Generalization Analysis for Deep Contrastive Representation Learning

Nong Minh Hieu [1, 2], **Antoine Ledent** [2], **Yunwen Lei** [3], **Cheng Yeaw Ku** [1]

[1]School of Physical and Mathematical Science, Nanyang Technological University, Singapore 639798
[2]School of Computing and Information Systems, Singapore Management University, Singapore 188065
[3]Department of Mathematics, University of Hong Kong, Pok Fu Lam, Hong Kong
mh.nong.2024@phdcs.smu.edu.sg, aledent@smu.edu.sg, leiyw@hku.hk, cyku@ntu.edu.sg

Abstract

In this paper, we present generalization bounds for the unsupervised risk in the Deep Contrastive Representation Learning framework, which employs deep neural networks as representation functions. We approach this problem from two angles. On the one hand, we derive a parameter-counting bound that scales with the overall size of the neural networks. On the other hand, we provide a norm-based bound that scales with the norms of neural networks' weight matrices. Ignoring logarithmic factors, the bounds are independent of k, the size of the tuples provided for contrastive learning. To the best of our knowledge, this property is only shared by one other work, which employed a different proof strategy and suffers from very strong exponential dependence on the depth of the network which is due to a use of the peeling technique. Our results circumvent this by leveraging powerful results on covering numbers with respect to uniform norms over samples. In addition, we utilize loss augmentation techniques to further reduce the dependency on matrix norms and the implicit dependence on network depth. In fact, our techniques allow us to produce many bounds for the contrastive learning setting with similar architectural dependencies as in the study of the sample complexity of ordinary loss functions, thereby bridging the gap between the learning theories of contrastive learning and DNNs.

Arxiv URL — https://arxiv.org/abs/2412.12014

Introduction

Contrastive Representation Learning (CRL) is a powerful framework that focuses on learning good data representations in an unsupervised learning manner. The CRL framework can be informally described as follows: given a dataset comprising of data tuples $\mathcal{S} = \{(x_j, x_j^+, x_{j1}^-, \ldots, x_{jk}^-)\}_{j=1}^n$ where each data instance belongs to an input space $\mathcal{X}$, the key idea of CRL is to pull similar pairs (x_j, x_j^+) closer together and to push apart dissimilar pairs (x_j, x_{ji}^-) in a representation space $\mathcal{R} \subset \mathbb{R}^d$. This is accomplished by training a representation function $f : \mathcal{X} \to \mathcal{R}$ that minimizes the empirical unsupervised risk:

$$\widehat{\mathrm{L}}_{\mathrm{un}}(f) = \frac{1}{n}\sum_{j=1}^{n} \ell\Big(\Big\{f(x_j)^\top (f(x_j^+) - f(x_{ji}^-))\Big\}_{i=1}^{k}\Big), \quad (1)$$

where k is the number of negative samples per input data tuple and $\ell : \mathbb{R}^k \to \mathbb{R}_+$ is a contrastive loss function for which popular choices include the hinge and logistic losses:

$$\begin{aligned} &\text{Hinge loss: } \ell(v) = \max\Big\{0, 1 + \max_{1\le i\le k}\{-v_i\}\Big\}, \\ &\text{Logistic loss: } \ell(v) = \log\left(1 + \sum_{i=1}^{k}\exp(-v_i)\right). \end{aligned} \quad (2)$$

The learned representations are then used for downstream tasks like classification, clustering or visualization.

Owing to its simplicity and effectiveness, CRL has been applied in a wide variety of machine learning tasks, ranging from computer vision (Chen et al. 2020; He et al. 2019; Gidaris, Singh, and Komodakis 2018), graph representation learning (Hassani and Khasahmadi 2020; Zhu et al. 2020; Velickovic et al. 2019), natural language models (Gao, Yao, and Chen 2021; Zhang et al. 2021; Reimers and Gurevych 2021) and time-series forecasting (Lee, Park, and Lee 2024; Yang, Zhang, and Cui 2022; Nie1 et al. 2023; Eldele et al. 2021). Despite the aforementioned successes, very few contributions have been made to explain the good performance of CRL. Even though there are several empirical studies that demonstrate the effectiveness of CRL (Chen et al. 2020; He et al. 2019), there are limited theoretical analyses conducted to explain its generalization behaviour.

In the work of Arora et al. (2019), a theoretical framework to study the generalization behaviour of CRL is proposed. Let $\mathcal{F} = \{f : \mathcal{X} \to \mathbb{R}^d | \|f(x)\|_2 \le B\}$ be a class of representation functions and assume that the loss function is ℓ^∞-Lipschitz with constant $\eta > 0$, where $\|\cdot\|_p$ denotes the ℓ^p norm for $p \ge 1$. The authors provided a bound that scales in the order of $\mathcal{O}(\eta B\sqrt{k}\mathcal{R}_{\mathcal{S}}(\mathcal{F})/n)$, where $\mathcal{R}_{\mathcal{S}}(\mathcal{F})$ is a measure of complexity for the function class $\mathcal{F}$. However, the $\sqrt{k}$ dependency on negative samples is inconsistent with some of the works that suggest large number of negative samples implicitly implies better generalization or at least does not degrade generalization capability (Awasthi, Dikkala, and Kamath 2022; Tian, Krishnan, and Isola 2020;

Henaff et al. 2020; Khosla et al. 2020). Therefore, the bound does not fully explain the good generalization behaviour in existing empirical works.

Later in Lei et al. (2023), an improvement is made by making the reliance on k at most logarithmic, obtaining the bound in the order of $\tilde{\mathcal{O}}(\eta B \mathcal{R}_{\mathcal{S}}(\mathcal{F})/n)$ (where the $\tilde{\mathcal{O}}$ notation hides logarithmic terms). However, in the case of Deep Contrastive Representation Learning (DCRL) where $\mathcal{F}$ is a class of neural networks with L layers, the authors made use of the peeling technique proposed by Golowich, Rakhlin, and Shamir (2018) to derive the following complexity order for the class $\mathcal{F}$: $\mathcal{R}_{\mathcal{S}}(\mathcal{F}) = \mathcal{O}\Big(B_x\sqrt{ndL}\prod_{l=1}^{L} B_{Fr}^{(l)}\Big)$, where B_x is the upper bound on the ℓ^2 norm of input vectors in the input space $\mathcal{X}$ and $B_{Fr}^{(l)}$ is the Frobenius norm of the weight matrix at the l^{th} layer. Due to the product of Frobenius norms in non-logarithmic terms, the bound suffers from a strong dependency on the neural networks' depth. Unfortunately, this downside is particularly unfavourable in practice when the network architectures are usually deep and the constraints on the weight matrices are not strict.

In terms of proof techniques, both Arora et al. (2019) and Lei et al. (2023) focus on general function classes and rely on vector contraction inequalities (Ledoux and Talagrand 2011; Maurer 2016) and inequalities between various complexity measures of the loss class and the feature mapping (Anthony and Bartlett 2002; Srebro, Sridharan, and Tewari 2010; Lei et al. 2019, 2023). This approach is prone to introducing architectural information (final layer's dimension) and dataset size information (number of negative samples) into the generalization bound. Even though the dependency on negative samples is resolved by Lei et al. (2023) using fat-shattering dimension and worst-case Rademacher complexity, the use of the peeling technique makes the bound scale impractically for (deep) neural networks.

In this work, we demonstrate how to achieve generalization bounds for the Contrastive Learning setting with more flexible tools such as covering numbers. This is achieved through the construction of *auxiliary datasets* consisting of all individual samples involved in any of the input tuples (for further details, cf. Appendix D, 'Basic Bounds'). This immediately allows us to prove generalization bounds for the DCRL setting with a spectral-type complexity term for the neural network component, a great improvement over the product of Frobenius norms present in the previous state-of-the-art results. Furthermore, by exploiting the ℓ^∞-Lipschitzness of popular losses such as the hinge loss and logistic loss, we show that this approach can naturally alleviate the strong reliance on the number of negative samples (with at most logarithmic dependency) without the need for other complexity measures such as fat-shattering dimension or peeling techniques. Moreover, we further tighten the complexity bound for neural networks by applying loss augmentation technique (Nagarajan and Kolter 2019; Wei and Ma 2019; Ledent et al. 2021b) to incorporate data-dependent terms in the bounds. Finally, we derive a parameter-counting bound that scales with the number of neurons in the network with no dependence on the number of negative samples.

Related Work

Arora et al. (2019) developed a framework to study the generalization behavior of CRL in terms of Rademacher complexity. The analysis there based on ℓ^2-Lipschitz continuity implies generalization bounds with a linear dependency on the number of negative examples. This linear dependency was recently improved to a logarithmic dependency in Lei et al. (2023) by arguments relying on ℓ^∞-Lipschitz continuity inspired from work on multi-class and multi-label classification (Lei et al. 2019; Mustafa et al. 2021; Wu et al. 2021). These discussions were later extended to CRL with adversarial training examples (Wen et al. 2024; Zou and Liu 2023). The above discussions are mainly based on Rademacher complexities. Other than this approach, there are also increasing discussions on CRL from the perspective of PAC-Bayesian analysis (Nozawa, Germain, and Guedj 2020), mutual information (Tsai et al. 2020), spectral clustering (HaoChen et al. 2021), gradient-descent dynamics (Tian et al. 2020), distributionally robust optimization (Wu et al. 2024) and causality (Mitrovic et al. 2021). There is also some work on the generalization analysis of pairwise or triplet wise loss functions in a similar i.i.d. setting as we consider (Lei, Ledent, and Kloft 2020; Alves and Ledent 2024; Yang et al. 2021; Lei, Liu, and Ying 2021). However, such works do not control the dependence on the number of samples in each input tuple. The benefit of representative learning to improve the generalization of downstream classification tasks were also studied extensively (Arora et al. 2019; Zou and Liu 2023; Chuang et al. 2020; Bao, Nagano, and Nozawa 2022).

CRL often learns nonlinear features by neural networks, and therefore one needs to study the complexity of neural networks to get the corresponding generalization bounds. Nearly tight VC dimension and pseudodimension bounds were developed (Bartlett et al. 2019). Rademacher complexity bounds were developed for neural networks under a norm constraint, which, however, exhibit an exponential dependency on the depth (Neyshabur, Tomioka, and Srebro 2015). This exponential dependency was improved to a square-root dependency by using the homogeneity of ReLU networks (Golowich, Rakhlin, and Shamir 2018). Spectrally-normalized margin bounds were developed based on induction arguments with covering numbers (Bartlett, Foster, and Telgarsky 2017; Hsu et al. 2021). The benefit of weight sharing in convolutional neural networks was also studied based on covering numbers (Ledent et al. 2021b; Lin et al. 2022) and parameter counting (Long and Sedghi 2020; Zhou and Huo 2024). The benefits of connection-sparsity in CNNs and related architectures was also ingeniously investigated in Galanti et al. (2024).

Problem Formulation

We begin by briefly describing the theoretical framework from Arora et al. (2019) for unsupervised learning task, which we will use to formulate our generalization bounds for unsupervised risk. Let $\mathcal{X}$ denote the space of all possible data points and let $\mathcal{C}$ denote the set of all latent classes. Let ρ be the discrete probability measure over $\mathcal{C}$ and for any

$c \in \mathcal{C}$, denote $\mathcal{D}_c$ as the class-conditional distribution such that for any $x \in \mathcal{X}$, $\mathcal{D}_c(x)$ quantifies the likelihood of x being relevant to class c. We also define the distribution $\bar{\mathcal{D}}_c$:

$$\bar{\mathcal{D}}_c(x) = \frac{\sum_{z \in \mathcal{C}, z \neq c} \rho(z) \mathcal{D}_z(x)}{\sum_{z \in \mathcal{C}, z \neq c} \rho(z)}, \tag{3}$$

which quantifies the conditional distribution of $x \in \mathcal{X}$, conditionally given that the class is not equal to c. Then, we can define the population unsupervised risk for a representation function as follows.

Definition 1. (Unsupervised risk). Let $f : \mathcal{X} \to \mathcal{R} \subset \mathbb{R}^d$ be a representation function and $\ell : \mathbb{R}^k \to \mathbb{R}_+$ be a loss function. The population unsupervised risk of f is:

$$\mathrm{L}_{\mathrm{un}}(f) = \tag{4}$$
$$\mathbb{E}_{\substack{c \sim \rho, (x, x^+) \sim \mathcal{D}_c^2 \\ (x_1^-, \ldots, x_k^-) \sim \bar{\mathcal{D}}_c^k}} \left[\ell\Big(\Big\{ f(x)^\top (f(x^+) - f(x_i^-)) \Big\}_{i=1}^k \Big) \right].$$

A natural way to find a representation function with low expected unsupervised risk is via empirical risk minimization. Specifically, given a hypothesis class $\mathcal{F}$ and a dataset of the form $\mathcal{S} = \Big\{ (x_j, x_j^+, x_{j1}^-, \ldots, x_{jk}^-) \Big\}_{j=1}^n$, the best representation function is then determined as the empirical risk minimizer $\widehat{f}_n = \arg\min_{f \in \mathcal{F}} \widehat{\mathrm{L}}_{\mathrm{un}}(f)$.

In this paper, we are interested in the performance of $\widehat{f}_n$ on testing dataset. More precisely, we are concerned with its capability to generalize to unseen data. This is often quantified by the generalization gap between the expected unsupervised risk and the empirical unsupervised risk $\mathrm{L}_{\mathrm{un}}(\widehat{f}_n) - \widehat{\mathrm{L}}_{\mathrm{un}}(\widehat{f}_n)$. We bound this gap by controlling the Rademacher complexity $\hat{\mathfrak{R}}_{\mathcal{S}}(\mathcal{G})$ of the loss function class, which we define for a general loss function $\ell : \mathbb{R}^k \to \mathbb{R}_+$ as follows:

$$\mathcal{G} = \Big\{ (x, x^+, x_1^-, \ldots, x_k^-) \mapsto \ell\Big(\Big\{ f(x)^\top (f(x^+) - f(x_i^-)) \Big\}_{i=1}^k \Big) : f \in \mathcal{F} \Big\}. \tag{5}$$

More specifically, we are interested in the case where $\mathcal{F}$ is a class of multi-layered deep neural networks and the loss function ℓ is ℓ^∞-Lipscthiz.

Main Results

Contributions

We aim to establish a solid theoretical foundation for DCRL using the flexibility of covering number arguments. The advantages of our bounds are two-fold. Firstly, we manage to alleviate the strong reliance on the number of negative samples and the product of spectral norms using only covering numbers without introducing complexity measures other than Rademacher complexity. Secondly, through loss function augmentation schemes, we are able to further alleviate implicit depth dependency by incorporating data-dependent properties in the bounds. We summarize our key contributions as follows:

1. **Basic generalization bound (Thm 1)**: Using a pure covering number approach, we establish a bound for ℓ^∞-Lipschitz loss functions with logarithmic dependency on the number of negative samples, which involves a spectral-type complexity measure for the neural network component, but features the square of the spectral norms.
2. **Loss function augmentation (Thms 2 & 3)**: We improve the basic bound through loss function augmentation: Theorem 2 replaces the extra factor of the product of spectral norms by an empirical maximum output norm, whilst Theorem 3 further improves depth dependency by introducing empirical estimates of intermediate norm activations.
3. **Parameter counting bound (Thm 4)**: In a different style from the above results, we derive a bound that scales with the overall size of the neural networks, i.e. the total number of neurons.

In table 1, we provide a comprehensive summary of our main results as well as the results from the previous works.

Remark 1. We note that in table 1, the original bounds from Arora et al. (2019) and Lei et al. (2023) involve an upper bound B on the output's ℓ^2 norm that is assumed to hold for *any* representation function in the class. Since we are dealing with a class of neural networks, the upper bound B is expanded to $B_x \prod_{l=1}^L \rho_l s_l$.

Notations

Let $L \geq 1$, $d_0, d_1, \ldots, d_L$ be known natural numbers and $M^{(l)} \in \mathbb{R}^{d_l \times d_{l-1}}$ be fixed reference matrices. Let $\{a_l\}_{l=1}^L$, $\{s_l\}_{l=1}^L$ be sequences of positive real numbers. We define the following matrix spaces:

$$\mathcal{B}_l = \Big\{ A^{(l)} \in \mathbb{R}^{d_l \times d_{l-1}} : \|A^{(l)}\|_\sigma \leq s_l, \ \|(A^{(l)} - M^{(l)})^\top\|_{2,1} \leq a_l \Big\}, \tag{6}$$

where $\|\cdot\|_\sigma$ denotes the spectral norm and $\|\cdot\|_{2,1}$ denotes the entry-wise matrix norm quantified by the sum of matrix columns' ℓ^2 norms. The reference matrices $M^{(l)}$ are fixed before training and often interpreted as initializations of weight matrices (Bartlett, Foster, and Telgarsky 2017; Ledent et al. 2021b). We define the product $\mathcal{A} = \prod_{l=1}^L \mathcal{B}_l$ as the parameters space for the class of neural networks $\mathcal{F}_{\mathcal{A}}$:

$$\mathcal{F}_{\mathcal{A}} = \mathcal{F}_L \circ \mathcal{F}_{L-1} \circ \cdots \circ \mathcal{F}_1, \tag{7}$$

where $\mathcal{F}_l = \sigma_l \circ \mathcal{V}_l$ such that:

- $\sigma_l : \mathbb{R}^{d_l} \to \mathbb{R}^{d_l}$ are ℓ^2-Lipschitz activation functions with constants ρ_l chosen a priori.
- $\mathcal{V}_l = \Big\{ z \mapsto A^{(l)} z : A^{(l)} \in \mathcal{B}_l \Big\}$ are classes of linear maps corresponding to pre-activated linear layers.

For a given set of weights $\mathbf{A} = (A^{(L)}, \ldots, A^{(1)})$ where for each $1 \leq l \leq L$, $A^{(l)} \in \mathcal{B}_l$, we denote $F_{\mathbf{A}} \in \mathcal{F}_{\mathcal{A}}$ as the corresponding neural network parameterized by $\mathbf{A}$. To be specific, for any $x \in \mathcal{X}$:

$$F_{\mathbf{A}}(x) = \sigma_L\Big(A^{(L)} \sigma_{L-1}\Big(\ldots \sigma_1\Big(A^{(1)} x \Big) \ldots \Big) \Big).$$

References	Analysis Technique	Generalization Bound	Result
Arora et al. (2019)	Peeling technique	$\tilde{\mathcal{O}}\Big(\frac{\eta B_x^2\sqrt{kdL}}{\sqrt{n}}\prod_{l=1}^{L}\rho_l s_l B_{Fr}^{(l)}\Big)$	–
Lei et al. (2023)	Peeling technique	${}^*\tilde{\mathcal{O}}\Big(\frac{\eta B_x^2\sqrt{dL}}{\sqrt{n}}\prod_{l=1}^{L}\rho_l s_l B_{Fr}^{(l)}\Big)$	–
Ours	Covering number	${}^*\tilde{\mathcal{O}}\Big(\frac{\eta B_x^2}{\sqrt{n}}\prod_{m=1}^{L}\rho_m^2 s_m^2\Big[\sum_{l=1}^{L}\frac{a_l^{2/3}}{s_l^{2/3}}\Big]^{3/2}\Big)$	Thm. 1
Ours	Covering number & augmentation	${}^*\tilde{\mathcal{O}}\Big(\frac{\eta R B_x}{\sqrt{n}}\prod_{m=1}^{L}\rho_m s_m\Big[\sum_{l=1}^{L}\frac{a_l^{2/3}}{s_l^{2/3}}\Big]^{3/2}\Big)$	Thm. 2
Ours	Covering number & augmentation	${}^*\tilde{\mathcal{O}}\Big(\frac{\eta b_L^2}{\sqrt{n}}\Big[\sum_{l=1}^{L}(a_l b_{l-1}\hat{\rho}_l)^{2/3}\Big]^{3/2}\Big)$	Thm. 3
Ours	Parameter counting	$\mathcal{O}\Big(\sqrt{\frac{\mathcal{W}}{n}}\log\Big(\eta L n B_x^2\prod_{l=1}^{L}\rho_l^2 s_l^2\Big)\Big)$	Thm. 4

Table 1: Summary of main results for Deep Contrastive Representation Learning (DCRL). We assume that the loss function of concern is ℓ^∞-Lipschitz with constant $\eta \geq 1$. The $\tilde{\mathcal{O}}$ notation hides poly-logarithmic terms of ALL variables and $({}^*)$ marks the bounds that have hidden logarithmic dependency on k.

In the results that follow, we present generalization bounds for unsupervised risk applied for neural networks in the hypothesis class $\mathcal{F}_{\mathcal{A}}$. However, we note that our results can be easily made post-hoc to apply for any neural network.

Basic Bound

In this section, we present the basic generalization bound without applying loss augmentation. We begin by stating the definition for ℓ^∞-Lipschitz continuity (Lei et al. 2019).

Definition 2 (Lipschitz continuity)**.** We say that a function $\ell : \mathbb{R}^k \to \mathbb{R}_+$ is Lipschitz continuous with respect to the ℓ^∞ norm with a constant $\eta > 0$ if and only if:

$$|\ell(v) - \ell(\bar{v})| \leq \eta \cdot \|v - \bar{v}\|_\infty, \quad \forall v, \bar{v} \in \mathbb{R}^k. \tag{8}$$

Theorem 1. *Let $\ell : \mathbb{R}^k \to [0, M]$ be a loss function that is ℓ^∞-Lipschitz with constant $\eta > 0$. Then, for any $F_{\mathbf{A}} \in \mathcal{F}_{\mathcal{A}}$ and $\delta \in (0, 1)$, the following bound holds with probability of at least $1 - \delta$:*

$$\begin{aligned} &\mathrm{L}_{\mathrm{un}}(F_{\mathbf{A}}) - \widehat{\mathrm{L}}_{\mathrm{un}}(F_{\mathbf{A}}) \leq 3M\sqrt{\frac{\log 2/\delta}{2n}} + \\ &\tilde{\mathcal{O}}\left(\frac{\eta B_x^2}{\sqrt{n}}\log(W)\prod_{m=1}^{L}\rho_m^2 s_m^2\left[\sum_{l=1}^{L}\frac{a_l^{2/3}}{s_l^{2/3}}\right]^{3/2}\right), \end{aligned} \tag{9}$$

where $W = \max_{1\leq l\leq L} d_l$ (maximum hidden width), $B_x = \sup_{x\in\mathcal{X}}\|x\|_2$, and the $\tilde{\mathcal{O}}$ notation hides logarithmic factors in all relevant quantities.

Remark 2. Whilst it is standard practice to assume that the loss function is bounded by a fixed constant (Long and Sedghi 2020; Bartlett, Foster, and Telgarsky 2017; Ledent and Alves 2024; Ledent et al. 2021a; Shamir and Shalev-Shwartz 2014), even in the case where the loss function is not bounded, we can still find an upper bound M for the loss owing to the fact that the weight matrices have bounded norms. Specifically, for all $F_{\mathbf{A}} \in \mathcal{F}_{\mathcal{A}}$, we can make the following estimation $M = \mathcal{O}\Big(\eta B_x^2\prod_{l=1}^{L}\rho_l^2 s_l^2\Big)$. Furthermore, no additional dependence on the number of classes is introduced implicitly through η when working with the logistic and hinge losses as we do: indeed, the L^∞ Lipschitz constant is bounded by $\eta = 1$ in both cases, as shown in Appendix H. Furthermore, we discuss the relationship between the cross-entropy loss from standard classification and its analogues used in CRL. Among the most common analogues are the N-pair loss (which is the logistic loss) and the InfoNCE loss (van den Oord, Li, and Vinyals. 2018). The results in this paper extend naturally to the InfoNCE loss.

The above bound is the result of directly using covering number to bound the Rademacher complexity. Unlike the vector contraction inequality approach in Arora et al. (2019) where the $\sqrt{k}$ dependency creeps into the bound, we immediately observe an absence of significant reliance on the number of negative samples in this result.

However, the bound also features a factor of the *square* of the product of spectral norms of all layers. This is in contrast to existing norm-based generalization bounds for ordinary neural networks, which typically feature a single product of spectral norms (Bartlett, Foster, and Telgarsky 2017). Roughly speaking, this new square dependency in the Contrastive Learning Setting is a byproduct of the presence of multiplicative interactions between $f(x)$ and $f(x^+)$ or $f(x^-)$, which means that the errors propagate through the network twice. In the next section, we discuss how we can make use of data-dependent properties to alleviate this issue with simple loss augmentation techniques.

Loss Augmentation

In previous works dedicated to multi-class classification problem (Nagarajan and Kolter 2019; Wei and Ma 2019; Ledent et al. 2021b), it has been shown that we can obtain tighter Rademacher complexity bound by incorporating data-dependent quantities. Informally, this is accomplished by augmenting the original loss function in a way that the augmented loss collapses to a large value if certain data-dependent well-behaved-ness properties do not hold. For instance, given the list of data-dependent properties $\{\gamma_l\}_{l=1}^m$ and their corresponding desired bounds $\mathbf{B} = \{b_l\}_{l=1}^m$, Wei

and Ma (2019) employ an augmentation scheme involving products of soft indicators of the data-dependent properties:

$$\tilde{\ell}(x) = 1 + (\ell(x) - 1) \prod_{l=1}^{m} \lambda_{b_l}(\gamma_l(x)), \tag{10}$$

where $\ell : \mathbb{R}^k \to [0, 1]$ is the original loss function and λ_{b_l} are soft indicators with margins b_l (whose definition is identical to that of the ramp loss), defined as follows:

$$\lambda_\gamma(r) = \begin{cases} 0 & r < -\gamma \\ 1 + r/\gamma & r \in [-\gamma, 0] \\ 1 & r > 0 \end{cases}. \tag{11}$$

Another example of loss augmentation is the work of Ledent et al. (2021b) where the augmented loss is the maximum value between the original loss and the maximum of the soft indicators themselves:

$$\tilde{\ell}(x) = \max \Big[\ell(x), \max_{1 \le l \le m} \lambda_{b_l}(\gamma_l(x)) \Big]. \tag{12}$$

These soft indicators act as validation filters for the intended data-dependent properties. Specifically, the value of $\tilde{\ell}$ will coincide with the original loss value if all bound conditions are met. On the other hand, when $\gamma_l(x) \ge 2b_l$ for any $1 \le l \le m$, $\tilde{\ell}$ will collapse to the upper bound of ℓ, making the augmented loss uniformly larger than the original loss. As a result, we can bound the excess risk of the original loss indirectly via the augmented loss. To be more precise, let $\mathcal{D}$ be a distribution over an input space $\mathcal{X}$ and $S = \{x_1, \ldots, x_n\} \in \mathcal{X}^n$ be a dataset drawn i.i.d from $\mathcal{D}$. Define the excess risk for a particular loss $\ell : \mathbb{R}^k \to \mathbb{R}_+$ as $\mathcal{E}[\ell; S] = \mathbb{E}_{x \sim \mathcal{D}}[\ell(x)] - \frac{1}{n} \sum_{j=1}^{n} \ell(x_j)$, we have:

$$\mathcal{E}[\ell; S] \le \mathcal{E}[\tilde{\ell}; S] + \frac{\mathcal{I}_{\mathbf{B}}}{n}, \tag{13}$$

where $\mathcal{I}_{\mathbf{B}} = \Big| \Big\{ x_j \in S : \exists l \text{ s.t } \gamma_l(x_j) > b_l \Big\} \Big|$, which is the count of data points that do not satisfy all bound conditions. Notice that we can bound the augmented generalization gap $\mathbb{E}_{x \sim \mathcal{D}}[\tilde{\ell}(x)] - \frac{1}{n} \sum_{j=1}^{n} \tilde{\ell}(x_j)$ by controlling the Rademacher complexity of the augmented loss class (which we denote by default as $\tilde{\mathcal{G}}$). The difficulty of this approach is that the Rademacher complexity of the augmented loss class can be much more complex than the original class.

In this section, we consider augmentation schemes that tighten the generalization bound and improve on the result in Theorem 1. By default, we consider the original loss function $\ell : \mathbb{R}^k \to \mathbb{R}_+$ to be ℓ^∞-Lipschitz with constant $\eta > 0$ and, without loss of generality, we assume that ℓ is bounded by 1. Our first attempt is through imposing bound conditions on the representation output of the neural networks. To be more precise, let $R > 0$ be a fixed real constant intended to be the upper bound for the output representation's ℓ^2 norm, we consider the following augmented loss function class:

$$\tilde{\mathcal{G}} = \Big\{ \mathbf{X}^{(\text{in})} = (x, x^+, x_1^-, \ldots, x_k^-) \mapsto \max \Big[\ell\Big(\Big\{ F_{\mathbf{A}}(x)^\top (F_{\mathbf{A}}(x^+) - F_{\mathbf{A}}(x_i^-)) \Big\}_{i=1}^{k} \Big), \max_{\tilde{x} \in \mathbf{X}^{(\text{in})}} \lambda_R(\|F_{\mathbf{A}}(\tilde{x})\|_2) \Big] : F_{\mathbf{A}} \in \mathcal{F}_{\mathcal{A}} \Big\}. \tag{14}$$

Bounding the Rademacher complexity of the above class results in the following theorem, which is our second main contribution.

Theorem 2. *Let $\ell : \mathbb{R}^k \to [0, 1]$ be a loss function that is ℓ^∞-Lipschitz with constant $\eta \ge 1$ and let $R \ge 1$ be given. Then, for any $F_{\mathbf{A}} \in \mathcal{F}_{\mathcal{A}}$ and $\delta \in (0, 1)$, the following bound holds with probability of at least $1 - \delta$:*

$$\mathrm{L}_{\text{un}}(F_{\mathbf{A}}) - \widehat{\mathrm{L}}_{\text{un}}(F_{\mathbf{A}}) \le \frac{\mathcal{I}_{\mathbf{A},R}}{n} + 3\sqrt{\frac{\log 2/\delta}{2n}} + \tilde{\mathcal{O}}\left(\frac{\eta R B_x}{\sqrt{n}} \log(W) \prod_{m=1}^{L} \rho_m s_m \left[\sum_{l=1}^{L} \frac{a_l^{2/3}}{s_l^{2/3}} \right]^{3/2} \right), \tag{15}$$

where $W = \max_{1 \le l \le L} d_l$, $B_x = \sup_{x \in \mathcal{X}} \|x\|_2$, and $\mathcal{I}_{\mathbf{A},R}$ is defined as:

$$\mathcal{I}_{\mathbf{A},R} = \Big| \Big\{ \mathbf{X}_j^{(\text{in})} \in \mathcal{S} : \max_{\tilde{x} \in \mathbf{X}_j^{(\text{in})}} \|F_{\mathbf{A}}(\tilde{x})\|_2 > R \Big\} \Big|,$$

where $\mathbf{X}_j^{(\text{in})} = (x_j, x_j^+, x_{j1}^-, \ldots, x_{jk}^-)$ is the j^{th} input tuple from the dataset $\mathcal{S}$.

Remark 3. Note that if all the samples satisfy $\max_{\tilde{x} \in \mathbf{X}_j^{(\text{in})}} \|F_{\mathbf{A}}(\tilde{x})\|_2 \le R$, then $\frac{\mathcal{I}_{\mathbf{A},R}}{n} = 0$. Furthermore, by a union bound, it is not difficult to show that a similar result holds even if R is selected from the data by observing the value of $\max_{\tilde{x} \in \mathbf{X}_j^{(\text{in})}} \|F_{\mathbf{A}}(\tilde{x})\|_2$. For further details, we refer the reader to Appendix G, 'Post Hoc Analysis'. This remark also applies to Theorem 3.

Further, although Theorem 2 assumes the original loss function is bounded by 1, we can easily generalize to any loss function $\ell : \mathbb{R}^k \to [0, M]$ by considering a slightly different augmentation scheme where the original loss is normalized to $M^{-1}\ell$ inside the max function. Then, an equivalent bound to theorem 2 for loss functions bounded by an arbitrary $M > 0$ is:

$$\mathrm{L}_{\text{un}}(F_{\mathbf{A}}) - \widehat{\mathrm{L}}_{\text{un}}(F_{\mathbf{A}}) \le \frac{M\mathcal{I}_{\mathbf{A},R}}{n} + 3M\sqrt{\frac{\log 2/\delta}{2n}} + \tilde{\mathcal{O}}\left(\frac{\eta R B_x}{\sqrt{n}} \log(W) \prod_{m=1}^{L} \rho_m s_m \left[\sum_{l=1}^{L} \frac{a_l^{2/3}}{s_l^{2/3}} \right]^{3/2} \right). \tag{16}$$

Compared to theorem 1, we have successfully reduced the product of squared spectral norms dependency down to a single product of spectral norms at the cost of a multiplicative factor of the more well-behaved empirical quantity R and an additive term of $\mathcal{I}_{\mathbf{A},B}/n$, which is the proportion of inputs in $\mathcal{S}$ that do not satisfy the output bound condition. In the following, we illustrate that the bound can be improved further by considering an augmentation scheme that enforces bounds on all the hidden layers' activations. Specifically, given $\mathbf{B} = \{b_0, b_1, \ldots, b_L\}$ a sequence of known positive constants, we consider the following augmented class:

$$\tilde{\mathcal{G}} = \Big\{ \mathbf{X}^{(\text{in})} = (x, x^+, x_1^-, \ldots, x_k^-) \mapsto \max \Big[\ell\Big(\Big\{ F_{\mathbf{A}}(x)^\top (F_{\mathbf{A}}(x^+) - F_{\mathbf{A}}(x_i^-)) \Big\}_{i=1}^{k} \Big), \max_{1 \le l \le L} \max_{\tilde{x} \in \mathbf{X}^{(\text{in})}} \lambda_{b_l}(\|F_{\mathbf{A}}^{1 \to l}(\tilde{x})\|_2) \Big] : F_{\mathbf{A}} \in \mathcal{F}_{\mathcal{A}} \Big\}. \tag{17}$$

where $F_{\mathbf{A}}^{1\to l}$ denotes the sub-network that consists of the first l layers of $F_{\mathbf{A}} \in \mathcal{F}_{\mathcal{A}}$. Bounding the above class Rademacher complexity yields the following result:

Theorem 3. *Let $\ell : \mathbb{R}^k \to [0,1]$ be a loss function that is ℓ^∞-Lipschitz with constant $\eta \geq 1$. Let $\mathbf{B} = \{b_0, b_1, \ldots, b_L\}$ be a sequence of known positive constants such that $b_l \geq 1$ for all $0 \leq l \leq L$. Then, for any $F_{\mathbf{A}} \in \mathcal{F}_{\mathcal{A}}$ and $\delta \in (0,1)$, the following bound holds with probability of at least $1-\delta$:*

$$\mathrm{L}_{\mathrm{un}}(F_{\mathbf{A}}) - \widehat{\mathrm{L}}_{\mathrm{un}}(F_{\mathbf{A}}) \leq \tilde{\mathcal{O}}\left(\frac{\eta b_L^2 \widehat{\mathcal{R}}_{\mathcal{A}}}{\sqrt{n}}\log(W)\right) + \frac{\mathcal{I}_{\mathbf{A},\mathbf{B}}}{n} + 3\sqrt{\frac{\log 2/\delta}{2n}}, \tag{18}$$

where $\widehat{\mathcal{R}}_{\mathcal{A}}$ is defined as follows:

$$\widehat{\mathcal{R}}_{\mathcal{A}}^{2/3} = \sum_{l=1}^{L}(a_l b_{l-1}\hat{\rho}_l)^{2/3},$$

where $\hat{\rho}_l = \rho_l \sup_{u\geq l} b_u^{-1} \prod_{m=l+1}^{u} s_m \rho_m$ and

$$\mathcal{I}_{\mathbf{A},\mathbf{B}} = \left|\left\{\mathbf{X}_j^{(\mathrm{in})} \in \mathcal{S} : \exists l \text{ s.t } \max_{\tilde{x}\in\mathbf{X}_j^{(\mathrm{in})}} \|F_{\mathbf{A}}^{1\to l}(\tilde{x})\|_2 > b_l\right\}\right|.$$

Again, without loss of generality, we can derive an analogous bound to the above result for loss functions bounded by any $M > 0$. Unlike the previous result which depends on the full L layers product of spectral norms, the l^{th} term in the above result only involves spectral norms of layers $l+1$ up to L (but not necessarily all the way to L).

Parameter Counting Bound

Inspired by previous works developed for neural networks used in multi-class classification (Long and Sedghi 2020; Graf et al. 2022; Srebro 2004; Mohri, Rostamizadeh, and Talwalkar 2018), our result below scales with network's size rather than the magnitude of weight matrices norms like the bounds presented in the previous section. The advantage of this type of bounds is the absence of a product of spectral norms (outside logarithmic factors), which effectively eliminates the strong dependency on neural network's depth.

Theorem 4. *Let $\ell : \mathbb{R}^k \to [0,M]$ be a loss function that is ℓ^∞-Lipschitz with constant $\eta > 0$ and $\mathcal{W} = \sum_{l=1}^{L} d_l$. Then, for any $F_{\mathbf{A}} \in \mathcal{F}_{\mathcal{A}}$ and $\delta \in (0,1)$, the following bound holds with probability of at least $1-\delta$:*

$$\mathrm{L}_{\mathrm{un}}(F_{\mathbf{A}}) - \widehat{\mathrm{L}}_{\mathrm{un}}(F_{\mathbf{A}}) \leq 3M\sqrt{\frac{\log 2/\delta}{2n}} + \mathcal{O}\left(M\sqrt{\frac{\mathcal{W}}{n}}\log\left(1 + 24\eta L n B_x^2 \prod_{l=1}^{L}\rho_l^2 s_l^2\right)\right). \tag{19}$$

Essentially, the above result scales with the total number of parameters of the neural networks. This characteristic can be disadvantageous compared to the previous norm-based results because (1) the bound can become unreasonably large for massive architectures and (2) the bound will still scale with $\mathcal{W}$ even if the weight matrices are arbitrarily close to the reference matrices. Even though the above bound might scale unfavourably in the case of large neural networks, we note that it has no reliance on the number of negative samples. Hence, it can be particularly useful in cases when the networks are small and we have a large amount of negative samples.

Downstream Classification

In this section, we discuss the application of the generalization bounds for unsupervised risk in the downstream classification task. We begin with the following definition of a classifier's population supervised risk.

Definition 3. (Supervised risk). Fixing a $(K+1)$-way supervised task $\mathcal{T} = \{c_1, \ldots, c_{K+1}\} \subseteq \mathcal{C}$ (where $\mathcal{C}$ is the set of latent classes defined in the previous section). Let $g : \mathcal{X} \to \mathbb{R}^{K+1}$ be a multi-class classifier and $\ell : \mathbb{R}^K \to \mathbb{R}_+$ be a loss function. The population supervised risk of g is defined as follows:

$$\mathrm{L}_{\mathrm{sup}}(\mathcal{T}, g) = \mathbb{E}_{(x,c)\sim\mathcal{D}_{\mathcal{T}}}\left[\ell\left(\left\{g(x)_c - g(x)_{c'}\right\}_{c'\neq c}\right)\right],$$

where $\mathcal{D}_{\mathcal{T}}$ is the joint distribution over $\mathcal{X}\times\mathcal{T}$.

In particular, we are interested in the class of mean classifiers from Arora et al. (2019). Let $\mathcal{T} \subseteq \mathcal{C}$ such that $|\mathcal{T}| = K+1$ and $f : \mathcal{X} \to \mathbb{R}^d$ be a representation function. A mean classifier $g : \mathcal{X} \to \mathcal{T}$ is defined as $g(x) = W^\mu f(x)$, where $W^\mu \in \mathbb{R}^{(K+1)\times d}$ is a weight matrix such that for each $c \in \mathcal{T}$, the c^{th} row of W^μ is the expected representation of $x \in \mathcal{X}$ given that x is relevant to class c. Specifically, $W_c^\mu = \mathbb{E}_{x\sim\mathcal{D}_c}[f(x)]$. Consider the average supervised loss:

$$\mathrm{L}_{\mathrm{sup}}^{\mu}(f) = \mathbb{E}_{\mathcal{T}\sim\rho^{K+1}}\left[\mathrm{L}_{\mathrm{sup}}(\mathcal{T}, W^\mu f)\middle| c_i \neq c_j\right],$$

which is the expectation of the mean classifier's supervised loss taken over $(K+1)$-way supervised tasks (with unique classes). In a general sense, the average supervised loss can be translated to a performance metric for the representation f when it is used to build a mean classifier. In the following lemma from Arora et al. (2019), it is shown the average supervised loss can be upper bounded by the population unsupervised risk:

Lemma 1. Fixing a class of representation functions $\mathcal{F}$ and let $\widehat{f}_n = \arg\min_{f\in\mathcal{F}} \widehat{\mathrm{L}}_{\mathrm{un}}(f)$. There exists a function $\rho : \mathcal{C}^{K+1} \to \mathbb{R}_+$[1] such that:

$$\mathbb{E}_{\mathcal{T}\sim\mathcal{D}}\left[\rho(\mathcal{T})\mathrm{L}_{\mathrm{sup}}^{\mu}(\widehat{f}_n)\right] \leq \mathrm{L}_{\mathrm{un}}(\widehat{f}_n), \tag{20}$$

where $\mathcal{D}$ is a distribution over $(K+1)$-way supervised tasks $\mathcal{T} \in \mathcal{C}^{K+1}$ such that there are no repeated classes in $\mathcal{T}$.

In the following results, we directly apply the generalization bounds obtained in the previous sections into lemma 1.

Corollary 1. (Norm-based bound). Let $\ell : \mathbb{R}^k \to [0,M]$ be a an ℓ^∞-Lipschitz loss with $\eta \geq 1$. Let $\mathbf{B} = \{b_0, b_1, \ldots, b_L\}$ be a sequence of known positive constants such that $b_l \geq 1$

[1]This lemma is an intermediate step in the proof of theorem B.1 from Arora et al. (2019, equation 26). For the exact form of the function ρ, we refer readers to their proof. For the formal definition of the distribution $\mathcal{D}$, please refer to Arora et al. (2019, section 6.1).

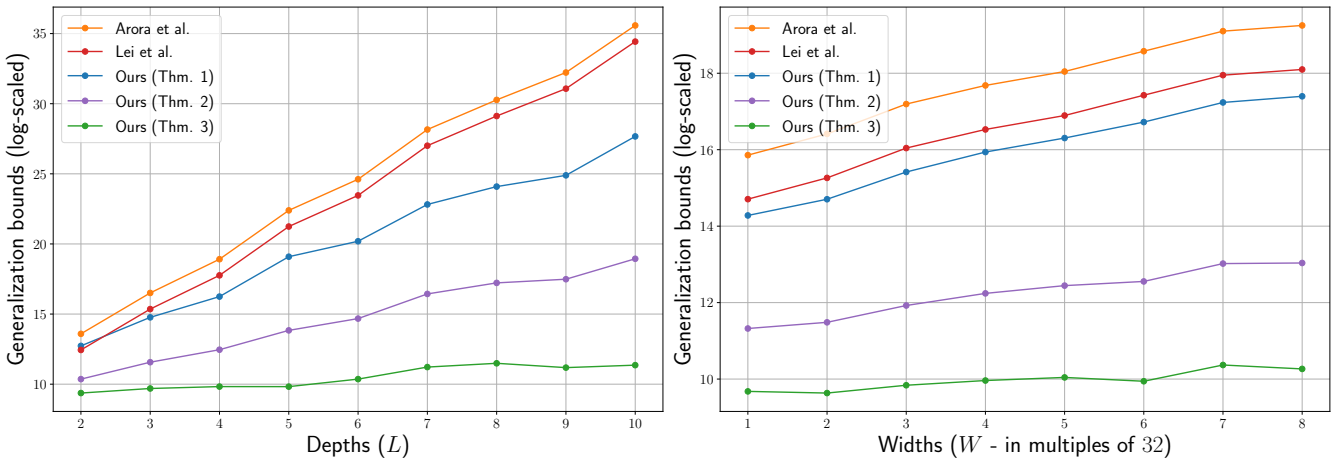

Figure 1: Graphical comparison of our results to that of previous works (Arora et al. 2019; Lei et al. 2023). The generalization bounds for all results have their logarithmic terms, constants ($\eta, \rho_i, \dots$) and $\mathcal{O}(\sqrt{\log 1/\delta})$ terms truncated. We present the comparison at varying depths (Left) and hidden layer's dimensions (Right).

for all $0 \leq l \leq L$. Let $\widehat{F}_{\mathbf{A}}$ be the empirical unsupervised risk minimizer, then, for any $\delta \in (0,1)$, we have:

$$\mathbb{E}_{\mathcal{T}\sim\mathcal{D}}\left[\rho(\mathcal{T})\mathrm{L}^{\mu}_{\mathrm{sup}}(\widehat{F}_{\mathbf{A}})\right] \leq \widehat{\mathrm{L}}_{\mathrm{un}}(\widehat{F}_{\mathbf{A}}) + 3M\sqrt{\frac{\log 2/\delta}{2n}} + \tilde{\mathcal{O}}\left(\frac{\eta b_L^2 \widehat{\mathcal{R}}_{\mathcal{A}}}{\sqrt{n}}\log(W)\right) + \frac{M\mathcal{I}_{\mathbf{A},\mathbf{B}}}{n}, \quad (21)$$

where $\widehat{\mathcal{R}}_{\mathcal{A}}$ and $\mathcal{I}_{\mathbf{A},\mathbf{B}}$ are defined in Theorem 3.

Corollary 2. (Parameter-counting bound). Let $\ell : \mathbb{R}^k \to [0, M]$ be a loss function that is ℓ^∞-Lipschitz with constant $\eta \geq 1$. Let $\widehat{F}_{\mathbf{A}}$ be the empirical unsupervised risk minimizer, then, for any $\delta \in (0,1)$, we have:

$$\mathbb{E}_{\mathcal{T}\sim\mathcal{D}}\left[\rho(\mathcal{T})\mathrm{L}^{\mu}_{\mathrm{sup}}(\widehat{F}_{\mathbf{A}})\right] \leq \widehat{\mathrm{L}}_{\mathrm{un}}(\widehat{F}_{\mathbf{A}}) + 3M\sqrt{\frac{\log 2/\delta}{2n}} + \mathcal{O}\left(M\sqrt{\frac{\mathcal{W}}{n}}\log\left(1 + 24\eta L n B_x^2 \prod_{l=1}^{L}\rho_l^2 s_l^2\right)\right). \quad (22)$$

Experiments

To compare our results with previous works, we conducted experiments by training fully-connected deep neural networks with the MNIST digits dataset (LeCun, Cortes, and Burges 2010) with a train-test ratio of $75\%/25\%$. We ran two ablation studies to test how our bounds vary with network depth and hidden layer dimension compared to the bounds proposed by Arora et al. (2019) and Lei et al. (2023). For the first experiment, we fixed the hidden layer dimensions to 64 and trained deep neural networks at different depths in the $[2, 10]$ range. For the second experiment, we fixed the depth to $L = 3$ and trained deep neural networks at different hidden layer dimensions of $32, 64, 128, \dots$ (in multiples of 32). In both experiments, we fixed the output dimension to $d = 64$ and the number of negative samples to $k = 10$ (furthermore, additional experiments with $k = 64$ are provided in Appendix J of the full ArXiv version). For all the neural networks trained in both experiments, we set the maximum number of training iterations to 1000 and stopped until the empirical unsupervised loss reached $1\mathrm{e}{-4}$ to ensure that all networks roughly converge to the empirical risk minimizers. A summary of our experiment results is presented in figure 1: the y axis shows the main factor in our and competing bounds, ignoring constants and logarithmic terms in all cases. The results demonstrate that our generalization bounds outperform the competing ones, especially for larger widths and depths.

Conclusion and Further Works

There is very limited amount of theoretical work explaining the impressive real world performance that CRL has achieved. Existing works focus on the case of general classes of representation functions through direct arguments on the Rademacher complexity. In the case of neural networks, this introduces strong depth dependency through a product of Frobenius norms of the weight matrices. In this work, we provided bounds relied on applying covering number arguments to carefully constructed auxiliary datasets to provide bounds with a milder dependency on depth. We also illustrate that with such a technique, the bounds automatically admit a weak dependency on the number of negative samples. Moreover, through loss augmentation, we improve our results by introducing data-dependent terms into the bounds, lessening the effect of residual exponential growth with the neural network's depth. In further work, it would be interesting to generalize our work to other architectures such as CNNs, GNNs, ResNets. Furthermore, a particularly tantalizing direction would be to study the generalization properties of CRL in the more realistic and challenging setting where the input tuples are formed from a fixed pool of reusable labeled examples. This would be much more challenging due to the violation of the i.i.d. assumption.

Acknowledgements

YL acknowledges support by the Research Grants Council of Hong Kong [Project No. 22303723].

References

Alves, R.; and Ledent, A. 2024. Context-Aware REpresentation: Jointly Learning Item Features and Selection From Triplets. *IEEE Transactions on Neural Networks and Learning Systems*.

Anthony; and Bartlett. 2002. *Neural Network Learning - Theoretical Foundations*. Cambridge University Press.

Arora; Khandeparkar; Khodak; Plevrakis; and Saunshi. 2019. A Theoretical Analysis of Contrastive Unsupervised Representation Learning. In *International Conference on Machine Learning*.

Awasthi, P.; Dikkala, N.; and Kamath, P. 2022. Do More Negative Samples Necessarily Hurt In Contrastive Learning? In *International Conference on Machine Learning*.

Bao, H.; Nagano, Y.; and Nozawa, K. 2022. On the Surrogate Gap between Contrastive and Supervised Losses. In *International Conference on Machine Learning*, 1585–1606. PMLR.

Bartlett; Foster; and Telgarsky. 2017. Spectrally-normalized Margin Bounds for Neural Networks. In *Advances in Neural Information Processing Systems*.

Bartlett, P. L.; Harvey, N.; Liaw, C.; and Mehrabian, A. 2019. Nearly-tight VC-dimension and Pseudodimension Bounds for Piecewise Linear Neural Networks. *Journal of Machine Learning Research*, 20(63): 1–17.

Chen, T.; Kornblith, S.; Norouz, M.; and Hinton, G. 2020. A Simple Framework for Contrastive Learning of Visual Representations. In *International Conference on Machine Learning*.

Chuang, C.-Y.; Robinson, J.; Lin, Y.-C.; Torralba, A.; and Jegelka, S. 2020. Debiased Contrastive Learning. *Advances in Neural Information Processing Systems*, 33: 8765–8775.

Eldele, E.; Ragab, M.; Chen, Z.; Wu, M.; Kwoh, C. K.; Li, X.; and Guan, C. 2021. Time-Series Representation Learning via Temporal and Contextual Contrasting. *International Joint Conference on Artificial Intelligence*.

Galanti, T.; Xu, M.; Galanti, L.; and Poggio, T. 2024. Norm-based Generalization Bounds for Sparse Neural Networks. *Advances in Neural Information Processing Systems*, 36.

Gao, T.; Yao, X.; and Chen, D. 2021. SimCSE: Simple Contrastive Learning of Sentence Embeddings. In *Conference on Empirical Methods in Natural Language Processing*.

Gidaris, S.; Singh, P.; and Komodakis, N. 2018. Unsupervised Representation Learning by Predicting Image Rotation. In *International Conference on Learning Representation*.

Golowich, N.; Rakhlin, A.; and Shamir, O. 2018. Size-independent Sample Complexity of Neural Networks. In *Conference On Learning Theory*, 297–299. PMLR.

Graf, F.; Zeng, S.; Rieck, B.; Niethammer, M.; and Kwitt, R. 2022. On Measuring Excess Capacity in Neural Networks. In *Advances in Neural Information Processing Systems*.

HaoChen, J. Z.; Wei, C.; Gaidon, A.; and Ma, T. 2021. Provable Guarantees for Self-supervised Deep Learning with Spectral Contrastive Loss. *Advances in Neural Information Processing Systems*, 34: 5000–5011.

Hassani, K.; and Khasahmadi, A. H. 2020. Contrastive Multi-View Representation Learning on Graphs. In *Advances in Neural Information Processing Systems*.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2019. Momentum Contrast for Unsupervised Visual Representation Learning. In *Computer Vision and Pattern Recognition*.

Henaff, O. J.; Srinivas, A.; Fauw, J. D.; Razavi, A.; Doersch, C.; Eslami, S. M. A.; and van den Oord, A. 2020. Data-Efficient Image Recognition with Contrastive Predictive Coding. In *International Conference on Machine Learning*.

Hsu, D.; Ji, Z.; Telgarsky, M.; and Wang, L. 2021. Generalization Bounds via Distillation. In *International Conference on Learning Representations*.

Khosla, P.; Teterwak, P.; Wang, C.; Sarna, A.; Tian, Y.; Isola, P.; Maschinot, A.; Liu, C.; and Krishnan, D. 2020. Supervised Contrastive Learning. In *Advances in Neural Information Processing Systems*.

LeCun, Y.; Cortes, C.; and Burges, C. 2010. MNIST handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2.

Ledent, A.; and Alves, R. 2024. Generalization Analysis of Deep Non-linear Matrix Completion. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, 26290–26360.

Ledent, A.; Alves, R.; Lei, Y.; and Kloft, M. 2021a. Fine-grained Generalization Analysis of Inductive Matrix Completion. In *Advances in Neural Information Processing Systems*, volume 34, 25540–25552.

Ledent, A.; Mustafa, W.; Lei, Y.; and Kloft, M. 2021b. Norm-based generalisation bounds for deep multi-class convolutional neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 8279–8287.

Ledoux, M.; and Talagrand, M. 2011. *Probability in Banach Spaces*. Springer Berlin, Heidelberg.

Lee, S.; Park, T.; and Lee, K. 2024. Soft Contrastive Learning for Time Series. In *International Conference on Learning Representation*.

Lei, Y.; Dogan, Ü.; Zhou, D.-X.; and Kloft, M. 2019. Data-dependent Generalization Bounds for Multi-class Classification. *IEEE Transactions on Information Theory*, 65(5): 2995–3021.

Lei, Y.; Ledent, A.; and Kloft, M. 2020. Sharper Generalization Bounds for Pairwise Learning. *Advances in Neural Information Processing Systems*, 33: 21236–21246.

Lei, Y.; Liu, M.; and Ying, Y. 2021. Generalization Guarantee of SGD for Pairwise Learning. *Advances in Neural Information Processing Systems*, 34: 21216–21228.

Lei, Y.; Yang, T.; Ying, Y.; and Zhou, D.-X. 2023. Generalization Analysis for Contrastive Representation Learning.

In *International Conference on Machine Learning*, 19200–19227.

Lin, S.-B.; Wang, K.; Wang, Y.; and Zhou, D.-X. 2022. Universal Consistency of Deep Convolutional Neural Networks. *IEEE Transactions on Information Theory*, 68(7): 4610–4617.

Long; and Sedghi. 2020. Generalization Bounds for Deep Convolutional Neural Networks. In *International Conference on Learning Representations*.

Maurer, A. 2016. A Vector-contraction Inequality for Rademacher Complexities. In *Algorithmic Learning Theory*.

Mitrovic, J.; McWilliams, B.; Walker, J. C.; Buesing, L. H.; and Blundell, C. 2021. Representation Learning via Invariant Causal Mechanisms. In *International Conference on Learning Representations*.

Mohri, M.; Rostamizadeh, A.; and Talwalkar, A. 2018. *Foundations of Machine Learning*. The MIT Press.

Mustafa, W.; Lei, Y.; Ledent, A.; and Kloft, M. 2021. Fine-grained Generalization Analysis of Structured Output Prediction. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI)*. International Joint Conferences on Artificial Intelligence.

Nagarajan, V.; and Kolter, J. Z. 2019. Deterministic PAC-Bayesian generalization bounds for deep networks via generalizing noise-resilience. In *International Conference on Learning Representation*.

Neyshabur, B.; Tomioka, R.; and Srebro, N. 2015. Norm-based Capacity Control in Neural Networks. In *Conference on Learning Theory*, 1376–1401. PMLR.

Nie1, Y.; Nguyen, N. H.; Sinthong, P.; and Kalagnanam, J. 2023. A Time Series is Worth 64 Words: Long-term Forecasting with Transformers. In *International Conference on Learning Representation*.

Nozawa, K.; Germain, P.; and Guedj, B. 2020. PAC-Bayesian Contrastive Unsupervised Representation Learning. In *Uncertainty in Artificial Intelligence*, 21–30. PMLR.

Reimers, N.; and Gurevych, I. 2021. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networkss. In *Conference on Empirical Methods in Natural Language Processing*.

Shamir, O.; and Shalev-Shwartz, S. 2014. Matrix Completion with the Trace Norm: Learning, Bounding, and Transducing. *Journal of Machine Learning Research*, 15: 3401–3423.

Srebro, N. 2004. *Learning with Matrix Factorizations*. Ph.D. thesis, Massachusetts Institute of Technology.

Srebro, N.; Sridharan, K.; and Tewari, A. 2010. Smoothness, Low Noise and Fast rates. *Advances in Neural Information Processing Systems*, 23.

Tian, Y.; Krishnan, D.; and Isola, P. 2020. Contrastive Multiview Coding. In *European Conference on Computer Vision*.

Tian, Y.; Yu, L.; Chen, X.; and Ganguli, S. 2020. Understanding Self-supervised Learning with Dual Deep Networks. *arXiv preprint arXiv:2010.00578*.

Tsai, Y.-H. H.; Wu, Y.; Salakhutdinov, R.; and Morency, L.-P. 2020. Demystifying Self-supervised Learning: An Information-theoretical Framework. *arXiv preprint arXiv:2006.05576*, 2.

van den Oord, A.; Li, Y.; and Vinyals., O. 2018. Representation Learning with Contrastive Predictive Coding. In *Advances in Neural Information Processing Systems*.

Velickovic, P.; Fedus, W.; Hamilton, W. L.; Lio, P.; Bengio, Y.; and Hjelm, R. D. 2019. Deep Graph Informax. In *International Conference on Learning Representation*.

Wei; and Ma. 2019. Data-dependent Sample Complexity of Deep Neural Networks via Lipschitz Augmentation. In *Advances in Neural Information Processing Systems*.

Wen, W.; Li, H.; Gong, T.; and Chen, H. 2024. Towards Sharper Generalization Bounds for Adversarial Contrastive Learning. In Larson, K., ed., *International Joint Conference on Artificial Intelligence*, 5190–5198.

Wu, J.; Chen, J.; Wu, J.; Shi, W.; Wang, X.; and He, X. 2024. Understanding Contrastive Learning via Distributionally Robust Optimization. *Advances in Neural Information Processing Systems*, 36.

Wu, L.; Ledent, A.; Lei, Y.; and Kloft, M. 2021. Fine-grained Generalization Analysis of Vector-valued Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 10338–10346.

Yang, X.; Zhang, Z.; and Cui, R. 2022. TimeCLR: A self-supervised contrastive learning framework for univariate time series representation. *Knowledge-Based Systems*.

Yang, Z.; Lei, Y.; Wang, P.; Yang, T.; and Ying, Y. 2021. Simple Stochastic and Online Gradient Descent Algorithms for Pairwise Learning. *Advances in Neural Information Processing Systems*, 34: 20160–20171.

Zhang, D.; Li, S.-W.; Xiao, W.; Zhu, H.; Nallapati, R.; Arnold, A. O.; and Xiang, B. 2021. Pairwise Supervised Contrastive Learning of Sentence Representations. In *Conference on Empirical Methods in Natural Language Processing*.

Zhou, T.-Y.; and Huo, X. 2024. Learning Ability of Interpolating Deep Convolutional Neural Networks. *Applied and Computational Harmonic Analysis*, 68: 101582.

Zhu, Y.; Xu, Y.; Yu, F.; Liu, Q.; Wu, S.; and Wang, L. 2020. Deep Graph Contrastive Representation Learning. In *ICML Workshop on Graph Representation Learning and Beyond*.

Zou, X.; and Liu, W. 2023. Generalization Bounds for Adversarial Contrastive Learning. *Journal of Machine Learning Research*, 24(114): 1–54.