the batch size) randomly selected data point $\{\boldsymbol{x}_{b_j}|\boldsymbol{x}_{b_j} \in \mathcal{X}, b_j \in B\}^{n+1}_{j=1}$, the NCE loss defined by Eq. (1) already has a stochastic form [4], so here we only need to demonstrate the stochastic regularizer in a mini-batch, *i.e.*,

$$\begin{aligned}
&\mathcal{R}_1(\boldsymbol{\varphi}) \\
&= \frac{2}{\binom{N}{n}} \sum\nolimits_{\boldsymbol{b}\in B} \sum\nolimits_{j=1}^{n+1} |\min((\mathcal{D}^{\boldsymbol{\varphi}}_{b_i b_j} - \delta^+) \odot (\mathcal{D}^{\boldsymbol{\varphi}}_{b_i b_j} - \delta^-), 0)| \\
&= \frac{1}{\binom{N}{n+1}} \sum\nolimits_{\boldsymbol{b}\in B} r(\boldsymbol{\varphi}; \{\boldsymbol{x}_{b_j}\}^{n+1}_{j=1}), \quad (10)
\end{aligned}$$

and thus $\mathcal{F}(\boldsymbol{\varphi})$ in Eq. (9) has the stochastic form $f(\boldsymbol{\varphi}; \{\boldsymbol{x}_{b_j}\}^{n+1}_{j=1}) = \ell(\boldsymbol{\varphi}; \{\boldsymbol{x}_{b_j}\}^{n+1}_{j=1}) + \lambda r(\boldsymbol{\varphi}; \{\boldsymbol{x}_{b_j}\}^{n+1}_{j=1})$. Based on such a stochastic loss, we further provide the Adam iteration steps to solve Eq. (9) in Algorithm 1.

In summary, introducing the DP regularizer merely incurs

Algorithm 1 Solving Eq. (9) via Adam.

Input: Training Data $\mathcal{X} = \{\boldsymbol{x}_i\}^N_{i=1}$; Step Size $\eta > 0$; Regularization Parameter $\lambda > 0$; Batch Size $n \in \mathbb{N}_+$.
Initialize: Momentum Vectors $\boldsymbol{m}_{(0)} = \boldsymbol{v}_{(0)} = \boldsymbol{0}$; Decay Rates $\alpha_1, \alpha_2 \in (0, 1)$; Iteration Number $t = 0$.
For t **from** 1 **to** T**:**

1). Uniformly pick $(n+1)$ data points $\{\boldsymbol{x}_{b_j}\}^{n+1}_{j=1}$ from $\mathcal{X}$;

2). Compute the stochastic gradient via Eq. (10):

$$\boldsymbol{g}_{(t)} \leftarrow \nabla_{\boldsymbol{\varphi}}(\ell(\boldsymbol{\varphi}; \{\boldsymbol{x}_{b_j}\}^{n+1}_{j=1}) + \lambda r(\boldsymbol{\varphi}; \{\boldsymbol{x}_{b_j}\}^{n+1}_{j=1})); \quad (11)$$

3). Compute moment vectors: $\boldsymbol{m}_{(t+1)} \leftarrow \alpha_1 \boldsymbol{m}_t + (1 - \alpha_1)\boldsymbol{g}_{(t)}$, and $\boldsymbol{v}_{(t+1)} \leftarrow \alpha_2 \boldsymbol{v}_t + (1 - \alpha_2)\boldsymbol{g}_{(t)} \odot \boldsymbol{g}_{(t)}$;

4). Update the learning parameter:

$$\boldsymbol{\varphi}_{(t+1)} \leftarrow \boldsymbol{\varphi}_{(t)} - \eta \frac{\boldsymbol{m}_{(t+1)}/(1 - \alpha_1^{t+1})}{\sqrt{\boldsymbol{v}_{(t+1)}/(1 - \alpha_2^{t+1}) + \epsilon}}; \quad (12)$$

End.
Output: The converged $\widetilde{\boldsymbol{\varphi}}$.

[2] Any distance $\mathcal{D}^{\boldsymbol{\varphi}}_{ij}$ fallen into the margin region $(\delta^+,\, \delta^-)$ will incur the negative product $(\mathcal{D}^{\boldsymbol{\varphi}}_{ij} - \delta^+)(\mathcal{D}^{\boldsymbol{\varphi}}_{ij} - \delta^-)$, and thereby leading to that $\min((\mathcal{D}^{\boldsymbol{\varphi}}_{ij} - \delta^+)(\mathcal{D}^{\boldsymbol{\varphi}}_{ij} - \delta^-), 0) \neq 0$ which increases the value of ℓ_0-norm as well as the value of regularizer $\mathcal{R}_0(\boldsymbol{\varphi})$.

[3] Here $\mathcal{R}_1(\boldsymbol{\varphi}) = \|\min((\boldsymbol{\mathcal{D}}^{\boldsymbol{\varphi}} - \boldsymbol{\Delta}^+) \odot (\boldsymbol{\mathcal{D}}^{\boldsymbol{\varphi}} - \boldsymbol{\Delta}^-), 0)\|_1$.

[4] Here the NCE loss $\mathcal{L}_{\text{NCE}}(\boldsymbol{\varphi}) = \mathbb{E}[\ell(\boldsymbol{\varphi}; \{\boldsymbol{x}_{b_j}\}^{n+1}_{j=1})]$, and the corresponding stochastic loss $\ell(\boldsymbol{\varphi}; \{\boldsymbol{x}_{b_j}\}^{n+1}_{j=1}) = -\log(\exp(\boldsymbol{\varphi}(\boldsymbol{x}_{b_{n+1}})^\top \boldsymbol{\varphi}(\boldsymbol{x}^+_{b_{n+1}}))/(\exp(\boldsymbol{\varphi}(\boldsymbol{x}_{b_{n+1}}))^\top \boldsymbol{\varphi}(\boldsymbol{x}^+_{b_{n+1}})) + \sum_{j=1}^n \exp(\boldsymbol{\varphi}(\boldsymbol{x}_{b_j}))^\top \boldsymbol{\varphi}(\boldsymbol{x}^-_{b_j}))))$. The index vector set $B = \{\boldsymbol{b} = (b_1, \ldots, b_{n+1})^\top | b_i, b_j = 1, \ldots, N,\, b_i \neq b_j,\, i, j = 1, \ldots, n+1\}$.

an additional stochastic gradient in Eq. (11). It means that our method can be easily implemented in most existing CL methods and only introduces very little computational overheads. Furthermore, the convergence of Adam has been well studied in previous works (Zaheer et al., 2018). It can be verified that $\ell(\boldsymbol{\varphi}; \{\boldsymbol{x}_{b_j}\}_{j=1}^{n+1})$ and $r(\boldsymbol{\varphi}; \{\boldsymbol{x}_{b_j}\}_{j=1}^{n+1})$ are both Lipschitz-smooth and gradient-bounded, as long as the embedding $\boldsymbol{\varphi}$ is Lipschitz-smooth and gradient-bounded. In this case, the iteration sequence $\boldsymbol{\varphi}_{(1)}, \ldots, \boldsymbol{\varphi}_{(T)}$ in Algorithm 1 converges to a stationary point of the learning objective $\mathcal{F}$ with a convergence rate $\mathcal{O}(1/\sqrt{T})$, where T is the number of iterations (Huang et al., 2019; 2020).

4. Theoretical Analyses

In this section, we further provide in-depth theoretical analyses for our proposed method. We first investigate the reliability of our method for similarity measure. After that, we demonstrate the generalizability of our method on the downstream classification task.

4.1. Error Bound for Similarity Measure

In general, CL usually considers the similarity between pairwise instances, so the reliability of CL algorithms depends on whether the pairwise similarity can be faithfully measured. Here we follow the common practice in learning theory (Xie et al., 2017) to study the error bound determined by the minimizer of our learning objective in Eq. (9). Specifically, we investigate the correctness of pairwise distances $\mathcal{D}_{ij}^{\boldsymbol{\varphi}^*}$ by building the expectations $\mathbb{E}_{y_i \neq y_j}[\max(\delta_\mu^- - \mathcal{D}_{ij}^{\boldsymbol{\varphi}^*}, 0)]$ and $\mathbb{E}_{y_k = y_l}[\max(\mathcal{D}_{kl}^{\boldsymbol{\varphi}^*} - \delta_\mu^+, 0)]$ to evaluate the false negatives and false positives, respectively. The corresponding error bound is provided in Theorem 3.

Theorem 3. *Assume that* $\boldsymbol{\varphi}^* \in \arg\min_{\boldsymbol{\varphi} \in \mathcal{H}} \mathcal{L}_{\text{NCE}}(\boldsymbol{\varphi}) + \lambda \mathcal{R}_1(\boldsymbol{\varphi})$*, and the underling class labels of training data* $\{\boldsymbol{x}_i\}_{i=1}^N$ *are* $\{y_i\}_{i=1}^N$*. Then we have that*

$$\begin{aligned} &\mathbb{E}_{y_i \neq y_j}[\max(\delta_\mu^- - \mathcal{D}_{ij}^{\boldsymbol{\varphi}^*}, 0)] + \mathbb{E}_{y_k = y_l}[\max(\mathcal{D}_{kl}^{\boldsymbol{\varphi}^*} - \delta_\mu^+, 0)] \\ &\leq (\delta^- - \delta^+)\mathcal{R}_1(\boldsymbol{\varphi}^*) + (K_{\max}/K_{\min})/C \\ &\leq 4(\delta^- - \delta^+)/\lambda + (K_{\max}/K_{\min})/C, \end{aligned} \tag{13}$$

where the constants $\delta_\mu^- = \delta^- - \mu$*,* $\delta_\mu^+ = \delta^+ + \mu$*,* $\mu \in (0, \delta^- - \delta^+)$*,* $K_{\min} = \min_{1 \leq k \leq C} \|\boldsymbol{y} - k \cdot \mathbf{1}_{N \times 1}\|_0$*, and* $K_{\max} = \max_{1 \leq k \leq C} \|\boldsymbol{y} - k \cdot \mathbf{1}_{N \times 1}\|_0$*.*

The above Eq. (13) clearly reveals that the error bound of the similarities measured by our method will gradually converge to 0 with the increasing of class number C and the decreasing of the regularizer value $\mathcal{R}_1(\boldsymbol{\varphi}^*)$. Firstly, it implies that the diversity of data (*i.e.*, a large C) will benefit the reliability of the similarity measured by CL algorithms. This conclusion is consistent with existing theoretical findings that the larger C leads to the better generalizability (Saunshi et al., 2019). Secondly, such an error bound also relies on a small regularizer value $\mathcal{R}_1(\boldsymbol{\varphi}^*)$. This demonstrates the necessity and usefulness of our proposed DP regularizer, because increasing the regularization parameter λ would assist the error bound in converging to zero.

4.2. Error Bound for Downstream Classification

The experimental performance of most CL algorithms is usually evaluated by a downstream classification task. Therefore, here we provide the generalization error bound (GEB) of our method for the classification task which trains a softmax classifier by minimizing the traditional cross entropy loss (Zhang & Sabuncu, 2018), *i.e.*, $\mathcal{L}_{\text{SM}}(\boldsymbol{\varphi}; \mathcal{X}) = \inf_{\boldsymbol{W} \in \mathbb{R}^{C \times d}} \mathcal{L}_{\text{CEP}}(\boldsymbol{W}\boldsymbol{\varphi}; \mathcal{X})$. For a feature embedding $\boldsymbol{\varphi}$, the generalization error is defined by $\mathcal{L}_{\text{SM}}^{\mathcal{T}}(\boldsymbol{\varphi}) = \mathbb{E}_{\mathcal{X} \sim \mathcal{T}}[\mathcal{L}_{\text{SM}}(\boldsymbol{\varphi}; \mathcal{X})]$, where $\mathcal{T}$ is the underlying distribution of the training data $\mathcal{X}$. Then we investigate how such a generalization error $\mathcal{L}_{\text{SM}}^{\mathcal{T}}(\boldsymbol{\varphi})$ is far from the learning objective $\mathcal{L}_{\text{NCE}}(\boldsymbol{\varphi})$ of contrastive learning.

Theorem 4. *Let* $\boldsymbol{\varphi}^* \in \arg\min_{\boldsymbol{\varphi} \in \mathcal{H}} \mathcal{L}_{\text{NCE}}(\boldsymbol{\varphi}) + \lambda \mathcal{R}_1(\boldsymbol{\varphi})$*. Then with probability at least* $1 - \delta$*, we have that*

$$\left|\mathcal{L}_{\text{SM}}^{\mathcal{T}}(\boldsymbol{\varphi}^*) - \mathcal{L}_{\text{NCE}}(\boldsymbol{\varphi}^*)\right| \leq \mathcal{O}\left(\frac{Q_1 \mathfrak{R}_{\mathcal{H}}(\lambda)}{N} + \sqrt{\frac{Q_2}{N}}\right), \tag{14}$$

where $Q_1 = \sqrt{1 + 1/n}$*,* $Q_2 = \log(1/\delta) \cdot \log^2(n)$*, and*[5] $\mathfrak{R}_{\mathcal{H}}(\lambda)$ *is monotonically decreasing w.r.t.* λ*.*

We can observe that the error bound in Eq. (14) gradually decreases with the increase of the training sample size N, and this is consistent with the traditional supervised learning method (Niu et al., 2016). Then, we find that the negative pair size n in the error term $\sqrt{Q_2/N}$ is negligible for the large sample size N. In this case, the relative large negative pair size n will effectively reduce the first error term $Q_1(\mathfrak{R}_{\mathcal{H}}(\lambda)/N)$, and thereby tightening the error bound. This conclusion is also in line with the empirical observations in existing works (He et al., 2020; Kim et al., 2020). Finally, when we enlarge the regularization parameter λ, the rademacher complexity $\mathfrak{R}_{\mathcal{H}}(\lambda)$ will also be decreased, and thus further reducing the error bound and improving the generalizability of contrastive learning algorithm.

5. Experimental Results

In this section, we show experimental results on both synthetic and real-world datasets to validate the effectiveness of our proposed method. In detail, we first give visualization results on synthetic data to demonstrate the efficacy of DP regularizer. Then, we compare our proposed learning algorithm with existing state-of-the-art models on vision and

[5] To be specific, here the Rademacher Complexity $\mathfrak{R}_{\mathcal{H}}(\lambda) = \mathbb{E}_{\boldsymbol{\sigma} \in \{\pm 1\}^{3dN}}[\sup_{\boldsymbol{\varphi} \in \mathcal{H}(\lambda)} \langle \boldsymbol{\sigma}, \boldsymbol{f} \rangle]$, in which the restricted hypothesis space $\mathcal{H}(\lambda) = \{\boldsymbol{\varphi} | \boldsymbol{\varphi} \in \mathcal{H}, \text{ and } \mathcal{R}_1(\boldsymbol{\varphi}) \leq 4/\lambda\}$.

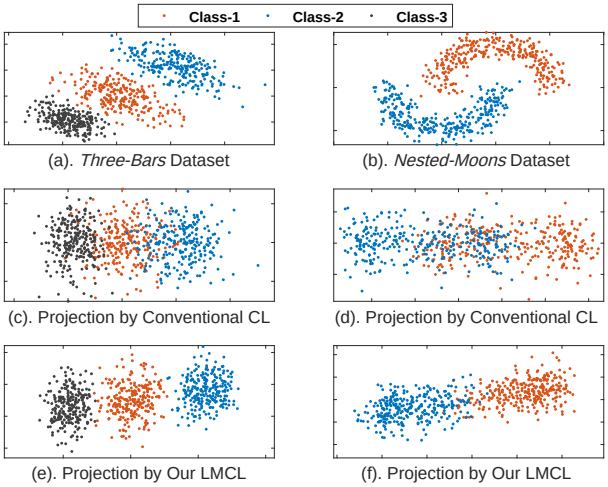

Figure 3. Visualization results of the conventional CL method and our proposed LMCL method on the two toy datasets.

Table 1. K-means clustering accuracy rates (mean $\pm$ std) of baseline methods and our proposed method on the toy datasets.

METHOD	*Three-Bars*	*Nested-Moons*	t-test
Euclidean Space	75.2 ± 1.2	77.3 ± 2.3	✓
Conventional CT	78.3 ± 2.2	77.5 ± 1.2	✓
LMCT (Ours)	$\mathbf{84.2 \pm 0.2}$	$\mathbf{85.2 \pm 2.3}$	–

language tasks. Finally, we test our method on the CL based reinforcement learning task. The regularization parameter λ of our method is fixed to 0.1. The thresholds δ^+ and δ^- are fixed to 0.1 and 0.5, respectively. The hyper-parameters of compared methods are set to the recommended values according to their original papers.

5.1. Experiments on Synthetic Data

We first consider learning a linear embedding $\varphi(\boldsymbol{x}) = \boldsymbol{P}\boldsymbol{x}$ on two-dimensional synthetic data, where the matrix $\boldsymbol{P} \in \mathbb{R}^{2\times 2}$ is the learning parameter. Here we employ the *Three-Bars* and *Nested-Moons* datasets (Chen et al., 2018) to evaluate the performance of the conventional CL algorithm and our proposed LMCL algorithm. For each data point in the two datasets (see Fig. 3(a) and (b)), we build its data augmentation by adding Gaussian noise on the original data point. Then, we simply regard each data point and its augmentation as a positive pair, and sampling every two data points as a negative pair. For these positive pairs and negative pairs, we use the Adam optimizer (learning rate = 0.001) for both the conventional CL (*i.e.*, Eq. (1)) and our proposed LMCL (*i.e.*, Eq. (9) with $\lambda = 0.1$). Both the projection matrices of conventional method and our method (*i.e.*, $\boldsymbol{P}_{\text{CL}}, \boldsymbol{P}_{\text{LM}} \in \mathbb{R}^{2\times 2}$) are initialized by 0. After obtaining the learned matrices $\boldsymbol{P}^*_{\text{CL}}$ and $\boldsymbol{P}^*_{\text{LM}}$, we record the projected points $\boldsymbol{P}^*_{\text{CL}}\boldsymbol{x}$ and $\boldsymbol{P}^*_{\text{LM}}\boldsymbol{x}$ to visualize the distribution of data points in embedding space.

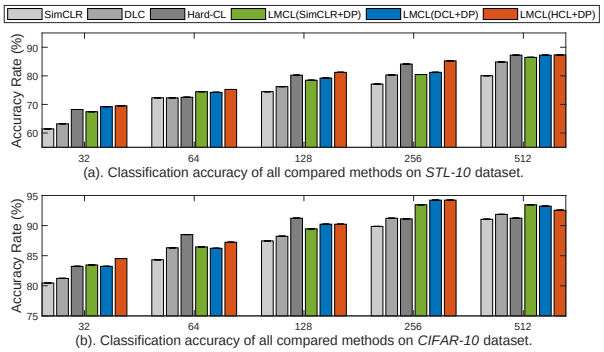

Figure 4. Classification accuracy of all methods on *STL-10* and *CIFAR-10* datasets. The negative sample size is from 32 to 512.

We can clearly obverse that although the conventional CL algorithm finds out the projection matrix $\boldsymbol{P}^*_{\text{CL}}$ to roughly distinguish each class of data points (as shown in Fig. 3(c) and (d)), it still yields many ambiguous points between each two classes in the embedding (projection) space. In comparison, when the DP regularizer is employed, our method LMCL could further improve the separability of data points and successfully obtain unambiguous projected points between each of the two classes (Fig. 3(e) and (f)). Furthermore, the K-means (Bradley & Fayyad, 1998) clustering accuracy (mean $\pm$ std, 20 random trials) of conventional CL and our LMCL are reported in Tab. 1, and we can obverse that our LMCL consistently outperforms the conventional CL algorithm. We also perform the t-test at significance level 0.05 in the last column, and "✓" indicates that our method is significantly better than the baseline method.

5.2. Experiments on Image Classification

In this subsection, we validate the effectiveness of our method on the image classification task. Here we select SimCLR (Chen et al., 2020a) and contrastive multiview coding (CMC) (Tian et al., 2020a) as baseline methods, and implement our method LMCL under such two classical frameworks. We also compare our method with three additional state-of-the-art methods including debiased contrastive learning (DCL) (Chuang et al., 2020), hard negative based contrastive learning (HCL) (Robinson et al., 2020), and the clustering based method (SwAV) (Caron et al., 2020) on *STL-10* (Coates et al., 2011), *CIFAR-10* (Krizhevsky et al., 2009), and *ImageNet-100* (Russakovsky et al., 2015) datasets. All methods are fairly implemented by the *ResNet50* with the same training epoch 100.

For *STL-10* and *CIFAR-10* datasets, we record the classification accuracy of all compared methods with varying numbers of negative sample. From Fig. 4, we can clearly observe that our method LMCL (DP+SimCLR) successfully improves the baseline for at least 1% and 2% on *CIFAR-10* dataset and *STL-10* dataset, respectively. Similar experi-

Table 2. Classification accuracy (%) of all methods on *ImageNet-100* dataset with negative sample size 1024 and 4096.

METHOD	1024		4096	
	Top1	Top5	Top1	Top5
CMC	60.23	79.23	73.58	92.06
SwAV	60.93	79.43	75.78	92.86
DCL	61.01	78.99	74.60	92.08
HCL	60.89	79.33	74.66	92.32
LMLC(CMC+DP)	**61.23**	**79.44**	75.67	**93.02**
LMLC(DCL+DP)	61.12	79.20	**75.89**	92.89
LMLC(HCL+DP)	60.92	79.43	74.94	92.39

Table 3. Parametric sensitivities of λ and τ. Here λ and τ are changed in $[0.01, 5]$ and $[0.1, 0.4]$, respectively.

λ \ τ	**0.1**	**0.2**	**0.25**	**0.3**	**0.4**
0.01	80.4	81.3	81.2	81.2	80.8
0.1	81.5	81.9	81.7	81.8	81.9
0.5	81.6	81.6	80.7	81.7	81.9
5	80.9	81.9	80.9	80.6	80.5

ments are conducted on *ImageNet-100* dataset, and Tab. 2 shows that our method improves the baseline method CMC from 73.58% to 75.88%. For different negative sample sizes, the accuracy rates of our method are competitive or superior to the compared methods DCL and HCL, which clearly demonstrates the effectiveness of our method. Furthermore, our method can also be incorporated by the two existing methods (*i.e.*, DP+DCL and DP+HCL) to achieve the improved recognition accuracy. Therefore, our method has good compatibility with existing CL algorithms on the image classification task.

Parametric Sensitivity. Here we further investigate the parametric sensitivities on λ and τ. Specifically, we change λ and τ in $[0.01, 5]$ and $[0.1, 0.4]$ respectively, and record the classification accuracy of our method on *STL-10* dataset (BatchSize=256). Tab. 3 shows that the accuracy variation of our method is smaller than 1.5, so the hyper-parameters of our method can be easily tuned in practice use.

5.3. Experiments on Sentence Representation

In this subsection, we employ the *BookCorpus* dataset (Kiros et al., 2015) to evaluate the performance of all compared methods on six text classification tasks, including movie review sentiment (*MR*), product reviews (*CR*), subjectivity classification (*SUBJ*), opinion polarity (*MPQA*), question type classification (*TREC*), and paraphrase identification (*MSRP*). We follow the experimental settings in the

Table 4. Classification accuracy (%) of all methods on *BookCorpus* dataset including six text classification tasks.

METHOD	*MR*	*CR*	*SUBJ*	*MPQA*	*TREC*	*MSRP*
QT	76.8	81.3	86.6	93.4	89.8	73.6
DCL	76.2	82.9	86.9	93.7	89.1	74.7
HCL	77.4	83.6	86.8	93.4	88.7	73.5
LMCL(QT+DP)	77.3	82.3	86.9	93.7	**90.2**	74.1
LMCL(DCL+DP)	77.2	**83.7**	**87.2**	93.8	90.1	**75.1**
LMCL(HCL+DP)	**78.1**	83.5	**87.2**	**94.0**	89.1	74.2

Table 5. 100K Scores (mean $\pm$ std, 3 random trials) achieved by all methods on the six control tasks.

METHOD	*Spin*	*Swingup*	*Easy*	*Run*	*Walk*	*Catch*
CURL	413±53	680±32	908±86	**298±38**	621±121	826±42
DCL	422±23	672±52	878±96	248±98	**626±98**	836±12
HCL	420±61	678±82	869±116	268±42	623±26	819±62
LMCL(CURL+DP)	**423±63**	682±13	**926±73**	296±32	625±53	842±27
LMCL(DCL+DP)	**423±33**	**683±93**	909±87	287±67	625±93	**843±37**
LMCL(HCL+DP)	421±51	**681±83**	910±95	292±78	**626±89**	832±83

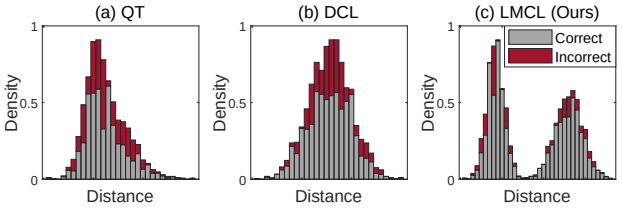

Figure 5. Distance histograms obtained by different methods (QT, DCL, and our proposed LMCL) on *BookCorpus* dataset.

baseline method quick-thought (QT) (Logeswaran & Lee, 2018), which chooses the neighboring sentences as positive pairs. Here the 10-fold cross validation is adopted, and the average classification accuracy is listed in Tab. 4.

For the six classification tasks, our method improves the classification accuracy of baseline method QT for at least one percentage on most classification benchmarks. The distance histograms of QT, DCL, and our LMCL are shown in Fig. 5. We clearly observe that our method obtains the more accurate distance determination than baseline methods, and this reveals that our method is effective for the text classification task.

5.4. Experiments on Reinforcement Learning

This subsection further extends our experiments on reinforcement learning task, which is another application scenario of contrastive learning. Here the contrastive unsupervised representations for reinforcement learning (CURL) (Laskin et al., 2020) method is employed to perform image-based policy control on representation learned by the CL algorithm. All methods are tested on the DeepMind control suite (Tassa et al., 2018), which consists of six control tasks listed in Tab. 5. By following the experimental settings in CURL, the positive pair is built by simply cropping a single image, and the negative pair is composed of each two images in the control sequence. All methods are retrained for 3 times, and the corresponding means and standards of 100K scores are shown in Tab. 5.

For the six control tasks, our method consistently outperforms the baseline method CURL with higher means. When compared to DCL and HCL methods, our method achieves better results in most cases. Although our method LMCL (DP+CURL) has slightly lower scores than DCL or HCL on the last two control tasks, our method shows smaller variance. Moreover, when we incorporate our DP regularizer

to DCL and HCL, our method could further improve the overall scores of compared methods on the six tasks. This also reveals that our method is compatible with existing CL algorithms on the reinforcement learning task.

6. Conclusion

In this paper, we first revealed that existing CL algorithms fail to maintain a margin region in the distance space to discriminate the semantically similar and dissimilar data pairs. To overcome such an issue, we proposed a distance polarization (DP) regularizer, which encourages the polarized distances and thus obtaining a large margin in the distance space in an unsupervised way. To the best of our knowledge, this is the first work in CL that considers introducing a margin region in the distance space. We conducted intensive theoretical analyses to guarantee the effectiveness of our method. Visualization experiments on toy data and comparison experiments on real-world datasets across multiple domains indicate that our learning algorithm acquires more reliable feature embedding than state-of-the-art methods. Considering the uncertainty of similarity determination in the distance polarization would be interesting future work.

Acknowledgment

SC, GN, and MS were supported by JST AIP Acceleration Research Grant Number JPMJCR20U3, Japan. MS was also supported by the Institute for AI and Beyond, UTokyo. CG, JL, and JY were supported by NSFC 62072242, 61973162, 61836014, U19B2034, and U1713208, Program for Changjiang Scholars, China Postdoctoral Science Foundation (No: 2020M681606), the Fundamental Research Funds for the Central Universities (No: 30920032202), and CCF-Tencent Open Fund (No: RAGR20200101).

References

Arpit, D., Zhou, Y., Ngo, H., and Govindaraju, V. Why regularized auto-encoders learn sparse representation? In *International Conference on Machine Learning (ICML)*, pp. 136–144, 2016. 2.3

Bradley, P. S. and Fayyad, U. M. Refining initial points for k-means clustering. In *International Conference on Machine Learning (ICML)*, volume 98, pp. 91–99, 1998. 5.1

Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., and Joulin, A. Unsupervised learning of visual features by contrasting cluster assignments. In *Advances in neural information processing systems (NeurIPS)*, pp. 1401–1413, 2020. 5.2

Chen, R. T., Rubanova, Y., Bettencourt, J., and Duvenaud, D. K. Neural ordinary differential equations. In *Advances in neural information processing systems (NeurIPS)*, pp. 6571–6583, 2018. 5.1

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning (ICML)*, pp. 1597–1607, 2020a. 1, 2.1, 3.1, 5.2

Chen, X., Chen, W., Chen, T., Yuan, Y., Gong, C., Chen, K., and Wang, Z. Self-pu: Self boosted and calibrated positive-unlabeled training. In *International Conference on Machine Learning (ICML)*, pp. 1510–1519, 2020b. 1, 2.1

Chu, X., Lin, Y., Wang, Y., Wang, X., Yu, H., Gao, X., and Tong, Q. Distance metric learning with joint representation diversification. In *International Conference on Machine Learning (ICML)*, pp. 1962–1973, 2020. 2.1, 2.2

Chuang, C.-Y., Robinson, J., Yen-Chen, L., Torralba, A., and Jegelka, S. Debiased contrastive learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 33, 2020. 1, 2.1, 5.2

Coates, A., Ng, A., and Lee, H. An analysis of single-layer networks in unsupervised feature learning. In *International conference on artificial intelligence and statistics (AISTATS)*, pp. 215–223, 2011. 5.2

Davis, J. V., Kulis, B., Jain, P., Sra, S., and Dhillon, I. S. Information-theoretic metric learning. In *International Conference on Machine learning (ICML)*, pp. 209–216, 2007. 2.2

Dong, W., Shi, G., Li, X., Ma, Y., and Huang, F. Compressive sensing via nonlocal low-rank regularization. *IEEE Transactions on Image Processing*, 23(8):3618–3632, 2014. 2.3

Guo, Z.-C., Shi, L., and Wu, Q. Learning theory of distributed regression with bias corrected regularization kernel network. *The Journal of Machine Learning Research*, 18(1):4237–4261, 2017. 2.3

Gutmann, M. and Hyvärinen, A. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 297–304, 2010. 2.1

He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. Momentum contrast for unsupervised visual representation learning. In *IEEEConference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9729–9738, 2020. 2.1, 4.2

Huang, F., Chen, S., and Huang, H. Faster stochastic alternating direction method of multipliers for nonconvex optimization. In *International Conference on Machine Learning (ICML)*, pp. 2839–2848, 2019. 3.3

Huang, F., Gao, S., Pei, J., and Huang, H. Accelerated zeroth-order momentum methods from mini to minimax optimization. *arXiv preprint arXiv:2008.08170*, 2020. 3.3

Huynh, T., Kornblith, S., Walter, M. R., Maire, M., and Khademi, M. Boosting contrastive self-supervised learning with false negative cancellation. *arXiv preprint arXiv:2011.11765*, 2020. 1

Jing, L. and Tian, Y. Self-supervised visual feature learning with deep neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 1

Kim, M., Tack, J., and Hwang, S. J. Adversarial self-supervised contrastive learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 33, 2020. 4.2

Kiros, R., Zhu, Y., Salakhutdinov, R. R., Zemel, R., Urtasun, R., Torralba, A., and Fidler, S. Skip thought vectors. *Advances in neural information processing systems (NeurIPS)*, 28:3294–3302, 2015. 5.3

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. 2009. 3.1, 5.2

Krogh, A. and Hertz, J. A. A simple weight decay can improve generalization. In *Advances in neural information processing systems (NeurIPS)*, pp. 950–957, 1992. 2.3

Laskin, M., Srinivas, A., and Abbeel, P. Curl: Contrastive unsupervised representations for reinforcement learning. In *International Conference on Machine Learning (ICML)*, pp. 5639–5650, 2020. 5.4

Li, J., Zhou, P., Xiong, C., Socher, R., and Hoi, S. C. Prototypical contrastive learning of unsupervised representations. *arXiv preprint arXiv:2005.04966*, 2020. 2.1

Liu, G., Lin, Z., and Yu, Y. Robust subspace segmentation by low-rank representation. In *International conference on machine learning (ICML)*, pp. 663–670, 2010. 3.2

Logeswaran, L. and Lee, H. An efficient framework for learning sentence representations. In *International Conference on Learning Representations (ICLR)*, 2018. 1, 5.3

Niu, G., du Plessis, M. C., Sakai, T., Ma, Y., and Sugiyama, M. Theoretical comparisons of positive-unlabeled learning against positive-negative learning. In *Advances in neural information processing systems (NeurIPS)*, pp. 1199–1207, 2016. 4.2

Oord, A. v. d., Li, Y., and Vinyals, O. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 1

Reddi, S. J., Kale, S., and Kumar, S. On the convergence of adam and beyond. In *International Conference on Learning Representations (ICLR)*, 2018. 3.1

Robinson, J., Chuang, C.-Y., Sra, S., and Jegelka, S. Contrastive learning with hard negative samples. *arXiv preprint arXiv:2010.04592*, 2020. 1, 2.1, 5.2

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3): 211–252, 2015. 5.2

Saunshi, N., Plevrakis, O., Arora, S., Khodak, M., and Khandeparkar, H. A theoretical analysis of contrastive unsupervised representation learning. In *International Conference on Machine Learning (ICML)*, pp. 5628–5637, 2019. 1, 1, 4.1

Scholkopf, B. and Smola, A. J. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. Adaptive Computation and Machine Learning series, 2018. 2.3

Sohn, K. Improved deep metric learning with multi-class n-pair loss objective. *Advances in neural information processing systems (NeurIPS)*, 29:1857–1865, 2016. 2.2

Song, J. and Ermon, S. Multi-label contrastive predictive coding. *Advances in Neural Information Processing Systems (NeurIPS)*, 33, 2020. 1

Tassa, Y., Doron, Y., Muldal, A., Erez, T., Li, Y., Casas, D. d. L., Budden, D., Abdolmaleki, A., Merel, J., Lefrancq, A., et al. Deepmind control suite. *arXiv preprint arXiv:1801.00690*, 2018. 5.4

Tian, Y., Krishnan, D., and Isola, P. Contrastive multiview coding. In *European Conference on Computer Vision (ECCV)*, pp. 1–18, 2020a. 1, 5.2

Tian, Y., Sun, C., Poole, B., Krishnan, D., Schmid, C., and Isola, P. What makes for good views for contrastive learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 33, 2020b. 1

Weinberger, K. Q., Blitzer, J., and Saul, L. K. Distance metric learning for large margin nearest neighbor classification. In *Advances in neural information processing systems (NeurIPS)*, pp. 1473–1480, 2006. 1

Wu, Z., Xiong, Y., Yu, S. X., and Lin, D. Unsupervised feature learning via non-parametric instance discrimination.

In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3733–3742, 2018. 1, 1, 2.1

Xie, P., Deng, Y., Zhou, Y., Kumar, A., Yu, Y., Zou, J., and Xing, E. P. Learning latent space models with angular constraints. In *International Conference on Machine Learning (ICML)*, pp. 3799–3810, 2017. 4.1

Yang, Y., Shen, H. T., Ma, Z., Huang, Z., and Zhou, X. L2, 1-norm regularized discriminative feature selection for unsupervised learning. In *International joint conference on artificial intelligence (IJCAI)*, 2011. 2.3

Yu, B. and Tao, D. Deep metric learning with tuplet margin loss. In *IEEE International Conference on Computer Vision (ICCV)*, pp. 6490–6499, 2019. 2.2

Zaheer, M., Reddi, S., Sachan, D., Kale, S., and Kumar, S. Adaptive methods for nonconvex optimization. In *Advances in neural information processing systems (NeurIPS)*, pp. 9793–9803, 2018. 3.3

Zhang, Z. and Sabuncu, M. Generalized cross entropy loss for training deep neural networks with noisy labels. *Advances in neural information processing systems (NeurIPS)*, 31:8778–8788, 2018. 4.2

Zhong, H., Chen, C., Jin, Z., and Hua, X.-S. Deep robust clustering by contrastive learning. *arXiv preprint arXiv:2008.03030*, 2020. 1

Supplementary Material for "Large-Margin Contrastive Learning with Distance Polarization Regularizer"

Abstract

This supplementary document contains all technical proofs for **Theorem 1**, **Theorem 2**, **Theorem 3**, and **Theorem 4** in the ICML-2021 paper entitled "Large-Margin Contrastive Learning with Distance Polarization Regularizer".

A. Proof for Theorem 1

We first introduce the following Lemmas to prove our Theorem 1.

Lemma 1. *The function* $g(\boldsymbol{t}) = \log(1 + \gamma \sum_{i=1}^{n} \mathrm{e}^{t_i})$ *is strictly convex for* $\boldsymbol{t} = (t_1, t_2, \ldots, t_n)^\top \in \mathbb{R}^n$*, where the constant* $\gamma > 0$*.*

Proof. For the gradient of g, we have that

$$\nabla g(\boldsymbol{t}) = \left(\frac{\gamma \mathrm{e}^{t_1}}{1 + \gamma \sum_{i=1}^{n} \mathrm{e}^{t_i}}, \frac{\gamma \mathrm{e}^{t_2}}{1 + \gamma \sum_{i=1}^{n} \mathrm{e}^{t_i}}, \cdots, \frac{\gamma \mathrm{e}^{t_n}}{1 + \gamma \sum_{i=1}^{n} \mathrm{e}^{t_i}} \right)^\top, \tag{1}$$

and the corresponding *Hensen matrix* is

$$\begin{aligned}
\nabla^2 g(\boldsymbol{t}) &= \begin{pmatrix}
\frac{\gamma \mathrm{e}^{t_1}(1+\gamma \sum_{i=1}^{n} \mathrm{e}^{t_i}) - \gamma \mathrm{e}^{t_1} \gamma \mathrm{e}^{t_1}}{(1+\gamma \sum_{i=1}^{n} \mathrm{e}^{t_i})^2} & \frac{-\gamma \mathrm{e}^{t_1} \gamma \mathrm{e}^{t_2}}{(1+\gamma \sum_{i=1}^{n} \mathrm{e}^{t_i})^2} & \cdots & \frac{-\gamma \mathrm{e}^{t_1} \gamma \mathrm{e}^{t_n}}{(1+\gamma \sum_{i=1}^{n} \mathrm{e}^{t_i})^2} \\
\frac{-\gamma \mathrm{e}^{t_2} \gamma \mathrm{e}^{t_1}}{(1+\gamma \sum_{i=1}^{n} \mathrm{e}^{t_i})^2} & \frac{\gamma \mathrm{e}^{t_2}(1+\gamma \sum_{i=1}^{n} \mathrm{e}^{t_i}) - \gamma \mathrm{e}^{t_2} \gamma \mathrm{e}^{t_2}}{(1+\gamma \sum_{i=1}^{n} \mathrm{e}^{t_i})^2} & \cdots & \frac{-\gamma \mathrm{e}^{t_2} \gamma \mathrm{e}^{t_n}}{(1+\gamma \sum_{i=1}^{n} \mathrm{e}^{t_i})^2} \\
\cdots & \cdots & \cdots & \cdots \\
\frac{-\gamma \mathrm{e}^{t_n} \gamma \mathrm{e}^{t_1}}{(1+\gamma \sum_{i=1}^{n} \mathrm{e}^{t_i})^2} & \frac{-\gamma \mathrm{e}^{t_n} \gamma \mathrm{e}^{t_2}}{(1+\gamma \sum_{i=1}^{n} \mathrm{e}^{t_i})^2} & \cdots & \frac{\gamma \mathrm{e}^{t_n}(1+\gamma \sum_{i=1}^{n} \mathrm{e}^{t_i}) - \gamma \mathrm{e}^{t_n} \gamma \mathrm{e}^{t_n}}{(1+\gamma \sum_{i=1}^{n} \mathrm{e}^{t_i})^2}
\end{pmatrix} \\
&= \frac{\gamma}{(1 + \gamma \sum_{i=1}^{n} \mathrm{e}^{t_i})^2} \begin{pmatrix}
\mathrm{e}^{t_1}(1 + \gamma \sum_{i \neq 1} \mathrm{e}^{t_i}) & -\gamma \mathrm{e}^{t_1} \mathrm{e}^{t_2} & \cdots & -\gamma \mathrm{e}^{t_1} \mathrm{e}^{t_n} \\
-\gamma \mathrm{e}^{t_2} \mathrm{e}^{t_1} & \mathrm{e}^{t_2}(1 + \gamma \sum_{i \neq 2} \mathrm{e}^{t_i}) & \cdots & -\gamma \mathrm{e}^{t_2} \mathrm{e}^{t_n} \\
\cdots & \cdots & \cdots & \cdots \\
-\gamma \mathrm{e}^{t_n} \mathrm{e}^{t_1} & -\gamma \mathrm{e}^{t_n} \mathrm{e}^{t_2} & \cdots & \mathrm{e}^{t_n}(1 + \gamma \sum_{i \neq n} \mathrm{e}^{t_i})
\end{pmatrix}.
\end{aligned} \tag{2}$$

Then for any $\Delta \boldsymbol{t} \in \mathbb{R}^n \backslash \mathbf{0}$, we have that

$$\begin{aligned}
&\Delta \boldsymbol{t}^\top [((1 + \gamma \sum_{i=1}^{n} \mathrm{e}^{t_i})^2 / \gamma) \nabla^2 g(\boldsymbol{t})] \Delta \boldsymbol{t} \\
&= \frac{1}{\gamma} \Delta \boldsymbol{t}^\top \begin{pmatrix}
\mathrm{e}^{t_1}(1/\gamma + \sum_{i \neq 1} \mathrm{e}^{t_i}) & -\mathrm{e}^{t_1} \mathrm{e}^{t_2} & \cdots & -\mathrm{e}^{t_1} \mathrm{e}^{t_n} \\
-\mathrm{e}^{t_2} \mathrm{e}^{t_1} & \mathrm{e}^{t_2}(1/\gamma + \sum_{i \neq 2} \mathrm{e}^{t_i}) & \cdots & -\mathrm{e}^{t_2} \mathrm{e}^{t_n} \\
\cdots & \cdots & \cdots & \cdots \\
-\mathrm{e}^{t_n} \mathrm{e}^{t_1} & -\mathrm{e}^{t_n} \mathrm{e}^{t_2} & \cdots & \mathrm{e}^{t_n}(1/\gamma + \sum_{i \neq n} \mathrm{e}^{t_i})
\end{pmatrix} \Delta \boldsymbol{t} \\
&> \frac{1}{\gamma} \Delta \boldsymbol{t}^\top \begin{pmatrix}
\mathrm{e}^{t_1} \sum_{i \neq 1} \mathrm{e}^{t_i} & -\mathrm{e}^{t_1} \mathrm{e}^{t_2} & \cdots & -\mathrm{e}^{t_1} \mathrm{e}^{t_n} \\
-\mathrm{e}^{t_2} \mathrm{e}^{t_1} & \mathrm{e}^{t_2} \sum_{i \neq 2} \mathrm{e}^{t_i} & \cdots & -\mathrm{e}^{t_2} \mathrm{e}^{t_n} \\
\cdots & \cdots & \cdots & \cdots \\
-\mathrm{e}^{t_n} \mathrm{e}^{t_1} & -\mathrm{e}^{t_n} \mathrm{e}^{t_2} & \cdots & \mathrm{e}^{t_n} \sum_{i \neq n} \mathrm{e}^{t_i}
\end{pmatrix} \Delta \boldsymbol{t} \\
&= \frac{1}{\gamma} \sum\nolimits_{1 \leq k < l \leq n} \mathrm{e}^{t_k + t_l} (\Delta t_k - \Delta t_l)^2 \\
&\geq 0,
\end{aligned} \tag{3}$$

which implies that the Hensen matrix of g is strictly positive definite, and thus g is strictly convex. □

Furthermore, we have the following Lemma 2 to reveal the distribution of embedding results in the spherical coordinate.

Lemma 2. *Assume that the optimal feature embedding $\widehat{\boldsymbol{\varphi}} \in \arg\min_{\boldsymbol{\varphi}\in\mathcal{H}} \mathcal{L}_{\text{NCE}}(\boldsymbol{\varphi})$ and the spherical coordinate of the embedding result $\boldsymbol{z}_i = \widehat{\boldsymbol{\varphi}}(\boldsymbol{x}_i) \in \mathbb{R}^d$ is denoted as $(\theta_i^{(1)}, \theta_i^{(2)}, \ldots, \theta_i^{(d-1)})^\top$ where $i = 1, 2, \ldots, N$ [1]. Then for any given $\alpha \in [0, \pi]$ and $\beta \in [0, 2\pi]$, we have that*

$$\lim_{N\to+\infty} \min_{i=1,2,\ldots,N} |\theta_i^{(k)} - \alpha| = \lim_{N\to+\infty} \min_{i=1,2,\ldots,N} |\theta_i^{(d-1)} - \beta| = 0, \tag{4}$$

where $k = 1, 2, \ldots, d-2$.

Proof. We prove the above Eq. (4) via contradiction. Without loss of generality, we assume that there exists $\widehat{\alpha} \in [0, \pi]$ such that $\min_{i=1,2,\ldots,N} |\theta_i^{(k)} - \alpha| > \delta > 0$ for any N, and we have that there exists $q \in \{1, 2, \ldots, N-1\}$ such that

$$|\theta_q^{(k)} - \theta_{q+1}^{(k)}| > \delta > 0. \tag{5}$$

We let $\widetilde{\theta}_q^{(k)} = \theta_q^{(k)} + (\theta_{q+1}^{(k)} - \theta_q^{(k)})/2$ and $\widetilde{\theta}_{q+1}^{(k)} = \theta_{q+1}^{(k)} + (\theta_q^{(k)} - \theta_{q+1}^{(k)})/2$. Then we construct a new embedding $\widetilde{\boldsymbol{\varphi}}$ which satisfies

$$\widetilde{\boldsymbol{\varphi}}(\boldsymbol{x}_i) = \begin{cases} \boldsymbol{z}_i, & i = 1, 2, \ldots, q-1, q+2, \ldots, N \\ \widetilde{\boldsymbol{z}}_i, & \text{Otherwise,} \end{cases} \tag{6}$$

where the spherical coordinate of $\widetilde{\boldsymbol{z}}_q$ and $\widetilde{\boldsymbol{z}}_{q+1}$ are $(\theta_q^{(1)}, \theta_q^{(2)}, \ldots, \widetilde{\theta}_q^{(k)}, \ldots, \theta_q^{(d-1)})^\top$ and $(\theta_{q+1}^{(1)}, \theta_{q+1}^{(2)}, \ldots, \widetilde{\theta}_{q+1}^{(k)}, \ldots, \theta_{q+1}^{(d-1)})^\top$, respectively. By using the strict convexity of $\log(1 + \gamma \sum_{i=1}^n \mathrm{e}^{t_i})$ as revealed in Lemma 1, we have that

$$\begin{aligned}
&\mathcal{L}_{\text{NCE}}(\widetilde{\boldsymbol{\varphi}}) - \mathcal{L}_{\text{NCE}}(\widehat{\boldsymbol{\varphi}}) \\
&= \mathbb{E}_{\boldsymbol{x},\, \boldsymbol{x}_j^- \in \mathcal{X}} \left[-\log \frac{\mathrm{e}^{\widetilde{\boldsymbol{\varphi}}(\boldsymbol{x})^\top \widetilde{\boldsymbol{\varphi}}(\boldsymbol{x}^+)}}{\mathrm{e}^{\widetilde{\boldsymbol{\varphi}}(\boldsymbol{x})^\top \widetilde{\boldsymbol{\varphi}}(\boldsymbol{x}^+)} + \sum_{j=1}^n \mathrm{e}^{\widetilde{\boldsymbol{\varphi}}(\boldsymbol{x})^\top \widetilde{\boldsymbol{\varphi}}(\boldsymbol{x}_j^-)}} + \log \frac{\mathrm{e}^{\widehat{\boldsymbol{\varphi}}(\boldsymbol{x})^\top \widehat{\boldsymbol{\varphi}}(\boldsymbol{x}^+)}}{\mathrm{e}^{\widehat{\boldsymbol{\varphi}}(\boldsymbol{x})^\top \widehat{\boldsymbol{\varphi}}(\boldsymbol{x}^+)} + \sum_{j=1}^n \mathrm{e}^{\widehat{\boldsymbol{\varphi}}(\boldsymbol{x})^\top \widehat{\boldsymbol{\varphi}}(\boldsymbol{x}_j^-)}} \right] \\
&= \mathbb{E}_{\boldsymbol{x},\, \boldsymbol{x}_j^- \in \mathcal{X}} \left[\log \left(1 + \gamma \sum_{j=1}^n \mathrm{e}^{\widetilde{\boldsymbol{\varphi}}(\boldsymbol{x})^\top \widetilde{\boldsymbol{\varphi}}(\boldsymbol{x}_j^-)} \right) - \log \left(1 + \gamma \sum_{j=1}^n \mathrm{e}^{\widehat{\boldsymbol{\varphi}}(\boldsymbol{x})^\top \widehat{\boldsymbol{\varphi}}(\boldsymbol{x}_j^-)} \right) \right] \\
&= \mathbb{E}_{\boldsymbol{x}_j^-} \left[\log \left(\frac{1 + \gamma \sum_{j=1}^n \mathrm{e}^{\widetilde{\boldsymbol{z}}_q^\top \widetilde{\boldsymbol{\varphi}}(\boldsymbol{x}_j^-)}}{1 + \gamma \sum_{j=1}^n \mathrm{e}^{\widehat{\boldsymbol{z}}_q^\top \widetilde{\boldsymbol{\varphi}}(\boldsymbol{x}_j^-)}} \right) + \log \left(\frac{1 + \gamma \sum_{j=1}^n \mathrm{e}^{\widetilde{\boldsymbol{z}}_{q+1}^\top \widetilde{\boldsymbol{\varphi}}(\boldsymbol{x}_j^-)}}{1 + \gamma \sum_{j=1}^n \mathrm{e}^{\widehat{\boldsymbol{z}}_{q+1}^\top \widetilde{\boldsymbol{\varphi}}(\boldsymbol{x}_j^-)}} \right) \right] \\
&= \mathbb{E}_{\boldsymbol{x}_j^-} [(g(\widetilde{\boldsymbol{z}}_q^\top \widetilde{\boldsymbol{\varphi}}(\boldsymbol{x}_1^-), \widetilde{\boldsymbol{z}}_q^\top \widetilde{\boldsymbol{\varphi}}(\boldsymbol{x}_2^-), \ldots, \widetilde{\boldsymbol{z}}_q^\top \widetilde{\boldsymbol{\varphi}}(\boldsymbol{x}_n^-)) + g(\widetilde{\boldsymbol{z}}_{q+1}^\top \widetilde{\boldsymbol{\varphi}}(\boldsymbol{x}_1^-), \widetilde{\boldsymbol{z}}_{q+1}^\top \widetilde{\boldsymbol{\varphi}}(\boldsymbol{x}_2^-), \ldots, \widetilde{\boldsymbol{z}}_{q+1}^\top \widetilde{\boldsymbol{\varphi}}(\boldsymbol{x}_n^-))) \\
&\quad - (g(\widehat{\boldsymbol{z}}_q^\top \widetilde{\boldsymbol{\varphi}}(\boldsymbol{x}_1^-), \widehat{\boldsymbol{z}}_q^\top \widetilde{\boldsymbol{\varphi}}(\boldsymbol{x}_2^-), \ldots, \widehat{\boldsymbol{z}}_q^\top \widetilde{\boldsymbol{\varphi}}(\boldsymbol{x}_n^-)) + g(\widehat{\boldsymbol{z}}_{q+1}^\top \widetilde{\boldsymbol{\varphi}}(\boldsymbol{x}_1^-), \widehat{\boldsymbol{z}}_{q+1}^\top \widetilde{\boldsymbol{\varphi}}(\boldsymbol{x}_2^-), \ldots, \widehat{\boldsymbol{z}}_{q+1}^\top \widetilde{\boldsymbol{\varphi}}(\boldsymbol{x}_n^-)))] \\
&= \mathbb{E}_{\boldsymbol{x}_j^-} [(g((\boldsymbol{t}_1 + \boldsymbol{t}_2)/2) + g((\boldsymbol{t}_1 + \boldsymbol{t}_2)/2)) - (g(\boldsymbol{t}_1) + g(\boldsymbol{t}_2))] \\
&= (1/2)\mathbb{E}_{\boldsymbol{x}_j^-} [g((\boldsymbol{t}_1 + \boldsymbol{t}_2)/2) - (g(\boldsymbol{t}_1) + g(\boldsymbol{t}_2))/2] \\
&< 0,
\end{aligned} \tag{7}$$

which is contradictory to the optimality of $\widehat{\boldsymbol{\varphi}}$, and thereby the proof is completed. □

Now we prove the Theorem 1 based on the above Lemma 2.

Theorem 1. *Assume that the optimal feature embedding $\widehat{\boldsymbol{\varphi}} \in \arg\min_{\boldsymbol{\varphi}\in\mathcal{H}} \mathcal{L}_{\text{NCE}}(\boldsymbol{\varphi})$ and the corresponding distance value $\mathcal{D}_{ij}^{\widehat{\boldsymbol{\varphi}}} = (1 - \widehat{\boldsymbol{\varphi}}(\boldsymbol{x}_i)^\top \widehat{\boldsymbol{\varphi}}(\boldsymbol{x}_j))/2$. Then for any given $\mu \in [0, 1]$ and $\epsilon > 0$, there exists sufficiently large N such that $\min_{1\le i<j\le N} \{|\mathcal{D}_{ij}^{\widehat{\boldsymbol{\varphi}}} - \mu|\} < \epsilon$.*

Proof. For any given $\mu \in [0, 1]$, by invoking Lemma 2, we have that there exists $\widetilde{\boldsymbol{\theta}}_k = (\theta_k^{(1)}, \theta_k^{(2)}, \ldots, \theta_k^{(d-1)})^\top$ and $\widetilde{\boldsymbol{\theta}}_l = (\theta_l^{(1)}, \theta_l^{(2)}, \ldots, \theta_l^{(d-1)})^\top$ such that

$$\|\widetilde{\boldsymbol{\theta}}_k - \mathbf{0}\|_2 < \delta \text{ and } \|\widetilde{\boldsymbol{\theta}}_l - (\arccos(1 - 2\mu), 0, \ldots, 0)^\top\|_2 < \delta, \tag{8}$$

[1] Here $\theta_i^{(1)}, \theta_i^{(2)}, \ldots, \theta_i^{(d-2)} \in [0, \pi]$ and $\theta_i^{(d-1)} \in [0, 2\pi]$.

and thus we have

$$\begin{aligned}
&\left|\mathcal{D}_{kl}^{\widehat{\varphi}} - \mu\right| \\
&= \left|(1-\cos(\|\widetilde{\boldsymbol{\theta}}_k - \widetilde{\boldsymbol{\theta}}_l\|_2))/2 - \mu\right| \\
&\leq |(1-\cos(\arccos(1-2\mu)+2\delta))/2 - \mu| \\
&\leq |(1-(1-2\mu)+O(2\delta))/2 - \mu| \\
&\leq O(\delta),
\end{aligned} \tag{9}$$

which completes the proof by letting δ be sufficiently small. □

B. Proof for Theorem 2

Theorem 2. *For training data $\{\boldsymbol{x}_i\}_{i=1}^N$ with underling class labels $\{y_i\}_{i=1}^N$ and any given $\tau \in (0,\, 1/2)$, there exists a feature embedding $\overline{\boldsymbol{\varphi}} \in \mathcal{H}$ such that*

$$\max_{(i,\,j)\in I^+} \mathcal{D}_{ij}^{\overline{\boldsymbol{\varphi}}} \;\leq\; 1/2 - \tau \;<\; 1/2 \;\leq\; \min_{(k,\,l)\in I^-} \mathcal{D}_{kl}^{\overline{\boldsymbol{\varphi}}}, \tag{10}$$

where $y_i = 1, 2, \ldots, C$ for $i = 1, 2, \ldots, N$ and $C < d$. Here the bivariate index sets $I^+ = \{(i,\, j)|y_i = y_j,\, i, j = 1, 2, \ldots, N\}$ and $I^- = \{(i,\, j)|y_i \neq y_j,\, i, j = 1, 2, \ldots, N\}$.

Proof. For the training data examples $\{\boldsymbol{x}_i\}_{i=1}^N$, we construct the mapping $\overline{\boldsymbol{\varphi}} : \mathbb{R}^m \to \mathbb{R}^d$ satisfying that

$$\overline{\boldsymbol{\varphi}}(\boldsymbol{x}_i) = \boldsymbol{e}_{y_i}, \tag{11}$$

where $y_i = 1, 2, \ldots, C$ and $\{\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_d\}$ is the standard orthogonal basis in the d-dimensional space. Then we have that

$$\max_{(i,\,j)\in I^+} \mathcal{D}_{ij}^{\overline{\boldsymbol{\varphi}}} = (1 - \boldsymbol{e}_{y_i}^\top \boldsymbol{e}_{y_i})/2 = (1 - \|\boldsymbol{e}_{y_i}\|_2^2)/2 = 0, \tag{12}$$

and

$$\min_{(k,\,l)\in I^-} \mathcal{D}_{kl}^{\overline{\boldsymbol{\varphi}}} = (1 - \boldsymbol{e}_{y_k}^\top \boldsymbol{e}_{y_l})/2 = (1-0)/2 = 1/2, \tag{13}$$

which completes the proof. □

C. Proof for Theorem 3

Theorem 3. *Assume that $\boldsymbol{\varphi}^* \in \arg\min_{\boldsymbol{\varphi}\in\mathcal{H}} \mathcal{L}_{\text{NCE}}(\boldsymbol{\varphi}) + \lambda\mathcal{R}_1(\boldsymbol{\varphi})$, and the underling class labels of training data $\{\boldsymbol{x}_i\}_{i=1}^N$ are $\{y_i\}_{i=1}^N$. Then we have that*

$$\begin{aligned}
&\mathbb{E}_{y_i\neq y_j}[\max(\delta_\mu^- - \mathcal{D}_{ij}^{\boldsymbol{\varphi}^*}, 0)] + \mathbb{E}_{y_k = y_l}[\max(\mathcal{D}_{kl}^{\boldsymbol{\varphi}^*} - \delta_\mu^+, 0)] \\
&\leq (\delta^- - \delta^+)\mathcal{R}_1(\boldsymbol{\varphi}^*) + (K_{\max}/K_{\min})/C \\
&\leq 4(\delta^- - \delta^+)/\lambda + (K_{\max}/K_{\min})/C,
\end{aligned} \tag{14}$$

where the constants $\delta_\mu^- = \delta^- - \mu$, $\delta_\mu^+ = \delta^+ + \mu$, $\mu \in (0,\, \delta^- - \delta^+)$, $K_{\min} = \min_{1\leq k\leq C} \|\boldsymbol{y} - k\cdot\boldsymbol{1}_{N\times 1}\|_0$, and $K_{\max} = \max_{1\leq k\leq C} \|\boldsymbol{y} - k\cdot\boldsymbol{1}_{N\times 1}\|_0$.

Proof. We denote that $s_1 \leq s_2 \leq \cdots \leq s_{N(N-1)/2}$ is a ranking of the distances $\{\mathcal{D}_{ij}^{\boldsymbol{\varphi}^*} | 1 \leq i < j \leq N\}$. Let $N^+ = |\{(i,j)|1 \leq i < j \leq N, y_i = y_j\}|$ which is the number of intra-class pairs. Then for $\delta^+ = s_{N^+}$, we have that

$$\begin{aligned}
&\mathbb{E}_{y_i\neq y_j}[\max(\delta_\mu^- - \mathcal{D}_{ij}^{\boldsymbol{\varphi}^*}, 0)] + \mathbb{E}_{y_k = y_l}[\max(\mathcal{D}_{kl}^{\boldsymbol{\varphi}^*} - \delta_\mu^+, 0)] \\
&\leq (\delta_\mu^- - \delta_\mu^+)\mathcal{R}_0(\boldsymbol{\varphi}^*) + (N^+/N)[\max((1-\delta_\mu^+), \delta_\mu^-)] \\
&\leq (\delta^- - \delta^+)\mathcal{R}_1(\boldsymbol{\varphi}^*) + (K_{\max}/K_{\min})/C.
\end{aligned} \tag{15}$$

Furthermore, for any φ_0 satisfying $\mathcal{R}_1(\varphi_0)=0$, by the optimality of φ^*, we have that

$$\mathcal{L}_{\text{NCE}}(\varphi^*)+\lambda\mathcal{R}_1(\varphi^*) \le \mathcal{L}_{\text{NCE}}(\varphi_0)+\lambda\mathcal{R}_1(\varphi_0) = \mathcal{L}_{\text{NCE}}(\varphi_0), \tag{16}$$

and thus

$$\begin{aligned}
&\mathcal{R}_1(\varphi^*)\\
&\le (\mathcal{L}_{\text{NCE}}(\varphi_0)-\mathcal{L}_{\text{NCE}}(\varphi^*))/\lambda\\
&\le \frac{1}{\lambda}\left(-\log\frac{\mathrm{e}^{\varphi_0(\boldsymbol{x})^\top\varphi_0(\boldsymbol{x}^+)}}{\mathrm{e}^{\varphi_0(\boldsymbol{x})^\top\varphi_0(\boldsymbol{x}^+)}+\sum_{j=1}^n \mathrm{e}^{\varphi_0(\boldsymbol{x})^\top\varphi_0(\boldsymbol{x}_j^-)}}+\log\frac{\mathrm{e}^{\varphi^*(\boldsymbol{x})^\top\varphi^*(\boldsymbol{x}^+)}}{\mathrm{e}^{\varphi^*(\boldsymbol{x})^\top\varphi^*(\boldsymbol{x}^+)}+\sum_{j=1}^n \mathrm{e}^{\varphi^*(\boldsymbol{x})^\top\varphi^*(\boldsymbol{x}_j^-)}}\right)\\
&\le \frac{1}{\lambda}\left(-\log\frac{\mathrm{e}^{-1}}{\mathrm{e}^{-1}+\sum_{j=1}^n\mathrm{e}^{1}}+\log\frac{\mathrm{e}^{1}}{\mathrm{e}^{1}+\sum_{j=1}^n\mathrm{e}^{-1}}\right)\\
&\le \frac{1}{\lambda}\log\left(\frac{\mathrm{e}^{1}}{\mathrm{e}^{1}+\sum_{j=1}^n\mathrm{e}^{-1}}\frac{\mathrm{e}^{-1}+\sum_{j=1}^n\mathrm{e}^{1}}{\mathrm{e}^{-1}}\right)\\
&\le \frac{1}{\lambda}\log\left(\mathrm{e}^2\frac{\mathrm{e}^{-1}+\sum_{j=1}^n\mathrm{e}^{1}}{\mathrm{e}^{1}+\sum_{j=1}^n\mathrm{e}^{-1}}\right)\\
&\le \frac{1}{\lambda}\log\left(\mathrm{e}^2\cdot\mathrm{e}^2\right)\\
&\le \frac{4}{\lambda}.
\end{aligned} \tag{17}$$

By combining the above Eq. (15) and Eq. (17), we have that

$$\mathbb{E}_{y_i\neq y_j}[\max(\delta_\mu^- - \mathcal{D}_{ij}^{\varphi^*},0)]+\mathbb{E}_{y_k=y_l}[\max(\mathcal{D}_{kl}^{\varphi^*}-\delta_\mu^+,0)] \le 4(\delta^- - \delta^+)/\lambda + (K_{\max}/K_{\min})/C, \tag{18}$$

and the proof is completed. □

D. Proof for Theorem 4

We introduce the following Lemma to prove the Theorem 4

Lemma 3. (Saunshi et al., 2019) *Assume that $\varphi^* \in \arg\min_{\varphi\in\mathcal{H}}\ \mathcal{L}_{\text{NCE}}(\varphi)+\lambda\mathcal{R}_1(\varphi)$. Then with probability at least $1-\delta$ over the training data $\mathcal{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_N\}$, for any $\varphi\in\mathcal{H}$*

$$\mathcal{L}_{\text{NCE}}(\varphi^*) \le \mathcal{L}_{\text{NCE}}(\varphi)+\mathcal{O}\left(\frac{Q_1\mathfrak{R}_{\mathcal{H}}(\lambda)}{N}+\sqrt{\frac{Q_2}{N}}\right), \tag{19}$$

where $Q_1=\sqrt{1+1/n}$, $Q_2=\log(1/\delta)\cdot\log^2(n)$, the Rademacher Complexity $\mathfrak{R}_{\mathcal{H}}(\lambda)=\mathbb{E}_{\boldsymbol{\sigma}\in\{\pm1\}^{3dN}}[\sup_{\varphi\in\mathcal{H}(\lambda)}\langle\boldsymbol{\sigma},\boldsymbol{f}\rangle]$, and the restricted hypothesis space $\mathcal{H}(\lambda)=\{\varphi|\varphi\in\mathcal{H},\ \text{and}\ \mathcal{R}_1(\varphi)\le 4/\lambda\}$.

Theorem 4. *Let $\varphi^*\in\arg\min_{\varphi\in\mathcal{H}}\mathcal{L}_{\text{NCE}}(\varphi)+\lambda\mathcal{R}_1(\varphi)$. Then with probability at least $1-\delta$, we have that*

$$\left|\mathcal{L}_{\text{SM}}^{\mathcal{T}}(\varphi^*)-\mathcal{L}_{\text{NCE}}(\varphi)\right| \le \mathcal{O}\left(\frac{Q_1\mathfrak{R}_{\mathcal{H}}(\lambda)}{N}+\sqrt{\frac{Q_2}{N}}\right), \tag{20}$$

where $Q_1=\sqrt{1+1/n}$, $Q_2=\log(1/\delta)\cdot\log^2(n)$, and $\mathfrak{R}_{\mathcal{H}}(\lambda)$ is monotonically decreasing w.r.t. λ.

Proof. For the traditional cross entropy loss $\mathcal{L}_{\mathrm{SM}}^{\mathcal{T}}(\boldsymbol{\varphi})$, we have that

$$
\begin{aligned}
&\mathcal{L}_{\mathrm{SM}}^{\mathcal{T}}(\boldsymbol{\varphi}) \\
&= \mathbb{E}_{\mathcal{X}\sim\mathcal{T}}[\inf_{\boldsymbol{W}\in\mathbb{R}^{C\times d}} \mathcal{L}_{\mathrm{CEP}}(\boldsymbol{W}\boldsymbol{\varphi};\mathcal{X})] \\
&= \mathbb{E}_{\mathcal{X}\sim\mathcal{T}}\left[-\log\frac{\mathrm{e}^{\boldsymbol{\varphi}(\boldsymbol{x})^\top\boldsymbol{\mu}_c}}{\mathrm{e}^{\boldsymbol{\varphi}(\boldsymbol{x})^\top\boldsymbol{\mu}_c}+\sum\mathrm{e}^{\boldsymbol{\varphi}(\boldsymbol{x})^\top\mu_{c^-}}}\right] \\
&= \mathbb{E}_{\mathcal{X}\sim\mathcal{T}}\left[-\log\frac{\mathrm{e}^{\boldsymbol{\varphi}(\boldsymbol{x})^\top\mathbb{E}_{\boldsymbol{x}^+\sim p}[\boldsymbol{\varphi}(\boldsymbol{x}^+)]}}{\mathrm{e}^{\boldsymbol{\varphi}(\boldsymbol{x})^\top\mathbb{E}_{\boldsymbol{x}^+\sim p^+}[\boldsymbol{\varphi}(\boldsymbol{x}^+)]}+n\mathbb{E}[\mathrm{e}^{\boldsymbol{\varphi}(\boldsymbol{x})^\top\mathbb{E}_{\boldsymbol{x}^-\sim p^-}[\boldsymbol{\varphi}(\boldsymbol{x}^-)]}]}\right] \\
&\geq \mathbb{E}_{\mathcal{X}\sim\mathcal{T}}\left[-\log\frac{\mathrm{e}^{\boldsymbol{\varphi}(\boldsymbol{x})^\top\mathbb{E}_{\boldsymbol{x}^+\sim p}[\boldsymbol{\varphi}(\boldsymbol{x}^+)]}}{\mathrm{e}^{\boldsymbol{\varphi}(\boldsymbol{x})^\top\mathbb{E}_{\boldsymbol{x}^+\sim p^+}[\boldsymbol{\varphi}(\boldsymbol{x}^+)]}+n\mathbb{E}_{\boldsymbol{x}^-\sim p^-}[\mathrm{e}^{\boldsymbol{\varphi}(\boldsymbol{x})^\top\boldsymbol{\varphi}(\boldsymbol{x}^-)}]}\right] \\
&\geq \mathbb{E}_{\mathcal{X}\sim\mathcal{T}}\left[-\log\frac{\mathrm{e}^{\boldsymbol{\varphi}(\boldsymbol{x})^\top\mathbb{E}_{\boldsymbol{x}^+\sim p}[\boldsymbol{\varphi}(\boldsymbol{x}^+)]}}{\mathrm{e}^{\boldsymbol{\varphi}(\boldsymbol{x})^\top\mathbb{E}_{\boldsymbol{x}^+\sim p^+}[\boldsymbol{\varphi}(\boldsymbol{x}^+)]}+n\mathbb{E}_{\boldsymbol{x}^-\sim\mathcal{T}}[\mathrm{e}^{\boldsymbol{\varphi}(\boldsymbol{x})^\top\boldsymbol{\varphi}(\boldsymbol{x}^-)}]}\right] \\
&= \mathcal{L}_{\mathrm{NCE}}(\boldsymbol{\varphi}).
\end{aligned} \tag{21}
$$

By combining the above Eq. (21) and Lemma 3, we finally have that

$$
\mathcal{L}_{\mathrm{SM}}^{\mathcal{T}}(\boldsymbol{\varphi}^*) - \mathcal{L}_{\mathrm{NCE}}(\boldsymbol{\varphi}) \leq \mathcal{L}_{\mathrm{NCE}}(\boldsymbol{\varphi}^*) - \mathcal{L}_{\mathrm{NCE}}(\boldsymbol{\varphi}) \leq \mathcal{O}\left(\frac{Q_1\mathfrak{R}_{\mathcal{H}}(\lambda)}{N}+\sqrt{\frac{Q_2}{N}}\right), \tag{22}
$$

where $\mathfrak{R}_{\mathcal{H}}(\lambda)=\mathbb{E}_{\boldsymbol{\sigma}\in\{\pm1\}^{3dN}}[\sup_{\boldsymbol{\varphi}\in\mathcal{H}(\lambda)}\langle\boldsymbol{\sigma},\boldsymbol{f}\rangle]$ is monotonically decreasing *w.r.t.* λ. □

References

Saunshi, N., Plevrakis, O., Arora, S., Khodak, M., and Khandeparkar, H. A theoretical analysis of contrastive unsupervised representation learning. In *International Conference on Machine Learning (ICML)*, pp. 5628–5637, 2019. 3