Robust Learning from Uncurated Data

by

Ching-Yao Chuang

B.S., National Tsing Hua University (2017)
M.S., Massachusetts Institute of Technology (2021)

Submitted to the Department of Electrical Engineering and Computer Science in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2023

Authored By: Ching-Yao Chuang
Department of Electrical Engineering and Computer Science
August 31, 2023

Certified By: Stefanie Jegelka
Associate Professor of Electrical Engineering and Computer Science
Thesis Supervisor

Certified By: Antonio Torralba
Professor of Electrical Engineering and Computer Science
Thesis Supervisor

Accepted By: Leslie A. Kolodziejski
Professor of Electrical Engineering and Computer Science
Chair, Department Committee for Graduate Students

Robust Learning from Uncurated Data

by

Ching-Yao Chuang

Submitted to the Department of Electrical Engineering and Computer Science
on August 31, 2023, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Electrical Engineering and Computer Science

Abstract

The field of machine learning has witnessed a growing interest in learning from uncurated data, which involves training models from data that has not been carefully curated or labeled. However, this type of data is typically noisy, incomplete, and riddled with errors, making it challenging for machine learning algorithms to learn effectively. This thesis focuses on the development of robust learning methods that can effectively leverage uncurated data while being resilient to the inherent noise and errors in the data. Specifically, we investigate the robustness of contrastive learning, a prominent technique for self-supervised representation learning by contrasting semantically similar and dissimilar pairs of samples.

Firstly, we delve into the fundamental challenge inherent in learning from unlabeled data. We find that eliminating false negatives and encouraging hard negatives notably enhance downstream performance and training efficiency.

Subsequently, we shift our focus to the omnipresent noise within the dataset. We pay particular attention to the emergence of false positive pairs, a phenomenon particularly prevalent in multimodal contrastive learning settings.

In the final segment of our study, we contemplate the efficient eradication of biases from large-scale models. It is observed that, when models are pretrained on biased, uncurated data, they frequently inherit numerous inappropriate biases, which consequentially lead to skewed predictions. In an effort to rectify this, we devise a debiasing algorithm that operates independently of any data or training requirements.

Throughout the dissertation, the common thread tying these three components together is a robust and comprehensive approach to mitigating the unique error types associated with unlabeled, noisy, and biased data respectively, offering substantial contributions to the realm of machine learning research.

Thesis Supervisor: Stefanie Jegelka
Title: Associate Professor of Electrical Engineering and Computer Science

Thesis Supervisor: Antonio Torralba
Title: Professor of Electrical Engineering and Computer Science

Acknowledgments

I feel incredibly fortunate to have been admitted to MIT and to have had the chance to meet so many remarkable individuals here. Research is never easy without the support from the others. There are countless individuals I want to thank for being part of this journey, witnessing both my challenges and my growth.

Firstly, I would like to express my deepest gratitude to my advisors, Stefanie Jegelka and Antonio Torralba. Stefanie introduced me to the world of machine learning, and I have greatly benefited from her deep expertise in the field. I fondly recall the days when we derived theorems together on the whiteboard in her office. Antonio, without a doubt, is among the most creative professors I've had the privilege to know. Conversations with him consistently spark fresh ideas. Both granted me immense intellectual freedom, allowing me to delve into topics that truly fascinated me. Beyond research, their unwavering support carried me through my most challenging days. I was very lucky to have Antonio and Stefanie as my thesis advisors, and was grateful to their education.

I am fortunate to have Phillip Isola as a member of my committee. His work has been a significant source of inspiration throughout my PhD journey. This thesis would not be complete without his invaluable guidance.

I would also like to thank my amazing lab mates. Engaging in research discussions with these brilliant individuals is truly one of the highlights of my time at MIT. Additionally, I have had the pleasure of collaborating with incredible co-authors during internships. My mentors, Youssef Mroueh, Yale Song, and David Alvarez-Melis, have been instrumental in shaping my PhD journey and this thesis.

To the friends I have made here, thank you for enriching my life. I firmly believe that a fulfilling life outside of research contributes to better work. The ski trips, beer pong, and battles in the summoner's rift have played an indispensable role in my journey.

I would like to thank my family, they are the every reason that I can be here today. My grandparents Jung-Chou Chuang and Hang-Szu Chuang; my parents, Chen-Ming

Chuang and Man-Ti Hsu; and my sister Ching-Yun Chuang, have unconditionally supported me through my entire life. I love my family with my whole heart. Whenever I return to Taiwan, it is the first home-cooked dinner by Mom and Grandma, along with Grandpa's finest Oolong tea, that truly revives my spirits. Bedtime chats with my sister and parents invariably uplift me, especially when I face challenges in my research.

To my father, how I wish you were here to share in this joy. The memory of your proud smile when I received my MIT offer remains etched in my heart. When I gaze at the sky, I feel as if I glimpse your comforting smile, hearing your words, 'Son, don't worry.' Dad, I believe you're watching over me from above. Today, I stand as a Ph.D., following in your footsteps. I dedicate this thesis to you.

To my wife, Yi-Yi Chu, thank you for being my steadfast anchor all these years. 12,559 km represents not just the distance between Taiwan and Boston, but the space we've bridged in our hearts. Through countless nights, we can only chat with each other through Skype. Though oceans apart, our hearts never drifted. It is you who made me believe in true love.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Machine learning has made significant progress in recent years, particularly in the area of supervised learning where models are trained on labeled data. However, obtaining large amounts of labeled data can be a costly and time-consuming process, which has led to increasing interest in learning from *uncurated data*.

Uncurated data refers to data that has not been carefully curated or labeled, and is often noisy, incomplete, and contains errors. Learning from such data presents a significant challenge to machine learning algorithms, as they must be able to effectively leverage the available information while being resilient to the inherent noise and errors. This thesis is concerned with the following fundamental goal:

Robust learning algorithms that can learn from uncurated data.

To achieve the goal, I analyze *contrastive learning*, a prominent technique for self-supervised representation learning through the comparison of semantically similar and dissimilar sample pairs [24, 81, 147]. Traditionally, supervised learning has been the cornerstone of progress in artificial intelligence (AI), relying on large quantities of labeled data to train models. However, the process of collecting and labeling such vast amounts of data can be costly and time-consuming. Moreover, in real-world applications, labeled data may often be scarce or even unavailable. To overcome these obstacles, researchers have turned to unsupervised and self-supervised learning techniques that leverage unlabeled data to train models. Yet, these techniques have

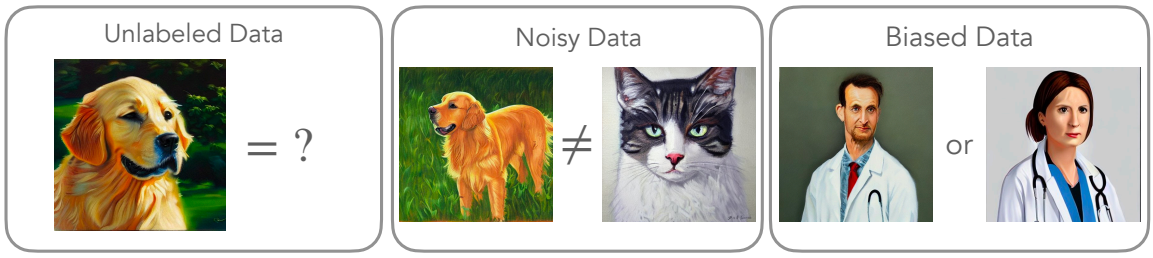


Figure 1-1: Problems of uncurated data. The uncurated data is often unlabeled, meaning it lacks explicit annotations for supervised learning. Moreover, the data can also be noisy, containing errors, outliers, or irrelevant information that can negatively impact performance. Lastly, the data can be biased, reflecting the inherent biases in the data sources, which, if not appropriately addressed, can lead to skewed model predictions and perpetuate existing inequalities or prejudices.

often lagged behind supervised methods in performance, primarily due to the difficulty in defining an objective that guides the model towards useful representations.

This is where contrastive learning enters the picture. By setting the objective to pull together semantically similar (positive) and dissimilar (negative) pairs of data points in a learned feature space, contrastive learning offers a way to construct useful and informative representations from unlabeled data. The power of contrastive learning has been demonstrated in several areas, including computer vision, natural language processing, graph representation learning, and reinforcement learning, achieving state-of-the-art performance in many benchmarks.

1.1 Challenges of Uncurated Data

While the framework of contrastive learning undeniably presents a promising avenue for harnessing the untapped potential of uncurated data, it is not without its set of unsolved complexities. The application of contrastive learning methodologies to raw, unfiltered data sources uncovers a myriad of challenges that remain unresolved and necessitate further research and innovation within this discipline. In this dissertation, my primary objective revolves around the development and implementation of strategies aimed at attenuating the errors associated with three specific types of data: *unlabeled data, noisy data, and biased data.*

Unlabeled Data Without access to the labels, the negative pairs are often chosen via a heuristic way. This introduces several key challenges in the context of contrastive learning. A significant disadvantage is the risk of encountering false negatives. A false negative occurs when a positive instance, i.e., a sample that should have been drawn closer to the target sample in the representation space, is mistakenly treated as a negative instance and pushed away. This can lead to the model developing a distorted representation of the data, which can affect downstream tasks that depend on these representations.

Similarly, the heuristic selection process can overlook hard negatives, which are negative instances that are hard to distinguish from positive ones. Hard negatives are crucial for improving the model's ability to discern subtle differences between different classes or types of data. When these are overlooked, the model may become overly generalized and less sensitive to critical dissimilarities, potentially decreasing its discriminative power.

Noisy Data Lastly, the issue of false positives poses another challenge. False positives emerge due to the systematic noise of the sampling process or the collection of pair data. These are instances that, despite being negative, are incorrectly treated as positive ones, leading to them being pulled closer to the target instance in the representation space. This erroneous pairing can blur the boundaries between different classes or clusters, making it more challenging for the model to accurately distinguish between them.

The challenge of mitigating the impact of noisy positive pairs in the framework of contrastive learning is significantly exacerbated within multimodal contexts, exemplified particularly in vision-language tasks. In the paradigm of large-scale vision-language training, for instance, the deployment of training models is often reliant on image-caption pairings, which are predominantly collated from the expansive Internet environment.

Given the vastness and inherent noise present in this environment, such pairings are inevitably plagued by noise, thus presenting a substantial complication to the

accuracy and efficiency of the learning process. Therefore, it becomes paramount to formulate robust methodologies and techniques to mitigate the effects of noisy data within this large-scale context. This issue of noise within the vast scale of the training setting highlights an urgent requirement for the development of more refined and noise-resistant models and learning strategies.

Biased Data Furthermore, the biases present in the collection of large-scale multimodal dataset processes pose another significant challenge. When positive pairs are collected from the internet, they often contain societal, cultural, or other types of inappropriate biases that reflect the inherent prejudices of the data sources. The danger here is that contrastive learning models can inadvertently learn and perpetuate these biases if they are present in the training data. Such biases can negatively affect the performance of the model and lead to unfair or unbalanced outcomes when the model is deployed in real-world applications.

As an example, in Chapter 7, we highlight several foundational models like Stable Diffusion [164] and CLIP [159] which, despite their sophisticated architecture and mechanisms, are not immune to detrimental predictive biases, encompassing gender and racial dimensions among others.

These instances underline the critical necessity for meticulous and ethically sound data curation practices. Alongside this, it also underscores the imperative for developing efficient de-biasing techniques within the pipeline of contrastive learning. The presence of such biases in prediction results not only impacts the model's performance but also raises significant ethical and societal considerations, which, if left unaddressed, can lead to potentially harmful real-world implications.

Therefore, our research places a significant emphasis on the exploration and advancement of data handling and model training practices that can effectively minimize the influence of such biases and ensure the deployment of more accurate, equitable, and socially responsible predictive models in the field of contrastive learning.

1.2 Outline

This thesis addresses the aforementioned bottlenecks with three parts: Part I: learning from unlabeled data, Part II: learning from noisy data, and Part III: learning from biased data.

In Part I, I address two key problems of the negative samples: false negative samples and hard negative samples. We start by giving background in Chapter 2 on problem formulation and contrastive learning. In Chapter 3, which is based on Chuang et al. [36], I address the problem of false negative samples with a new contrastive loss. Then in Chapter 4, based on Robinson et al. [163], we extend the proposed loss with hard negative sampling estimated via importance sampling.

In Part II, I will unveil the problem of false positive samples, especially in the multimodal setting and develop a robust loss function against it. In Chapter 5, based on Chuang et al. [39], I relate contrastive learning to binary classification and develop robust loss functions for contrastive loss. Chapter 6 provides a theoretical justification of the proposed loss with a variational lower bound on Wasserstein mutual information.

In Part III, I will discuss how to remove the biases in large-scale foundation models based on Chuang et al. [40]. Chapter 7 provides an overview of the rise of foundation models and proposes a debiasing algorithm to efficiently remove the biases of vision-language foundation models.

Chapter 8 establishes a nexus between representation learning and generalization theory using margin bounds, thereby offering a theoretical validation for contrastive learning. As an epilogue, Chapter 9 summarizes the thesis along with some discussions.

1.3 List of Publications

The contributions in this thesis mainly comprise work from the following publications [36, 39, 40, 163]:

- CY Chuang, J Robinson, L Yen-Chen, A Torralba, S Jegelka. **Debiased contrastive learning**. *Advances in Neural Information Processing Systems, 2020.*

- J Robinson, CY Chuang, S Sra, S Jegelka. **Contrastive learning with hard negative samples**. *In Proceedings of the 9th International Conference on Learning Representations, 2021.*[1]

- CY Chuang, Y Mroueh, K Greenewald, A Torralba, S Jegelka. **Measuring Generalization with Optimal Transport**. *Advances in Neural Information Processing Systems, 2021.*

- CY Chuang, RD Hjelm, X Wang, V Vineet, N Joshi, A Torralba, S Jegelka, Y Song. **Robust Contrastive Learning against Noisy Views**. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022.*

- CY Chuang, V Jampani, Y Li, A Torralba, S Jegelka. **Debiasing Vision-language Models via Biased Prompts**. *Preprint, 2023.*

The following publications [34, 35, 37, 38, 41] are a part of my PhD research. While these papers are strongly related to the thesis, they were left out to maintain a clearer story:

- CY Chuang, A Torralba, S Jegelka. **Estimating Generalization under Distribution Shifts via Domain-Invariant Representations**. *In Proceedings of the 37th International Conference on Machine Learning, 2020.*

- CY Chuang, Y Mroueh. **Fair Mixup: Fairness via Interpolation**. *In Proceedings of the 9th International Conference on Learning Representations, 2021.*

[1] I served as the second author in Robinson et al. [163], where I contributed to the discussions with co-authors and conducted experiments that included sentence embedding, CIFAR-10, and visualization.

- CY Chuang, S Jegelka. **Tree Mover's Distance: Bridging Graph Metrics and Stability of Graph Neural Networks**. *Advances in Neural Information Processing Systems, 2022.*

- CY Chuang, S Jegelka, D Alvarez-Melis. **InfoOT: Information Maximizing Optimal Transport**. *In Proceedings of the 40th International Conference on Machine Learning, 2023.*

Part I

Learning from Unlabeled Data

Chapter 2

Foundations of Contrastive Learning

Learning good representations without supervision has been a long-standing goal of machine learning. One such approach is *self-supervised learning*, where auxiliary learning objectives leverage labels that can be observed without a human labeler. The aim is to formulate learning tasks from the input data itself, making the process of model training less reliant on human supervision, thus reducing time and cost. This learning paradigm can discover intricate patterns and structures in the data, leading to a richer and more generalized feature representation. This high-quality representation is instrumental in enhancing the performance of downstream tasks such as classification, regression, or decision making, where traditionally labeled data would be required. For instance, in computer vision, representations can be learned from colorization [225], predicting transformations [47, 146], or generative modeling [25, 69, 101]. Remarkable success has also been achieved in the language domain [45, 104, 128].

Recently, self-supervised representation learning algorithms that use a contrastive loss have outperformed even supervised learning [24, 74, 80, 119]. The key idea of *contrastive loss* is to contrast semantically similar (positive) and dissimilar (negative) pairs of data points, encouraging the representations of similar pairs (x, x^+) to be close, and those of dissimilar pairs (x, x^-) to be more orthogonal. Figure 2-1 provides an illustration. In particular, contrastive loss is one of the earliest approaches used for deep metric learning [32], where various variants such as triplet loss [169] and multi-

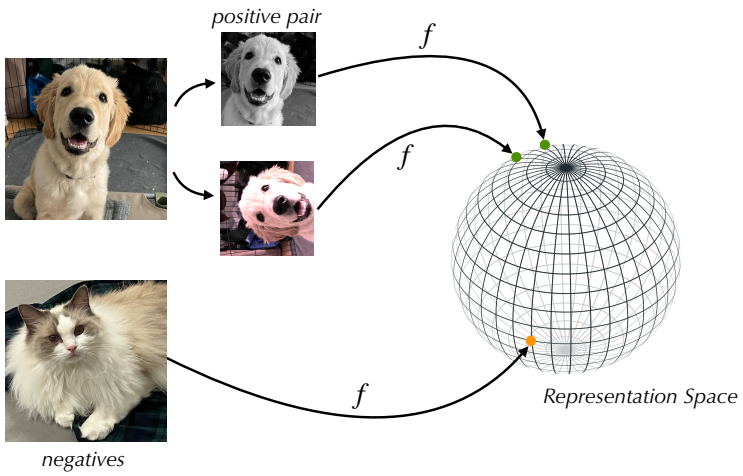


Figure 2-1: Illustration of contrastive learning. The key idea of contrastive learning is to contrast semantically similar (positive) and dissimilar (negative) pairs of data points, encouraging the representations of similar pairs to be close, and those of dissimilar pairs to be more orthogonal.

class N-pair loss are proposed [174]. In the current landscape of contrastive learning, a loss function that has gained widespread acceptance is the InfoNCE loss, as proposed by Oord et al. [147]. This approach takes its inspiration from the Noise Contrastive Estimation (NCE), a method first introduced by Gutmann and Hyvärinen [73].

2.1 InfoNCE Loss

The fundamental innovation behind InfoNCE is its use of categorical cross-entropy loss, a strategy that is effective in identifying the positive sample x^+ among a set of unrelated noise samples x_i^-, commonly referred to as negative samples:

$$\text{(InfoNCE Loss)} \quad \mathbb{E}_{x,x^+,\{x_i^-\}_{i=1}^N}\left[-\log \frac{e^{f(x)^T f(x^+)}}{e^{f(x)^T f(x^+)} + \sum_{i=1}^N e^{f(x)^T f(x_i^-)}}\right]. \tag{2.1}$$

In practice, it's common to draw negative samples, denoted as x_i^-, uniformly from the entirety of the training dataset. However, the strategy for deriving positive samples can significantly differ and is intrinsically tied to the unique characteristics of the domain in question. Whether it's computer vision, natural language processing,

or reinforcement learning, the method of identifying and generating positive samples is thoughtfully adjusted to reflect the nature of the data and the task at hand, demonstrating the adaptability of contrastive learning across diverse domains.

For instance, in computer vision, positive samples are often generated through data augmentation techniques. For a given image, different augmented versions - which could be transformations like rotation, scaling, color jittering, or cropping - of the same image serve as positive samples because they maintain the inherent visual information. Negative samples, on the other hand, typically come from other images in the batch that are unrelated to the target image.

For natural language processing, positive samples might be derived by using different sentences or phrases from the same document or contextually related documents, assuming that these share semantic meaning. Alternatively, one can use different embeddings of the same sentence (e.g., from different layers of a transformer model). Negative samples could be sentences or phrases from different contexts, different documents, or even randomly sampled from the dataset.

In reinforcement learning, the process is a bit more complex due to the temporal dimension of the data. Positive samples might come from the same or similar episodes or trajectories, as they share a similar series of actions and rewards. Negative samples, conversely, could be different episodes, different trajectories, or instances where the agent takes a different series of actions, leading to different outcomes.

Beyond single-modal contrastive learning, multimodal contrastive learning that involves learning representations from data across multiple modalities, such as text, images, audio, or video, has also gained significant attention. In this setting, positive pairs usually consist of different modalities of the same instance, like an image and its corresponding caption. Meanwhile, negative pairs could be different modalities from unrelated instances. The objective is still to bring the positive pairs closer and push negative pairs further apart in the learned representation space.

2.2 Related Works

Contrastive approaches have become prominent in unsupervised representation learning [32, 74, 83]. InfoNCE [147] and its variants [24, 70, 81, 220] achieve state-of-the-art across different modalities [12, 119, 120, 127, 132]. Modern approaches improve upon InfoNCE in different directions. A major line of work focuses on modifying training mechanisms, e.g., appending projection head (SimCLR) [24], momentum encoder with dynamic dictionary update (MoCo-v1/v2/v3) [27, 28, 81], siamese networks with stop gradient trick (e.g., BYOL) [26, 70], and online cluster assignment [23].

Contrastive learning can be adopted to various domains with different strategies of obtaining positive pairs. Examples in computer vision include random cropping and flipping [147], or different views of the same scene [186]. Chen et al. [24] extensively study verious data augmentation methods. For language, Logeswaran and Lee [119] treat the context sentences as positive samples to efficiently learn sentence representations. Srinivas et al. [178] improve the sample efficiency of reinforcement learning with representations learned via the contrastive loss. Computational efficiency has been improved by maintaining a dictionary of negative examples [27, 80].

2.3 Bottlenecks of Contrastive Learning

Contrastive learning, despite its potential for greatly improving machine learning tasks, faces significant drawbacks when dealing with negative samples. In particular, two problems arise when we uniformly draw the negative samples from the dataset: *false negatives* and *easy negatives*, as illustrated in Figure 2-2.

The problem of false negatives arises from the way negative samples are typically selected. By default, these samples are drawn uniformly from the dataset, which means that all examples have an equal chance of being chosen. However, the intricacies of real-world data sets mean that certain negative examples could be semantically similar to positive examples. False negatives can lead to a cascade of problems within the model's learning process. Mislabeling instances as dissimilar when they are, in

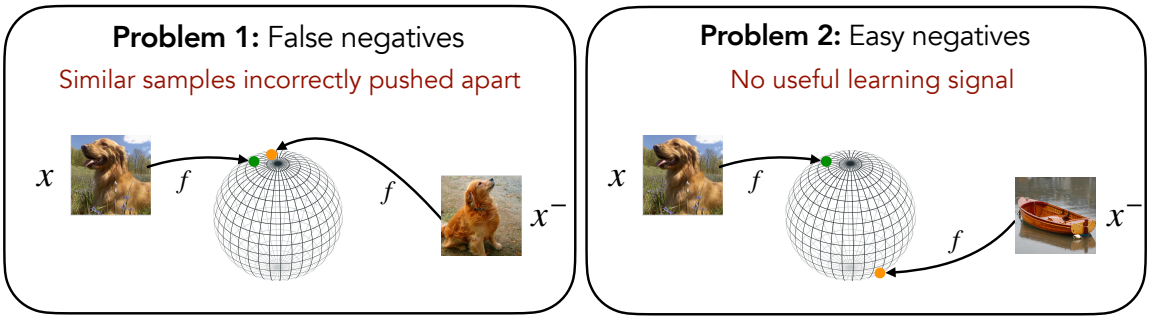


Figure 2-2: Problems of Uniform Negative Sampling. By default, the negative sample x^- is uniformly at random from the training set, which implies that x^- could be similar to x. Uniform sampling also treats hard negatives and easy negatives equally, reducing training efficiency.

fact, similar can result in the model developing a skewed understanding of the data. This can cause the model to make erroneous predictions when encountering similar examples in future tasks, resulting in reduced model performance and accuracy.

The second challenge relates to the model's tendency to overlook hard negative samples. Hard negatives are instances that, while dissimilar from the positive sample, are difficult to distinguish due to subtle differences. For example, images of two different dog breeds could act as hard negatives for each other due to their close visual resemblance despite belonging to different classes.

The issue with the uniform selection process is that it doesn't account for the complexity of these hard negatives. As a result, these samples can end up being underrepresented in the learning process. This is a significant limitation, as encountering and learning from these challenging examples is crucial for improving the model's robustness and ability to generalize to new, unseen data.

In the next two chapters, we will introduce new contrastive losses that elegantly address the aforementioned problems without significantly modifying the training pipeline.

Chapter 3

False Negative Samples

Much research effort has addressed sampling strategies for positive pairs, and has been a key driver of recent progress in multi-view and contrastive learning [11, 19, 24, 188, 210]. Surprisingly, the choice of negative pairs has drawn much less attention in contrastive learning. Often, given an "anchor" point x, a "negative" x^- is simply sampled uniformly from the training data, independent of how informative it may be for the learned representation.

But, this means it is possible that x^- is actually similar to x, as illustrated in Figure 3-1. This phenomenon, which we refer to as *sampling bias*, can empirically lead to significant performance drop. Figure 3-2 compares the accuracy for learning with this bias, and for drawing x_i^- from data with truly different labels than x; we refer to this method as *unbiased* (further details in Section 3.4).

However, the ideal unbiased objective is unachievable in practice since it requires knowing the labels, i.e., *supervised* learning. This dilemma poses the question whether it is possible to reduce the gap between the ideal objective and standard contrastive learning, without supervision. In this chapter, we demonstrate that this is indeed possible, while still assuming only access to unlabeled training data and positive examples. In particular, we develop a correction for the sampling bias that yields a new, modified loss which we call *debiased contrastive loss*.

The key idea underlying our approach is to indirectly approximate the distribution of negative examples. The new objective is easily compatible with any algorithm

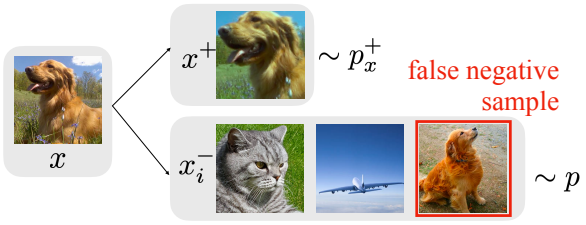


Figure 3-1: "Sampling bias": The common practice of drawing negative examples x_i^- from the data distribution $p(x)$ may result in x_i^- that are actually similar to x.

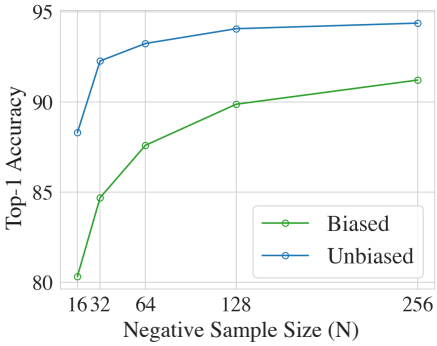


Figure 3-2: Sampling bias leads to performance drop: Results on CIFAR-10 for drawing x_i^- from $p(x)$ (biased) and from data with different labels, i.e., truly semantically different data (unbiased).

that optimizes the standard contrastive loss. Empirically, our approach improves over the state of the art in vision, language and reinforcement learning benchmarks. Our theoretical analysis relates the debiased contrastive loss to supervised learning: optimizing the debiased contrastive loss corresponds to minimizing an upper bound on a supervised loss. This leads to a generalization bound for the supervised task, when training with the debiased contrastive loss.

3.1 Setup of Contrastive Learning

Contrastive learning assumes access to semantically similar pairs of data points (x, x^+), where x is drawn from a data distribution $p(x)$ over $\mathcal{X}$. The goal is to learn an embedding $f : \mathcal{X} \to \mathbb{R}^d$ that maps an observation x to a point on a hypersphere with radius $1/t$, where t is the temperature scaling hyperparameter. Without loss of generality, we set $t = 1$ for all theoretical results.

Similar to [8], we assume an underlying set of discrete latent classes $\mathcal{C}$ that represent semantic content, i.e., similar pairs (x, x^+) have the same latent class. Denoting the distribution over classes by $\rho(c)$, we obtain the joint distribution $p_{x,c}(x, c) = p(x|c)\rho(c)$. Let $h : \mathcal{X} \to \mathcal{C}$ be the function assigning the latent class labels. Then $p_x^+(x') = p(x'|h(x') = h(x))$ is the probability of observing x' as a positive example for x and $p_x^-(x') = p(x'|h(x') \neq h(x))$ the probability of a negative example. We

assume that the class probabilities $\rho(c) = \tau^+$ are uniform, and let $\tau^- = 1 - \tau^+$ be the probability of observing any different class.

Note that to remain unsupervised in practice, our method and other contrastive losses only sample from the data distribution and a "surrogate" positive distribution, mimicked by data augmentations or context sentences [24, 119]. We will provide a more comprehensive analysis of positives in Chapter 5.

3.2 Problem of False Negative Samples

Intuitively the contrastive loss will provide most informative representations for downstream classification tasks if the positive and negative pairs correspond to the desired latent classes. Hence, the ideal loss to optimize would be

$$L^N_{\text{Unbiased}}(f) = \mathbb{E}_{\substack{x\sim p, x^+\sim p^+_x \\ x^-_i \sim p^-_x}} \left[-\log \frac{e^{f(x)^T f(x^+)}}{e^{f(x)^T f(x^+)} + \frac{Q}{N}\sum_{i=1}^N e^{f(x)^T f(x^-_i)}} \right], \tag{3.1}$$

which we will refer to as the *unbiased loss*. Here, we introduce a weighting parameter Q for the analysis. When the number N of negative examples is finite, we set $Q = N$, in agreement with the standard contrastive loss. However, $p^-_x(x^-_i) = p(x^-_i | h(x^-_i) \neq h(x))$ is not accessible in practice. The standard approach is thus to sample negative examples x^-_i from the (unlabeled) $p(x)$ instead. We refer to the resulting loss as the *biased* loss L^N_{Biased}. When drawn from $p(x)$ the sample x^-_i will come from the same class as x with probability τ^+.

Lemma 3.1 shows that in the limit, the standard loss L^N_{Biased} upper bounds the ideal, unbiased loss.

Lemma 3.1. *For any embedding f and finite N, we have*

$$L^N_{\text{Biased}}(f) \geq L^N_{\text{Unbiased}}(f) + \mathbb{E}_{x\sim p}\left[0 \wedge \log \frac{\mathbb{E}_{x^+\sim p^+_x} \exp f(x)^\top f(x^+)}{\mathbb{E}_{x^-\sim p^-_x} \exp f(x)^\top f(x^-)} \right] - e^{3/2}\sqrt{\frac{\pi}{2N}}. \tag{3.2}$$

where $a \wedge b$ denotes the minimum of two real numbers a and b.

Proof. I use the notation $h(x, \bar{x}) = \exp f(x)^\top f(\bar{x})$ for the critic. We will borrow Theorem 3.3 to prove this lemma. Setting $\tau^+ = 0$, Theorem 3.3 states that

$$\mathbb{E}_{\substack{x\sim p \\ x^+\sim p_x^+}}\left[-\log \frac{h(x,x^+)}{h(x,x^+)+N\mathbb{E}_{x^-\sim p}h(x,x_i^-)}\right] - \mathbb{E}_{\substack{x\sim p \\ x^+\sim p_x^+ \\ \{x_i^-\}_{i=1}^N\sim p^N}}\left[-\log \frac{h(x,x^+)}{h(x,x^+)+\sum_{i=1}^N h(x,x_i^-)}\right] \leq e^{3/2}\sqrt{\frac{\pi}{2N}}.$$

Equip with this inequality, the biased objective can be decomposed into the sum of the debiased objective and a second term as follows,

$$\begin{aligned}
&L^N_{\text{Biased}}(f) \\
&= \mathbb{E}_{\substack{x\sim p \\ x^+\sim p_x^+ \\ \{x_i^-\}_{i=1}^N\sim p^N}}\left[-\log \frac{h(x,x^+)}{h(x,x^+)+\sum_{i=1}^N h(x,x_i^-)}\right] \\
&\geq \mathbb{E}_{\substack{x\sim p \\ x^+\sim p_x^+}}\left[-\log \frac{h(x,x^+)}{h(x,x^+)+N\mathbb{E}_{x^-\sim p_x}h(x,x^-)}\right] - e^{3/2}\sqrt{\frac{\pi}{2N}} \\
&= \mathbb{E}_{\substack{x\sim p \\ x^+\sim p_x^+}}\left[-\log \frac{h(x,x^+)}{h(x,x^+)+N\mathbb{E}_{x^-\sim p_x^-}h(x,x^-)}\right] \\
&\qquad + \mathbb{E}_{\substack{x\sim p \\ x^+\sim p_x^+}}\left[\log \frac{h(x,x^+)+N\mathbb{E}_{x^-\sim p_x}h(x,x^-)}{h(x,x^+)+N\mathbb{E}_{x^-\sim p_x^-}h(x,x^-)}\right] - e^{3/2}\sqrt{\frac{\pi}{2N}} \\
&= L^N_{\text{Debiased}}(f) + \mathbb{E}_{\substack{x\sim p \\ x^+\sim p_x^+}}\left[\log \frac{h(x,x^+)+N\mathbb{E}_{x^-\sim p_x}h(x,x^-)}{h(x,x^+)+N\mathbb{E}_{x^-\sim p_x^-}h(x,x^-)}\right] - e^{3/2}\sqrt{\frac{\pi}{2N}} \\
&= L^N_{\text{Debiased}}(f) + \mathbb{E}_{\substack{x\sim p \\ x^+\sim p_x^+}}\left[\log \frac{h(x,x^+)+\tau^-N\mathbb{E}_{x^-\sim p_x^-}h(x,x^-)+\tau^+N\mathbb{E}_{x^-\sim p_x^+}h(x,x^-)}{h(x,x^+)+\tau^-N\mathbb{E}_{x^-\sim p_x^-}h(x,x^-)+\tau^+N\mathbb{E}_{x^-\sim p_x^-}h(x,x^-)}\right] \\
&\qquad - e^{3/2}\sqrt{\frac{\pi}{2N}}.
\end{aligned}$$

If $\mathbb{E}_{x^-\sim p_x^+}h(x,x^-) \geq \mathbb{E}_{x^-\sim p_x^-}h(x,x^-)$ then this expression can be lower bounded by $L^N_{\text{Debiased}}(f) + \log 1 = L^N_{\text{Debiased}}(f)$. Else, if $\mathbb{E}_{x^-\sim p_x^+}h(x,x^-) \leq \mathbb{E}_{x^-\sim p_x^-}h(x,x^-)$ we can use the elementary fact that $\frac{a+c}{b+c} \geq \frac{a}{b}$ for $a \leq b$ and $a, b, c \geq 0$. Bringing these

two possibilities together we conclude that

$$L^N_{\text{Biased}}(f) \geq L^N_{\text{Unbiased}}(f) + \mathbb{E}_{x\sim p}\left[0 \wedge \log \frac{\mathbb{E}_{x^+\sim p^+_x} \exp f(x)^\top f(x^+)}{\mathbb{E}_{x^-\sim p^-_x} \exp f(x)^\top f(x^-)}\right] - e^{3/2}\sqrt{\frac{\pi}{2N}}.$$

where we replaced the dumour variable x^- in the numerator by x^+. □

Recent works often use large N, e.g., $N = 65536$ in [80], making the last term negligible. While, in general, minimizing an upper bound on a target objective is a reasonable idea, two issues arise here: (1) the smaller the unbiased loss, the larger is the second term, widening the gap; and (2) the empirical results in Figure 3-2 and Section 3.4 show that minimizing the upper bound L^N_{Biased} and minimizing the ideal loss L^N_{Unbiased} can result in very different learned representations.

3.3 Debiased Contrastive Loss (DCL)

Next, we derive a loss that is closer to the ideal L^N_{Unbiased}, while only having access to positive samples and samples from p. Figure 3-2 shows that the resulting embeddings are closer to those learned with L^N_{Unbiased}. We begin by decomposing the data distribution as

$$p(x') = \tau^+ p^+_x(x') + \tau^- p^-_x(x').$$

This decomposition is inspired by *Positive-Unlabeled* (PU) learning, i.e., learning from only positive (P) and unlabeled (U) data. Common applications of PU learning are retrieval or outlier detection [49, 52, 105].

An immediate approach would be to replace p^-_x in L^N_{Unbiased} with $p^-_x(x') = (p(x') - \tau^+ p^+_x(x'))/\tau^-$ and then use the empirical counterparts for p and p^+_x. The resulting objective can be estimated with samples from only p and p^+_x, but is computationally

expensive for large N:

$$\frac{1}{(\tau^-)^N}\sum_{k=0}^{N}\binom{N}{k}(-\tau^+)^k\mathbb{E}_{\substack{x\sim p,x^+\sim p_x^+\\ \{x_i^-\}_{i=1}^k\sim p_x^+\\ \{x_i^-\}_{i=k+1}^N\sim p}}\left[-\log\frac{e^{f(x)^Tf(x^+)}}{e^{f(x)^Tf(x^+)}+\sum_{i=1}^N e^{f(x)^Tf(x_i^-)}}\right], \tag{3.3}$$

where $\{x_i^-\}_{i=k}^j=\emptyset$ if $k>j$. It also demands at least N positive samples. To give the derivation of this claim, let

$$\ell(x,x^+,\{x_i^-\}_{i=1}^N,f)=-\log\frac{e^{f(x)^Tf(x^+)}}{e^{f(x)^Tf(x^+)}+\sum_{i=1}^N e^{f(x)^Tf(x_i^-)}}.$$

We plug in the decomposition as follows:

$$\begin{aligned}
&\mathbb{E}_{\substack{x\sim p,x^+\sim p_x^+\\ \{x_i^-\}_{i=1}^N\sim p_x^-}}\left[\ell(x,x^+,\{x_i^-\}_{i=1}^N,f)\right]\\
&=\int p(x)p_x^+(x^+)\prod_{i=1}^N p_x^-(x_i^-)\ell(x,x^+,\{x_i^-\}_{i=1}^N,f)\mathrm{d}x\mathrm{d}x^+\prod_{i=1}^N\mathrm{d}x_i^-\\
&=\int p(x)p_x^+(x^+)\prod_{i=1}^N\frac{p(x_i^-)-\tau^+p_x^+(x_i^-)}{\tau^-}\ell(x,x^+,\{x_i^-\}_{i=1}^N,f)\mathrm{d}x\mathrm{d}x^+\prod_{i=1}^N\mathrm{d}x_i^-\\
&=\frac{1}{(\tau^-)^N}\int p(x)p_x^+(x^+)\prod_{i=1}^N\left(p(x_i^-)-\tau^+p_x^+(x_i^-)\right)\ell(x,x^+,\{x_i^-\}_{i=1}^N,f)\mathrm{d}x\mathrm{d}x^+\prod_{i=1}^N\mathrm{d}x_i^-
\end{aligned}$$

By the Binomial Theorem the product can be separated into $N+1$ groups corresponding to how many x_i^- are sampled from p.

$$\begin{aligned}
&(1)\qquad \prod_{i=1}^N p(x_i^-)\\
&(2)\qquad \binom{N}{1}(-\tau^+)p_x^+(x_1^-)\prod_{i=2}^N p(x_i^-)\\
&(3)\qquad \binom{N}{2}\prod_{j=1}^2(-\tau^+)p_x^+(x_j^-)\prod_{i=3}^N p(x_i^-)\\
&\dots
\end{aligned}$$

$$
\begin{aligned}
&(k+1) \qquad \binom{N}{k} \prod_{j=1}^{k} (-\tau^+) p_x^+(x_j^-) \prod_{i=k+1}^{N} p(x_i^-) \\
&\cdots \\
&(N+1) \qquad \prod_{i=1}^{N} (-\tau^+) p_x^+(x_i^-)
\end{aligned}
$$

In particular, the objective becomes

$$
\frac{1}{(\tau^-)^N} \sum_{k=0}^{N} \binom{N}{k} (-\tau^+)^k \mathbb{E}_{\substack{x\sim p, x^+\sim p_x^+ \\ \{x_i^-\}_{i=1}^k \sim p_x^+ \\ \{x_i^-\}_{i=k+1}^N \sim p}} \left[-\log \frac{e^{f(x)^T f(x^+)}}{e^{f(x)^T f(x^+)} + \sum_{i=1}^N e^{f(x)^T f(x_i^-)}} \right].
$$

Note that this is exactly *Inclusion–exclusion principle*.

3.3.1 Asymptotic Approximation

To obtain a more practical form, we consider the asymptotic form as the number N of negative examples goes to infinity.

Lemma 3.2. *For fixed Q and $N \to \infty$, it holds that*

$$
\mathbb{E}_{\substack{x\sim p, x^+\sim p_x^+ \\ \{x_i^-\}_{i=1}^N \sim {p_x^-}^N}} \left[-\log \frac{e^{f(x)^T f(x^+)}}{e^{f(x)^T f(x^+)} + \frac{Q}{N} \sum_{i=1}^N e^{f(x)^T f(x_i^-)}} \right] \tag{3.4}
$$

$$
\longrightarrow \mathbb{E}_{\substack{x\sim p \\ x^+\sim p_x^+}} \left[-\log \frac{e^{f(x)^T f(x^+)}}{e^{f(x)^T f(x^+)} + \frac{Q}{\tau^-} (\mathbb{E}_{x^-\sim p}[e^{f(x)^T f(x^-)}] - \tau^+ \mathbb{E}_{v\sim p_x^+}[e^{f(x)^T f(v)}])} \right]. \tag{3.5}
$$

Proof. Since the contrastive loss is bounded, applying Dominated Convergence Theorem completes the proof:

$$
\begin{aligned}
&\lim_{N\to\infty} \mathbb{E} \left[-\log \frac{e^{f(x)^T f(x^+)}}{e^{f(x)^T f(x^+)} + \frac{Q}{N} \sum_{i=1}^N e^{f(x)^T f(x_i^-)}} \right] \\
=&\mathbb{E} \left[\lim_{N\to\infty} -\log \frac{e^{f(x)^T f(x^+)}}{e^{f(x)^T f(x^+)} + \frac{Q}{N} \sum_{i=1}^N e^{f(x)^T f(x_i^-)}} \right] \quad \text{(Dominated Convergence Theorem)} \\
=&\mathbb{E} \left[-\log \frac{e^{f(x)^T f(x^+)}}{e^{f(x)^T f(x^+)} + Q \mathbb{E}_{x^-\sim p_x^-} e^{f(x)^T f(x^-)}} \right].
\end{aligned}
$$

Since $p_x^-(x') = (p(x') - \tau^+ p_x^+(x'))/\tau^-$ and by the linearity of the expectation, we have

$$\mathbb{E}_{x^- \sim p_x^-} e^{f(x)^T f(x^-)} = \tau^- (\mathbb{E}_{x^- \sim p}[e^{f(x)^T f(x^-)}] - \tau^+ \mathbb{E}_{x^- \sim p_x^+}[e^{f(x)^T f(x^-)}]),$$

which completes the proof. □

The limiting objective equation 3.5, which we denote by $\widetilde{L}_{\text{Debiased}}$, still samples examples x^- from p, but corrects for that with additional positive samples v. This essentially reweights positive and negative terms in the denominator.

The empirical estimate of $\widetilde{L}_{\text{Debiased}}$ is much easier to compute than the straightforward objective equation 3.4. With N samples $\{u_i\}_{i=1}^N$ from p and M samples $\{v_i\}_{i=1}^M$ from p_x^+, we estimate the expectation of the secong term in the denominator as

$$g(x, \{u_i\}_{i=1}^N, \{v_i\}_{i=1}^M) = \max\Big\{\frac{1}{\tau^-}\Big(\frac{1}{N}\sum_{i=1}^N e^{f(x)^T f(u_i)} - \tau^+ \frac{1}{M}\sum_{i=1}^M e^{f(x)^T f(v_i)}\Big), e^{-1/t}\Big\}. \tag{3.6}$$

We constrain the estimator g to be greater than its theoretical minimum $e^{-1/t} \leq \mathbb{E}_{x^- \sim p_x^-} e^{f(x)^T f(x_i^-)}$ to prevent calculating the logarithm of a negative number. The resulting population loss with fixed N and M per data point is

$$L_{\text{Debiased}}^{N,M}(f) = \mathbb{E}_{\substack{x \sim p; x^+ \sim p_x^+ \\ \{u_i\}_{i=1}^N \sim p^N \\ \{v_i\}_{i=1}^N \sim {p_x^+}^M}} \left[-\log \frac{e^{f(x)^T f(x^+)}}{e^{f(x)^T f(x^+)} + N g(x, \{u_i\}_{i=1}^N, \{v_i\}_{i=1}^M)}\right] \tag{3.7}$$

where, for simplicity, we set Q to the finite N. The class prior τ^+ can be estimated from data [33, 88] or treated as a hyperparameter. Theorem 3.3 bounds the error due to finite N and M as decreasing with rate $\mathcal{O}(N^{-1/2} + M^{-1/2})$.

Theorem 3.3. *For any embedding f and finite N and M, we have*

$$\left|\widetilde{L}_{\text{Debiased}}^N(f) - L_{\text{Debiased}}^{N,M}(f)\right| \leq \frac{e^{3/2}}{\tau^-}\sqrt{\frac{\pi}{2N}} + \frac{e^{3/2}\tau^+}{\tau^-}\sqrt{\frac{\pi}{2M}}. \tag{3.8}$$

Empirically, the experiments in Section 3.4 also show that larger N and M con-

sistently lead to better performance. In the implementations, we use a full empirical estimate for $L^{N,M}_{\text{Debiased}}$ that averages the loss over T points x, for finite N, M.

Proof. In order to prove Theorem 3.3, which shows that the empirical estimate of the asymptotic debiased objective is a good estimate, we first seek a bound on the tail probability that the difference of the integrands of the asymptotic and non-asymptotic objective functions. That is we wish to bound the probability that the following quantity is greater than ε,

$$\Delta = \left| -\log \frac{h(x, x^+)}{h(x, x^+) + Qg(x, \{u_i\}_{i=1}^N, \{v_i\}_{i=1}^M)} + \log \frac{h(x, x^+)}{h(x, x^+) + Q\mathbb{E}_{x^- \sim p_x^-} h(x, x^-)} \right|.$$

where we again write $h(x, \bar{x}) = \exp f(x)^\top f(\bar{x})$ for the critic. Note that implicitly Δ depends on x, x^+ and the collections $\{u_i\}_{i=1}^N$ and $\{v_i\}_{i=1}^M$. We achieve control over the tail using through the following theorem.

Theorem 3.4. *Let x and x^+ in $\mathcal{X}$ be fixed. Further, let $\{u_i\}_{i=1}^N$ and $\{v_i\}_{i=1}^M$ be collections of i.i.d. random variables sampled from p and p_x^+ respectively. Then for all $\varepsilon > 0$,*

$$\mathbb{P}(\Delta \geq \varepsilon) \leq 2\exp\left(-\frac{N\varepsilon^2(\tau^-)^2}{2e^3}\right) + 2\exp\left(-\frac{M\varepsilon^2(\tau^-/\tau^+)^2}{2e^3}\right).$$

We delay the proof of Theorem 3.4 to Appendix B.1. By Jensen's inequality we may push the absolute value inside the expectation to see that $|\widetilde{L}^N_{\text{Unbiased}}(f) - L^{N,M}_{\text{Debiased}}(f)| \leq \mathbb{E}\Delta$. All that remains is to exploit the exponential tail bound of Theorem 3.4. To do this we write the expectation of Δ for fixed x, x^+ as the integral of its tail probability,

$$\mathbb{E}\ \Delta = \mathbb{E}_{x,x^+}\left[\mathbb{E}[\Delta|x,x^+]\right] = \mathbb{E}_{x,x^+}\left[\int_0^\infty \mathbb{P}(\Delta \geq \varepsilon|x,x^+)\mathrm{d}\varepsilon\right]$$
$$\leq \int_0^\infty 2\exp\left(-\frac{N\varepsilon^2(\tau^-)^2}{2e^3}\right)\mathrm{d}\varepsilon + \int_0^\infty 2\exp\left(-\frac{M\varepsilon^2(\tau^-/\tau^+)^2}{2e^3}\right)\mathrm{d}\varepsilon.$$

The outer expectation disappears since the tail probably bound of Theorem 3.4 holds uniformly for all fixed x, x^+. Both integrals can be computed analytically using the classical identity

$$\int_0^\infty e^{-cz^2}\mathrm{d}z = \frac{1}{2}\sqrt{\frac{\pi}{c}}.$$

Applying the identity to each integral we finally obtain the claimed bound,

$$\sqrt{\frac{2e^3\pi}{(\tau^-)^2 N}} + \sqrt{\frac{2e^3\pi}{(\tau^-/\tau^+)^2 M}} = \frac{e^{3/2}}{\tau^-}\sqrt{\frac{2\pi}{N}} + \frac{e^{3/2}\tau^+}{\tau^-}\sqrt{\frac{2\pi}{M}}.$$

□

3.4 Experiments

In this section we evaluate our new objective L^N_{Debiased} empirically, and compare it to the standard loss L^N_{Biased} and the ideal loss L^N_{Unbiased}. In summary, we observe the following: (1) the new loss outperforms state of the art contrastive learning on vision, language and reinforcement learning benchmarks; (2) the learned embeddings are closer to those of the ideal, unbiased objective; (3) both larger N and large M improve the performance; even one more positive example than the standard $M = 1$ can help noticeably.

3.4.1 CIFAR10 and STL10

First, for CIFAR10 [108] and STL10 [42], we implement SimCLR [24] with ResNet-50 [79] as the encoder architecture and use the Adam optimizer [100] with learning rate

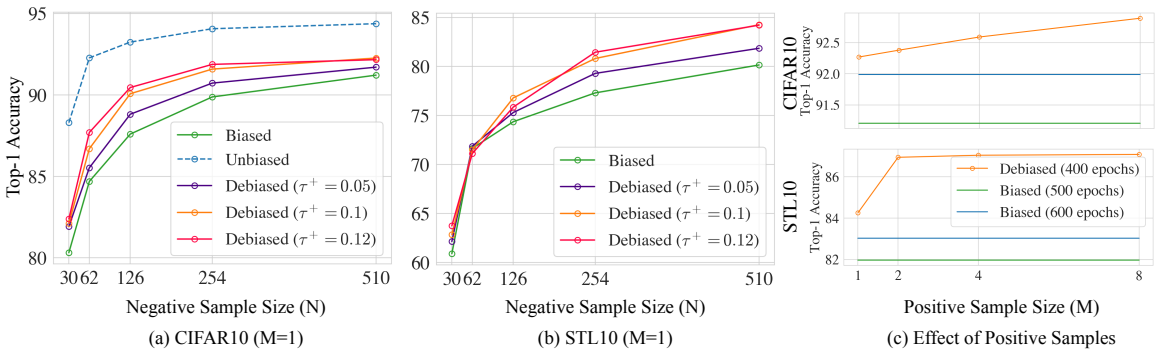


Figure 3-3: Classification accuracy on CIFAR10 and STL10. (a,b) Biased and Debiased ($M = 1$) SimCLR with different negative sample size N. (c) Comparison with biased SimCLR with 50% more training epochs (600 epochs) while fixing the training epoch for Debiased ($M \geq 1$) SimCLR to 400 epochs.

0.001. Following [24], we set the temperature $t = 0.5$ and the dimension of the latent vector to 128. All the models are trained for 400 epochs and evaluated by training a linear classifier after fixing the learned embedding.

To understand the effect of the sampling bias, we additionally consider an estimate of the ideal L^N_{Unbiased}, which is a *supervised* version of the standard loss, where negative examples x_i^- are drawn from the true p_x^-, i.e., using known classes. Since STL10 is not fully labeled, we only use the unbiased objective on CIFAR10.

Debiased Objective with $M = 1$. For a fair comparison, i.e., no possible advantage from additional samples, we first examine our debiased objective with positive sample size $M = 1$ by setting $v_1 = x^+$. Then, our approach uses exactly the same data batch as the biased baseline. The debiased objective can be implemented by a slight modification of code. The results with different τ^+ are shown in Figure 3-3(a,b). Increasing τ^+ in Objective equation 3.6 leads to more correction, and gradually improves the performance in both benchmarks for different N. Remarkably, with only a slight modification to the loss, we improve the accuracy of SimCLR on STL10 by 4.26%. The performance of the debiased objective also improves by increasing the negative sample size N.

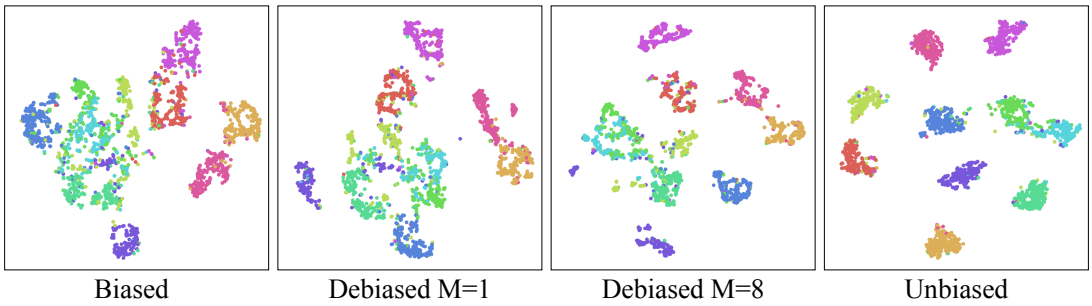


Figure 3-4: t-SNE visualization of learned representations on CIFAR10. Classes are indicated by colors. The debiased objective ($\tau^+ = 0.1$) leads to better data clustering than the (standard) biased loss; its effect is closer to the supervised unbiased objective.

Debiased Objective with $M \geq 1$. By Theorem 3.3, a larger M leads to a better estimate of the loss. To probe its effect, we sample M positive samples for each x (e.g., M times data augmentation) while fixing $N = 256$ and $\tau^+ = 0.1$. A larger M would require additional computation, therefore, we compare our debiased objective with biased SimCLR trained for 50% more epochs (600 epochs). The results for $M = 1, 2, 4, 8$ are shown in Figure 3-3(c), and indicate that the performance of the debiased objective can indeed be further improved by increasing the number of positive samples. Surprisingly, with only one additional positive sample, the top-1 accuracy on STL10 can be significantly improved. We can also see that the debiased objective ($M > 1$) still outperforms the biased baseline even it is trained with 50% more epochs.

Figure 3-4 shows t-SNE visualizations of the representations learned by the biased and debiased objectives ($N = 256$) on CIFAR10. The debiased contrastive loss leads to better class separation than the contrastive loss, and the result is closer to that of the ideal, unbiased loss.

Objective	Top-1	Top-5
Biased (CMC)	73.58	92.06
Debiased ($\tau^+ = 0.005$)	73.86	91.86
Debiased ($\tau^+ = 0.01$)	**74.6**	**92.08**

Table 3.1: ImageNet-100 Top-1 and Top-5 classification results.

3.4.2 ImageNet-100

Following [186], we test our approach on ImageNet-100, a randomly chosen subset of 100 classes of Imagenet. Compared to CIFAR10, ImageNet-100 has more classes and

hence smaller class probabilities τ^+. We use contrastive multiview coding (CMC) [186] as our contrastive learning baseline, and $M = 1$ for a fair comparison. The results in Table 3.1 show that, although τ^+ is small, our debiased objective still improves over the biased baseline.

3.4.3 Sentence Embeddings

Next, we test the debiased objective for learning sentence embeddings. We use the BookCorpus dataset [104] and examine six classification tasks: movie review sentiment (MR) [151], product reviews (CR) [85], subjectivity classification (SUBJ) [150], opinion polarity (MPQA) [206], question type classification (TREC) [196], and paraphrase identification (MSRP) [46]. Our experimental settings follow those for quick-thought (QT) vectors in [119]. In contrast to vision tasks, positive pairs here are chosen as neighboring sentences, which can form a different positive distribution than data augmentation. The minibatch of QT is constructed with a contiguous set of sentences, hence we can use the preceding and succeeding sentences as positive samples ($M = 2$). We retrain each model 3 times and show the mean in Table 3.2. The debiased objective improves over the baseline in 4 out of 6 downstream tasks, verifying that our objective also works for a different modality.

Objective	**MR**	**CR**	**SUBJ**	**MPQA**	**TREC**	**MSRP**	
						(Acc)	**(F1)**
Biased (QT)	**76.8**	81.3	86.6	93.4	**89.8**	73.6	81.8
Debiased ($\tau^+ = 0.005$)	76.5	81.5	86.6	93.6	89.1	74.2	82.3
Debiased ($\tau^+ = 0.01$)	76.2	**82.9**	**86.9**	**93.7**	89.1	**74.7**	**82.7**

Table 3.2: Classification accuracy on downstream tasks. we compare sentence representations on six classification tasks. 10-fold cross validation is used in testing the performance for binary classification tasks (MR, CR, SUBJ, MPQA)

3.4.4 Reinforcement Learning

Lastly, we consider reinforcement learning. We follow the experimental settings of Contrastive unsupervised representations for reinforcement learning (CURL) [178] to

perform image-based policy control on top of the learned contrastive representations. Similar to vision tasks, the positive pairs are two different augmentations of the same image. We again set $M = 1$ for a fair comparison. Methods are tested at 100k environment steps on the DeepMind control suite [184], which consists of several continuous control tasks. We retrain each model 3 times and show the mean and standard deviation in Table 3.3. Our method consistently outperforms the state-of-the-art baseline (CURL) in different control tasks, indicating that correcting the sampling bias also improves the performance and data efficiency of reinforcement learning. In several tasks, the debiased approach also has smaller variance. With more positive examples ($M = 2$), we obtain further improvements.

Objective	Finger Spin	Cartpole Swingup	Reacher Easy	Cheetah Run	Walker Walk	Ball in Cup Catch
Biased (CURL)	310±33	850±20	918±96	266±41	623±120	928±47
Debiased Objective with $M = 1$						
Debiased ($\tau^+ = 0.01$)	324±34	843±30	**927±99**	**310±12**	**626±82**	937±9
Debiased ($\tau^+ = 0.05$)	308±57	**866±7**	916±114	284±20	613±22	945±13
Debiased ($\tau^+ = 0.1$)	**364±36**	860±4	868±177	302±29	594±33	**951±11**
Debiased Objective with $M = 2$						
Debiased ($\tau^+ = 0.01$)	330±10	858±10	754±179	286±20	**746±93**	949±5
Debiased ($\tau^+ = 0.1$)	**381±24**	864±6	904±117	303±5	671±75	**957±5**

Table 3.3: Scores achieved by biased and debiased objectives. Our debiased objective outperforms the biased baseline (CURL) in all the environments, and often has smaller variance.

3.4.5 Discussion

Class Distribution: Our theoretical results assume that the class distribution ρ is close to uniform. In reality, this is often not the case, e.g., in our experiments, CIFAR10 and Imagenet-100 are the only two datasets with perfectly balanced class distributions. Nevertheless, our debiased objective still improves over the baselines even when the classes are not well balanced, indicating that the objective is robust to violations of the class balance assumption.

Positive Distribution: Even if we approximate the true positive distribution with a surrogate positive distribution, our debiased objective still consistently improves over the baselines. It is an interesting avenue of future work to adopt our debiased objective to a semi-supervised learning setting [221] where true positive samples are accessible.

3.5 Theoretical Analysis of Debiasing Loss

3.5.1 Generalization Implications for Classification Tasks

Next, we relate the debiased contrastive objective to a supervised loss, and show how our contrastive learning approach leads to a generalization bound for a downstream supervised learning task.

Our supervised task is a classification task $\mathcal{T}$ with K classes $\{c_1, \ldots, c_K\} \subseteq \mathcal{C}$. After contrastive representation learning, we fix the representations $f(x)$ and then train a linear classifier $q(x) = Wf(x)$ on task $\mathcal{T}$ with the standard multiclass softmax cross entropy loss $L_{\text{Softmax}}(\mathcal{T}, q)$. Hence, we define the supervised loss for the representation f as

$$L_{\text{Sup}}(\mathcal{T}, f) = \inf_{W \in \mathbb{R}^{K \times d}} L_{\text{Softmax}}(\mathcal{T}, Wf). \tag{3.9}$$

In line with the approach of [8] we analyze the supervised loss of a mean classifier [173], where for each class c, the rows of W are set to the mean of representations $\mu_c = \mathbb{E}_{x \sim p(\cdot|c)}[f(x)]$. we will use $L^{\mu}_{\text{Sup}}(\mathcal{T}, f)$ as shorthand for its loss. Note that $L^{\mu}_{\text{Sup}}(\mathcal{T}, f)$ is always an upper bound on $L_{\text{Sup}}(\mathcal{T}, f)$. To allow for uncertainty about the task $\mathcal{T}$, we will bound the average supervised loss for a uniform distribution $\mathcal{D}$ over K-way classification tasks with classes in $\mathcal{C}$.

$$L_{\text{Sup}}(f) = \mathbb{E}_{\mathcal{T} \sim \mathcal{D}} L_{\text{Sup}}(\mathcal{T}, f). \tag{3.10}$$

I begin by showing that the asymptotic unbiased contrastive loss is an upper bound on the supervised loss of the mean classifier.

Lemma 3.5. *For any embedding f, whenever $N > K$ we have*

$$L_{\text{Sup}}(f) \leq L^{\mu}_{\text{Sup}}(f) \leq \widetilde{L}_{\text{Debiased}}(f).$$

Proof. I first show that $N = K - 1$ is the smallest loss:

$$\begin{aligned}
&\widetilde{L}^{N}_{\text{Unbiased}}(f) \\
&= \mathbb{E}_{\substack{x\sim p \\ x^+\sim p_x^+}} \left[-\log \frac{e^{f(x)^T f(x^+)}}{e^{f(x)^T f(x^+)} + N\mathbb{E}_{x^-\sim p_x^-} e^{f(x)^T f(x^-)}} \right] \\
&\geq \mathbb{E}_{\substack{x\sim p \\ x^+\sim p_x^+}} \left[-\log \frac{e^{f(x)^T f(x^+)}}{e^{f(x)^T f(x^+)} + (K-1)\mathbb{E}_{x^-\sim p_x^-} e^{f(x)^T f(x^-)}} \right] \\
&= L^{K-1}_{\text{Unbiased}}(f)
\end{aligned}$$

To show that $L^{K-1}_{\text{Unbiased}}(f)$ is an upper bound on the supervised loss $L_{\text{sup}}(f)$, we additionally introduce a task specific class distribution $\rho_{\mathcal{T}}$ which is a uniform distribution over the classes in task $\mathcal{T}$.

$$\begin{aligned}
&L^{K-1}_{\text{Unbiased}}(f) \\
&= \mathbb{E}_{\substack{x\sim p \\ x^+\sim p_x^+}} \left[-\log \frac{e^{f(x)^T f(x^+)}}{e^{f(x)^T f(x^+)} + (K-1)\mathbb{E}_{x^-\sim p_x^-} e^{f(x)^T f(x^-)}} \right] \\
&= \mathbb{E}_{\mathcal{T}\sim\mathcal{D}} \mathbb{E}_{\substack{c\sim\rho_{\mathcal{T}}; x\sim p(\cdot|c) \\ x^+\sim p(\cdot|c)}} \left[-\log \frac{e^{f(x)^T f(x^+)}}{e^{f(x)^T f(x^+)} + (K-1)\mathbb{E}_{\mathcal{T}\sim\mathcal{D}}\mathbb{E}_{\rho_{\mathcal{T}}(c^-\sim|c^-\neq h(x))}\mathbb{E}_{x^-\sim p(\cdot|c^-)} e^{f(x)^T f(x^-)}} \right] \\
&\geq \mathbb{E}_{\mathcal{T}\sim\mathcal{D}} \mathbb{E}_{c\sim\rho_{\mathcal{T}}; x\sim p(\cdot|c)} \left[-\log \frac{e^{f(x)^T \mathbb{E}_{x^+\sim p(\cdot|c)} f(x^+)}}{e^{f(x)^T \mathbb{E}_{x^+\sim p^+_{x,\mathcal{T}}} f(x^+)} + (K-1)\mathbb{E}_{\mathcal{T}\sim\mathcal{D}}\mathbb{E}_{\rho_{\mathcal{T}}(c^-|c^-\neq h(x))}\mathbb{E}_{x^-\sim p(\cdot|c^-)} e^{f(x)^T f(x^-)}} \right] \\
&\geq \mathbb{E}_{\mathcal{T}\sim\mathcal{D}} \mathbb{E}_{c\sim\rho_{\mathcal{T}}; x\sim p(\cdot|c)} \left[-\log \frac{e^{f(x)^T \mathbb{E}_{x^+\sim p(\cdot|c)} f(x^+)}}{e^{f(x)^T \mathbb{E}_{x^+\sim p(\cdot|c)} f(x^+)} + (K-1)\mathbb{E}_{\rho_{\mathcal{T}}(c^-|c^-\neq h(x))}\mathbb{E}_{x^-\sim p(\cdot|c^-)} e^{f(x)^T f(x^-)}} \right] \\
&= \mathbb{E}_{\mathcal{T}\sim\mathcal{D}} \mathbb{E}_{c\sim\rho_{\mathcal{T}}; x\sim p(\cdot|c)} \left[-\log \frac{e^{f(x)^T \mathbb{E}_{x^+\sim p(\cdot|c)} f(x^+)}}{e^{f(x)^T \mathbb{E}_{x^+\sim p(\cdot|c)} f(x^+)} + (K-1)\mathbb{E}_{\rho_{\mathcal{T}}(c^-|c^-\neq h(x))}\mathbb{E}_{x^-\sim p(\cdot|c^-)} e^{f(x)^T f(x^-)}} \right] \\
&\geq \mathbb{E}_{\mathcal{T}\sim\mathcal{D}} \mathbb{E}_{c\sim\rho_{\mathcal{T}}; x\sim p(\cdot|c)} \left[-\log \frac{e^{f(x)^T \mathbb{E}_{x^+\sim p(\cdot|c)} f(x^+)}}{e^{f(x)^T \mathbb{E}_{x^+\sim p(\cdot|c)} f(x^+)} + (K-1)\mathbb{E}_{\rho_{\mathcal{T}}(c^-|c^-\neq h(x))} e^{f(x)^T \mathbb{E}_{x^-\sim p(\cdot|c^-)} f(x^-)}} \right]
\end{aligned}$$

$$
\begin{aligned}
&= \mathbb{E}_{\mathcal{T}\sim\mathcal{D}}\mathbb{E}_{c\sim\rho_{\mathcal{T}};x\sim p(\cdot|c)}\left[-\log\frac{\exp(f(x)^T\mu_c)}{\exp(f(x)^T\mu_c)+\sum_{c^-\in\mathcal{T},c^-\neq c}\exp(f(x)^T\mu_{c^-})}\right] \\
&= \mathbb{E}_{\mathcal{T}\sim\mathcal{D}}L^{\mu}_{\mathrm{Sup}}(\mathcal{T},f) \\
&= \bar{L}^{\mu}_{\mathrm{Sup}}(f)
\end{aligned}
$$

where three inequalities follows from Jensen's inequality. The first and third inequality shift the expectations $\mathbb{E}_{x^+\sim p^+_{x,\mathcal{T}}}$ and $\mathbb{E}_{x^-\sim p(\cdot|c^-)}$, respectively, via the convexity of the functions and the second move the expectation $\mathbb{E}_{\mathcal{T}\sim\mathcal{D}}$ out via the concavity. Note that $\bar{L}_{\mathrm{Sup}}(f) \leq \bar{L}^{\mu}_{\mathrm{Sup}}(f)$ holds trivially.

□

Lemma 3.5 uses the asymptotic version of the debiased loss. Together with Theorem 3.3 and a concentration of measure result, it leads to a generalization bound for debiased contrastive learning, as we show next.

3.5.2 Generalization Bound for Contrastive Learning

In practice, we use an empirical estimate $\widehat{L}^{N,M}_{\mathrm{Debiased}}$, i.e., an average over T data points x, with M positive and N negative samples for each x. Our algorithm learns an empirical risk minimizer $\hat{f} \in \arg\min_{f\in\mathcal{F}}\widehat{L}^{N,M}_{\mathrm{Debiased}}(f)$ from a function class $\mathcal{F}$. The generalization depends on the *empirical Rademacher complexity* $\mathcal{R}_{\mathcal{S}}(\mathcal{F})$ of $\mathcal{F}$ with respect to our data sample $\mathcal{S} = \{x_j, x_j^+, \{u_{i,j}\}_{i=1}^N, \{v_{i,j}\}_{i=1}^M\}_{j=1}^T$. Let $f_{|\mathcal{S}} = (f_k(x_j), f_k(x_j^+), \{f_k(u_{i,j})\}_{i=1}^N, \{f_k(v_{i,j})\}_{i=1}^M)_{j\in[T],k\in[d]} \in \mathbb{R}^{(N+M+2)dT}$ be the restriction of f onto $\mathcal{S}$, using $[T] = \{1,\ldots,T\}$. Then $\mathcal{R}_{\mathcal{S}}(\mathcal{F})$ is defined as

$$
\mathcal{R}_{\mathcal{S}}(\mathcal{F}) := \mathbb{E}_{\sigma}\sup_{f\in\mathcal{F}}\langle\sigma, f_{|\mathcal{S}}\rangle \tag{3.11}
$$

where $\sigma \sim \{\pm 1\}^{(N+M+1)dT}$ are Rademacher random variables. Combining Theorem 3.3 and Lemma 3.5 with a concentration of measure argument yields the final generalization bound for debiased contrastive learning.

Theorem 3.6. *With probability at least* $1-\delta$*, for all* $f \in \mathcal{F}$*,*

$$L_{\text{Sup}}(\hat{f}) \leq L^{N,M}_{\text{Debiased}}(f) + \mathcal{O}\left(\frac{1}{\tau^-}\sqrt{\frac{1}{N}} + \frac{\tau^+}{\tau^-}\sqrt{\frac{1}{M}} + \frac{\lambda \mathcal{R}_{\mathcal{S}}(\mathcal{F})}{T} + B\sqrt{\frac{\log\frac{1}{\delta}}{T}}\right) \quad (3.12)$$

where $\lambda = \sqrt{\frac{1}{(\tau^-)^2}(\frac{M}{N}+1) + (\tau^+)^2(\frac{N}{M}+1)}$ *and* $B = \log N\left(\frac{1}{\tau^-} + \tau^+\right)$.

Proof. We wish to derive a data dependent bound on the downstream supervised generalization error of the debiased contrastive objective. Recall that a sample $(x, x^+, \{u_i\}_{i=1}^N, \{v_i\}_{i=1}^M)$ yields loss

$$-\log\left\{\frac{e^{f(x)^\top f(x^+)}}{e^{f(x)^\top f(x^+)} + N g(x, \{u_i\}_{i=1}^N, \{v_i\}_{i=1}^M)}\right\} = \log\left\{1 + N\frac{g(x, \{u_i\}_{i=1}^N, \{v_i\}_{i=1}^M)}{e^{f(x)^\top f(x^+)}}\right\}$$

which is equal to $\ell\left(\left\{f(x)^\top\big(f(u_i) - f(x^+)\big)\right\}_{i=1}^N, \left\{f(x)^\top\big(f(v_i) - f(x^+)\big)\right\}_{i=1}^M\right)$ where we define,

$$\ell(\{a_i\}_{i=1}^N, \{b_i\}_{i=1}^M) = \log\left\{1 + N\max\left(\frac{1}{\tau^-}\frac{1}{N}\sum_{i=1}^N a_i - \tau^+\frac{1}{M}\sum_{i=1}^M b_i, e^{-1}\right)\right\}.$$

In order to derive our bound we will exploit a concentration of measure result due to [8]. They consider an objective of the form

$$L_{un}(f) = \mathbb{E}\left[\ell(\{f(x)^\top\big(f(x_i) - f(x^+)\big)\}_{i=1}^k)\right]$$

where $(x, x^+, x_1^-, \ldots, x_k^-)$ are sampled from any fixed distribution on $\mathcal{X}^{k+2}$ (they were particularly focused on the case where $x_i^- \sim p$, but the proof holds for arbitrary distributions). Let $\mathcal{F}$ be a class of representation functions $\mathcal{X} \to \mathbb{R}^d$ such that $\|f(\cdot)\| \leq R$ for $R > 0$. The corresponding empirical risk minimizer is,

$$\hat{f} \in \arg\min_{f\in\mathcal{F}} \frac{1}{T}\sum_{j=1}^T \ell\left(\{f(x_j)^\top\big(f(x_{ji}) - f(x^+)\big)\}_{i=1}^k\right)$$

over a training set $\mathcal{S} = \{(x_j, x_j^+, x_{j1}^-, \ldots, x_{jk}^-)\}_{j=1}^T$ of i.i.d. samples. Their result bounds the loss of the empirical risk minimizer as follows.

Lemma 3.7. (Arora et al. [8]) *Let $\ell : \mathbb{R}^k \to \mathbb{R}$ be η-Lipschitz and bounded by B. Then with probability at least $1-\delta$ over the training set $\mathcal{S} = \{(x_j, x_j^+, x_{j1}^-, \ldots, x_{jk}^-)\}_{j=1}^T$, for all $f \in \mathcal{F}$*

$$L_{un}(\hat{f}) \leq L_{un}(f) + \mathcal{O}\left(\frac{\eta R\sqrt{k}\mathcal{R}_{\mathcal{S}}(\mathcal{F})}{T} + B\sqrt{\frac{\log\frac{1}{\delta}}{T}}\right)$$

where

$$\mathcal{R}_{\mathcal{S}}(\mathcal{F}) = \mathbb{E}_{\sigma\sim\{\pm 1\}^{(k+2)dT}}\left[\sup_{f\in\mathcal{F}}\langle\sigma, f_{|\mathcal{S}}\rangle\right],$$

and $f_{|\mathcal{S}} = \Big(f_t(x_j), f_t(x_j^+), f_t(x_{j1}^-), \ldots, , f_t(x_{jk}^-)\Big)_{\substack{j\in[T]\\ t\in[d]}}$.

In our context we have $k = N + M$ and $R = e$. So, it remains to obtain constants η and B such that $\ell(\{a_i\}_{i=1}^N, \{b_i\}_{i=1}^M)$ is η-Lipschitz, and bounded by B. Note that since we consider normalized embeddings f, we have $\|f(\cdot)\| \leq 1$ and therefore only need to consider the domain where $e^{-1} \leq a_i, b_i \leq e$.

Lemma 3.8. *Suppose that $e^{-1} \leq a_i, b_i \leq e$. The function $\ell(\{a_i\}_{i=1}^N, \{b_i\}_{i=1}^M)$ is η-Lipschitz, and bounded by B for*

$$\eta = e \cdot \sqrt{\frac{1}{(\tau^-)^2 N} + \frac{(\tau^+)^2}{M}}, \qquad B = \mathcal{O}\left(\log N\left(\frac{1}{\tau^-} + \tau^+\right)\right).$$

Proof. First it is easily observed that ℓ is upper bounded by plugging in $a_i = e$ and $b_i = e^{-1}$, yielding a bound of,

$$\log\left\{1 + N\max\left(\frac{1}{\tau^-}e - \tau^+e^{-1}, e^{-1}\right)\right\} = \mathcal{O}\left(\log N\left(\frac{1}{\tau^-} + \tau^+\right)\right).$$

To bound the Lipschitz constant we view ℓ as a composition $\ell(\{a_i\}_{i=1}^N, \{b_i\}_{i=1}^M) = \phi\left(g\left(\ell(\{a_i\}_{i=1}^N, \{b_i\}_{i=1}^M)\right)\right.$ where[1],

[1]Note the definition of g is slightly modified in this context.

$$\phi(z) = \log\left(1 + N\max(z, e^{-1}\right)$$
$$g(\{a_i\}_{i=1}^N, \{b_i\}_{i=1}^M) = \frac{1}{\tau^-}\frac{1}{N}\sum_{i=1}^N a_i - \tau^+\frac{1}{M}\sum_{i=1}^M b_i.$$

If $z < e^{-1}$ then $\partial_z\phi(z) = 0$, while if $z \geq e^{-1}$ then $\partial_z\phi(z) = \frac{N}{1+Nz} \leq \frac{N}{1+Ne^{-1}} \leq e$. We therefore conclude that ϕ is e-Lipschitz. Meanwhile, $\partial_{a_i} g = \frac{1}{\tau^- N}$ and $\partial_{b_i} g = \frac{\tau^+}{M}$. The Lipschitz constant of g is bounded by the Forbenius norm of the Jacobian of g, which equals,

$$\sqrt{\sum_{i=1}^N \frac{1}{(\tau^- N)^2} + \sum_{j=1}^M \frac{(\tau^+)^2}{M^2}} = \sqrt{\frac{1}{(\tau^-)^2 N} + \frac{(\tau^+)^2}{M}}.$$

□

Now we have control on the bound on ℓ and its Lipschitz constant, we are ready to prove theorem 5 by combining several of our previous results with Lemma 3.7. In particular, by Lemma 3.8 and Theorem 3 we have

$$L_{\text{sup}}(\hat{f}) \leq \widetilde{L}^N_{\text{Unbiased}}(\hat{f}) \leq L^{N,M}_{\text{Debiased}}(\hat{f}) + \frac{e^{3/2}}{\tau^-}\sqrt{\frac{\pi}{2N}} + \frac{e^{3/2}\tau^+}{\tau^-}\sqrt{\frac{\pi}{2M}}$$

Combining Lemma 3.7 and Lemma 3.8, with probability at least $1 - \delta$, for all $f \in \mathcal{F}$ we have

$$L^{N,M}_{\text{Debiased}}(\hat{f}) \leq L^{N,M}_{\text{Debiased}}(f) + \mathcal{O}\left(\frac{\lambda\mathcal{R}_{\mathcal{S}}(\mathcal{F})}{T} + B\sqrt{\frac{\log\frac{1}{\delta}}{T}}\right)$$

where $\lambda = \eta\sqrt{k} = \sqrt{\frac{1}{\tau^{-2}}(\frac{M}{N}+1) + \tau^{+2}(\frac{N}{M}+1)}$ and $B = 2 + \log(N+1)$ □

The bound states that if the function class $\mathcal{F}$ is sufficiently rich to contain some embedding for which $L^{N,M}_{\text{Debiased}}$ is small, then the representation encoder $\hat{f}$, learned from a large enough dataset, will perform well on the downstream classification task.

The bound also highlights the role of the positive and unlabeled sample sizes M and N in the objective function, in line with the observation that a larger number of negative/positive examples in the objective leads to better results [24, 80]. The last two terms in the bound grow slowly with N, but the effect of this on the generalization error is small if the dataset size T is much larger than N and M, as is commonly the case. The dependence on on N and T in Theorem 3.6 is roughly equivalent to the result in [8], but the two bounds are not directly comparable since the proof strategies differ.

3.6 Conclusion

In this chapter, we propose *debiased contrastive learning*, a new unsupervised contrastive representation learning framework that corrects for the bias introduced by the common practice of sampling negative (dissimilar) examples for a point from the overall data distribution. Our debiased objective consistently improves the state-of-the-art baselines in various benchmarks in vision, language and reinforcement learning. The proposed framework is accompanied by generalization guarantees for the downstream classification task. Interesting directions of future work include (1) trying the debiased objective in semi-supervised learning or few shot learning, and (2) studying the effect of how positive (similar) examples are drawn, e.g., analyzing different data augmentation techniques.

Chapter 4

Hard Negative Samples

In the preceding chapter, we effectively established that rectifying the instances of false negative samples significantly enhances the performance across a wide array of downstream tasks. In this present chapter, our focus shifts towards examining and elucidating the impacts and potential optimization strategies related to what are categorized as *hard negative samples*.

In supervised and metric learning settings, "hard" (true negative) examples can help guide a learning method to correct its mistakes more quickly [169, 176]. For representation learning, informative negative examples are intuitively those pairs that are mapped nearby but should be far apart. This idea is successfully applied in metric learning, where true pairs of dissimilar points are available, as opposed to unsupervised contrastive learning.

With this motivation, this chapter will address the challenge of selecting informative negatives for contrastive representation learning. In response, we propose a solution that builds a tunable sampling distribution that prefers negative pairs whose representations are currently very similar. This solution faces two challenges: (1) we do not have access to any true similarity or dissimilarity information; (2) we need an efficient sampling strategy for this tunable distribution. We overcome (1) by building on ideas from positive-unlabeled learning [49, 52], and (2) by designing an efficient, easy to implement importance sampling technique that incurs no computational overhead.

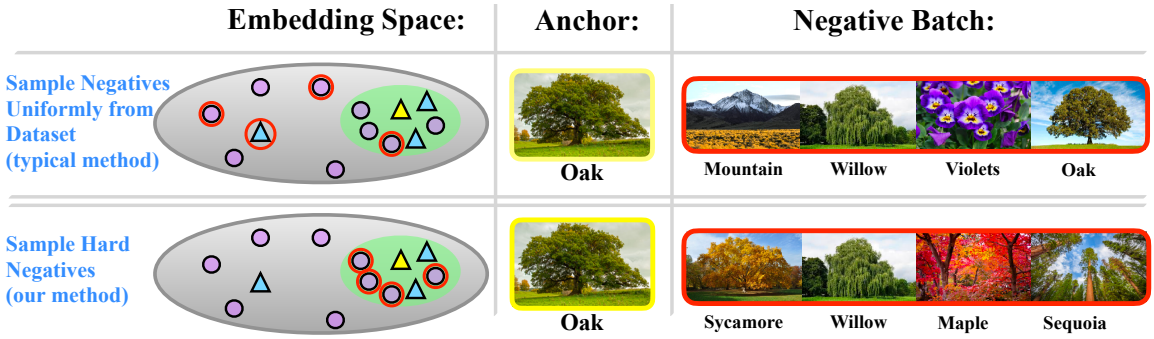


Figure 4-1: Schematic illustration of negative sampling methods for the example of classifying species of tree. Top row: uniformly samples negative examples (red rings); mostly focuses on very different data points from the anchor (yellow triangle), and may even sample examples from the same class (triangles, vs. circles). Bottom row: Hard negative sampling prefers examples that are (incorrectly) close to the anchor.

4.1 Hard Negative Contrastive Loss (HCL)

We will primarily follow the setup in Chapter 3, where we wish to learn an embedding $f : \mathcal{X} \to \mathbb{S}^{d-1}/t$ that maps an observation x to a point on a hypersphere $\mathbb{S}^{d-1}/t$ in $\mathbb{R}^d$ of radius $1/t$, where t is the temperature scaling hyperparameter.

As a quick recap, the InfoNCE objective for learning the representation f uses a *positive* example x^+ with the same label as x, and *negative* examples $\{x_i^-\}_{i=1}^N$ with (supposedly) different labels, $h(x_i^-) \neq h(x)$, sampled from q:

$$\mathbb{E}_{\substack{x\sim p,\ \ x^+\sim p_x^+ \\ \{x_i^-\}_{i=1}^N \sim q}} \left[-\log \frac{e^{f(x)^T f(x^+)}}{e^{f(x)^T f(x^+)} + \frac{Q}{N}\sum_{i=1}^N e^{f(x)^T f(x_i^-)}} \right]. \tag{4.1}$$

The negative sample distribution q is frequently chosen to be the marginal distribution p, or, in practice, an empirical approximation of it [24, 27, 27, 81, 82, 147, 186]. In this chapter we ask: is there a better way to choose the negative distribution q?

4.1.1 Principle of Negative Samples

We begin by asking *what makes a good negative sample?* To answer this question we adopt the following two guiding principles:

Principle 4.1. *q should only sample "true negatives" x_i^- whose labels differ from that of the anchor x.*

Principle 4.2. *The most useful negative samples are ones that the embedding currently believes to be similar to the anchor.*

In short, negative samples that have different label from the anchor, but that are embedded nearby are likely to be most useful and provide significant gradient information during training. In metric learning there is access to true negative pairs, automatically fulfilling the first principle.

In unsupervised contrastive learning there is no supervision, so upholding Principle 1 is impossible to do exactly. Chapter 3 introduces a method that upholds Principle 1 approximately, where in this section, we will simultaneously combine this idea with the key additional conceptual ingredient of "hardness" (encapsulated in Principle 2). The level of "hardness" in our method can be smoothly adjusted, allowing the user to select the hardness that best trades-off between an improved learning signal from hard negatives, and the harm due to the correction of false negatives being only approximate. This important since the hardest points are those closest to the anchor, and are expected to have a high propensity to have the same label. Therefore the damage from the approximation not removing all false negatives becomes larger for harder samples, creating the trade-off. As a special case our method, when the hardness level is tuned fully down, we obtain the debiased contrastive loss proposed in Chapter 3 that only upholds Principle 1 (approximately) but not Principle 2.

4.1.2 Hard Negatives via Importance Sampling

Our first goal is to design a distribution q on $\mathcal{X}$ that is allowed to depend on the embedding f and the anchor x. From q we sample a batch of negatives $\{x_i^-\}_{i=1}^N$ according to the principles noted above. We propose sampling negatives from the distribution q_β^- defined as

$$q_\beta^-(x^-) := q_\beta(x^-|h(x) \neq h(x^-)), \quad \text{where} \quad q_\beta(x^-) \propto e^{\beta f(x)^\top f(x^-)} \cdot p(x^-),$$

for $\beta \geq 0$. Note that q_β^- and q_β both depend on x, but we suppress the dependance from the notation. The exponential term in q_β is an unnormalized von Mises–Fisher

distribution with mean direction $f(x)$ and "concentration parameter" β [123]. There are two key components to q_β^-, corresponding to each principle: 1) conditioning on the event $\{h(x) \neq h(x^-)\}$ which guarantees that (x, x^-) correspond to different latent classes (Principle 1); 2) the concentration parameter β term controls the degree by which q_β up-weights points x^- that have large inner product (similarity) to the anchor x (Principle 2). Since f lies on the surface of a hypersphere of radius $1/t$, we have $\|f(x) - f(x')\|^2 = 2/t^2 - 2f(x)^\top f(x')$ so preferring points with large inner product is equivalent to preferring points with small squared Euclidean distance.

Although we have designed q_β^- to have all of the desired components, it is not clear how to sample efficiently from it. To work towards a practical method, note that we can rewrite this distribution by adopting a positive-unlabeled decomposition introduced in the previous chapter. That is, by conditioning on the event $\{h(x) = h(x^-)\}$ we can split $q_\beta(x^-)$ as

$$q_\beta(x^-) = \tau^- q_\beta^-(x^-) + \tau^+ q_\beta^+(x^-), \tag{4.2}$$

where $q_\beta^+(x^-) = q_\beta(x^-|h(x) = h(x^-)) \propto e^{\beta f(x)^\top f(x^-)} \cdot p^+(x^-)$. Rearranging equation 4.2 yields a formula $q_\beta^-(x^-) = \big(q_\beta(x^-) - \tau^+ q_\beta^+(x^-)\big)/\tau^-$ for the negative sampling distribution q_β^- in terms of two distributions that are tractable since we have samples from p and can approximate samples from p^+ using a set of semantics-preserving transformations, as is typical in contrastive learning methods.

It is possible to generate samples from q_β and (approximately from) q_β^+ using rejection sampling. However, rejection sampling involves an algorithmic complication since the procedure for sampling batches must be modified. To avoid this, we instead take an importance sampling approach. To obtain this, analogous to the previous chapter, first note that fixing the number Q and taking the limit $N \to \infty$ in the objective (4.1) yields,

$$\mathcal{L}(f, q) = \mathbb{E}_{\substack{x \sim p \\ x^+ \sim p_x^+}} \left[-\log \frac{e^{f(x)^T f(x^+)}}{e^{f(x)^T f(x^+)} + Q \mathbb{E}_{x^- \sim q}[e^{f(x)^T f(x^-)}]} \right]. \tag{4.3}$$

The original objective (4.1) can be viewed as a finite negative sample approximation to $\mathcal{L}(f,q)$ (note implicitly $\mathcal{L}(f,q)$ depends on Q) . Inserting $q = q_\beta^-$ and using the rearrangement of equation (4.2) we obtain the following hardness-biased objective:

$$\mathbb{E}_{\substack{x\sim p \\ x^+\sim p_x^+}}\left[-\log \frac{e^{f(x)^T f(x^+)}}{e^{f(x)^T f(x^+)} + \frac{Q}{\tau^-}(\mathbb{E}_{x^-\sim q_\beta}[e^{f(x)^T f(x^-)}] - \tau^+ \mathbb{E}_{v\sim q_\beta^+}[e^{f(x)^T f(v)}])}\right]. \quad (4.4)$$

This objective suggests that we need only to approximate *expectations* $\mathbb{E}_{x^-\sim q_\beta}[e^{f(x)^T f(x^-)}]$ and $\mathbb{E}_{v\sim q_\beta^+}[e^{f(x)^T f(v)}]$ over q_β and q_β^+ (rather than explicily sampling). This can be achieved using classical Monte-Carlo importance sampling techniques using samples from p and p^+ as follows:

$$\mathbb{E}_{x^-\sim q_\beta}[e^{f(x)^T f(x^-)}] = \mathbb{E}_{x^-\sim p}[e^{f(x)^T f(x^-)} q_\beta/p] = \mathbb{E}_{x^-\sim p}[e^{(\beta+1)f(x)^T f(x^-)}/Z_\beta],$$
$$\mathbb{E}_{v\sim q_\beta^+}[e^{f(x)^T f(v)}] = \mathbb{E}_{v\sim p^+}[e^{f(x)^T f(v)} q_\beta^+/p^+] = \mathbb{E}_{v\sim p^+}[e^{(\beta+1)f(x)^T f(v)}/Z_\beta^+],$$

where Z_β, Z_β^+ are the partition functions of q_β and q_β^+ respectively. The right hand terms readily admit empirical approximations by replacing p and p^+ with $\hat{p}(x) = \frac{1}{N}\sum_{i=1}^N \delta_{x_i^-}(x)$ and $\hat{p}^+(x) = \frac{1}{M}\sum_{i=1}^M \delta_{x_i^+}(x)$ respectively (δ_w denotes the Dirac delta function centered at w). The only unknowns left are the partition functions, $Z_\beta = \mathbb{E}_{x^-\sim p}[e^{\beta f(x)^T f(x^-)}]$ and $Z_\beta^+ = \mathbb{E}_{x^+\sim p^+}[e^{\beta f(x)^T f(x^+)}]$ which themselves are expectations over p and p^+ and therefore admit empirical estimates,

$$\widehat{Z}_\beta = \frac{1}{N}\sum_{i=1}^N e^{\beta f(x)^\top f(x_i^-)}, \qquad \widehat{Z}_\beta^+ = \frac{1}{M}\sum_{i=1}^M e^{\beta f(x)^\top f(x_i^+)}.$$

It is important to emphasize the simplicity of the implementation of our proposed approach. Since we propose to reweight the objective instead of modifying the sampling procedure, only two extra lines of code are needed to implement our approach, with no additional computational overhead.

4.2 Experiments

Next, we evaluate our hard negative sampling method empirically, and apply it as a modification to state-of-the-art contrastive methods on image, graph, and text data. For all experiments β is treated as a hyper-parameter (see ablations in Fig. 4-2 for more understanding of how to pick β). Values for M and τ^+ must also be determined. We fix $M = 1$ for all experiments, since taking $M > 1$ would increase the number of inputs for the forward-backward pass. Lemma 4.6 in the end of previous section gives a theoretical justification for the choice of $M = 1$. Choosing the class-prior τ^+ can be done in two ways: estimating it from data [33, 88], or treating it as a hyper-parameter. The first option requires the possession of labeled data *before* contrastive training.

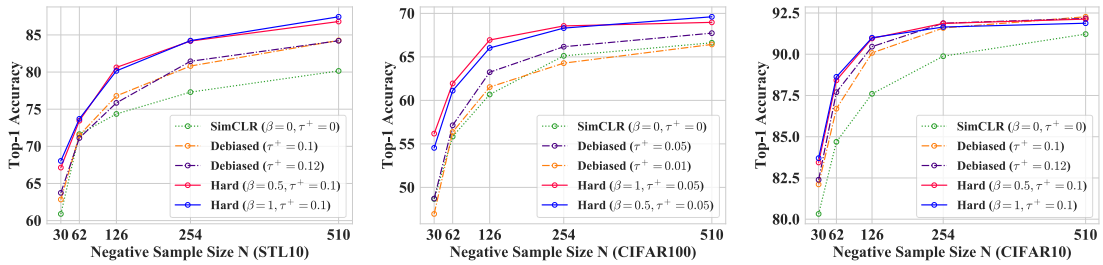


Figure 4-2: Classification accuracy on downstream tasks. Embeddings trained using hard, debiased, and standard ($\beta = 0, \tau^+ = 0$) versions of SimCLR, and evaluated using linear readout accuracy.

4.2.1 Image Representations

SimCLR	Debiased	Hard ($\beta = 1$)
53.4%	53.7%	57.0%

Table 4.1: Top-1 linear readout on tinyImageNet. Class prior is set to $\tau^+ = 0.01$.

Similar to Chapter 3, we begin by testing the hard sampling method on vision tasks using the STL10, CIFAR100 and CIFAR10 data. We use SimCLR [24] as the baseline method, and all models are trained for 400 epochs. The results in Fig. 4-2 show consistent improvement over SimCLR ($q = p$) and the particular case of our method with $\beta = 0$ proposed in [36] (called debiasing) on STL10 and CIFAR100. For $N = 510$ negative examples per data point we observe absolute

improvements of 3% and 7.3% over SimCLR on CIFAR100 and STL10 respectively, and absolute improvements over the best debiased baseline of 1.9% and 3.2%. On tinyImageNet (Tab. 4.1) we observe an absolute improvement of 3.6% over SimCLR, while on CIFAR10 there is a slight improvement for smaller N, which disappears at larger N. Figure 4-3 further demonstrates that hard negative sampling significantly improves the training efficiency, where it takes much fewer epochs to reach the same accuracy as SimCLR does in 400 epochs.

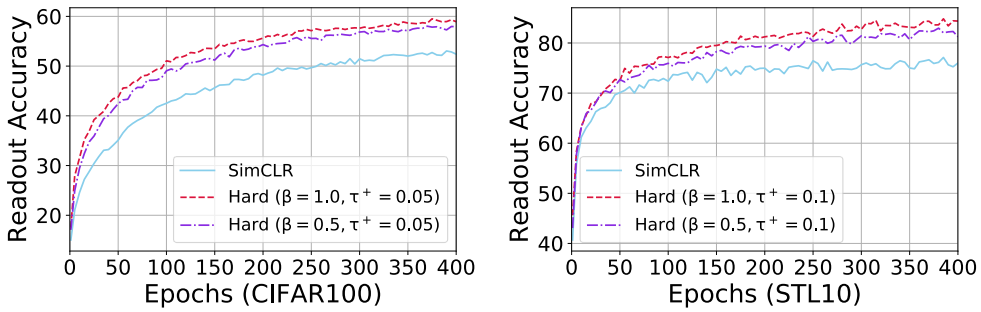


Figure 4-3: Hard sampling takes much fewer epochs to reach the same accuracy as SimCLR does in 400 epochs; for STL10 with $\beta = 1$ it takes only 60 epochs, and on CIFAR100 it takes 125 epochs (also with $\beta = 1$).

4.2.2 Hard Negatives with Large Batch Sizes

The vision experiments so far are all based off the SimCLR framework [24]. They use a relatively small batch size (up to 512). In order to test whether our hard negatives sampling method can help when the negative batch size is very large, we also run experiments using MoCo-v2 with standard negative memory bank size $N = 65536$ [27, 81]. We adopt the official MoCo-v2 code. Embeddings are trained for 200 epochs, with batch size 128. Figure 4-4 summarizes the results. We find that hard

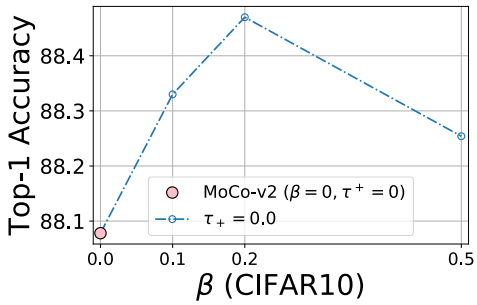


Figure 4-4: Hard negative sampling using MoCo-v2 framework. Results show that hard negative samples can still be useful when the negative memory bank is very large (in this case $N = 65536$).

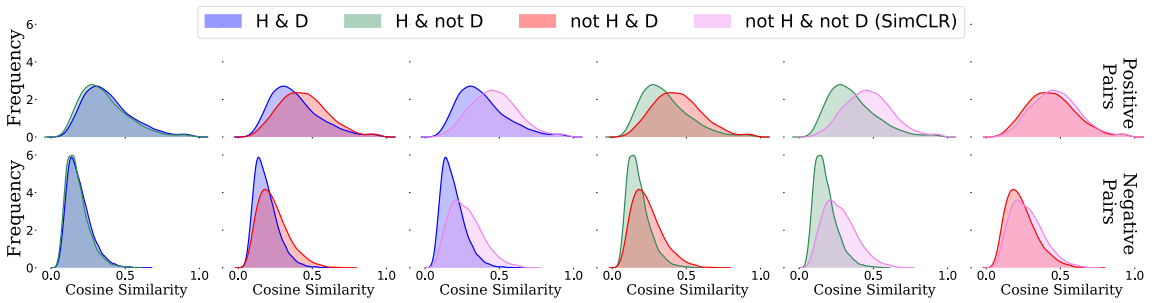


Figure 4-5: Histograms of cosine similarity of pairs of points with the same label (top) and different labels (bottom) for embeddings trained on CIFAR100 with four different objectives. H=Hard Sampling, D=Debiasing. Histograms overlaid pairwise to allow for convenient comparison.

Figure 4-6: Qualitative comparison of hard negatives and uniformly sampled negatives for embedding trained on STL10 for 400 epochs using SimCLR. Top row: selecting the 10 images with highest inner product with anchor in latent space from a batch of 128 inputs. Bottom row: a set of random samples from the same batch. Hard negatives are semantically much more similar to the anchor than uniformly sampled negatives - hard negatives possess many similar characteristics to the anchor, including texture, colors, animals vs machinery.

negative sampling can still improve the generalization of embeddings trained on CIFAR10: MoCo-v2 attains linear readout accuracy of 88.08%, and MoCo-v2 with hard negatives ($\beta = 0.2$, $\tau^+ = 0$) attains 88.47%.

Ablations To study the effect of varying the concentration parameter β on the learned embeddings Figure 4-5 plots cosine similarity histograms of pairs of similar and dissimilar points. The results show that for β moving from 0 through 0.5 to 2 causes both the positive and negative similarities to gradually skew left. In terms of downstream classification, an important property is the *relative* difference in similarity between positive and negative pairs. In this case $\beta = 0.5$ find the best balance (since it achieves the highest downstream accuracy). When β is taken very large ($\beta = 6$), we see a change in conditions. Both positive and negative pairs are assigned higher similarities in general. Visually it seems that the positive and negative histograms for

$\beta = 6$ overlap a lot more than for smaller values, which helps explain why the linear readout accuracy is lower for $\beta = 6$.

Figure 4-6 gives real examples of hard vs. uniformly sampled negatives. Given an anchor x (a monkey) and trained embedding f (trained on STL10 using standard SimCLR for 400 epochs), we sample a batch of 128 images. The top row shows the ten negatives x^- that have the largest inner product $f(x)^\top f(x^-)$, while the bottom row is a random sample from from the same batch. Negatives with the largest inner product with the anchor correspond to the items in the batch are the most important terms in the objective since they are given the highest weighting by q_β^-. Figure 4-6 shows that "real" hard negatives are conceptually similar to the idea as proposed in Figure 1: hard negatives are semantically similar to the anchor, possessing various similarities, including color (browns and greens), texture (fur), and objects (animals vs machinery).

4.2.3 Graph Representations

Second, we consider hard negative sampling in the context of learning graph representations. We use the state-of-the-art InfoGraph method introduced by [181] as the baseline, which is suitable for downstream graph-level classification. The objective is of a slightly different form from the NCE loss. Because of this we use a generalization of the formulation. In doing so, we illustrate that it is easy to adapt our hard sampling method to other contrastive frameworks.

Fig. 4-7 shows the results of fine-tuning an SVM [21, 43] on the fixed, learned embedding for a range of different values of β. Hard sampling does as well as InfoGraph in all cases, and better in 6 out of 8 cases. For ENZYMES and REDDIT, hard negative samples improve the accuracy by 3.2% and 2.4%, respectively, for DD and PTC by $1-2\%$, and for IMDB-B and MUTAG by at least 0.5%. Usually, multiple different choices of $\beta > 0$ were competitive with the InfoGraph baseline: 17 out of the 24 values of $\beta > 0$ tried (across all 8 datasets) achieve accuracy as high or better than InfoGraph ($\beta = 0$).

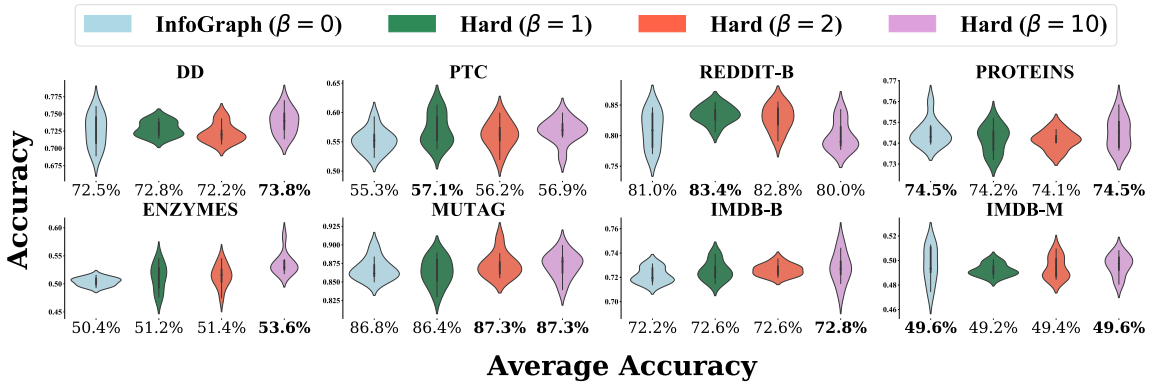


Figure 4-7: Classification accuracy on downstream tasks. We compare graph representations on four classification tasks. Accuracies are obtained by fine-tuning an SVM readout function, and are the average of 10 runs, each using 10-fold cross validation. Results in **bold** indicate best performer.

4.2.4 Sentence Embeddings

Third, we test hard negative sampling on learning representations of sentences using the *quick-thoughts* (QT) vectors framework introduced by [119], which uses adjacent sentences (before/after) as positive samples. Embeddings are trained using the unlabeled BookCorpus dataset [104], and evaluated following the protocol of [119] on six downstream tasks. The results are reported in Table 4.2. Hard sampling outperforms or equals the QT baseline in 5 out of 6 cases, the debiased baseline [36] in 4 out of 6, and both in 3 out of 6 cases. Setting $\tau^+ > 0$ led to numerical issues in optimization for hard sampling.

Objective	**MR**	**CR**	**SUBJ**	**MPQA**	**TREC**	**MSRP**	
						(Acc)	**(F1)**
QT ($\beta = 0, \tau^+ = 0$)	76.8	81.3	86.6	93.4	**89.8**	73.6	81.8
Debiased ($\tau^+ = 0.01$)	76.2	82.9	86.9	**93.7**	89.1	**74.7**	**82.7**
Hard ($\beta = 1, \tau^+ = 0$)	77.1	82.5	**87.0**	92.9	89.2	73.9	82.2
Hard ($\beta = 2, \tau^+ = 0$)	**77.4**	**83.6**	86.8	93.4	88.7	73.5	82.0

Table 4.2: Classification accuracy on downstream tasks. Sentence representations are learned using quick-thoughts (QT) vectors on the BookCorpus dataset and evaluated on six classification tasks. Evaluation of binary classification tasks (MR, CR, SUBJ, MPQA) uses 10-fold cross validation.

4.3 Theoretical Analysis of Hard Negative Sampling

4.3.1 Hard Sampling Interpolates between Marginal and Worst-case Negatives

Intuitively, the concentration parameter β in the proposed negative sample distribution q_β^- controls the level of "hardness" of the negative samples. As discussed earlier, the debiasing method in Chapter 3 can be recovered as a special case: taking $\beta = 0$ to obtain the distribution q_0^-. This case amounts to correcting for the fact that some samples in a negative batch sampled from p will have the same label as the anchor. But what interpretation does large β admit? Specifically, what does the distribution q_β^- converge to in the limit $\beta \to \infty$, if anything? We show that in the limit q_β^- approximates an inner solution to the following zero-sum two player game.

$$\inf_f \sup_{q \in \Pi} \left\{ \mathcal{L}(f, q) = \mathbb{E}_{\substack{x \sim p \\ x^+ \sim p_x^+}} \left[-\log \frac{e^{f(x)^T f(x^+)}}{e^{f(x)^T f(x^+)} + Q \mathbb{E}_{x^- \sim q}[e^{f(x)^T f(x^-)}]} \right] \right\}. \tag{4.5}$$

where $\Pi = \{q = q(\cdot; x, f) : \operatorname{supp}\big(q(\cdot; x, f)\big) \subseteq \{x' \in \mathcal{X} : x' \nsim x\}, \forall x \in \mathcal{X}\}$ is the set of distributions with support that is disjoint from points with the same class as x (without loss of generality we assume $\{x' \in \mathcal{X} : x' \nsim x\}$ is non-empty). Since $q = q(\cdot; x, f)$ depends on x and f it can be thought of as a family of distributions. The formal statement is as follows.

Proposition 4.3. *Let $\mathcal{L}^*(f) = \sup_{q \in \Pi} \mathcal{L}(f, q)$. Then for any $t > 0$ and $f : \mathcal{X} \to \mathbb{S}^{d-1}/t$ we observe the convergence $\mathcal{L}(f, q_\beta^-) \longrightarrow \mathcal{L}^*(f)$ as $\beta \to \infty$.*

To develop a better intuitive understanding of the worst case negative distribution objective $\mathcal{L}^*(f) = \sup_{q \in \Pi} \mathcal{L}(f, q)$, we note that the supremum can be characterized analytically. Indeed,

$$\begin{aligned}
\sup_{q \in \Pi} \mathcal{L}(f, q) &= -\mathbb{E}_{\substack{x \sim p \\ x^+ \sim p_x^+}} f(x)^T f(x^+) + \sup_{q \in \Pi} \mathbb{E}_{\substack{x \sim p \\ x^+ \sim p_x^+}} \log \left\{ e^{f(x)^T f(x^+)} + Q \mathbb{E}_{x^- \sim q}[e^{f(x)^T f(x^-)}] \right\} \\
&= -\mathbb{E}_{\substack{x \sim p \\ x^+ \sim p_x^+}} f(x)^T f(x^+) + \mathbb{E}_{\substack{x \sim p \\ x^+ \sim p_x^+}} \log \left\{ e^{f(x)^T f(x^+)} + Q \cdot \sup_{q \in \Pi} \mathbb{E}_{x^- \sim q}[e^{f(x)^T f(x^-)}] \right\}.
\end{aligned}$$

The supremum over q can be pushed inside the expectation since q is a family of distribution indexed by x, reducing the problem to maximizing $\mathbb{E}_{x^- \sim q}[e^{f(x)^T f(x^-)}]$, which is solved by any q^* whose support is a subset of $\arg\sup_{x^-: x^- \nsim x} e^{f(x)^T f(x^-)}$ if the supremum is attained. However, computing such points involves maximizing a neural network. Instead of taking this challenging route, using q_β^- defines a lower bound by placing higher probability on x^- for which $f(x)^T f(x^-)$ is large. This lower bound becomes tight as $\beta \to \infty$ (Proposition 4.3). We now give the proof of the proposition.

Proof. Consider the following essential supremum,

$$M(x) = \operatorname*{ess\,sup}_{x^- \in \mathcal{X}: x^- \nsim x} f(x)^T f(x^-) = \sup\{m > 0 : m \geq f(x)^T f(x^-) \text{ a.s. for } x^- \sim p^-\}.$$

The second inequality holds since $\operatorname{supp}(p) = \mathcal{X}$. We may rewrite

$$\mathcal{L}^*(f) = \mathbb{E}_{\substack{x \sim p \\ x^+ \sim p_x^+}} \left[-\log \frac{e^{f(x)^T f(x^+)}}{e^{f(x)^T f(x^+)} + Q e^{M(x)}} \right],$$

$$\mathcal{L}(f, q_\beta^-) = \mathbb{E}_{\substack{x \sim p \\ x^+ \sim p_x^+}} \left[-\log \frac{e^{f(x)^T f(x^+)}}{e^{f(x)^T f(x^+)} + Q \mathbb{E}_{x^- \sim q_\beta^-}[e^{f(x)^T f(x^-)}]} \right].$$

The difference between these two terms can be bounded as follows,

$$\begin{aligned}
&\left| \mathcal{L}^*(f) - \mathcal{L}(f, q_\beta^-) \right| \\
&\leq \mathbb{E}_{\substack{x \sim p \\ x^+ \sim p_x^+}} \left| -\log \frac{e^{f(x)^T f(x^+)}}{e^{f(x)^T f(x^+)} + Q e^{M(x)}} + \log \frac{e^{f(x)^T f(x^+)}}{e^{f(x)^T f(x^+)} + Q \mathbb{E}_{x^- \sim q_\beta^-}[e^{f(x)^T f(x^-)}]} \right| \\
&= \mathbb{E}_{\substack{x \sim p \\ x^+ \sim p_x^+}} \left| \log \left(e^{f(x)^T f(x^+)} + Q \mathbb{E}_{x^- \sim q_\beta^-}[e^{f(x)^T f(x^-)}] \right) - \log \left(e^{f(x)^T f(x^+)} + Q e^{M(x)} \right) \right| \\
&\leq \frac{e^{1/t}}{Q+1} \cdot \mathbb{E}_{\substack{x \sim p \\ x^+ \sim p_x^+}} \left| e^{f(x)^T f(x^+)} + Q \mathbb{E}_{x^- \sim q_\beta^-}[e^{f(x)^T f(x^-)}] - e^{f(x)^T f(x^+)} - Q e^{M(x)} \right| \\
&= \frac{e^{1/t} Q}{Q+1} \cdot \mathbb{E}_{x \sim p} \left| \mathbb{E}_{x^- \sim q_\beta^-}[e^{f(x)^T f(x^-)}] - e^{M(x)} \right| \\
&\leq e^{1/t} \cdot \mathbb{E}_{x \sim p} \mathbb{E}_{x^- \sim q_\beta^-} \left| e^{M(x)} - e^{f(x)^T f(x^-)} \right|
\end{aligned}$$

where for the second inequality we have used the fact that f lies on the hypersphere of radius $1/t$ to restrict the domain of the logarithm to values greater than $(Q+1)e^{-1/t}$. Because of this the logarithm is Lipschitz with parameter $e^{1/t}/(Q+1)$. Using again the fact that f lies on the hypersphere we know that $\left|f(x)^T f(x^-)\right| \leq 1/t^2$ and hence have the following inequality,

$$\mathbb{E}_{x\sim p}\mathbb{E}_{q_\beta^-}\left|e^{M(x)} - e^{f(x)^T f(x^-)}\right| \leq e^{1/t^2}\mathbb{E}_{x\sim p}\mathbb{E}_{q_\beta^-}\left|M(x) - f(x)^T f(x^-)\right|$$

Let us consider the inner expectation $E_\beta(x) = \mathbb{E}_{q_\beta^-}\left|M(x) - f(x)^T f(x^-)\right|$. Note that since f is bounded, $E_\beta(x)$ is uniformly bounded in x. Therefore, in order to show the convergence $\mathcal{L}(f, q_\beta^-) \to \mathcal{L}^*(f)$ as $\beta \to \infty$, it suffices by the dominated convergence theorem to show that $E_\beta(x) \to 0$ pointwise as $\beta \to \infty$ for arbitrary fixed $x \in \mathcal{X}$.

From now on we denote $M = M(x)$ for brevity, and consider a fixed $x \in \mathcal{X}$. From the definition of q_β^- it is clear that $q_\beta^- \ll p^-$. That is, since $q_\beta^- = c \cdot p^-$ for some (non-constant) c, it is absolutely continuous with respect to p^-. So $M(x) \geq f(x)^T f(x^-)$ almost surely for $x^- \sim q_\beta^-$, and we may therefore drop the absolute value signs from our expectation. Define the following event $\mathcal{G}_\varepsilon = \{x^- : f(x)^\top f(x^-) \geq M - \varepsilon\}$ where $\mathcal{G}$ is refers to a "good" event. Define its complement $\mathcal{B}_\varepsilon = \mathcal{G}_\varepsilon^c$ where $\mathcal{B}$ is for "bad". For a fixed $x \in \mathcal{X}$ and $\varepsilon > 0$ consider,

$$\begin{aligned}
E_\beta(x) &= \mathbb{E}_{x^-\sim q_\beta^-}\left|M(x) - f(x)^T f(x^-)\right| \\
&= \mathbb{P}_{x^-\sim q_\beta^-}(\mathcal{G}_\varepsilon) \cdot \mathbb{E}_{x^-\sim q_\beta^-}\left[\left|M(x) - f(x)^T f(x^-)\right| \,|\mathcal{G}_\varepsilon\right] \\
&\qquad + \mathbb{P}_{x^-\sim q_\beta^-}(\mathcal{B}_\varepsilon) \cdot \mathbb{E}_{x^-\sim q_\beta^-}\left[\left|M(x) - f(x)^T f(x^-)\right| \,|\mathcal{B}_\varepsilon\right] \\
&\leq \mathbb{P}_{x^-\sim q_\beta^-}(\mathcal{G}_\varepsilon) \cdot \varepsilon + 2\mathbb{P}_{x^-\sim q_\beta^-}(\mathcal{B}_\varepsilon) \\
&\leq \varepsilon + 2\mathbb{P}_{x^-\sim q_\beta^-}(\mathcal{B}_\varepsilon).
\end{aligned}$$

We need to control $\mathbb{P}_{x^- \sim q_\beta^-}(\mathcal{B}_\varepsilon)$. Expanding,

$$\mathbb{P}_{x^- \sim q_\beta^-}(\mathcal{B}_\varepsilon) = \int_{\mathcal{X}} \mathbf{1}\left\{ f(x)^T f(x^-) < M(x) - \varepsilon \right\} \frac{e^{\beta f(x)^T f(x^-)} \cdot p^-(x^-)}{Z_\beta} \mathrm{d}x^-$$

where $Z_\beta = \int_{\mathcal{X}} e^{\beta f(x)^T f(x^-)} p^-(x^-) \mathrm{d}x^-$ is the partition function of q_β^-. We may bound this expression by,

$$\begin{aligned}
&\int_{\mathcal{X}} \mathbf{1}\left\{ f(x)^T f(x^-) < M - \varepsilon \right\} \frac{e^{\beta(M-\varepsilon)} \cdot p^-(x^-)}{Z_\beta} \mathrm{d}x^- \\
&\leq \frac{e^{\beta(M-\varepsilon)}}{Z_\beta} \int_{\mathcal{X}} \mathbf{1}\left\{ f(x)^T f(x^-) < M - \varepsilon \right\} p^-(x^-) \mathrm{d}x^- \\
&= \frac{e^{\beta(M-\varepsilon)}}{Z_\beta} \mathbb{P}_{x^- \sim p^-}(\mathcal{B}_\varepsilon) \\
&\leq \frac{e^{\beta(M-\varepsilon)}}{Z_\beta}
\end{aligned}$$

Note that

$$Z_\beta = \int_{\mathcal{X}} e^{\beta f(x)^T f(x^-)} p^-(x^-) \mathrm{d}x^- \geq e^{\beta(M-\varepsilon/2)} \mathbb{P}_{x^- \sim p^-}(f(x)^T f(x^-) \geq M - \varepsilon/2).$$

By the definition of $M = M(x)$ the probability $\rho_\varepsilon = \mathbb{P}_{x^- \sim p^-}(f(x)^T f(x^-) \geq M - \varepsilon/2) > 0$, and we may therefore bound,

$$\begin{aligned}
\mathbb{P}_{x^- \sim q_\beta^-}(\mathcal{B}_\varepsilon) &= \frac{e^{\beta(M-\varepsilon)}}{e^{\beta(M-\varepsilon/2)} \rho_\varepsilon} \\
&= e^{-\beta\varepsilon/2} / \rho_\varepsilon \\
&\longrightarrow 0 \ \text{ as } \ \beta \to \infty.
\end{aligned}$$

We may therefore take β to be sufficiently big so as to make $\mathbb{P}_{x^- \sim q_\beta^-}(\mathcal{B}_\varepsilon) \leq \varepsilon$ and therefore $E_\beta(x) \leq 3\varepsilon$. In other words, $E_\beta(x) \longrightarrow 0$ as $\beta \to \infty$.

□

4.3.2 Optimal Embedding for Worst-case Negatives

What desirable properties does an optimal contrastive embedding (global minimizer of $\mathcal{L}$) possess that make the representation generalizable? To study this question, we first analyze the distribution of an optimal embedding f^* on the hypersphere when negatives are sampled from the adversarial worst-case distribution. We consider a different limiting viewpoint of objective (4.1) as the number of negative samples $N \to \infty$. Following the formulation of [200] we take $Q = N$ in (4.1), and subtract $\log N$. This changes neither the set of minimizers, nor the geometry of the loss surface. Taking the number of negative samples $N \to \infty$ yields the limiting objective,

$$\mathcal{L}_\infty(f, q) = \mathbb{E}_{\substack{x \sim p \\ x^+ \sim p_x^+}} \left[-\log \frac{e^{f(x)^T f(x^+)}}{\mathbb{E}_{x^- \sim q}\left[e^{f(x)^T f(x^-)}\right]} \right]. \tag{4.6}$$

Theorem 4.4. *Suppose the downstream task is classification (i.e. $\mathcal{C}$ is finite), and let $\mathcal{L}_\infty^*(f) = \sup_{q \in \Pi} \mathcal{L}_\infty(f, q)$. The infimum $\inf_{f:\, measurable} \mathcal{L}_\infty^*(f)$ is attained, and any f^* achieving the global minimum is such that $f^*(x) = f^*(x^+)$ almost surely. Furthermore, letting $\mathbf{v}_c = f^*(x)$ for any x such that $h(x) = c$ (so $\mathbf{v}_c$ is well defined up to a set of x of measure zero), f^* is characterized as being any solution to the following ball-packing problem,*

$$\max_{\{\mathbf{v}_c \in \mathbb{S}^{d-1}/t\}_{c \in \mathcal{C}}} \sum_{c \in \mathcal{C}} \rho(c) \cdot \min_{c' \neq c} \|\mathbf{v}_c - \mathbf{v}_{c'}\|^2. \tag{4.7}$$

Proof. In order to study properties of global optima of the contrastive objective using the adversarial worst case hard sampling distribution recall that we have the following limiting objective,

$$\mathcal{L}_\infty(f, q) = \mathbb{E}_{\substack{x \sim p \\ x^+ \sim p_x^+}} \left[-\log \frac{e^{f(x)^T f(x^+)}}{\mathbb{E}_{x^- \sim q_\beta}\left[e^{f(x)^T f(x^-)}\right]} \right]. \tag{4.8}$$

We may separate the logarithm of a quotient into the sum of two terms plus a constant,

$$\mathcal{L}_\infty(f, q) = \mathcal{L}_{\text{align}}(f) + \mathcal{L}_{\text{unif}}(f, q) - 1/t^2$$

where $\mathcal{L}_{\text{align}}(f) = \mathbb{E}_{x,x^+}\|f(x)-f(x^+)\|^2/2$ and $\mathcal{L}_{\text{unif}}(f,q) = \mathbb{E}_{x\sim p}\log \mathbb{E}_{x^-\sim q} e^{f(x)^\top f(x^-)}$. Here we have used the fact that f lies on the boundary of the hypersphere of radius $1/t$, which gives us the following equivalence between inner products and squared Euclidean norm,

$$2/t^2 - 2f(x)^\top f(x^+) = \|f(x)\|^2 + \|f(x^+)\|^2 - 2f(x)^\top f(x^+) = \|f(x) - f(x^+)\|^2. \quad (4.9)$$

Taking supremum to obtain $\mathcal{L}^*_\infty(f) = \sup_{q\in\Pi} \mathcal{L}_\infty(f,q)$ we find that the second expression simplifies to,

$$\mathcal{L}^*_{\text{unif}}(f) = \sup_{q\in\Pi} \mathcal{L}_{\text{unif}}(f,q) = \mathbb{E}_{x\sim p}\log \sup_{x^-\nsim x} e^{f(x)^\top f(x^-)} = \mathbb{E}_{x\sim p} \sup_{x^-\nsim x} f(x)^\top f(x^-).$$

Using Eqn. (4.9), this can be re-expressed as,

$$\mathbb{E}_{x\sim p} \sup_{x^-\nsim x} f(x)^\top f(x^-) = -\mathbb{E}_{x\sim p} \inf_{x^-\nsim x} \|f(x) - f(x^-)\|^2/2 + 1/t^2. \quad (4.10)$$

Any minimizer of $\mathcal{L}_{\text{align}}(f)$ has the property that $f(x) = f(x^+)$ almost surely. So, in order to prove the first claim, it suffices to show that there exist functions $f \in \arg\inf_f \mathcal{L}^*_{\text{unif}}(f)$ for which $f(x) = f(x^+)$ almost surely. This is because, at that point, we have shown that $\arg\min_f \mathcal{L}_{\text{align}}(f)$ and $\arg\min_f \mathcal{L}^*_{\text{unif}}(f)$ intersect, and therefore any solution of $\mathcal{L}^*_\infty(f) = \mathcal{L}_{\text{align}}(f) + \mathcal{L}^*_{\text{unif}}(f)$ must lie in this intersection.

To this end, suppose that $f \in \arg\min_f \mathcal{L}^*_{\text{unif}}(f)$ but that $f(x) \neq f(x^+)$ with non-zero probability. We shall show that we can construct a new embedding $\hat{f}$ such that $f(x) = f(x^+)$ almost surely, and $\mathcal{L}^*_{\text{unif}}(\hat{f}) \leq \mathcal{L}^*_{\text{unif}}(f)$. Due to Eqn. (4.10) this last condition is equivalent to showing,

$$\mathbb{E}_{x\sim p} \inf_{x^-\nsim x} \|\hat{f}(x) - \hat{f}(x^-)\|^2 \geq \mathbb{E}_{x\sim p} \inf_{x^-\nsim x} \|f(x) - f(x^-)\|^2. \quad (4.11)$$

Fix a $c \in \mathcal{C}$, and let $x_c \in \arg\max_{x:h(x)=c} \inf_{x^-\nsim x} \|f(x) - f(x^-)\|^2$. The maximum is guaranteed to be attained, as we explain now. Indeed we know the maximum is attained at some point in the closure $\partial\{x : h(x) = c\} \cup \{x : h(x) = c\}$. Since $\mathcal{X}$ is

compact and connected, any point $\bar{x} \in \partial\{x : h(x) = c\} \setminus \{x : h(x) = c\}$ is such that $\inf_{x^- \nsim \bar{x}} \|f(\bar{x}) - f(x^-)\|^2 = 0$ since $\bar{x}$ must belong to $\{x : h(x) = c'\}$ for some other c'. Such an $\bar{x}$ cannot be a solution unless all points in $\{x : h(x) = c\}$ also achieve 0, in which case we can simply take x_c to be a point in the interior of $\{x : h(x) = c\}$.

Now, define $\hat{f}(x) = f(x_c)$ for any x such that $h(x) = c$ and $\hat{f}(x) = f(x)$ otherwise. Let us first aim to show that Eqn. (4.11) holds for this $\hat{f}$. Let us begin to expand the left hand side of Eqn. (4.11),

$$
\begin{aligned}
&\mathbb{E}_{x\sim p} \inf_{x^- \nsim x} \|\hat{f}(x) - \hat{f}(x^-)\|^2 \\
&\qquad = \mathbb{E}_{\hat{c}\sim\rho}\mathbb{E}_{x\sim p(\cdot|\hat{c})} \inf_{x^- \nsim x} \|\hat{f}(x) - \hat{f}(x^-)\|^2 \\
&\qquad = \rho(c)\mathbb{E}_{x\sim p(\cdot|c)} \inf_{x^- \nsim x} \|\hat{f}(x) - \hat{f}(x^-)\|^2 \\
&\qquad\qquad + (1-\rho(c))\mathbb{E}_{\hat{c}\sim\rho(\cdot|\hat{c}\neq c)}\mathbb{E}_{x\sim p(\cdot|\hat{c})} \inf_{x^- \nsim x} \|\hat{f}(x) - \hat{f}(x^-)\|^2 \\
&\qquad = \rho(c)\mathbb{E}_{x\sim p(\cdot|c)} \inf_{x^- \nsim x} \|f(x_c) - f(x^-)\|^2 \\
&\qquad\qquad + (1-\rho(c))\mathbb{E}_{\hat{c}\sim\rho(\cdot|\hat{c}\neq c)}\mathbb{E}_{x\sim p(\cdot|\hat{c})} \inf_{x^- \nsim x} \|\hat{f}(x) - \hat{f}(x^-)\|^2 \\
&\qquad = \rho(c) \inf_{x^- \nsim x_c} \|f(x_c) - f(x^-)\|^2 \\
&\qquad\qquad + (1-\rho(c))\mathbb{E}_{\hat{c}\sim\rho(\cdot|\hat{c}\neq c)}\mathbb{E}_{x\sim p(\cdot|\hat{c})} \inf_{h(x^-)\neq\hat{c}} \|\hat{f}(x) - \hat{f}(x^-)\|^2 \qquad (4.12)
\end{aligned}
$$

By construction, the first term can be lower bounded by $\inf_{x^- \nsim x_c} \|f(x_c) - f(x^-)\|^2 \geq \mathbb{E}_{x\sim p(\cdot|c)} \inf_{h(x^-)\neq c} \|f(x) - f(x^-)\|^2$ for any x such that $h(x) = c$. To lower bound the second term, consider any fixed $\hat{c} \neq c$ and $x \sim p(\cdot|\hat{c})$ (so $h(x) = \hat{c}$). Define the following two subsets of the input space $\mathcal{X}$

$$
\mathcal{A} = \{f(x^-) : f(x^-) \neq \hat{c} \text{ for } x^- \in \mathcal{X}\} \qquad \widehat{\mathcal{A}} = \{f(x^-) \in \mathcal{X} : \hat{f}(x^-) \neq \hat{c} \text{ for } x^- \in \mathcal{X}\}.
$$

Since by construction the range of $\hat{f}$ is a subset of the range of f, we know that $\widehat{\mathcal{A}} \subseteq \mathcal{A}$. Combining this with the fact that $\hat{f}(x) = f(x)$ whenever $h(x) = \hat{c} \neq c$ we

see,

$$\begin{aligned}\inf_{h(x^-)\neq\hat{c}} \|\hat{f}(x) - \hat{f}(x^-)\|^2 &= \inf_{h(x^-)\neq\hat{c}} \|f(x) - \hat{f}(x^-)\|^2 \\ &= \inf_{u\in\widehat{\mathcal{A}}} \|f(x) - u\|^2 \\ &\geq \inf_{u\in\mathcal{A}} \|f(x) - u\|^2 \\ &= \inf_{h(x^-)\neq\hat{c}} \|f(x) - f(x^-)\|^2\end{aligned}$$

Using these two lower bounds we may conclude that Eqn. (4.12) can be lower bounded by,

$$\rho(c)\mathbb{E}_{x\sim p(\cdot|c)} \inf_{h(x^-)\neq c} \|f(x) - f(x^-)\|^2 + (1-\rho(c))\mathbb{E}_{\hat{c}\sim\rho(\cdot|\hat{c}\neq c)}\mathbb{E}_{x\sim p(\cdot|\hat{c})} \inf_{h(x^-)\neq\hat{c}} \|f(x) - f(x^-)\|^2$$

which equals $\mathbb{E}_{x\sim p} \inf_{x^-\nsim x} \|f(x) - f(x^-)\|^2$. We have therefore proved Eqn. (4.11). To summarize the current progress; given an embedding f we have constructed a new embedding $\hat{f}$ that attains lower $\mathcal{L}_{\text{unif}}$ loss and which is constant on x such that $\hat{f}$ is constant on $\{x : h(x) = c\}$. Enumerating $\mathcal{C} = \{c_1, c_2 \ldots, c_{|\mathcal{C}|}\}$, we may repeatedly apply the same argument to construct a sequence of embeddings $f_1, f_2, \ldots, f_{|\mathcal{C}|}$ such that f_i is constant on each of the following sets $\{x : h(x) = c_j\}$ for $j \leq i$. The final embedding in the sequence $f^* = f_{|\mathcal{C}|}$ is such that $\mathcal{L}^*_{\text{unif}}(f^*) \leq \mathcal{L}^*_{\text{unif}}(f)$ and therefore f^* is a minimizer. This embedding is constant on each of $\{x : h(x) = c_j\}$ for $j = 1, 2, \ldots, |\mathcal{C}|$. In other words, $f^*(x) = f^*(x^+)$ almost surely. We have proved the first claim.

Obtaining the second claim is a matter of manipulating $\mathcal{L}^*_\infty(f^*)$. Indeed, we know that $\mathcal{L}^*_\infty(f^*) = \mathcal{L}^*_{\text{unif}}(f^*) - 1/t^2$ and defining $\mathbf{v}_c = f^*(x) = f(x_c)$ for each $c \in \mathcal{C}$, this expression is minimized if and only if f^* attains,

$$\begin{aligned}\max_f \mathbb{E}_{x\sim p} \inf_{x^-\nsim x} \|f(x) - f(x^-)\|^2 &= \max_f \mathbb{E}_{c\sim\rho}\mathbb{E}_{x\sim p(\cdot|c)} \inf_{h(x^-)\neq c} \|f(x) - f(x^-)\|^2 \\ &= \max_f \sum_{c\in\mathcal{C}} \rho(c) \cdot \inf_{h(x^-)\neq c} \|f(x) - f(x^-)\|^2 \\ &= \max_{\{\mathbf{v}_c\in\mathbb{S}^{d-1}/t\}_{c\in\mathcal{C}}} \sum_{c\in\mathcal{C}} \rho(c) \cdot \min_{c'\neq c} \|\mathbf{v}_c - \mathbf{v}_{c'}\|^2\end{aligned}$$

where the final equality inserts f^* as an optimal f and reparameterizes the maximum to be over the set of vectors $\{\mathbf{v}_c \in \mathbb{S}^{d-1}/t\}_{c\in\mathcal{C}}$.

□

Interpretation: The first component of the result is that $f^*(x) = f^*(x^+)$ almost surely for an optimal f^*. That is, an optimal embedding f^* must be invariant across pairs of similar inputs x, x^+. The second component is characterizing solutions via the classical geometrical Ball-Packing Problem of [183] (Eq. 4.7) that has only been solved exactly for uniform ρ, for specific of $|\mathcal{C}|$ and typically for $\mathbb{S}^2$ [134, 170, 183]. When the distribution ρ over classes is uniform this problem is solved by a set of $|\mathcal{C}|$ points on the hypersphere such that the average squared-ℓ_2 distance from a point to the nearest other point is as large as possible. In other words, suppose we wish to place $|\mathcal{C}|$ number of balls[1] on $\mathbb{S}^{d-1}$ so that they do not intersect. Then solutions to Tammes' Problem (4.7) expresses (twice) the largest possible average squared radius that the balls can have. So, we have a ball-packing problem where instead of trying to pack as many balls as possible of a fixed size, we aim to pack a fixed number of balls (one for each class) to have as big radii as possible. Non-uniform ρ adds importance weights to each fixed ball. In summary, solutions of the problem $\min_f \mathcal{L}^*_\infty(f)$ are a maximum margin clustering.

This understanding of global minimizers of $\mathcal{L}^*_\infty(f) = \sup_{q\in\Pi} \mathcal{L}_\infty(f, q)$ can further developed into a better understanding of generalization on downstream tasks. The next result shows that representations that achieve small excess risk on the objective

[1]For a manifold $\mathcal{M} \subseteq \mathbb{R}^d$, we say $C \subset \mathcal{M}$ is a ball if it is connected, and there exists a Euclidean ball $\mathcal{B} = \{x \in \mathbb{R}^d : \|x\|_2 \leq R\}$ for which $C = \mathcal{M} \cap \mathcal{B}$.

$\mathcal{L}^*_\infty$ still separate clusters well in the sense that a simple 1-nearest neighbor classifier achieves low classification error.

Theorem 4.5. *Suppose ρ is uniform on $\mathcal{C}$ and f is such that $\mathcal{L}^*_\infty(f) - \inf_{\bar{f}\ measurable} \mathcal{L}^*_\infty(\bar{f}) \leq \varepsilon$ with $\varepsilon \leq 1$. Let $\{\mathbf{v}^*_c \in \mathbb{S}^{d-1}/t\}_{c\in\mathcal{C}}$ be a solution to Problem 4.7, and define the constant $\xi = \min_{c,c^-:c\neq c^-} \left\|\mathbf{v}^*_c - \mathbf{v}^*_{c-}\right\| > 0$. Then there exists a set of vectors $\{\mathbf{v}_c \in \mathbb{S}^{d-1}/t\}_{c\in\mathcal{C}}$ such that the 1-nearest neighbor classifier $\hat{h}(x) = \arg\min_{\bar{c}\in\mathcal{C}} \left\|f(x) - \mathbf{v}_{\bar{c}}\right\|$ (ties broken arbitrarily) achieves misclassification risk,*

$$\mathbb{P}_{x,c}(\hat{h}(x) \neq c) \leq \frac{8\varepsilon}{(\xi^2 - 2|\mathcal{C}|\,(1+1/t)\varepsilon^{1/2})^2}$$

Proof. To begin, using the definition of $\hat{h}$ we know that for any $0 < \delta < \xi$,

$$\begin{aligned}
\mathbb{P}_{x,c}(\hat{h}(x) = c) &= \mathbb{P}_{x,c}\left(\left\|f(x) - \mathbf{v}_c\right\| \leq \min_{c^-:c^-\neq c} \left\|f(x) - \mathbf{v}_{c^-}\right\|\right) \\
&\geq \mathbb{P}_{x,c}\left(\left\|f(x) - \mathbf{v}_c\right\| \leq \delta, \quad \text{and} \quad \delta \leq \min_{c^-:c^-\neq c} \left\|f(x) - \mathbf{v}_{c^-}\right\|\right) \\
&\geq 1 - \mathbb{P}_{x,c}\left(\left\|f(x) - \mathbf{v}_c\right\| > \delta\right) - \mathbb{P}_{x,c}\big(\min_{c^-:c^-\neq c} \left\|f(x) - \mathbf{v}_{c^-}\right\| < \delta\big)
\end{aligned}$$

So to prove the result, our goal is now to bound these two probabilities. To do so, we use the bound on the excess risk. Indeed, combining the fact $\mathcal{L}^*_\infty(f) - \inf_{\bar{f}\ \text{measurable}} \mathcal{L}^*_\infty(\bar{f}) \leq \varepsilon$ with the notational rearrangements before Theorem **??** we observe that $\mathbb{E}_{x,x^+}\left\|f(x) - f(x^+)\right\|^2 \leq 2\varepsilon$.

We have,

$$2\varepsilon \geq \mathbb{E}_{x,x^+}\left\|f(x) - f(x^+)\right\|^2 = \mathbb{E}_{c\sim\rho}\mathbb{E}_{x^+\sim p(\cdot|c)}\mathbb{E}_{x\sim p(\cdot|c)}\left\|f(x) - f(x^+)\right\|^2.$$

For fixed c, x^+, let $x_c \in \arg\min_{\overline{\{x^+:h(x^+)=c\}}} \mathbb{E}_{x\sim p(\cdot|c)}\left\|f(x) - f(x^+)\right\|^2$ where we extend the minimum to be over the closure, a compact set, to guarantee it is attained. Then we have

$$2\varepsilon \geq \mathbb{E}_{c\sim\rho}\mathbb{E}_{x^+\sim p(\cdot|c)}\mathbb{E}_{x\sim p(\cdot|c)}\big\|f(x)-f(x^+)\big\|^2 \geq \mathbb{E}_{c\sim\rho}\mathbb{E}_{x\sim p(\cdot|c)}\big\|f(x)-\mathbf{v}_c\big\|^2$$

where we have now defined $\mathbf{v}_c = f(x_c)$ for each $c \in \mathcal{C}$. Note in particular that $\mathbf{v}_c$ lies on the surface of the hypersphere $\mathbb{S}^{d-1}/t$. This enables us to obtain the follow bound using Markov's inequality,

$$\begin{aligned}\mathbb{P}_{x,c}\left(\big\|f(x)-\mathbf{v}_c\big\| > \delta\right) &= \mathbb{P}_{x,c}\left(\big\|f(x)-\mathbf{v}_c\big\|^2 > \delta^2\right)\\ &\leq \frac{\mathbb{E}_{x,c}\big\|f(x)-\mathbf{v}_c\big\|^2}{\delta^2}\\ &\leq \frac{2\varepsilon}{\delta^2}.\end{aligned}$$

so it remains still to bound $\mathbb{P}_{x,c}\big(\min_{c^-:c^-\neq c}\big\|f(x)-\mathbf{v}_{c^-}\big\| < \delta\big)$. Defining $\xi' = \min_{c,c^-:c\neq c^-}\|\mathbf{v}_c-\mathbf{v}_{c-}\|$, we have the following fact (proven in Appendix B.2).

Fact (see lemma B.1): $\xi' \geq \sqrt{\xi^2 - 2|\mathcal{C}|\,(1+1/t)\sqrt{\varepsilon}}$.

Using this fact we are able to get control over the tail probability as follows,

$$\begin{aligned}\mathbb{P}_{x,c}\left(\min_{c^-:c^-\neq c}\big\|f(x)-\mathbf{v}_{c^-}\big\| < \delta\right) &\leq \mathbb{P}_{x,c}\left(\big\|f(x)-\mathbf{v}_c\big\| > \xi' - \delta\right)\\ &\leq \mathbb{P}_{x,c}\left(\big\|f(x)-\mathbf{v}_c\big\| > \xi - \sqrt{\xi^2 - 2|\mathcal{C}|\,(1+1/t)\varepsilon^{1/2}} - \delta\right)\\ &= \mathbb{P}_{x,c}\left(\big\|f(x)-\mathbf{v}_c\big\|^2 > (\sqrt{\xi^2 - 2|\mathcal{C}|\,(1+1/t)\varepsilon^{1/2}} - \delta)^2\right)\\ &\leq \frac{2\varepsilon}{(\sqrt{\xi^2 - 2|\mathcal{C}|\,(1+1/t)\varepsilon^{1/2}} - \delta)^2}.\end{aligned}$$

where this inequality holds for for any $0 \leq \delta \leq \sqrt{\xi^2 - 2|\mathcal{C}|\,(1+1/t)\varepsilon^{1/2}}$.

Gathering together our tail probability bounds we find that $\mathbb{P}_{x,c}(\hat{h}(x) = c) \geq 1 - \frac{2\varepsilon}{\delta^2} - \frac{2\varepsilon}{(\sqrt{\xi^2-2|\mathcal{C}|(1+1/t)\varepsilon^{1/2}}-\delta)^2}$ for any $0 \leq \delta \leq \sqrt{\xi^2 - 2|\mathcal{C}|\,(1+1/t)\varepsilon^{1/2}}$. That is,

$$\mathbb{P}_{x,c}(\hat{h}(x) \neq c) \leq \frac{2\varepsilon}{\delta^2} + \frac{2\varepsilon}{(\sqrt{\xi^2 - 2|\mathcal{C}|\,(1+1/t)\varepsilon^{1/2}} - \delta)^2}$$

Since this holds for any $0 \leq \delta \leq \sqrt{\xi^2 - 2|\mathcal{C}|\,(1+1/t)\varepsilon^{1/2}}$,

$$\mathbb{P}_{x,c}(\hat{h}(x) \neq c) \leq \min_{0 \leq \delta \leq \sqrt{\xi^2 - 2|\mathcal{C}|\varepsilon}} \left\{ \frac{2\varepsilon}{\delta^2} + \frac{2\varepsilon}{(\sqrt{\xi^2 - 2|\mathcal{C}|\,(1+1/t)\varepsilon^{1/2}} - \delta)^2} \right\}.$$

Elementary calculus shows that the minimum is attained at $\delta = \frac{\sqrt{\xi^2 - 2|\mathcal{C}|(1+1/t)\varepsilon^{1/2}}}{2}$. Plugging this in yields the final bound,

$$\mathbb{P}(\hat{h}(x) \neq c) \leq \frac{8\varepsilon}{(\xi^2 - 2|\mathcal{C}|\,(1+1/t)\varepsilon^{1/2})^2}.$$

□

In particular, $\mathbb{P}(\hat{h}(x) \neq c) = \mathcal{O}(\varepsilon)$ as $\varepsilon \to 0$, and in the limit $\varepsilon \to 0$ we recover the invariance claim of Theorem 4.4 as a special case. The result can be generalized to arbitrary ρ by replacing $|\mathcal{C}|$ in the bound by $1/\min_c \rho(c)$. The result also implies that it is possible to build simple classifiers for tasks that involve only a subset of classes from $\mathcal{C}$, or classes that are a union of classes from $\mathcal{C}$. The constant $\xi = \min_{c,c^-:c\neq c^-} \left\| \mathbf{v}_c^* - \mathbf{v}_{c^-}^* \right\| > 0$ is a purely geometrical property of spheres, and describes the minimum separation distance between a set of points that solves the Tammes' ball-packing problem.

4.3.3 Bias-variance of Empirical Estimation

We close the theoretical section with a bias-variance analysis of our objective. Recall the final hard negative samples objective we derive is,

$$\mathbb{E}_{\substack{x\sim p\\ x^+\sim p_x^+}}\left[-\log\frac{e^{f(x)^Tf(x^+)}}{e^{f(x)^Tf(x^+)}+\frac{Q}{\tau^-}\big(\mathbb{E}_{x^-\sim q_\beta}\big[e^{f(x)^Tf(x^-)}\big]-\tau^+\mathbb{E}_{v\sim q_\beta^+}\big[e^{f(x)^Tf(v)}\big]\big)}\right]. \quad (4.13)$$

This objective admits a practical counterpart by using empirical approximations to $\mathbb{E}_{x^-\sim q_\beta}[e^{f(x)^Tf(x^-)}]$ and $\mathbb{E}_{v\sim q_\beta^+}[e^{f(x)^Tf(v)}]$. In practice we use a fairly large number of samples (e.g. $N = 510$) to approximate the first expectation, and only $M = 1$ samples to approximate the second. Clearly in both cases the resulting estimator is unbiased. Further, since the first expectation is approximated using many samples, and the integrand is bounded, the resulting estimator is well concentrated (e.g. apply Hoeffding's inequality out-of-the-box). But what about the second expectation? This might seem uncontrolled since we use only one sample, however it turns out that the random variable $X = e^{f(x)^Tf(v)}$ where $x \sim p$ and $v \sim q_\beta^+$ has variance that is bounded by $\mathcal{L}_{\text{align}}(f)$.

Lemma 4.6. *Consider the random variable $X = e^{f(x)^Tf(v)}$ where $x \sim p$ and $v \sim q_\beta^+$. Then $Var(X) \leq \mathcal{O}\big(\mathcal{L}_{align}(f)\big)$.*

Recall that $\mathcal{L}_{\text{align}}(f) = \mathbb{E}_{x,x^+}\|f(x)-f(x^+)\|^2/2$ is termed *alignment*, and [200] show that the contrastive objective jointly optimize *alignment* and *uniformity*. Lemma 4.6 therefore shows that as training evolves, the variance of the $X = e^{f(x)^Tf(v)}$ where $x \sim p$ and $v \sim q_\beta^+$ is bounded by a term that we expect to see becoming small, suggesting that using a single sample ($M = 1$) to approximate this expectation is not unreasonable. We cannot, however, say more than this since we have no guarantee that $\mathcal{L}_{\text{align}}(f)$ goes to zero.

Proof. Fix an x and recall that we are considering $q_\beta^+(\cdot) = q_\beta^+(\cdot; x)$. First let X' be an i.i.d. copy of X, and note that, conditioning on x, we have $2\text{Var}(X|x) = \text{Var}(X|x) + \text{Var}(X'|x) = \text{Var}(X - X'|x) \leq \mathbb{E}\big[(X - X')^2|x\big]$. Bounding this difference,

$$
\begin{aligned}
\mathbb{E}\big[(X - X')^2 | x\big] &= \mathbb{E}_{v,v' \sim q_\beta^+} \left(e^{f(x)^\top f(v)} - e^{f(x)^\top f(v')} \right)^2 \\
&\leq \mathbb{E}_{v,v' \sim q_\beta^+} \left(e^{1/t^2} \left[f(x)^\top f(v) - f(x)^\top f(v') \right] \right)^2 \\
&\leq e^{1/t^4} \mathbb{E}_{v,v' \sim q_\beta^+} \left(\left[\left\| f(x) \right\| \left\| f(v) - f(v') \right\| \right] \right)^2 \\
&= \frac{e^{1/t^4}}{t^2} \mathbb{E}_{v,v' \sim q_\beta^+} \left\| f(v) - f(v') \right\|^2 \\
&\leq \mathcal{O}\left(\mathbb{E}_{v,v' \sim p^+} \left\| f(v) - f(v') \right\|^2 \right)
\end{aligned}
$$

where the first inequality follows since f lies on the sphere of radius $1/t$, the second inequality by Cauchy–Schwarz, the third again since f lies on the sphere of radius $1/t$, and the fourth since q_β^+ is absolutely continuous with respect to p^+ with bounded ratio.

Since $p^+(x^+) = p(x^+|h(x))$ only depends on $c = h(x)$, rather than x itself, taking expectations over $x \sim p$ is equivalent to taking expectations over $c \sim \rho$. Further, $\rho(c)p(v|c)p(v'|c) = p(v)p(v'|c) = p(v)p_v^+(v')$. So $\mathbb{E}_{c \sim \rho} \mathbb{E}_{v,v' \sim p^+} \left\| f(v) - f(v') \right\|^2 = \mathbb{E}_{x,x^+} \left\| f(x) - f(x^+) \right\|^2 = 2\mathcal{L}_{\text{align}}(f)$, where $x \sim p$ and $x^+ \sim p_x^+$. Thus we obtain the lemma. □

4.4 Conclusion

We argue for the value of hard negatives in unsupervised contrastive representation learning, and introduce a simple hard negative sampling method. Our work connects two major lines of work: contrastive learning, and negative mining in metric learning. Doing so requires overcoming an apparent roadblock: negative mining in metric learning uses pairwise similarity information as a core component, while contrastive learning is unsupervised. Our method enjoys several nice aspects: having desirable theoretical properties, a very simple implementation that requires modify-

ing only a couple of lines of code, not changing anything about the data sampling pipeline, introducing zero extra computational overhead, and handling false negatives in a principled way.

Part II

Learning from Noisy Data

Chapter 5

False Positive Samples

The central idea of contrastive learning is to learn representations that capture the underlying information shared between different "views" of data [147, 188]. For images, the views, which are also known as positive samples, are typically constructed by applying common data augmentation techniques, such as jittering, cropping, resizing and rotation [24], and for video the views are often chosen as adjacent frames [171] or co-occurring multimodal signals, such as video and the corresponding optical flow [78], audio [132] and transcribed speech [127].

Designing the *right* contrasting views has shown to be a key ingredient of contrastive learning [24, 154]. This often requires domain knowledge, intuition, trial-and-error (and luck!). What would happen if the views are wrongly chosen and do not provide meaningful shared information? Prior work has reported deteriorating effects of such *noisy* views in contrastive learning under various scenarios, e.g., unrelated image patches due to extreme augmentation [188], irrelevant video-audio pairs due to overdubbing [131], and misaligned video-caption pairs [127]. The major issue with noisy views is that representations of different views are forced to align with each other even if there is no meaningful shared information. This often leads to sub-optimal representations that merely capture spurious correlations [6] or make them collapse to a trivial solution [94]. Worse yet, when we attempt to learn from large-scale unlabeled data – i.e., the scenario where self-supervised learning is particularly expected to shine – the issue is only aggravated because of the increased noise in the

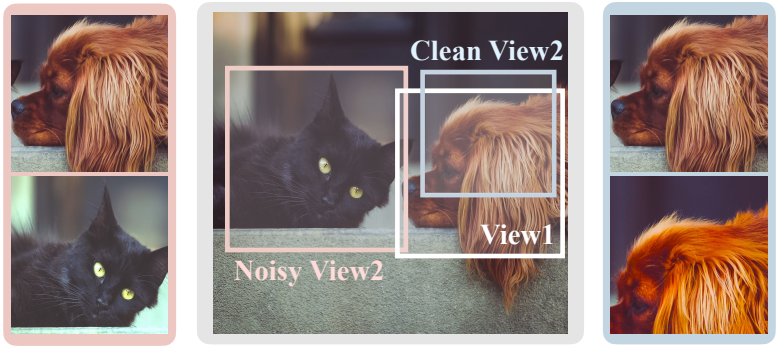


Figure 5-1: Problem of Noisy Views (False Positive Pairs). Noisy views, e.g., unrelated image patches due to extreme augmentation, can deteriorate the downstream performance.

real-world data [114], hindering the ultimate success of contrastive learning.

Consequently, a few attempts have been made to design contrastive approaches that are noise-tolerant. For example, Morgado et al. [131] optimize a soft instance discrimination loss to weaken the impact of noisy views. Miech et al. [127] address the misalignment between video and captions by aligning multiple neighboring segments of a video. However, existing approaches are often tied to specific modalities or make assumptions that may not hold for general scenarios, e.g., MIL-NCE [127] is not designed to address the issues of irrelevant audio-visual signals.

In this chapter, we develop a principled approach to make contrastive learning robust against noisy views. We start by making connections between contrastive learning and the classical noisy binary classification in supervised learning [64, 140]. This allows us to explore the wealth of literature on learning with noisy labels [65, 115, 194]. In particular, we focus on a family of robust loss functions that has the *symmetric* property [64], which provides strong theoretical guarantees against noisy labels in binary classification. We then show a functional form of contrastive learning that can satisfy the symmetry condition if given a proper symmetric loss function, motivating the design of new contrastive loss functions that provide similar theoretical guarantees.

This leads us to propose **R**obust **I**nfo**NCE** (RINCE), a contrastive loss function that satisfies the symmetry condition. RINCE can be understood as a generalized

form of the contrastive objective that is robust against noisy views. Intuitively, its symmetric property provides an implicit means to reweight sample importance in the gradient space *without requiring* an explicit form of noise estimator. It also provides a simple "knob" (a real-valued scalar $q \in (0, 1]$) that controls the behavior of the loss function balancing the exploration-exploitation trade-off (i.e., from being conservative to playing adventures on potentially noisy samples). Implementing RINCE requires only a few lines code and can be a simple drop-in replacement for the InfoNCE loss to make contrastive learning robust against noisy views. Since InfoNCE sets the basis for many modern contrastive methods such as SimCLR [24] and MoCo-v1/v2/v3 [27, 28, 81], our construction can be easily applied to many existing frameworks.

5.1 From Noisy Labels to Noisy Views

To address the problem of noisy views, we start by connecting two seemingly different but related frameworks: supervised binary classification with noisy labels and self-supervised contrastive learning with noisy views. We then introduce a family of symmetric loss functions that is noise-tolerant and show how we can transform contrastive objectives to a symmetric form.

5.1.1 Robust Loss Functions against Noise

Before the era of self-supervised learning, The problem of noise has been actively explored explored in supervised learning setting, especially in supervised classification with noisy labels [64, 76, 89, 115, 117, 140, 155, 162, 179, 194, 209]. One line of work attempts to develop robust loss functions that are noise-tolerant [64, 65, 201, 226]. Ghosh et al. [64] prove that symmetric loss functions are robust against noisy labels, e.g., Mean Absolute Error (MAE) [65], while commonly used Cross Entropy (CE) loss is not. Based on this idea, Zhang and Sabuncu [226] propose the generalized cross entropy loss to combine MAE and CE loss functions. A similar idea is adopted in [201] by combining the reversed cross entropy loss with CE loss. In the section 5.2, we will relate noisy views to noisy labels by interpreting contrastive learning as

binary classification, and developed a robust symmetric contrastive loss that enjoys the similar theoretical guarantees.

5.1.2 Symmetric Loss Function

Denoting the input space by $\mathcal{X}$ and the binary output space by $\mathcal{Y} = \{-1, 1\}$, let $\mathcal{S} = \{x_i, y_i\}_{i=1}^m$ be the unobserved clean dataset that is drawn i.i.d. from the data distribution $\mathcal{D}$. In the noisy setting, the learner obtains a noisy dataset $\mathcal{S}_\eta = \{x_i, \hat{y}_i\}_{i=1}^m$, where $\hat{y}_i = y_i$ with probability $1-\eta_{x_i}$ and $\hat{y}_i = -y_i$ with probability η_{x_i}. Note that the noise rate η_x is data point-dependent. For a classifier $f \in \mathcal{F} : \mathcal{X} \to \mathbb{R}$, the expected risk under the noise-free scenario is $R_\ell(f) = \mathbb{E}_\mathcal{D}[\ell(f(x), y)]$ where $\ell : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}$ is a binary classification loss function. When the noise exists, the learner minimizes the noisy expected risk $R_\ell^\eta(f) = \mathbb{E}_{\mathcal{D}_\eta}[\ell(f(x), \hat{y})]$.

Ghosh et al. [64] show that *symmetric* loss functions are robust against noisy labels in binary classification. In particular, a loss function ℓ is symmetric if it sums to a constant:

$$\ell(s, 1) + \ell(s, -1) = c, \quad \forall s \in \mathbb{R}, \tag{5.1}$$

where s is the prediction score from f. Note that the symmetry condition should also hold with the gradients w.r.t. s. They show that if the noise rate is $\eta_x \leq \eta_{\max} < 0.5, \forall x \in \mathcal{X}$ and if the loss is symmetric and non-negative, the minimizer of the noisy risk $f_\eta^* = \arg\inf_{f\in\mathcal{F}} R^\eta(f)$ approximately minimizes the clean risk:

$$R(f_\eta^*) \leq \epsilon/(1 - 2\eta_{\max}),$$

where $\epsilon = \inf_{f\in\mathcal{F}} R(f)$ is the optimal clean risk. This implies that the noisy risk under symmetric loss is a good surrogate of the clean risk. We further relax the non-negative constraint on the loss with a corollary[1]:

[1]This is important for our proposed RINCE loss that involves an exponential function $\ell(s, y) = -ye^s$, which can produce negative values.

Corollary 5.1. *Consider the setting of Ghosh et al. [64] and the exponential loss function* $\mathcal{L}(s, y) = -ye^s$. *Let* $f_\eta^* = \arg\inf_{f\in\mathcal{F}} R_\mathcal{L}^\eta(f)$ *be the minimizer of the noisy risk and* $\epsilon = \inf_{f\in\mathcal{F}} R_\mathcal{L}(f)$ *be the optimal risk. If* $\eta_x \leq \eta_{\max} < 0.5$ *for all* $x \in \mathcal{X}$. *If the prediction score is bounded by* $s_{\max}$, *we have* $R(f_\eta^*) \leq (\epsilon + 2\eta_{\max}e^{s_{\max}})/(1 - 2\eta_{\max})$.

Proof. Consider a binary classification loss with the following form:

$$\tilde{\mathcal{L}}_x(f(x), y) = B + \mathcal{L}_x(f(x), y) = B - y \cdot e^{f(x)} \geq 0,$$

where the prediction score $f(x)$ is bounded by $s_{\max} = \log(B)$. Note that the boundedness assumption holds for general representation learning on hypersphere, where the prediction score is the inner product between normalized feature vectors. Importantly, the loss satisfies

$$\tilde{\mathcal{L}}(f(x), 1) + \tilde{\mathcal{L}}(f(x), -1) = 2B.$$

By construction, the optimal risk takes the following value:

$$\inf_{f\in\mathcal{F}} R_{\tilde{\mathcal{L}}}(f) = \inf_{f\in\mathcal{F}} \mathbb{E}_{x\sim\mu}[\tilde{\mathcal{L}}(f(x), y_x)] = \epsilon + B := \tilde{\epsilon},$$

and $f^* = \arg\inf_{f\in\mathcal{F}} R_{\tilde{\mathcal{L}}}(f)$. Note that f^* is also a minimizer w.r.t. the original loss $\mathcal{L}$($f^* = \arg\inf_{f\in\mathcal{F}} R_\mathcal{L}(f)$), as an additive constant will not change the optimum solutions. Expanding the noisy risk gives

$$\begin{aligned}
R_{\tilde{\mathcal{L}}}^\eta(f) &= \mathbb{E}_{(x,y)\sim\mu}[(1 - \eta_x)\tilde{\mathcal{L}}(f(x), y_x) + \eta_x\tilde{\mathcal{L}}(f(x), -y_x)] \\
&= \mathbb{E}_{x\sim\mu}[(1 - \eta_x)\tilde{\mathcal{L}}(f(x), y_x) + \eta_x(2B - \tilde{\mathcal{L}}(f(x), y_x))] \qquad \text{(Symmetry)} \\
&= \mathbb{E}_{x\sim\mu}[(1 - 2\eta_x)\tilde{\mathcal{L}}(f(x), y_x)] + 2B\mathbb{E}_{x\sim\mu}[\eta_x].
\end{aligned}$$

Let $f_\eta^* = \arg\inf R_\mathcal{L}^\eta(f_\eta^*) = \arg\inf R_{\tilde{\mathcal{L}}}^\eta(f_\eta^*)$, we have

$$R_{\tilde{\mathcal{L}}}^\eta(f^*) - R_{\tilde{\mathcal{L}}}^\eta(f_\eta^*) = \mathbb{E}_{x\sim\mu}[(1 - 2\eta_x)(\tilde{\mathcal{L}}(f^*(x), y_x) - \tilde{\mathcal{L}}(f_\eta^*(x), y_x))] \geq 0$$

since f_η^* is the minimizer of $R_{\tilde{\mathcal{L}}}^\eta$, which implies that

$$E_{x\sim\mu}[(1-2\eta_x)\tilde{\mathcal{L}}(f_\eta^*(x),y_x)] \leq E_{x\sim\mu}[(1-2\eta_x)\tilde{\mathcal{L}}(f^*(x),y_x)] \leq \tilde{\epsilon},$$

since $0 < 1-2\eta_x \leq 1$ by assumption. Let $\eta_{max} = \sup_{x\in\mathcal{X}} \eta_x$, we have

$$(1-2\eta_{\max})E_{x\sim\mu}[\tilde{\mathcal{L}}(f_\eta^*(x),y_x)] \leq \epsilon,$$

since the loss is non-negative, which implies

$$R_{\tilde{\mathcal{L}}}(f_\eta^*) \leq \frac{\tilde{\epsilon}}{1-2\eta_{\max}}.$$

Finally, we recover the original exponential loss without the additive term B. Plugging the form we have

$$B + R_{\mathcal{L}}(f_\eta^*) \leq \frac{\epsilon + B}{1-2\eta_{max}},$$

which implies

$$R_{\mathcal{L}}(f_\eta^*) \leq \frac{\epsilon + B}{1-2\eta_{max}} - B = \frac{\epsilon + 2B\eta_{max}}{1-2\eta_{max}}.$$

For exponential loss, setting B to $e^{s_{\max}}$ completes the proof. □

For instance, when the noise level is 40%, we have $R_{\mathcal{L}}(f_\eta^*) \leq 5\epsilon + 4B$. Note that the prediction score is bounded by $1/t$ in our case as the representations are projected onto the unit hypersphere.

5.1.3 Towards Symmetric Contrastive Objectives

The results above suggest that we can achieve robustness against noisy views if a contrastive objective can be expressed in a form that satisfies the symmetry condition in the binary classification framework. To this end, we first relate contrastive learning to binary classification, and then express it in a form where symmetry can be achieved.

Contrastive learning as binary classification. Given two views X and V, we can interpret contrastive learning as noisy binary classification operating over pairs of samples (x, v) with a label 1 if it is sampled from the joint distribution, $(x, v) \sim P_{XV}$, and -1 if it comes from the product of marginals, $(x, v') \sim P_X P_V$. In the presence of noisy views, some negative pairs $(x, v') \sim P_X P_V$ could be mislabeled as positive, introducing noisy labels.

To see this more concretely, we again consider the InfoNCE loss:

$$\begin{aligned}\mathcal{L}_{\text{InfoNCE}}(\mathbf{s}) &= -\log \frac{e^{s^+}}{e^{s^+} + \sum_{i=1}^{K} e^{s_i^-}} \\ &:= -\log \frac{e^{f(x)^T g(v)/t}}{e^{f(x)^T g(v)/t} + \sum_{i=1}^{K} e^{f(x)^T g(v_i)/t}},\end{aligned} \tag{5.2}$$

where $\mathbf{s} = \{s^+, \{s_i^-\}_{i=1}^K\}$, s^+ and s_i^- are the scores of related (positive) and unrelated (negative) pairs and t is the temperature parameter introduced to avoid gradient saturation. Note that different from Part I, we use two different encoders, f and g for different views, which makes InfoNCE loss adaptable for multimodal scenarios such as image-text or video-audio contrastive learning.

The expectation of the loss is taken over $(x, v) \sim P_{XV}$ and K independent samples $v_i \sim P_V$, where P_{XV} denotes the joint distribution over pairs of views such as transformations of the same image or co-occurring multimodal signals. Although InfoNCE has a functional form of the $(K+1)$-way softmax cross entropy loss, the model ultimately learns to classify whether a pair (x, v) is positive or negative by maximizing/minimizing the positive score s^+/negative scores s_i^-. Therefore, InfoNCE under noisy views can be seen as binary classification with noisy labels. We acknowledge that similar interpretations have been made in prior works under different contexts [73, 187, 208].

Symmetric form of contrastive learning. Now we turn to a functional form of contrastive learning that can achieve the symmetric property. Assume that we have a noise-tolerant loss function ℓ that satisfies the symmetry condition of equation 5.1.

We say a contrastive learning objective is symmetric if it accepts the following form

$$\mathcal{L}(\mathbf{s}) = \underbrace{\ell(s^+, 1)}_{\text{Positive Pair}} + \lambda \underbrace{\sum_{i=1}^{K} \ell(s_i^-, -1)}_{K \text{ Negative Pairs}} \tag{5.3}$$

which consists of a collection of $(K+1)$ binary classification losses; $\lambda > 0$ is a density weighting term controlling the ratio between classes 1 (positive pairs) and -1 (negative pairs). Reducing λ places more weight on the positive score s^+, while setting λ to zero recovers the negative-pair-free contrastive loss such as BYOL [70].

Contrastive objectives that satisfy the symmetric form enjoy strong theoretical guarantees against noisy labels as described in Ghosh et al. [64], as long as we plug in the right contrastive loss function ℓ that satisfies the symmetry condition. Unfortunately, the InfoNCE loss [147] does not satisfy the symmetry condition in the gradients w.r.t. $s^{+/-}$. To see this, note that by taking the derivative with respect to the prediction score s, the definition is equivalent to $\frac{\partial \mathcal{L}(s,1)}{\partial s} + \frac{\partial \mathcal{L}(s,-1)}{\partial s} = 0 \ \forall s \in \mathbb{R}$. Applying the equation to the InfoNCE loss, we have

$$\begin{aligned}
\frac{\partial \mathcal{L}_{\text{InfoNCE}}(\mathbf{s})}{\partial s^+} &= \frac{-1}{\mathcal{L}_{\text{InfoNCE}}(\mathbf{s})} \cdot \frac{e^{s^+} \cdot \sum_{i=1}^{K} e^{s_i^-}}{(e^{s^+} + \sum_{i=1}^{K} e^{s_i^-})^2} \\
\frac{\partial \mathcal{L}_{\text{InfoNCE}}(\mathbf{s})}{\partial s_i^-} &= \frac{-1}{\mathcal{L}_{\text{InfoNCE}}(\mathbf{s})} \cdot \frac{e^{s^+}(1 - e^{s_i^-}) + \sum_{i=1}^{K} e^{s_i^-}}{(e^{s^+} + \sum_{i=1}^{K} e^{s_i^-})^2}.
\end{aligned}$$

Within a batch of data, the gradients with respect to s^+ and s^- are entangled and do not sum to a constant, which fail to meet the symmetry condition. This motivates us to develop a new contrastive loss function that satisfies the symmetry condition, described next.

5.2 Robust InfoNCE Loss (RINCE)

Based on the idea of robust symmetric classification loss, we present the following Robust InfoNCE (RINCE) loss:

$$\mathcal{L}_{\text{RINCE}}^{\lambda,q}(\mathbf{s}) = \frac{-e^{q \cdot s^+}}{q} + \frac{(\lambda \cdot (e^{s^+} + \sum_{i=1}^{K} e^{s_i^-}))^q}{q},$$

where $q, \lambda \in (0, 1]$. When $q = 1$, RINCE becomes a contrastive loss that fully satisfies the symmetry property in the form of equation 5.3 with $\ell(s, y) = -ye^s$:

$$\mathcal{L}_{\text{RINCE}}^{\lambda, q=1}(\mathbf{s}) = -(1-\lambda)e^{s^+} + \lambda \sum_{i=1}^{K} e^{s_i^-}.$$

Notice that the exponential loss $-ye^s$ satisfies the symmetric condition defined in equation 5.1 with $c = 0$. Therefore, when $q \to 1$, we achieve robustness against noisy views in the same manner as binary classification with noisy labels.

In the limit of $q \to 0$, RINCE becomes asymptotically equivalent to InfoNCE, as the following lemma describes:

Lemma 5.2. *For any $\lambda > 0$, it holds that*

$$\lim_{q \to 0} \mathcal{L}_{\text{RINCE}}^{\lambda, q}(\boldsymbol{s}) = \mathcal{L}_{\text{InfoNCE}}(\boldsymbol{s}) + \log(\lambda);$$
$$\lim_{q \to 0} \frac{\partial}{\partial \boldsymbol{s}} \mathcal{L}_{\text{RINCE}}^{\lambda, q}(\boldsymbol{s}) = \frac{\partial}{\partial \boldsymbol{s}} \mathcal{L}_{\text{InfoNCE}}(\boldsymbol{s}).$$

Proof. We first prove the convergence in the function space with the L'Hôpital's rule:

$$\begin{aligned}
\lim_{q \to 0} \mathcal{L}_{\text{RINCE}}^{\lambda, q}(\mathbf{s}) &= \lim_{q \to 0} \frac{-e^{q \cdot s^+}}{q} + \frac{\left(\lambda \cdot (e^{s^+} + \sum_{i=1}^{K} e^{s_i^-})\right)^q}{q} \\
&= \lim_{q \to 0} \frac{1 - e^{q \cdot s^+}}{q} + \frac{-1 + \left(\lambda \cdot (e^{s^+} + \sum_{i=1}^{K} e^{s_i^-})\right)^q}{q} \\
&= \lim_{q \to 0} \frac{1 - e^{q \cdot s^+}}{q} + \lim_{q \to 0} \frac{-1 + \left(\lambda \cdot (e^{s^+} + \sum_{i=1}^{K} e^{s_i^-})\right)^q}{q} \\
&= -\log(e^{s^+}) + \log\left(\lambda(e^{s^+} + \sum_{i=1}^{K} e^{s_i^-})\right) \qquad \text{(L'Hôpital's rule)} \\
&= -\log \frac{e^{s^+}}{\lambda\left(e^{s^+} + \sum_{i=1}^{K} e^{s_i^-}\right)} \\
&= \mathcal{L}_{\text{InfoNCE}}(\mathbf{s}) + \log(\lambda).
\end{aligned}$$

To prove the convergence in its derivative, we analyze the derivative with respect

to the positive score s^+ and the negative score s_i^-. We begin with RINCE:

$$\text{(positive score)} \qquad \lim_{q\to 0}\frac{\partial}{\partial s^+}\mathcal{L}^{\lambda,q}_{\text{RINCE}}(\mathbf{s}) = \lim_{q\to 0}\frac{\partial}{\partial s^+}\frac{-e^{q\cdot s^+}}{q} + \frac{\partial}{\partial s^+}\frac{\left(\lambda\cdot(e^{s^+}+\sum_{i=1}^K e^{s_i^-})\right)^q}{q}$$
$$= \lim_{q\to 0} -e^{q\cdot s^+} + (\lambda\cdot(e^{s^+}+\sum_{i=1}^K e^{s_i^-}))^{q-1}\cdot\lambda\cdot e^{s^+}$$
$$= -1 + \frac{e^{s^+}}{e^{s^+}+\sum_{i=1}^K e^{s_i^-}};$$

$$\text{(negative score)} \qquad \lim_{q\to 0}\frac{\partial}{\partial s_i^-}\mathcal{L}^{\lambda,q}_{\text{RINCE}}(\mathbf{s}) = \lim_{q\to 0}\frac{\partial}{\partial s_i^-}\frac{\left(\lambda\cdot(e^{s^+}+\sum_{i=1}^K e^{s_i^-})\right)^q}{q}$$
$$= \lim_{q\to 0}(\lambda\cdot(e^{s^+}+\sum_{i=1}^K e^{s_i^-}))^{q-1}\cdot\lambda\cdot e^{s_i^-}$$
$$= \frac{e^{s_i^-}}{e^{s^+}+\sum_{i=1}^K e^{s_i^-}}.$$

We can see that the derivatives match the ones of InfoNCE

$$\text{(positive score)} \qquad \frac{\partial}{\partial s^+}\mathcal{L}_{\text{InfoNCE}}(\mathbf{s}) = \frac{\partial}{\partial s^+} - \log\frac{e^{s^+}}{e^{s^+}+\sum_{i=1}^K e^{s_i^-}}$$
$$= -\frac{e^{s^+}}{e^{s^+}+\sum_{i=1}^K e^{s_i^-}}\cdot\frac{(e^{s^+}+\sum_{i=1}^K e^{s_i^-})\cdot e^{s^+} - e^{2s^+}}{(e^{s^+}+\sum_{i=1}^K e^{s_i^-})^2}$$
$$= -1 + \frac{e^{s^+}}{e^{s^+}+\sum_{i=1}^K e^{s_i^-}};$$

$$\text{(negative score)} \qquad \frac{\partial}{\partial s_i^-}\mathcal{L}_{\text{InfoNCE}}(\mathbf{s}) = -\frac{e^{s^+}+\sum_{i=1}^K e^{s_i^-}}{e^{s_i^-}}\cdot\frac{-e^{s^+}\cdot e^{s_i^-}}{(e^{s^+}+\sum_{i=1}^K e^{s_i^-})^2}$$
$$= \frac{e^{s_i^-}}{e^{s^+}+\sum_{i=1}^K e^{s_i^-}}.$$

□

Note that the convergence also holds for the derivatives: optimizing RINCE in the limit of $q \to 0$ is mathematically equivalent to optimizing InfoNCE. Therefore,

by controlling $q \in (0, 1]$ we smoothly interpolate between the InfoNCE loss ($q \to 0$) and the RINCE loss in its fully symmetric form ($q \to 1$).

5.2.1 Intuition behind RINCE

We now analyze the behavior of RINCE through the lens of exploration-exploitation trade-off. In particular, we reveal an implicit easy/hard positive mining scheme by inspecting the gradients of RINCE under different q values, and show that we achieve stronger robustness (more exploitation) with larger q at the cost of potentially useful clean hard positive samples (less exploration).

To simplify the analysis, we consider InfoNCE and RINCE with a single negative pair ($K = 1$):

$$\mathcal{L}_{\text{InfoNCE}}(\mathbf{s}) = -\log(e^{s^+}/(e^{s^+} + e^{s^-}));$$
$$\mathcal{L}^{\lambda,q}_{\text{RINCE}}(\mathbf{s}) = \frac{-e^{q \cdot s^+}}{q} + \frac{(\lambda \cdot (e^{s^+} + e^{s^-}))^q}{q}.$$

We visualize the loss and the scale of the gradients with respect to positive scores s^+ in Figure 5-2. Although the loss values are different for each q, they follow the same principle: The loss achieves its minimum when the positive score s^+ is maximized and the negative score s^- is minimized.

The interesting bit lies in the gradients. The InfoNCE loss ($q \to 0$) places more emphasis on *hard* positive pairs, i.e., the pairs with *low* positive scores s^+ (the left-most part in the plot). In contrast, the fully symmetric RINCE loss ($q = 1$) places more weights on *easy* positive pairs (the right-most part). This reveals an implicit trade-off between exploration (convergence) and exploitation (robustness). When $q \to 0$, the loss performs hard positive mining, providing faster convergence in the noise-free setting. But in the presence of noise, exploration is harmful; it wrongly puts higher weights to false positive pairs because noisy samples tend to induce larger losses [9, 76, 131, 217], and this could hinder convergence. In contrast, when $q \to 1$, we perform easy positive mining. This provides robustness especially against false positives; but this is done at the cost of exploration with clean hard positives. An important

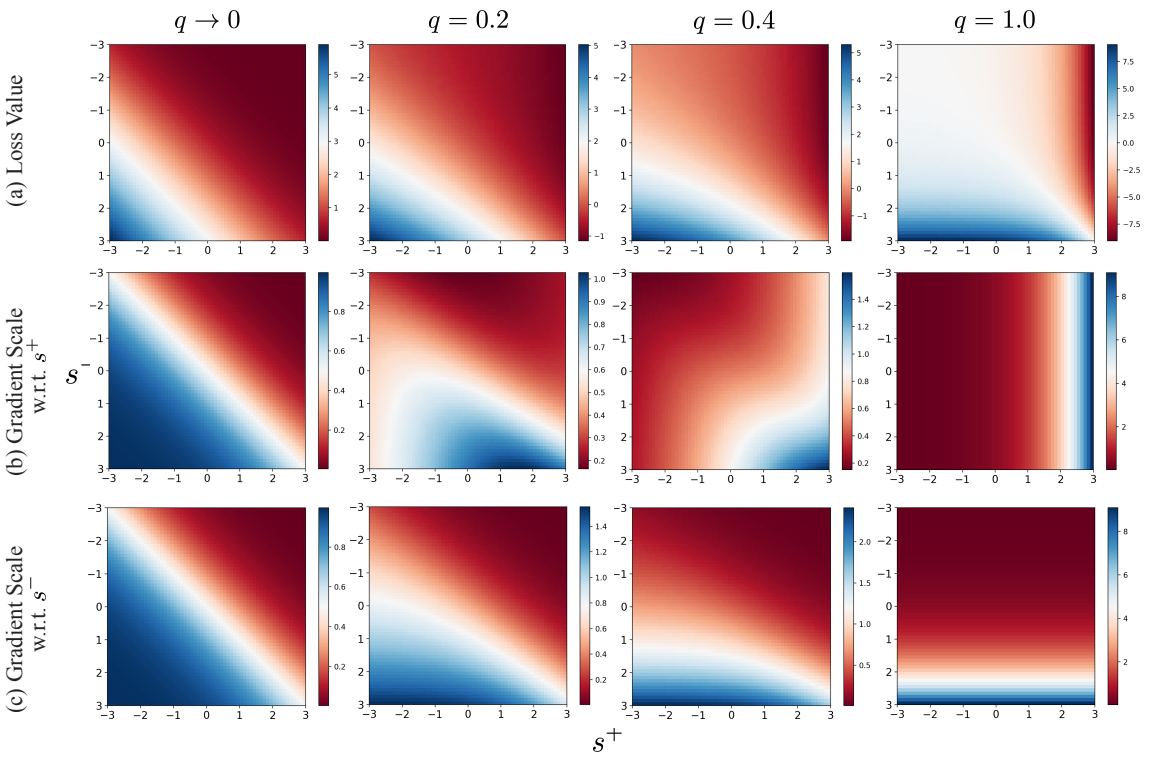


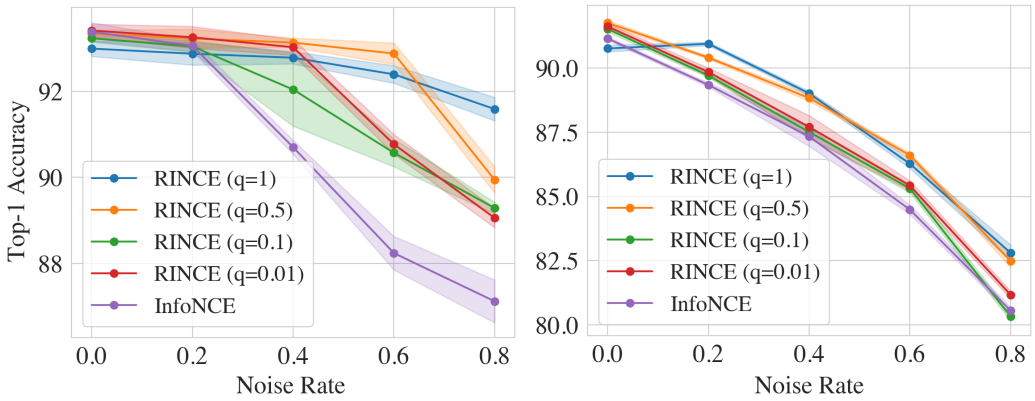
Figure 5-2: Loss Visualization. We visualize the (a) loss value and the (b) gradient scale with respect to the positive score s^+ and (c) negative score s^- for different q while setting $\lambda = 0.5$. The gradient scale of InfoNCE ($q \to 0$) is larger when the positive score is smaller (hard positive pair). In contrast, for fully symmetric RINCE ($q = 1$), the gradient is larger when positive score is large (easy positive pair).

aspect here is that RINCE does not require an explicit form of noise estimator: the scores s^+ and s^-, and the relationship between the two (which is what the loss function measures) act as noise estimates. In practice, we set $q \in [0.1, 0.5]$ to strike the balance between exploration and exploitation.

Note that both $q \to 0$ and $q \to 1$ naturally perform hard negative mining; both their derivatives put exponentially more weights on hard negative pairs.

5.3 Experiments

We evaluate RINCE on various contrastive learning scenarios involving images (CIFAR-10 [108], ImageNet [44]), videos (ACAV100M [114], Kinetics400 [98]) and graphs (TU-Dataset [133]). Empirically, we find that RINCE is insensitive to the choice of λ; we



(a) Label Noise (b) Augmentation Noise

Figure 5-3: Noisy CIFAR-10. We show the top-1 accuracy of RINCE with different values of q across different noise rate η. Large q ($q = 0.5, 1$) leads to better robustness, while smaller q ($q = 0.01$) performs similar to InfoNCE ($q \to 0$).

simply set $\lambda = 0.01$ for all vision experiments and $\lambda = 0.025$ for graph experiments.

5.3.1 Noisy CIFAR-10

We begin with controlled experiments on CIFAR-10 to verify the robustness of RINCE against synthetic noise by controlling the noise rate η. We consider two noise types:

Label noise. We start with the case of supervised contrastive learning [99] where positive pairs are different images of the same label. This allows us to control noise in the traditional sense, i.e., learning with noisy labels. Similar to [226], we flip the true labels to semantically related ones, e.g., CAT $\leftrightarrow$ DOG with probability $\eta/2$. This is commonly referred to as class-dependent noise [76, 201, 226].

Augmentation noise. We consider the self-supervised learning scenario and vary the crop size during data augmentation similar to [188], i.e., after applying all the transformations as in SimCLR [24], images are further cropped into 1/5 of their original size with probability η. This effectively controls the noise rate as cropped patches will most likely to be too small to contain any shared information.

Figure 5-3 shows the results of SimCLR trained with InfoNCE and RINCE with different choices of q and λ. When the augmentation noise is present, e.g., $\eta = 0.4$, the accuracy of InfoNCE drops from 91.14% to 87.33%. In contrast, the robustness

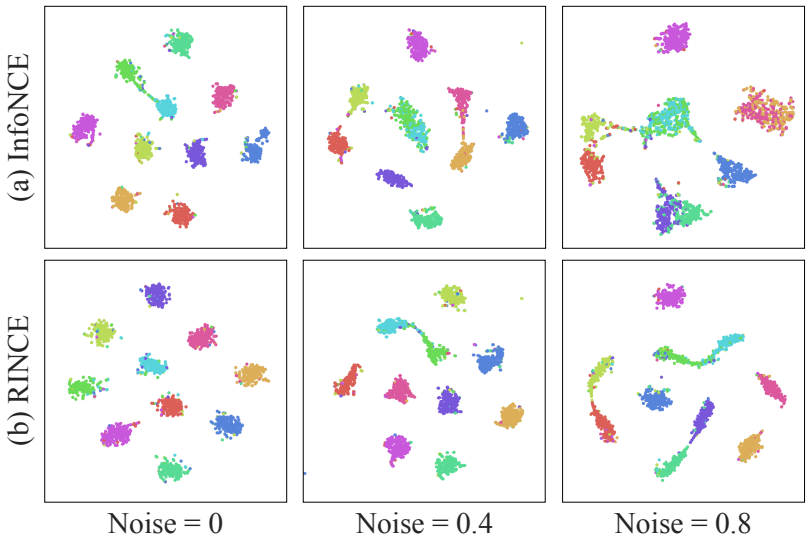


Figure 5-4: t-SNE Visualization on CIFAR-10 with label noise. Colors indicate classes. RINCE leads to better class-wise separation than the InfoNCE loss in both noise-less and noisy cases.

of RINCE is enhanced by increasing q, achieving 89.01% when $q = 1.0$. InfoNCE also fails to address label noise and suffers from significant performance drop (93.38% $\rightarrow$ 87.11% when $\eta = 0.8$). In comparison, RINCE retains the performance even when the noise rate is large (91.59% for $q = 1.0$). In both cases, reducing the value of q makes the performance of RINCE closer to InfoNCE, verifying our analysis in Lemma 5.2.

Figure 5-4 shows t-SNE visualization [192] of representations learned with InfoNCE and RINCE ($q = 1.0$) under different label noise. As the noise rate increases, representations of different classes start to tangle up for InfoNCE, while RINCE still achieves decent class-wise separation.

5.3.2 Image Contrastive Learning

We verify our approach on the well-established ImageNet benchmark [44]. We adopt the same training protocol and hyperparameter settings of SimCLR [24] and MoCo-v3 [28] and simply replace the InfoNCE with our RINCE loss ($q = 0.1$ and $q = 0.6$, respectively). Table 5.1 shows that RINCE improves InfoNCE (SimCLR and MoCo-v3) by a non-trivial margin. We also include results from the SOTA baselines, where

Method	Δ to SimCLR [24]	Top 1	Top 5
Supervised [79]	N/A	76.5	-
SimSiam [26]	No negative pairs	71.3	-
BYOL [70]	No negative pairs	74.3	91.6
Barlow Twins [220]	Redundancy reduction	73.2	91.0
SwAV [23]	Cluster discrimination	75.3	-
SimCLR [24]	None	69.3	89.0
+RINCE (Ours)	Symmetry controller q	**70.0**	**89.8**
MoCo [81]	Momentum encoder	60.6	-
MoCo-v2 [27]	Momentum encoder	71.1	90.1
MoCo-v3 [28]	Momentum encoder	73.8	-
+RINCE (Ours)	Symmetry controller q	**74.2**	**91.8**

Table 5.1: Linear Evaluation on ImageNet. All the methods use ResNet-50 [79] as backbone architecture with 24M parameters.

they improve SimCLR by introducing dynamic dictionary plus momentum encoder (MoCo-v1/v2/v3 [27, 28, 81]), removing negative pairs plus the stop-gradient trick (SimSiam [26], BYOL [70]), or online cluster assignment (SwAV [23]). In comparison, our work is orthogonal to the recent developments, and the existing tricks can be applied along with RINCE.

Figure 5-5 shows the positive pairs from SimCLR augmentations and the corresponding positive scores $s^+ = f(x)^T g(v)$ output by trained RINCE model. Examples with lower positive scores contain pairs that is less informative to each other, while semantically meaningful pairs often have higher scores. This implies that positive scores are good noise detectors, and down-weighting the samples with lower positive score brings robustness during training, verifying our analysis in section 5.2.1.

5.3.3 Video Contrastive Learning

We examine our approach in the audio-visual learning scenario using two video datasets: Kinetics400 [98] and ACAV100M [114]. Here, we find that simple *q-warmup* improves the stability of RINCE, i.e., q starts at 0.01 and linearly increases to 0.4 until the last epoch. We apply this to all RINCE models in this section. As we show

Method	Backbone	Finetune Input Size	HMDB	UCF
3D-RotNet [95]	R3D-18	16×112^2	33.7	62.9
ClipOrder [211]	R3D-18	16×112^2	30.9	72.4
DPC [77]	R3D-18	25×128^2	35.7	75.7
CBT [180]	S3D	16×112^2	44.6	79.5
AVTS [107]	MC3-18	25×224^2	56.9	85.8
SeLaVi [10]	R(2+1)D-18	32×112^2	47.1	83.1
XDC [5]	R(2+1)D-18	32×224^2	52.6	86.8
Robust-xID [131]	R(2+1)D-18	32×224^2	55.0	85.6
Cross-AVID [132]	R(2+1)D-18	32×224^2	59.9	86.9
AVID+CMA [132]	R(2+1)D-18	32×224^2	60.8	87.5
InfoNCE (Ours)	R(2+1)D-18	32×224^2	57.8	88.6
RINCE (Ours)	R(2+1)D-18	32×224^2	**61.6**	**88.8**
GDT [154]	R(2+1)D-18	30×112^2	62.3*	90.9*

**Based on an advanced hierarchical data augmentation during pretraining.*

Table 5.2: Kinetics400-pretrained performance on UCF101 and HMDB51 (top-1 accuracy). Ours use the same data augmentation approach as Cross-AVID and AVID+CMA, while GDT uses a hierarchical sampling process to obtain good performance.

below, RINCE outperforms SOTA noise-robust contrastive methods [131, 132] on Kinetics400, while also providing scalability and computational efficiency compared to InfoNCE.

Kinetics400 For fair comparison to SOTA, we follow the same experimental protocol and hyperparameter settings of [132] and simply replace their InfoNCE loss functions with RINCE. We use the same network architecture, i.e., 18-layer R(2+1)D video encoder [189], 9-layer VGG-like audio encoder, and 3-layer MLP projection head producing 128-dim embeddings. We use the ADAM optimizer [100] for 400 epochs with 4,096 batch size, 1e-4 learning rate and 1e-5 weight decay. The pretrained encoders are finetuned on UCF-101 [177] and HMDB-51 [109] with clips composed of 32 frames of size 224×224.

Table 5.2 shows RINCE outperforming most of the baselines, including Robust-xID [131] and AVID+CMA [132] which are recent InfoNCE-based SOTA methods proposed to address the noisy view issues in audio-visual contrastive learning. Con-

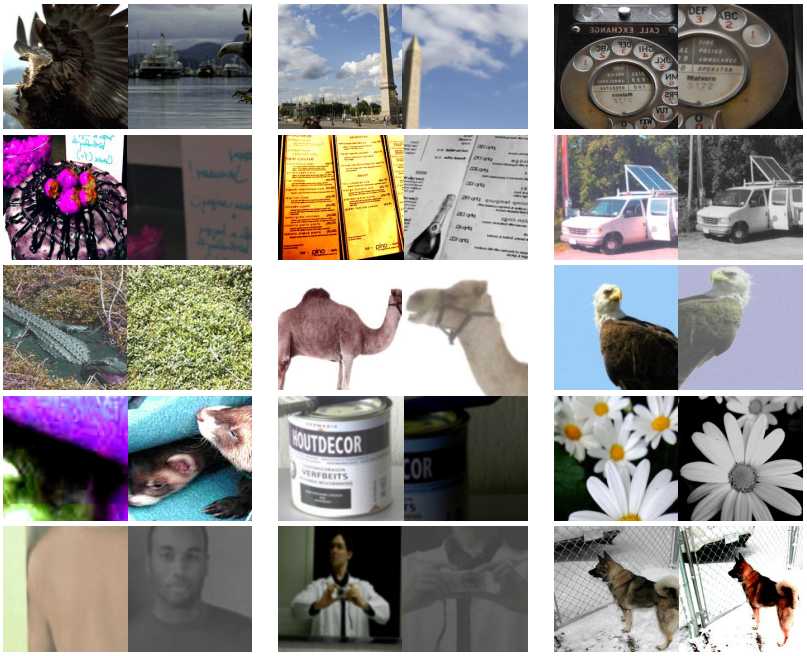


Figure 5-5: Positive pairs and their scores. The positive scores $s^+ \in [-1, 1]$ are output by the trained RINCE model (temperature = 1). Pairs that have lower scores are visually noisy, while informative pairs often have higher scores.

sidering the only change required is the simple replacement of the InfoNCE with our RINCE loss, the results clearly show the effectiveness of our approach. The simplicity means we can easily apply RINCE to a variety of InfoNCE-based approaches, such as GDT [154] that uses advanced data augmentation to achieve SOTA results.

ACAV100M We conduct an in-depth analysis of RINCE on ACAV100M [114], a recent large-scale video dataset for self-supervised learning. Compared to Kinetics400 which is limited to human actions, ACAV100M contains videos "in the wild" exhibiting a wide variety of audio-visual patterns. The unconstrained nature of the dataset makes it a good benchmark to investigate the robustness of RINCE to various types of real-world noise, e.g., background music, overdubbed audio, studio narrations, etc.

We focus on evaluating the (a) scalability and (b) convergence rate of RINCE, thereby answering the question: *Will it retrain its edge over InfoNCE (a) even in the large-scale regime and (b) with a longer training time?* We follow the same experimental setup as described above, but reduce the batch size to 512 and report

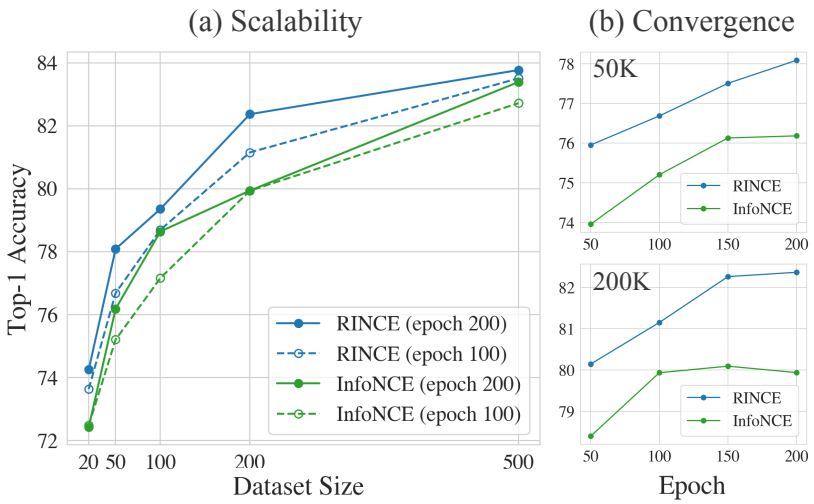


Figure 5-6: RINCE outperforms InfoNCE with fewer epochs across different scales The results are based on ACAV100M-pretrained models transferred to UCF-101.

the results only on the first split of UCF-101 to make our experiments tractable.

Figure 5-6 (a) shows the top-1 accuracy across different data scales and training epochs. RINCE outperforms InfoNCE by a large margin at every data scale. In terms of the convergence rate, RINCE is comparable to or even outperforms fully-trained (200 epochs) InfoNCE models with only 100 or fewer epochs. Figure 5-6 (b) gives a closer look at the convergence at 50K and 200K scales. Interestingly, InfoNCE saturates and even degenerates after epoch 150, while RINCE keeps improving. This verifies our analysis in section 5.2.1: InfoNCE can overfit noisy samples due to its exploration property, while RINCE downweights them and continue to obtain the learning signal from clean ones, achieving robustness against noise.

5.3.4 Graph Contrastive Learning

To see whether the modality-agnostic nature of RINCE applies beyond image and video data, we examine our approach on TUDataset [133], a popular benchmark suite for graph inference on molecules (BZR, NCI1), bioinformatics (PROTEINS), and social network (RDT-B, IMDB-B). Unlike vision datasets, data augmentation for graphs requires careful engineering with domain knowledge, limiting the applicability of InfoNCE-type contrastive objectives.

Methods	RDT-B	NCI1	PROTEINS	DD
node2vec [71]	-	54.9±1.6	57.5±3.6	-
sub2vec [2]	71.5±0.4	52.8±1.5	53.0±5.6	-
graph2vec [139]	75.8±1.0	73.2±1.8	73.3±2.1	-
InfoGraph [195]	82.5±1.4	76.2±1.1	74.4±0.3	72.9±1.8
GraphCL [215]	89.5±0.8	77.9±0.4	74.4±0.5	78.6±0.4
JOAO [216]	85.3±1.4	78.1±0.5	74.6±0.4	77.3±0.5
JOAOv2 [216]	86.4±1.5	78.4±0.5	74.1±1.1	77.4±1.2
InfoNCE* (Ours)	89.9±0.4	78.2±0.8	74.4±0.5	78.6±0.8
RINCE (Ours)	**90.9±0.6**	**78.6±0.4**	**74.7±0.8**	**78.7±0.4**

** GraphCL [215] but uses the same data augmentation as RINCE.*

Table 5.3: Self-supervised representation learning on TUDataset: The baseline results are excerpted from the published papers.

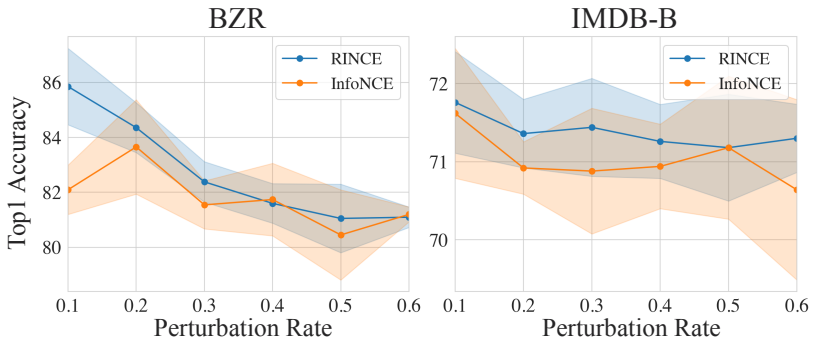


Figure 5-7: Performance v.s. Perturbation Rate: We increase the perturbation rate of node dropping, edge perturbation, and attribute masking from 10% to 60%. RINCE outperforms InfoNCE in terms of accuracy and variance when perturbation enhances.

For fair comparison, we follow the protocol of [215] and train graph isomorphism networks [213] with four types of data augmentation: node dropout, edge perturbation, attribute masking, and subgraph sampling. We train models using ADAM [100] for 20 epochs with a learning rate 0.01 and report mean and standard deviation over 5 independent trials. We set $q = 0.1$ for all the experiments in this section.

Table 5.3 shows that RINCE outperforms SOTA InfoNCE-based contrastive methods, GraphCL and JOAO /JOAOv2, setting the new records on all four datasets. GraphCL applies different augmentations for different datasets, while JOAO/JOAOv2 require solving bi-level optimization to choose optimal augmentation per dataset. In contrast, we apply the same augmentation across all four datasets and achieve competitive performance, demonstrating its generality and robustness. In Figure 5-7, we

control perturbation rate by applying three augmentation types (node dropout, edge perturbation, attribute masking) to different % of nodes/edges. We show results on two datasets most sensitive to augmentation. Again, RINCE consistently outperform InfoNCE and has relatively smaller variances when the noise rate increases.

Chapter 6

InfoNCE and Mutual Information

In this chapter, we will provide a theoretical analysis of the proposed RINCE objective and show that it extends the analyses by Ghosh et al. [64] to the self-supervised contrastive learning regime. Furthermore, we relate the proposed loss function to dependency measurement. Analogous to InfoNCE loss, which is a lower bound of mutual information between two views [147], we show that RINCE is a lower bound of Wassersein Dependency Measure (WDM) [149], even in the noisy setting. By replacing the KL divergence in the mutual information estimator with the Wasserstein distance, WDM is able to capture the geometry of the representation space and provides robustness against noisy views better than the KL divergence, both in theory and practice. In particular, the features learned with RINCE achieve better class-wise separation, which is proved to be crucial to improve generalization [38].

6.1 Variational Lower Bounds of InfoNCE

Without loss of generality, let $f = g$ and consider $f = f' \circ \phi$, where ϕ is a representation encoder and f' is a projection head [24]. Also, let $P^\phi = \phi_\# P$ be the pushforward measure of P with respect to ϕ. It has been shown that InfoNCE is a variational lower-

bound of MI in the representation space expressed with KL-divergence [157, 188]:

$$-\mathbb{E}\left[\mathcal{L}_{\text{InfoNCE}}(\mathbf{s})\right] + \log(K) \leq I(\phi(X), \phi(V)) \\ = D_{\text{KL}}(P^{\phi}_{XV}, P^{\phi}_{X} P^{\phi}_{V}).$$

Intuitively, maximizing MI can be interpreted as maximizing the discrepancy between positive and negative pairs. However, prior works [124, 149] have identified theoretical limitations of maximizing MI using the KL divergence: Because KL divergence is not a metric, it is sensitive to small differences in data samples regardless of the geometry of the underlying data distributions. Therefore, the encoder ϕ can capture limited information shared between X and V as long as the differences are sufficient to maximize the KL divergence. Note that this can be especially detrimental in the presence of noisy views, as the learner can quickly settle on spurious correlations in false positive pairs due to the absence of the actual shared information.

6.2 Variational Lower Bounds of RINCE

6.2.1 Wasserstein Mutual Information

We now establish RINCE as a lower bound of WDM [149], which is proposed as a replacement for the KL divergence in MI estimation.

WDM is based on the Wasserstein distance, a distance metric between probability distributions defined via an optimal transport cost. Letting μ and $\nu \in \text{Prob}(\mathbb{R}^d \times \mathbb{R}^d)$ be two probability measures, we define the Wasserstein-1 distance with a Euclidean cost function as

$$\mathcal{W}(\mu, \nu) = \inf_{\pi \in \Pi(\mu, \nu)} \mathbb{E}_{\substack{(X,V) \\ (X',V') \sim \pi}} \left[\left\| X - X' \right\| + \left\| V - V' \right\| \right]$$

where $\Pi(\mu, \nu)$ denotes the set of measure couplings whose marginals are μ and ν, respectively. By replacing the KL divergence in MI with Wasserstein distance, we obtain the following definition of Wasserstein dependency measure, or sometimes

referred as Wasserstein mutual information:

$$I_{\mathcal{W}}(\phi(X), \phi(V)) := \mathcal{W}_1(P^{\phi}_{XV}, P^{\phi}_X P^{\phi}_V)$$

By virtue of symmetry when $q = 1$, if $\lambda > 1/(K+1)$, the Kantorovich-Rubinstein duality [84] implies the following result:

Theorem 6.1. *If $\lambda K > 1-\lambda$ and f projects the representation to a unit hypersphere, we have*

$$-\mathbb{E}\left[\mathcal{L}^{\lambda,q=1}_{\text{RINCE}}(\boldsymbol{s})\right] \leq \frac{\text{Lip}(f)\cdot(1-\lambda)\cdot e^{1/t}}{t} I_{\mathcal{W}}(\phi(X), \phi(V)). \tag{6.1}$$

Proof. By the additivity of expectation, we can bound the negative symmetric loss as follows

$$\begin{aligned}
&-\mathbb{E}\left[\mathcal{L}^{\lambda,q=1}_{\text{RINCE}}(\mathbf{s})\right] \\
&= \mathbb{E}_{\substack{x\sim P_X \\ v\sim P_{V|X=x} \\ v_i\sim P_V}}\left[(1-\lambda)e^{f(\phi(x))^T f(\phi(v)/t} - \lambda\sum_{i=1}^{K} e^{f(\phi(x))^T f(\phi(v_i)/t}\right] \\
&= \mathbb{E}_{(x,v)\sim P_{XV}}\left[(1-\lambda)e^{f(\phi(x))^T f(\phi(v)/t}\right] - \mathbb{E}_{\substack{x\sim P_X \\ v_i\sim P_V}}\left[\lambda\sum_{i=1}^{K} e^{f(\phi(x))^T f(\phi(v_i)/t}\right] \\
&= \mathbb{E}_{(x,v)\sim P_{XV}}\left[(1-\lambda)e^{f(\phi(x))^T f(\phi(v)/t}\right] - \mathbb{E}_{\substack{x\sim P_X \\ v_i\sim P_V}}\left[\lambda\sum_{i=1}^{K} e^{f(\phi(x))^T f(\phi(v_i)/t}\right] \\
&= \mathbb{E}_{(x,v)\sim P_{XV}}\left[(1-\lambda)e^{f(\phi(x))^T f(\phi(v)/t}\right] - \lambda K\mathbb{E}_{\substack{x\sim P_X \\ v\sim P_V}}\left[e^{f(\phi(x))^T f(\phi(v)/t}\right] \\
&\leq (1-\lambda)\cdot(\mathbb{E}_{(x,v)\sim P_{XV}}\left[e^{f(\phi(x))^T f(\phi(v)/t}\right] - \mathbb{E}_{\substack{x\sim P_X \\ v\sim P_V}}\left[e^{f(\phi(x))^T f(\phi(v)/t}\right]),
\end{aligned}$$

where the last equality follows by $\lambda K > 1-\lambda$. Note that for $\frac{-1}{t} \leq s \leq \frac{1}{t}$, which

implies $|\nabla_s e^s| \leq e^{1/t}$. Therefore, by the mean value theorem, we have

$$
\begin{aligned}
&|e^{f(\phi(x))^T f(\phi(v))/t} - e^{f(\phi(x'))^T f(\phi(v'))/t}| \\
&\leq \frac{e^{1/t}}{t} |\langle f(\phi(x)), f(\phi(v))\rangle - \langle f(\phi(x')), f(\phi(v'))\rangle| && \text{(Mean Value Theorem)} \\
&= \frac{e^{1/t}}{t} |\langle f(\phi(x)) - f(\phi(x')), f(\phi(v))\rangle + \langle f(\phi(x')), f(\phi(v) - f(\phi(v'))\rangle| \\
&\leq \frac{e^{1/t}}{t} (|\langle f(\phi(x)) - f(\phi(x')), f(\phi(v))\rangle| + |\langle f(\phi(x')), f(\phi(v) - f(\phi(v'))\rangle|) \\
&\leq \frac{e^{1/t}}{t} (\|f(\phi(x)) - f(\phi(x'))\| \|f(\phi(v))\| + \|f(\phi(v) - f(\phi(v'))\| \|f(\phi(x'))\|) \\
& && \text{(Cauchy–Schwarz Ineq.)} \\
&= \frac{e^{1/t}}{t} (\|f(\phi(x)) - f(\phi(x'))\| + \|f(\phi(v) - f(\phi(v'))\|) && (f(\phi(x)) \text{ is unit norm}) \\
&\leq \frac{\mathrm{Lip}(f) \cdot e^{1/t}}{t} (\|\phi(x) - \phi(x')\| + \|\phi(v) - \phi(v')\|) \\
&= \frac{\mathrm{Lip}(f) \cdot e^{1/t}}{t} d((\phi(x), \phi(v)), (\phi(x'), \phi(v'))).
\end{aligned}
$$

We can see that the Lipschitz constant of $\exp(f(\cdot, \cdot))$ with respect to the metric d is bounded by $\frac{\mathrm{Lip}(f) \cdot e^{1/t}}{t}$. Therefore, by Kantorovich-Rubinstein duality, we have

$$
\begin{aligned}
&- \mathbb{E}\left[\mathcal{L}_{\mathrm{RINCE}}^{\lambda, q=1}(\mathbf{s})\right] \\
&\leq (1 - \lambda) \cdot (\mathbb{E}_{(x,v) \sim P_{XV}}[e^{f(\phi(x))^T f(\phi(v)/t}] \\
&\qquad\qquad - \mathbb{E}_{\substack{x \sim P_X \\ v \sim P_V}}[e^{f(\phi(x))^T f(\phi(v)/t}]), \\
&\leq \frac{\mathrm{Lip}(f) \cdot (1 - \lambda) \cdot e^{1/t}}{t} \mathcal{W}_1(\phi_\# P_{XV}, \phi_\# P_X \cdot \phi_\# P_V)
\end{aligned}
$$

□

The $I_{\mathcal{W}}(\phi(X), \phi(V))$ is the WDM defined in [149] and L is a constant that depends on t, λ, and the Lispchitz constant of the projection head f. Note that we are not aware of any work that showed it is possible to establish a similar bound with WDM for the InfoNCE loss.

This provides another explanation of what makes RINCE robust against noisy

views. Unlike InfoNCE which maximizes the KL divergence, optimizing RINCE is equivalent to maximizing the WDM with a Lipschitz function. Equipped with a proper metric, this allows RINCE to measure the divergence between two distributions P^{ϕ}_{XV} and $P^{\phi}_{X}P^{\phi}_{V}$ without being overly sensitive to individual sample noise, as long as the noise does not alter the geometry of the distributions. This also allows the encoder ϕ to learn more complete representations, as maximizing the Wasserstein distance requires the encoder to not only model the density ratio between the two distributions but also the optimal cost of transporting one distribution to another.

6.2.2 Lower Bounds with Noise

Finally, we show that RINCE still maximizes the noise-less WDM under additive noise, corroborating the robustness of RINCE. Let's consider a simple mixture noise model:

$$P^{\eta}_{XV} = (1-\eta)P_{XV} + \eta P_X P_V,$$

where η is the noise rate and the noisy joint distribution P^{η}_{XV} is a weighted sum between the noise-less positive distribution P_{XV} and negative distribution $P_X P_V$. Note that the marginals of P^{η}_{XV} are still P_X and P_V by construction. The intuition behind mixture noise model is that when we draw positive pairs from P^{η}_{XV}, we obtain false positives from $P_X P_V$ with probability η. Via the symmetry of the contrastive loss, we can extend bound (6.1) as follows.

Theorem 6.2. *If $\lambda \geq \frac{\eta K - \eta + 1}{\eta K - \eta + 1 + K}$ and f projects the representation to a unit hypersphere, we have*

$$-\mathbb{E}\left[\mathcal{L}^{\lambda, q=1}_{\text{RINCE}}(\boldsymbol{s})\right] \leq (1-\eta) \cdot \frac{\text{Lip}(f) \cdot (1-\lambda) \cdot e^{1/t}}{t} I_{\mathcal{W}}(\phi(X), \phi(V)).$$

Proof. The result is a simple combination of Corollary 5.1 and Theorem 6.1. If $\lambda \geq \frac{\eta K - \eta + 1}{\eta K - \eta + 1 + K}$, by the assumption of additive noisy models and the symmetry of loss,

we have

$$
\begin{aligned}
&-\mathbb{E}\left[\mathcal{L}^{\lambda,q=1}_{\text{RINCE}}(\mathbf{s})\right] \\
&= \mathbb{E}_{\substack{(x,v)\sim P^{\eta}_{XV} \\ v_i\sim P_V}}[(1-\lambda)e^{f(\phi(x))^T f(\phi(v)/t} - \lambda\sum_{i=1}^{K} e^{f(\phi(x))^T f(\phi(v_i)/t}] \\
&= \mathbb{E}_{(x,v)\sim P^{\eta}_{XV}}\left[(1-\lambda)e^{f(\phi(x))^T f(\phi(v)/t}\right] - \mathbb{E}_{\substack{x\sim P_X \\ v_i\sim P_V}}\left[\lambda\sum_{i=1}^{K} e^{f(\phi(x))^T f(\phi(v_i)/t}\right] \\
&= (1-\lambda)(1-\eta)\mathbb{E}_{(x,v)\sim P_{XV}}\left[e^{f(\phi(x))^T f(\phi(v)/t}\right] - K\cdot(\lambda-\eta+\eta\lambda)\mathbb{E}_{\substack{x\sim P_X \\ v\sim P_V}}\left[e^{f(\phi(x))^T f(\phi(v)/t}\right] \\
&\hspace{30em}\text{(symmetry)} \\
&\leq (1-\lambda)(1-\eta)\cdot(\mathbb{E}_{(x,v)\sim P_{XV}}\left[e^{f(\phi(x))^T f(\phi(v)/t}\right] - \mathbb{E}_{\substack{x\sim P_X \\ v_i\sim P_V}}\left[e^{f(\phi(x))^T f(\phi(v_i)/t}\right]) \\
&\hspace{30em}(\lambda \geq \tfrac{\eta K-\eta+1}{\eta K-\eta+1+K}) \\
&\leq (1-\eta)\cdot\frac{\text{Lip}(f)\cdot(1-\lambda)\cdot e^{1/t}}{t}\mathcal{W}_1(\phi_{\#}P_{XV}, \phi_{\#}P_X\cdot\phi_{\#}P_V)
\end{aligned}
$$

□

Comparing to the bound (6.1), the right hand side is rewieghted with $(1-\eta)$. This implies that minimizing RINCE with noisy views still maximizes a lower bound of noise-less WDM. Despite the simplicity of the analysis, it intuitively relates dependency measures and the noisy views with interpretable bounds. It would be an interesting future direction to extend the analysis to more complicated noise models, e.g., $P^{\eta}_{XV} = (1-\eta)P_{XV} + \eta Q_{XV}$, where Q is an unknown perturbation on positive distribution.

6.3 Conclusion

In the second part, We presented RINCE as a simple drop-in replacement for the InfoNCE loss in contrastive learning. Despite its simplicity, it comes with strong theoretical justifications and guarantees against noisy views. Empirically, we provided extensive results across image, video, and graph contrastive learning scenarios demonstrating its robustness against a variety of realistic noise patterns.

Part III

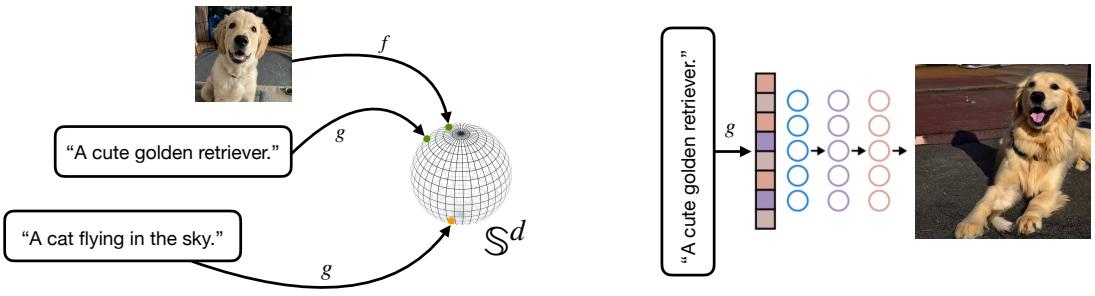
Learning from Biased Data

Chapter 7

Debiasing Foundation Models

Foundation Vision-Language (VL) models, such as CLIP [159], DALLE-2 [161], Imagen [166], and Stable Diffusion [164], which are trained on extensive multimodal data at a massive scale, have led to a significant shift in the landscape of machine learning systems. Specifically, contrastive vision-language encoders like CLIP have the ability to perform zero-shot inferences without fine-tuning, and language embeddings can be used to train high-quality text-to-image models [164]. Figure 7-1 provides an illustration.

While vision-language models demonstrate impressive capabilities, it is important to recognize that they may also exacerbate biases [3, 126, 197]. Recent studies [17] have shown that the datasets these models are trained on can contain inappropriate image-text pairs with stereotypes, racist content, and ethnic slurs. The biases are then propagated to downstream applications [3, 197], resulting in biased predictions. Figure 7-2 provides an example of biases within Stable Diffusion [164]. In addition to social biases, zero-shot models derived from vision-language models can also suffer from more general forms of spurious correlations such as image background, leading to poor group robustness [224]. Biases also exist in generative models, where generated images may exhibit bias towards certain genders and races [29, 129]. Substantial progress has been made recently toward mitigating biases in vision-language models [16, 152, 224]. However, many current approaches for addressing bias in models require training or fine-tuning the models using resampled datasets or modified



(a) Discriminative VL Model (b) Generative VL Model

Figure 7-1: Discriminative and Generative Vision-language Models. Discriminative VL models typically leverage multimodal contrastive learning to map image and language to the same representation space. Generative VL models learn to generate images conditioned on a text embedding.

objectives, which can be computationally intensive for foundation models.

In this chapter, we propose a general approach for self-debiasing foundation vision-language models by projecting out biased directions in the text embedding. Given a vision-language encoder such as CLIP, we define a set of biased directions in the embedding using prompts that describe the biases. For instance, prompts like "a photo of a male/female" define a biased subspace in the latent space. One approach to mitigating these biases is to construct a projection matrix, a linear transformation of the text embedding that projects out the biased directions [20]. However, solely relying on prompts to define biased directions may be unstable and noisy [68]. To address this issue, we propose a calibration loss that minimizes the discrepancy of a pair of prompt embeddings. For example, given a projection matrix that removes gender information, the projected vectors of prompts "a photo of a male doctor" and "a photo of a female doctor" should be similar. Based on this principle, we design an objective to calibrate the projection matrix, which has an easily solvable closed-form solution. This allows for the construction of the projection matrix to be *training-free and requires no downstream dataset or labels*, making it suitable for large-scale models. Empirically, we find that debiasing only the text embedding with a calibrated projection matrix suffices to improve the group robustness of zero-shot models on well-established benchmarks.

We then extend our approach to generative models such as Stable Diffusion [164],

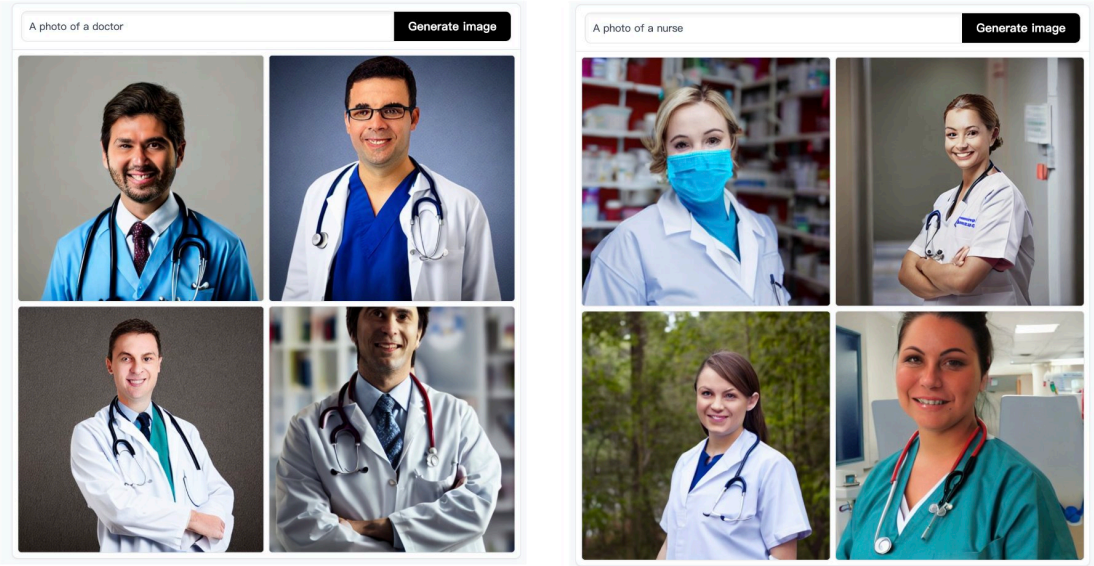


Figure 7-2: Biases of Diffusion Models. We generate images with prompt "a photo of a doctor" and "a photo of a nurse" using online Stable Diffusion model [164]. The current model has clear gender biases.

a widely adopted text-to-image model conditioned on text embeddings from CLIP [159]. The inherent challenge lies in the fact that generative models are distinctly dissimilar from zero-shot classification, where the target classes are explicitly defined. With generative models, our objective is to develop a debiasing matrix that is universally applicable to every prompt. This matrix can then be employed as a standard preprocessing stage prior to feeding the text embedding into the generative model. To accomplish this, we solve the calibration matrix with a set of positive pairs which comprise various prompts from the training dataset, and debias the unseen prompts with the obtained matrix. Similar to debiasing zero-shot models, the projection matrix improves the diversity of generated images from text-to-image models without altering the model parameters.

7.1 Biases of Foundation Models

Vision-Language models [159, 161, 164, 166] have become increasingly widespread in recent years. However, these models are known to suffer from spurious correlations and can be biased towards certain races and gender. Birhane et al. [17] study the

datasets these models are trained on and show that their biases can be inherited by the models. Various methods have been proposed to address biases, but many of them only address single-modality models.

Biases in Language Models Large-scale language models have been shown to contain harmful or misrepresentative biases [18, 135, 205]. Previous research has demonstrated the presence of gender bias in natural language processing systems [20, 228] as well as racial bias [62, 122]. Bolukbasi et al. [20] first proposed the use of orthogonal projections to remove gender biases in word embeddings. This approach was later extended to debiasing sentence embeddings [116]. Alternative methods include regularizing the models with constraints on training data [86, 227] or directly modifying the dataset [182, 228]. However, scaling these approaches to large foundation models can be challenging as they often require retraining the backbone encoders.

Biases in Vision Models Gender and racial biases have also been widely explored in computer vision [4, 198], in terms of dicsriminative models [199] and generative models [29, 72, 212]. Many debiasing approaches aim to learn good representations via adversarial training [121, 202], or augmenting the biased dataset [35, 160]. Beyond social bias, many works study spurious correlations, a more general form of bias that can include features such as image background or other non-target attributes that are correlated with labels. This problem of spurious correlations is often studied and tackled as a group robustness problem [87, 165]. Kirichenko et al. [102] show that last layer re-training is sufficient for robustness to spurious correlations, which aligns with our finding that debiasing the zero-shot weights suffices to yield robust classifiers.

Biases in Vision-Language Models Recently, biases in multimodal settings have gained significant attention [3, 75]. Wang et al. [197] propose to remove dimensions in the CLIP embedding that are highly correlated with gender attributes. Berg et al. [16] debias the CLIP models with prompt learning via an adversarial approach. Seth et al. [172] learn additive residual image representations to offset the biased

representations. Recently, Zhang and Ré [224] address the group robustness of vision-language models with contrastive learning. These previous works are data-oriented, where models are trained or finetuned on labeled datasets. In contrast, our approach is fully zero-shot, which does not require any downstream dataset and model training. To debias generative models, a recent work [59] pre-defines a look-up table to provide fair guidance for text-to-image diffusion models. Nevertheless, this method encounters limitations when faced with previously unseen classes that are absent from the look-up table, while our approach generalizes well to new concepts.

7.2 Biases and Spurious Correlations

We consider a dataset in which each input $x \in \mathcal{X}$ is associated with multiple attributes, including the target class $y \in \mathcal{Y}$ and a spurious attribute $a \in \mathcal{A}$. We focus on the case where biases are present and the attribute a is spuriously correlated with the label y. For instance, the class "doctor" could be correlated with the spurious attribute "gender" in the datasets foundation models are trained on [17]. Importantly, these biases can be transferred to downstream tasks, both discriminative and generative.

Discriminative Models In this work, we examine the biases present in zero-shot classifiers obtained via a vision-language encoder such as CLIP. These classifiers are built by assigning each row of the linear classifier weight $\beta \in \mathbb{R}^{K \times d}$ to be the embedding of a "class prompt", for example, "a photo of a [class name]" [159]. Importantly, it does not require any data or training to construct these zero-shot classifiers. However, it is possible for these zero-shot classifiers to inherit biases from the dataset used to train the vision-language models. To study these biases, we utilize the group robustness framework proposed by Sagawa et al. [165]. In this setting, groups are defined by a combination of the labels and spurious attributes: $\mathcal{G} \in \mathcal{Y} \times \mathcal{A}$. Given a distribution P_g conditioned on $g \in \mathcal{G}$ and a loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, group robustness requires that the classifier $f : \mathcal{X} \to \mathcal{Y}$ achieves a small gap between its

worst-group error and average error:

$$\max_{g \in \mathcal{G}} \mathbb{E}_{x,y \sim P_g} \left[\ell(f(x), y) \right] - \mathbb{E}_{x,y \sim P} \left[\ell(f(x), y) \right]. \tag{7.1}$$

The definition of metrics for text-image retrievals, such as maximal skewness [63], will be deferred to the experiment section.

Generative Models A text-to-image model learns a conditional distribution $\hat{P}(X|Z = z)$, where z is the embedding of the prompt. However, the biased nature of the dataset used to train the generative model can affect the distribution $\hat{P}$. To measure the bias present in generative models, recent works [30, 185] propose using statistical parity. Specifically, given a classifier $h : \mathcal{X} \to \mathcal{A}$ for the spurious attribute, the discrepancy of the generative distribution $\hat{P}$ is defined as the L2 norm between empirical and uniform distributions [30]:

$$\sqrt{\sum_{a \in \mathcal{A}} \left(\mathbb{E}_{x \sim \hat{P}} \left[\mathbb{1}_{h(x)=a} \right] - 1/|\mathcal{A}| \right)^2} \tag{7.2}$$

In practice, the expectation is estimated with empirical samples. A fair generative model minimizes the discrepancy by ensuring that each attribute $a \in \mathcal{A}$ has an equal probability (uniformly distributed).

7.3 Debiasing Discriminative Models

It is essential for a robust classifier to evade dependence on irrelevant features present in images. This necessitates the classifier to be invariant to image backgrounds and insensitive to attributes such as race or gender. Prior research has employed datasets with target labels and spurious attributes to quantify and eliminate biases [165, 224]. However, this approach is not feasible in a zero-shot setting, where data and training are prohibitive.

	CelebA		Waterbird	
	Male	Female	Land	Water
$\beta_{y=0}$	0.83	0.78	0.75	0.66
$\beta_{y=1}$	0.77	0.85	0.65	0.70

Table 7.1: Cosine similarity between classifier weights and spurious directions. In both datasets, the classifier weights are biased toward certain spurious attributes.

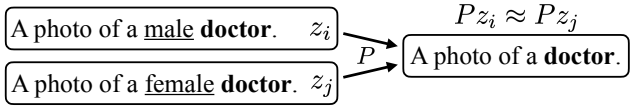


Figure 7-3: Calibration with Positive Pairs. Upon projecting out irrelevant features (such as gender), the embeddings of group prompts should exhibit similarity and contain only information pertaining to the target class (e.g. doctor).

7.3.1 Measuring Biases with Prompts

In contrast to previous approaches, our proposed method for measuring biases utilizes prompts, drawing inspiration from studies on debiasing word embeddings [20]. The use of vision-language contrastive training allows for the description of irrelevant features through natural language. As such, embeddings of prompts such as "a photo of a [irrelevant attribute]" can capture these spurious features in the visual embedding. Consequently, the bias of a classifier can be quantified by computing the cosine similarity between its weights and the corresponding spurious feature. Table 7.1 illustrates the cosine similarity between the embeddings of prompts that describe the target classes and irrelevant attributes, using two popular group robustness benchmarks: Waterbird [165] and CelebA [118]. The details of datasets and the specific prompts can be found in section 7.4 and appendix A.4. The results demonstrate that the classifier weights are inclined towards certain irrelevant attributes (gender or image background), implicitly implying that the classifiers are using these spurious directions to make predictions.

7.3.2 Debiasing via Orthogonal Projection

As the zero-shot weights can also be viewed as natural language embeddings, a straightforward approach is to follow the debiasing pipeline employed in word and sentence embeddings [20, 116]. In particular, to make the classifier invariant to these irrelevant features, we align the classifier weights with the orthogonal complement of

these embeddings. Let $A \in \mathbb{R}^{d \times m}$ be a matrix whose columns are the embeddings of spurious prompts. The orthogonal projection matrix is then:

$$P_0 = I - A(A^T A)^{-1} A^T.$$

We can use the projection matrix to eliminate spurious directions in a text embedding z as $P_0 z$.

7.3.3 Calibrating the Projection Matrix

It is essential to acknowledge that the estimation of the irrelevant feature directions may introduce an approximation error in the projection matrix [68]. Additionally, in certain scenarios, it may be challenging to thoroughly describe the irrelevant attribute using a limited number of prompts, resulting in increased uncertainty in the projection matrix estimation. This issue is also evident in our empirical results (Table 7.2 and 7.4), where the use of orthogonal projection fails to enhance performance.

To improve the estimation of the projection matrix, we leverage *positive pairs* of prompts that are expected to have the same semantic meaning after projection. In particular, the embedding of prompts such as "a photo of a [class name] with [spurious attribute]" should only contain information about "[class name]" after projecting out the spurious information, as Figure 7-3 illustrates. Motivated by this intuition, we propose to regularize the difference between the projected embeddings using a set of positive pairs S:

$$\min_P \ \|P - P_0\|^2 + \frac{\lambda}{|\mathcal{S}|} \sum_{(i,j) \in \mathcal{S}} \left\| P z_i - P z_j \right\|^2, \tag{7.3}$$

where (z_i, z_j) is the embedding of pair (i, j) in $\mathcal{S}$ and (i, j) are prompts that describe the same class but different spurious attributes. The loss encourages the linear projection P to be invariant to the difference between (i, j), i.e., the spurious attributes. The optimization problem has a convenient closed-form solution, as demonstrated in Lemma 7.1.

Lemma 7.1. *The minimizer of the calibration loss is*

$$P^* = P_0\Big(I + \frac{\lambda}{|\mathcal{S}|}\sum_{(i,j)\in\mathcal{S}}(z_i - z_j)(z_i - z_j)^T\Big)^{-1}.$$

Proof. We will leverage the first order optimality criteria to derive the solution.

$$\min_P \ \|P - P_0\|^2 + \frac{\lambda}{|\mathcal{S}|}\sum_{(i,j)\in\mathcal{S}}\|Pz_i - Pz_j\|^2$$

The loss can be written as

$$\begin{aligned}\mathcal{L}(P) &= \frac{\lambda}{|\mathcal{S}|}(Pz_i - Pz_j)^T(Pz_i - Pz_j) + (P - P_0)^T(P - P_0)\\ &= \frac{\lambda}{|\mathcal{S}|}({z_i}^T P^T P z_i + {z_j}^T P^T P z_j - {z_i}^T P^T P z_j - {z_j}^T P^T P z_i)\\ &\quad + (P^T P + P_0^T P_0 - P_0^T P - P^T P_0)\end{aligned}$$

Setting the derivate w.r.t. P to zero yields:

$$\begin{aligned}\frac{\partial\mathcal{L}(P)}{\partial P} = \frac{\lambda}{|\mathcal{S}|}\sum_{(i,j)\in\mathcal{S}} 2(Pz_i{z_i}^T + Pz_j{z_j}^T - Pz_i{z_j}^T - Pz_j{z_i}^T) + (2P - 2P_0) = 0\\ P(I + \frac{\lambda}{|\mathcal{S}|}\sum_{(i,j)\in\mathcal{S}}(z_i{z_i}^T + z_j{z_j}^T - z_i{z_j}^T - z_j{z_i}^T)) = P_0\\ P = P_0(I + \frac{\lambda}{|\mathcal{S}|}\sum_{(i,j)\in\mathcal{S}}(z_i{z_i}^T + z_j{z_j}^T - z_i{z_j}^T - z_j{z_i}^T))^{-1}\\ = P_0\Big(I + \frac{\lambda}{|\mathcal{S}|}\sum_{(i,j)\in\mathcal{S}}(z_i - z_j)(z_i - z_j)^T\Big)^{-1}\end{aligned}$$

One can also rewrite the objective as

$$\begin{aligned}&\min_P \ \|P - P_0\|^2 + \frac{\lambda}{|\mathcal{S}|}\sum_{(i,j)\in\mathcal{S}}\|P(z_i - z_j)\|^2\\ =&\min_P \ \|P - P_0\|^2 + \frac{\lambda}{|\mathcal{S}|}\|PZ_{\text{diff}}\|^2\end{aligned}$$

We then have

$$\frac{\partial \mathcal{L}(P)}{\partial P} = 2(P - P_0) + 2\frac{\lambda}{|\mathcal{S}|} P Z_{\text{diff}} Z_{\text{diff}}^T = 0$$
$$P(I + \frac{\lambda}{|\mathcal{S}|} Z_{\text{diff}} Z_{\text{diff}}^T) = P_0$$
$$P = P_0(I + \frac{\lambda}{|\mathcal{S}|} Z_{\text{diff}} Z_{\text{diff}}^T)^{-1}.$$

Note that two optimums are equivalent, where the second one is simply the matrix form of the first. □

We can obtain an interpretation of the minimizer by relating it to singular value decomposition (SVD). Let $Z_{\text{diff}} \in \mathbb{R}^{d \times |S|}$, where the columns of Z_{diff} enumerate the pairwise difference $z^i - z^j$ for all $(i, j) \in S$. The matrix Z_{diff} defines a subspace that represents the variation in the embedding when the irrelevant feature is changed. Using Z_{diff}, the minimizer can be written as $P^* = P_0(I + \lambda' Z_{\text{diff}} Z_{\text{diff}}^T)^{-1}$, where we define $\lambda' = \lambda/|\mathcal{S}|$ to simplify the notation. Assume that the SVD of Z_{diff} is $U\Sigma V^T$. Then we have $Z_{\text{diff}} Z_{\text{diff}}^T = U\Sigma^2 U^T$. The optimal solution P^* can then be rewritten as

$$P^* = P_0(U(I + \lambda'\Sigma^2)U^T)^{-1} = P_0 \underbrace{U(I + \lambda'\Sigma^2)^{-1}U^T}_{\text{Calibration Matrix}}.$$

We can see that $U(I + \lambda'\Sigma^2)^{-1}U^T$ acts as a calibration term. Before multiplying the text embedding with the projection matrix P_0, variation due to the change of the spurious feature, namely, the eigenvectors with large squared singular value in Z_{diff} (spurious direction) will be down-weighted due to the inverse $(I + \lambda'\Sigma^2)^{-1}$. Therefore, varying the spurious attributes should result in similar embeddings after multiplying the calibration matrix.

7.3.4 Relation to an Equalization Loss

Finally, we provide an equivalent form of the calibrated projection and relate it to an equalization loss. Ideally, we want each row of the classifier weight $\beta \in \mathbb{R}^{K \times d}$ to have similar cosine similarity to pairs of embeddings in $\mathcal{S}$. For instance, the embedding

of "a photo of a doctor" should be equally similar to "a photo of a male doctor" and "a photo of a female doctor". In this section, we will show that the optimum of the calibration loss does satisfy this criterion.

We consider the following objective for obtaining a debiased text embedding $z \in \mathbb{R}^d$ of a prompt given its initialization $z_0 \in \mathbb{R}^d$ from the text encoder:

$$\min_z \ \|z - z_0\|^2 + \frac{\lambda}{|\mathcal{S}|} \sum_{(i,j)\in\mathcal{S}} (z^T z_i - z^T z_j)^2. \tag{7.4}$$

The loss encourages the embedding z to have similar cosine similarity to embeddings in positive pairs while maintaining proximity to the initialization z_0. Objective (7.4) has the same optimal solution as the calibration loss (7.3).

Lemma 7.2. *The minimizer of objective (7.4) reads*

$$z^* = \underbrace{\left(I + \frac{\lambda}{|\mathcal{S}|} \sum_{(i,j)\in\mathcal{S}} (z_i - z_j)(z_i - z_j)^T\right)^{-1}}_{\text{Calibration Matrix}} z_0$$

In particular, we have $P_0 z^* = P^* z_0$ where P^* is the minimizer of the calibration loss (7.3).

Proof. Similarly, the objective can be rewritten as

$$\begin{aligned} &\min_z \ \|z - z_0\|^2 + \frac{\lambda}{|\mathcal{S}|} \sum_{(i,j)\in\mathcal{S}} (z^T(z_i - z_j))^2 \\ =&\min_z \ \|z - z_0\|^2 + \frac{\lambda}{|\mathcal{S}|} \|z^T Z_{\text{diff}}\|^2. \end{aligned}$$

The derivative is:

$$\begin{aligned} \frac{\partial \mathcal{L}(z)}{\partial z} &= 2z - 2z_0 + 2\frac{\lambda}{|\mathcal{S}|} Z_{\text{diff}} Z_{\text{diff}}^T z = 0 \\ z &= (I + \frac{\lambda}{|\mathcal{S}|} Z_{\text{diff}} Z_{\text{diff}}^T)^{-1} z_0 \\ &= \left(I + \frac{\lambda}{|\mathcal{S}|} \sum_{(i,j)\in\mathcal{S}} (z_i - z_j)(z_i - z_j)^T\right)^{-1} z_0 \end{aligned}$$

□

Lemma 7.2 shows that the optimal solution of (7.4) is equivalent to multiplying the original embedding z with the calibration matrix defined before. Applying the projection P_0 to z^* leads to the same weight in Lemma 7.1. This interpretation is particularly useful in cases where the ideal solution does not lie in the middle of z_i and z_j, as will be shown in section 7.5 where we address biases in generative models.

The equalization objective has a similar motivation as the equalization step proposed by Bolukbasi et al. [20] in their work on removing gender bias from word embeddings. Similar to the idea of positive pairs, given a set of word embeddings that has the same semantic meaning except for gender, their approach centers these embeddings by setting them to the average embedding of the set. After centering, any word in the dictionary will be equidistant to all words in the set. However, our approach differs in that we modify the embedding of the target prompt z, rather than the embedding of positive pairs, making it more suitable for debiasing zero-shot classifiers as we are primarily concerned with the embedding of z.

7.4 Experiments: Discriminative Models

7.4.1 Group Robustness against Spurious Correlations

By following the setting of Zhang and Ré [224], we evaluate our approach on two popular benchmarks for evaluating spurious correlations, Waterbird [165] and CelebA [118]. On Waterbird, a water/land background is a confounding factor for the waterbirds/landbirds class, while on CelebA the binary gender is the spurious feature for blond/dark hair. As such, both datasets contain four groups defined by the labels and the spurious attributes.

We evaluate our approach against several baselines, including zero-shot classification [159], empirical risk minimization (ERM) with linear probing [110], and ERM with non-linear adapter [60]. Additionally, we also consider three recent methods designed to improve the group robustness of vision-language foundation classifiers:

Backbone	CLIP ResNet-50						CLIP ViT-L/14					
Dataset	**Waterbird**			**CelebA**			**Waterbird**			**CelebA**		
	WG	Avg	Gap	WG	Avg	Gap	WG	Avg	Gap	WG	Avg	Gap
methods using data and labels												
ERM Linear	7.9	93.5	85.6	11.9	94.7	82.8	65.9	97.6	31.7	28.3	94.7	66.4
ERM Adapter	60.8	96.0	35.2	36.1	94.2	58.1	78.4	97.8	19.4	36.7	94.2	57.5
WiSE-FT	49.8	91.0	41.2	85.6	88.6	3.0	65.9	97.6	31.7	80.0	87.4	7.4
DFR (Sub)	63.9	91.8	27.9	76.9	92.5	15.6	51.9	95.7	43.8	76.3	92.1	15.8
DFR (Up)	51.3	92.4	41.1	89.6	91.8	2.2	65.9	96.1	30.2	83.7	91.2	7.5
CA	83.7	89.4	5.7	90.0	90.7	0.7	86.9	96.2	9.3	84.6	90.4	5.8
methods without data and labels												
Zero-shot	39.6	77.3	37.7	75.9	82.3	6.4	45.3	84.4	39.1	72.8	87.6	14.9
Orth-Proj (Ours)	48.1	83.6	35.4	61.4	86.4	25.0	61.4	86.4	25.0	71.1	87.0	15.9
Orth-Cali (Ours)	**74.0**	78.7	**4.7**	**82.2**	84.4	**2.2**	**68.8**	84.5	**15.7**	**76.1**	86.2	**10.1**

Table 7.2: Group Robustness of Vision-Language Models. For each backbone, the first blocks contain methods that require data and labels, while the second blocks contain zero-shot methods. The numbers for the first block are adopted from Zhang and Ré [224]. The proposed calibration loss achieves comparable or even smaller gaps between average and worst group accuracy without the need for any data or labels.

- Weight Space Ensembling (WiSE-FT) [207], which trains a linear classifier first using ERM and then combines the classifier outputs with the initial zero-shot predictions;
- Deep Feature Reweighting (DFR) [102], which trains a linear probe on embeddings obtained from a pre-trained model using group-balanced data. Following Zhang and Ré [224], the group labels are replaced with zero-shot predictions;
- Contrastive Adapter (CA) [224], which trains adapters using contrastive learning to bring embeddings in the same class closer.

It is important to note that **all of the baselines** except the zero-shot classifier **require at least training data and class labels**, while our debiasing approach does not require access to any input data, labels, or group labels, which follows the principles of zero-shot learning.

We evaluate the performance of our proposed approach using two CLIP back-

	Waterbird			CelebA		
	WG	Avg	Gap	WG	Avg	Gap
$\lambda = 200$	71.8	80.8	9.0	80.7	83.9	3.2
$\lambda = 400$	72.9	79.5	6.6	81.6	84.2	2.6
$\lambda = 600$	73.5	79.2	5.7	81.9	84.3	2.4
$\lambda = 1000$	74.0	78.7	4.7	82.2	84.4	2.2

Table 7.3: Sensitivity to λ. We vary the weighting parameter λ and evaluate group robustness with ResNet-50 backbone.

bones: ResNet-50 [79] and ViT-L/14 [48]. The results are presented in Table 7.2. The results indicate that a simple application of the orthogonal projection (Orth-Proj) by itself only yields limited improvement of the worst group accuracy, whereas the calibration loss (Orth-Cali) significantly improves robustness across datasets and base models. The proposed Orth-Cali method achieves comparable or even smaller gaps between average and worst group accuracy compared to the state-of-the-art contrastive adapter [224], without the need for any data or labels. Note that the baselines generally achieve better average accuracy as they require fine-tuning on the target datasets.

Empirically, we found that gradually increasing the parameter λ improves the worst group accuracy and leads to a stable solution as shown in Table 7.3. Therefore, for all the experiments on discriminative models, we set λ to 1000 by default. To investigate the importance of orthogonal projection and calibration, we present an ablation study in Table 7.4. The results indicate that the calibration loss alone ($P_0 = I$) performs well on the CelebA dataset, as the spurious feature (gender) is relatively easy to describe with prompts. However, performance drops on the Waterbird dataset without a good initialization from the orthogonal projection. More ablation studies can also be found in Appendix A.4, where we demonstrate the importance of class names in positive pairs.

	Waterbird			CelebA		
	WG	Avg	Gap	WG	Avg	Gap
Proj only	48.1	83.6	35.4	61.4	86.4	25.0
Cali only	55.6	81.6	26.0	81.6	84.7	3.1
Proj + Cali	**74.0**	78.7	**4.7**	**82.2**	84.4	**2.2**

Table 7.4: Dissecting Orthogonal Projections. We evaluate variants of orthogonal projection with ResNet-50 backbone.

	CLIP ViT-B/32			CLIP ViT-L/14		
	Gen	Race	Age	Gen	Race	Age
Zero-shot	.206	.743	.797	.206	.768	.703
Orth-Proj	.146	.755	**.635**	.349	.605	.706
Orth-Cali	**.102**	**.638**	.641	**.200**	**.461**	**.662**

Table 7.5: Measuring biases on FairFace. We MaxSkew@1000 (the smaller the better) on FairFace validation set.

7.4.2 Debiased Information Retrieval

Fairness in text-image retrieval has gained increasing attention in recent years. Building on the work of Berg et al. [16], we propose to utilize the MaxSkew metric, introduced by Geyik et al. [63], to evaluate the level of fairness in the retrieval results. Specifically, we conduct our analysis on the FairFace dataset [97], which is specifically designed to address issues of fairness in facial recognition systems. Given a ranked list of images in response to a text query, let $r_{a,k}$ be the ratio of the top k images that are labeled with attribute a. Then MaxSkew@k is defined as $\max_{a \in \mathcal{A}} \log \frac{r_{a,k}}{1/|\mathcal{A}|}$. It quantifies the maximal discrepancy between the ratio of top k images labeled with a specific sensitive attribute, denoted as $r_{a,k}$, and the uniform weight $1/|\mathcal{A}|$, where $\mathcal{A}$ represents the set of sensitive attributes. The MaxSkew metric provides a useful measure of fairness in text-image retrieval systems, by assessing the degree to which the retrieval results are evenly distributed across sensitive attributes. A small MaxSkew value indicates that the distribution of retrieved images across different sensitive attributes is close to being uniform.

To measure the bias, we query the validation set of FairFace based on 10 prompts that are uncorrelated with facial expressions or sensitive attributes, e.g., “a photo of a [concept] person”, where the [concept] is a neutral concept such as evil or smart. The detailed prompts are described in Appendix A.4. We measure the MaxSkew based on three labeled attributes of FairFace: gender, race, and age. Table 7.5 shows the average MaxSkew@1000 over concepts, demonstrating that our approach significantly reduces the MaxSkew across different attributes and backbones.

7.5 Debiasing Generative Models

We now explore the possibility of extending the methodology developed for discriminative models to generative models. Our primary focus is on addressing social group biases, specifically gender and race discrepancy, as measured by metric (7.1). In particular, the main experiment is to query the generative model using profession-related prompts, specifically “a photo of a [profession]". Empirically, the generated images were found to exhibit a strong bias towards certain gender and race, and we attempt to improve the diversity of generated images with the proposed equalization loss in this section. We also demonstrate that our approach can also address spurious correlations beyond social biases.

A Single Matrix for Comprehensive Debiasing Unlike the well-defined targets prevalent in zero-shot classification, the nature of generative models requires a more universal solution. Specifically, we seek to derive a debiasing matrix capable of accommodating any prompt. This matrix could subsequently be treated as a standardized preprocessing step, applied prior to the introduction of the embedding into the generator.

To achieve this, we optimize the equalization loss with positive pairs consisting of an enumeration of “a photo of a [attribute] [profession]” where the [attribute] is a member of the set of gender or races and the [profession] is a job title sampled from a training set. For instance, to mitigate gender bias, we adopt $\mathcal{S} = \{$(“a photo of

a male doctor", "a photo of a female doctor"), $\cdots$, ("a photo of a male engineer", "a photo of a female engineer") }. By solving the calibration matrix with professions in the training set, we expect the obtained matrix can also mitigate the biases in unseen professions. Note that we optimize the equalization loss (7.4) without applying the initial orthogonal projection matrix P_0. This is because our goal is to balance rather than completely eliminate biased information in the generated images.

7.6 Experiments: Generative Models

To evaluate the effectiveness of our approach in the context of generative models, we conducted experiments using the Stable Diffusion (SD) v2.1 framework [164]. We construct a list of professions that consists of 100 job titles with GPT-4 [148] and randomly separate them into 80 training and 20 testing professions. In alignment with the framework proposed by Kärkkäinen and Joo [97], we consider the gender attributes of male and female, and racial attributes of White, Asian, Black, Indian, and Latino[1].

Evaluating generative models can be challenging without the use of human labels. Inspired by Cho et al. [29], we used sensitive attribute classifiers to predict the sensitive attributes of the generated images. The discrepancy, as defined in (7.2), was then calculated. In particular, we generate 100 images for each train / test profession for evaluation, resulting in 10000 images for each model. We then leverage the CLIP classifier to predict the sensitive attributes to calculate the discrepancy. An alternative to CLIP is the FairFace classifier [97]; however, we found that the domain shift between the FairFace dataset and the generated images significantly impairs its performance. The debiased and biased models share the same random seed for fair comparison. We set $\lambda = 500$ for all the experiments in this section.

[1]It is essential to recognize gender and race are complex social constructs that cannot be simply reduced to binary or discrete categories. The choice of using binary gender and discrete race attributes in our work was primarily based on the existing literature and benchmark datasets that have commonly adopted this setting for evaluation purposes.

Figure 7-4: Improving Gender Diversity of Stable Diffusion. We fix the random seed of initial latent noise of Stable Diffusion [164] and generate the images with the training / testing prompt "a photo of a doctor / firefighter". The results demonstrate that applying the calibration matrix to the prompt embedding improves the balance between male and female in the generated images.

7.6.1 Measuring the Generalization of Calibration

We first examine whether minimizing the calibration loss with training prompts can yield a calibration matrix that also works for unseen (testing) professions. In particular, we measure the average L2 difference between the projected embedding $\sum_{(i,j)\in\mathcal{S}_{\text{test}}} \left\|Pz_i - Pz_j\right\| / |\mathcal{S}_{\text{test}}|$ for testing prompts and show the results in Table 7.6. We can see that the calibration matrix successfully minimizes the difference after projection, even for unseen professions.

	Gender		**Race**	
	before	after	before	after
train	0.56	0.14	0.70	0.23
test	0.54	0.16	0.69	0.26

Table 7.6: Difference between embeddings ($\lambda = 500$).

7.6.2 Quantitative and Qualitative Results

The results presented in Table 7.7 demonstrate a significant reduction in both gender and race discrepancy after debiasing. Importantly, the improvements are observed for both training and testing professions, implying that the obtained debiasing matrix

can generalize beyond training prompts. To further illustrate the effectiveness of our approach, we present quantitative results for mitigating gender bias in Figure 7-4. By applying the calibration matrix to balance the male and female directions, the gender diversity of the generated images significantly improved. Additional examples can be found in appendix A.4.

		Gender	**Race**
Train	SD	0.472±0.225	0.485±0.160
	Ours	**0.395±0.205**	**0.434±0.163**
Test	SD	0.412±0.255	0.528±0.184
	Ours	**0.354±0.253**	**0.455±0.169**

Table 7.7: Discrepancy between Groups: Calibration matrix reduces the discrepancy over gender and race. The calibration matrix derived from the training set generalizes well to testing set.

Compared to gender bias, we found that addressing racial bias is a more challenging task. One source of complexity is the ambiguity of ethnicity, as individuals may identify with multiple races. Nevertheless, as Figure 7-5 and Table 7.7 demonstrate, the diversity in the output images is improved by simply debiasing the prompt embedding with the calibration matrix.

7.6.3 Human Evaluation

Despite the scalability, the prediction from a trained classifier could be erroneous. Therefore, we also evaluate our approach with human evaluation, where we invite annotators of different genders, races, and nationalities to label the sensitive attributes of the generated images. Details and the interface are included in appendix A.4.3. For human evaluation, we generate 25 images for each test profession, re-

	Gender	**Race**
SD	0.472±0.257	0.723±0.185
Ours	**0.372±0.253**	**0.589±0.188**

Table 7.8: Human Evaluation. We calculate the discrepancy on testing professions with human annotations. Our approach improves the diversity of Stable Diffusion by a non-trivial margin.

Figure 7-5: Improving Racial Diversity of Stable Diffusion. We again generate the images with Stable Diffusion. After applying the calibration matrix, the race attributes are more diverse in the generated images.

sulting in 500 images for each model. As Table 7.8 shows, our approach greatly improves the diversity of the generated images, corroborating the previous results.

7.6.4 Beyond Social Biases

Our approach can also be applied to address general spurious attributes beyond social biases. As an example, we draw inspiration from the WaterBird dataset [165] and debias the prompt "a photo of a waterbird" by using {"a photo of a [animal] with water background" and "a photo of a [animal] with land background" } as positive pairs, where we construct a list of 100 names of animals with GPT-4 [148].

As Figure 7-6 illustrates, our approach successfully generates images of water birds in both land and water backgrounds, whereas the original models only generated images with water background.

7.7 Conclusion

In this chapter, we present a new approach to debiasing vision-language foundation models by utilizing prompts to mitigate biases. The proposed calibrated projec-

tion effectively mitigates biases in both discriminative and generative vision-language models without any additional training or data.

Figure 7-6: Generation against Non-social Biases. The results demonstrate the ability of the proposed method to generate images of waterbirds in both land and water backgrounds.

Chapter 8

Generalization Theory of Representation Learning

In this chapter, we culminate our discussions from previous sections by delving into the intrinsic generalization theory of representation learning. Our focus lies in understanding how the cultivation of superior representations can enhance generalization across a multitude of downstream tasks. To this end, we propose a new generalization theory that accentuates the importance of the representations encoded by the model. Furthermore, this theory elucidates why contrastive learning bolsters generalization, offering a comprehensive explanation.

Motivated by the remarkable empirical success of deep learning, there has been significant effort in statistical learning theory toward deriving generalization error bounds for deep learning, i.e complexity measures that predict the gap between training and test errors. Recently, substantial progress has been made, e.g., [7, 13, 22, 50, 67, 144, 204]. Nevertheless, many of the current approaches lead to generalization bounds that are often vacuous or not consistent with empirical observations [51, 91, 137]. Furthermore, conventional constraints primarily limit the generalization error by considering the complexity of the hypothesis, while neglecting the characteristics of the input or feature space. This approach renders them less than optimal for the study of representation learning's generalization potential.

To empirically study the generalization theory, Jiang et al. [91] present a large scale

study of generalization in deep networks and show that many existing approaches, e.g., norm-based bounds [13, 143, 144], are not predictive of generalization in practice. Recently, the *Predicting Generalization in Deep Learning (PGDL)* competition described in [92] sought complexity measures that are predictive of generalization error given training data and network parameters. To achieve a high score, the predictive measure of generalization had to be robust to different hyperparameters, network architectures, and datasets. The participants [112, 141, 168] achieved encouraging improvement over the classic measures such as VC-dimension [193] and weight norm [13]. Unfortunately, despite the good empirical results, these proposed approaches are not yet supported by rigorous theoretical bounds [93].

In this chapter, we attempt to decrease this gap between theory and practice with margin bounds based on optimal transport. In particular, we show that the expected optimal transport cost of matching two independent random subsets of the training distribution is a natural alternative to Rademacher complexity. Interestingly, this optimal transport cost can be interpreted as the k-variance [175], a generalized notion of variance that captures the structural properties of the data distribution. Applied to latent space, it captures important properties of the learned feature distribution. The resulting k-variance normalized margin bounds can be easily estimated and correlate well with the generalization error on the PGDL datasets [92]. In addition, our formulation naturally encompasses the gradient normalized margin proposed by Elsayed et al. [53], further relating our bounds to the decision boundary of neural networks and their robustness.

Theoretically, our bounds reveal that the *concentration* and *separation* of learned features are important factors for the generalization of multiclass classification. In particular, the downstream classifier generalizes well if (1) the features within a class are well clustered, and (2) the classes are separable in the feature space in the Wasserstein sense. This aligns with the driving principle behind contrastive learning, where representations are learned to concentrate via positive pairs while segregating negative pairs.

8.1 Background of Generalization Bounds

Margin-based Generalization Bounds Classic approaches in learning theory bound the generalization error with the complexity of the hypothesis class [14, 193]. Nevertheless, previous works show that these uniform convergence approaches are not able to explain the generalization ability of deep neural networks given corrupted labels [222] or on specific designs of data distributions as in [137]. Recently, substantial progress has been made to develop better data-dependent and algorithm-dependent bounds [7, 15, 22, 37, 145, 191, 204]. Among them, we will focus on margin-based generalization bounds for multi-class classification [106, 111]. Bartlett et al. [13] show that margin normalized with the product of spectral norms of weight matrices is able to capture the difficulty of the learning task, where the conventional margin struggles. Concurrently, Neyshabur et al. [144] derive spectrally-normalized margin bounds via weight perturbation within a PAC-Bayes framework [125]. However, empirically, spectral norm-based bounds can correlate negatively with generalization [91, 137]. Elsayed et al. [53] present a gradient-normalized margin, which can be interpreted as the first order approximation to the distance to the decision boundary. Jiang et al. [90] further show that gradient-normalized margins, when combined with the total feature variance, are good predictors of the generalization gap. Despite the empirical progress, gradient-normalized margins are not yet supported by theoretical bounds.

Empirical Measures of Generalization Large scale empirical studies have been conducted to study various proposed generalization predictors [90, 91]. In particular, Jiang et al. [91] measure the average correlation between various complexity measures and generalization error under different experimental settings. Building on their study, Dziugaite et al. [51] emphasize the importance of the robustness of the generalization measures to the experimental setting. These works show that well-known complexity measures such as weight norm [136, 143], spectral complexity [13, 144], and their variants are often negatively correlated with the generalization gap. Recently, Jiang et al. [92] hosted the *Predicting Generalization in Deep Learning (PGDL)* competition, encouraging the participants to propose robust and general

complexity measures that can rank networks according to their generalization errors. Encouragingly, several approaches [112, 141, 168] outperformed the conventional baselines by a large margin. Nevertheless, none of these are theoretically motivated with rigorous generalization bounds. Our k-variance normalized margins are good empirical predictors of the generalization gap, while also being supported with strong theoretical bounds.

8.2 Optimal Transport and k-Variance

Before presenting our generalization bounds with optimal transport, we first give a brief introduction to the Wasserstein distance, a distance function between probability distributions defined via an optimal transport cost. Letting μ and $\nu \in \text{Prob}(\mathbb{R}^d)$ be two probability measures, the p-Wasserstein distance with Euclidean cost function is defined as

$$\mathcal{W}_p(\mu, \nu) = \inf_{\pi \in \Pi(\mu,\nu)} \left(\mathbb{E}_{(X,Y)\sim\pi} \|X - Y\|^p\right)^{1/p},$$

where $\Pi(\mu, \nu) \subseteq \text{Prob}(\mathbb{R}^d \times \mathbb{R}^d)$ denotes the set of measure couplings whose marginals are μ and ν, respectively. The 1-Wasserstein distance is also known as the Earth Mover distance. Intuitively, Wasserstein distances measure the minimal cost to transport the distribution μ to ν.

Based on the Wasserstein distance, Solomon et al. [175] propose the *k-variance*, a generalization of variance, to measure structural properties of a distribution beyond variance.

Definition 8.1 (Wasserstein-p k-variance)**.** *Given a probability measure $\mu \in \text{Prob}(\mathbb{R}^d)$ and a parameter $k \in \mathbb{N}$, the* Wasserstein-p k-variance *is defined as*

$$\text{Var}_{k,p}(\mu) = c_p(k, d) \cdot \mathbb{E}_{S,\tilde{S}\sim\mu^k} \left[\mathcal{W}_p^p(\mu_S, \mu_{\tilde{S}})\right],$$

where $\mu_S = \frac{1}{k}\sum_{i=1}^k \delta_{x_i}$ for $x_i \overset{\text{i.i.d.}}{\sim} \mu$ and $c_p(k, d)$ is a normalization term described in [175].

When $k = 1$ and $p = 2$, the k-variance is equivalent to the variance $\mathrm{Var}[X]$ of the random variable $X \sim \mu$. For $k > 1$ and $d \geq 3$, Solomon et al. [175] show that when $p = 2$, the k-variance provides an intuitive way to measure the average intra-cluster variance of clustered measures. In this work, we use the unnormalized version ($c_p(k, d) = 1$) of k-variance and $p = 1$, and drop the p in the notation:

$$\text{(Wasserstein-1 } k\text{-variance):} \quad \mathrm{Var}_k(\mu) = \mathbb{E}_{S,\tilde{S}\sim\mu^k}\left[\mathcal{W}_1(\mu_S, \mu_{\tilde{S}})\right].$$

Note that setting $c_p(k, d) = 1$ is not an assumption, but instead an alternative definition of k-variance. The change in constant has no effect on any part of our paper, as we could reintroduce the constant of Solomon et al. [175] and simply include a pre-multiplication term in the generalization bounds to cancel it out. In Section 8.5, we will show that this unnormalized Wasserstein-1 k-variance captures the concentration of learned features. Next, we use it to derive generalization bounds.

8.3 Generalization Bounds with Optimal Transport

We present our generalization bounds in the multi-class setting. Let $\mathcal{X}$ denote the input space and $\mathcal{Y} = \{1, \ldots, K\}$ denote the output space. We will assume a compositional hypothesis class $\mathcal{F} \circ \Phi$, where the hypothesis $f \circ \phi$ can be decomposed as a feature (representation) encoder $\phi \in \Phi$ and a predictor $f \in \mathcal{F}$. This includes dividing multilayer neural networks at an intermediate layer.

We consider the score-based classifier $f = [f_1, \ldots, f_K]$, $f_c \in \mathcal{F}_c$, where the prediction for $x \in \mathcal{X}$ is given by $\arg\max_{y\in\mathcal{Y}} f_y(\phi(x))$. The margin of f for a datapoint (x, y) is defined by

$$\rho_f(\phi(x), y) := f_y(\phi(x)) - \max_{y'\neq y} f_{y'}(\phi(x)), \tag{8.1}$$

where f misclassifies if $\rho_f(\phi(x), y) \leq 0$. The dataset $S = \{x_i, y_i\}_{i=1}^m$ is drawn i.i.d. from distribution μ over $\mathcal{X} \times \mathcal{Y}$. Define m_c as the number of samples in class c, yielding $m = \sum_{c=1}^K m_c$. We denote the marginal over a class $c \in \mathcal{Y}$ as μ_c and the

distribution over classes by $p(c)$. The pushforward measure of μ with respect to ϕ is denoted as $\phi_{\#}\mu$. We are interested in bounding the expected zero-one loss of a hypothesis $f \circ \phi$: $R_\mu(f \circ \phi) = \mathbb{E}_{(x,y)\sim\mu}[\mathbb{1}_{\rho_f(\phi(x),y)\leq 0}]$ by the corresponding empirical γ-margin loss $\hat{R}_{\gamma,m}(f \circ \phi) = \mathbb{E}_{(x,y)\sim S}[\mathbb{1}_{\rho_f(\phi(x),y)\leq \gamma}]$.

8.3.1 Feature Learning and Generalization: Margin Bounds with k-Variance

Our theory is motivated by recent progress in feature learning, which suggests that imposing certain structure on the feature distribution improves generalization [24, 36, 99, 214, 219]. The participants [112, 141] of the PGDL competition [92] also demonstrate nontrivial correlation between feature distribution and generalization.

To study the connection between learned features and generalization, we derive generalization bounds based on the k-variance of the feature distribution. In particular, we first derive bounds for a fixed encoder and discuss the generalization error of the encoder at the end of the section. Theorem 8.2 provides a generalization bound for neural networks via the concentration of μ_c in each class.

Theorem 8.2. *Let $f = [f_1, \cdots, f_K] \in \mathcal{F} = \mathcal{F}_1 \times \cdots \times \mathcal{F}_K$ where $\mathcal{F}_i : \mathcal{X} \to \mathbb{R}$. Fix $\gamma > 0$. The following bound holds for all $f \in \mathcal{F}$ with probability at least $1 - \delta > 0$:*

$$R_\mu(f \circ \phi) \leq \hat{R}_{\gamma,m}(f \circ \phi) + \mathbb{E}_{c\sim p}\left[\frac{\mathrm{Lip}(\rho_f(\cdot, c))}{\gamma}\mathrm{Var}_{m_c}(\phi_{\#}\mu_c)\right] + \sqrt{\frac{\log(1/\delta)}{2m}},$$

where $\mathrm{Lip}(\rho_f(\cdot, c)) = \sup_{x,x'\in\mathcal{X}} \frac{|\rho_f(\phi(x),c) - \rho_f(\phi(x'),c)|}{||\phi(x)-\phi(x')||_2}$ *is the margin Lipschitz constant w.r.t ϕ.*

We give a proof sketch here and defer the full proof later. For a given class c and a given feature map ϕ, let $\mathcal{H}_c = \{h(x) = \rho_f(\phi(x), c) | f = (f_1 \dots f_K), f_y \in \mathcal{F}\}$. The last step in deriving Rademacher-based generalization bounds [14] amounts to bounding

for each class c:

$$\Delta_c = \mathbb{E}_{S,\tilde{S}\sim\mu_c^{m_c}}\left[\sup_{h\in\mathcal{H}_c}\frac{1}{m_c}\sum_{i=1}^{m_c}h(\tilde{x}_i) - \frac{1}{m_c}\sum_{i=1}^{m_c}h(x_i)\right], \tag{8.2}$$

where $S, \tilde{S} \sim \mu_c^{m_c}$. Typically we would plug in the Rademacher variable and arrive at the standard Rademacher generalization bound. Instead, our key observation is that the Kantorovich-Rubinstein duality [84] implies

$$\mathcal{W}_1(\mu,\nu) = \sup_{\mathrm{Lip}(h)\leq 1}\mathbb{E}_{x\sim\mu}h(x) - \mathbb{E}_{x\sim\nu}h(x),$$

where the supremum is over the 1-Lipschitz functions $h : \mathbb{R}^d \to \mathbb{R}$. Suppose $\mathcal{H}_c$ is a subset of L-Lipschitz functions. By definition of the supremum, the duality result immediately implies that equation 8.2 can be bounded with k-variance for $k = m_c$:

$$\Delta_c \leq L\cdot\mathbb{E}_{S,\tilde{S}\sim\mu_c^{m_c}}[W_1(\phi_{\#}\mu_S, \phi_{\#}\mu_{\tilde{S}})] = L\cdot\mathrm{Var}_{m_c}(\phi_{\#}\mu_c). \tag{8.3}$$

Now we provide the full proof here.

Proof of Theorem 8.2. Recall the margin definition:

$$\rho_f(\phi(x), y) = f_y(\phi(x)) - \max_{y'\neq y} f_{y'}(\phi(x))$$

Let $\mu_c(x) = \mathbb{P}(x|y=c)$, and let $p(y) = \mathbb{P}(Y=y) = \pi_y$. Given $f \in \mathcal{F}$ and $\phi \in \Phi = \{\phi : \mathcal{X} \to \mathcal{Z}, ||\phi(x)|| \leq R\}$, we are interested in bounding the class-average zero-one loss of a hypothesis $f \circ \phi$:

$$R_\mu(f\circ\phi) = \sum_{c=1}^{K}\pi_k R_{\mu_c}(f\circ\phi) = \sum_{c=1}^{K}\pi_k\mathbb{E}_{x\sim\mu_c}[\mathbb{1}_{\rho_f(\phi(x),c)\leq 0}],$$

where we will bound the error of each class $c \in \mathcal{Y}$ separately. To do so, the margin loss defined by L_γ by $L_\gamma(u) = \mathbb{1}_{u\leq 0} + (1-\frac{u}{\gamma})\mathbb{1}_{0<u\leq\gamma}$ would be handy.

Note that :

$$R_\mu(f \circ \phi) \leq \mathbb{E}_{(x,y)} L_\gamma(\rho_f(\phi(x), y)),$$

(see for example Lemma A.4 in [13] for a proof of this claim.)

By McDiarmid Inequality, we have with probability at least $1 - \delta$,

$$R_\mu(f \circ \phi) \leq \mathbb{E}_{(x,y)} L_\gamma(\rho_f(\phi(x), y)) \tag{8.4}$$

$$\leq \sum_{c=1}^{K} \pi_c \hat{\mathbb{E}}_{S \sim \mu_c^m} L_\gamma(\rho_f(\phi(x), c)) + \mathbb{D}(f \circ \phi, \mu) + \sqrt{\frac{\log(1/\delta)}{2m}}. \tag{8.5}$$

where

$$\mathbb{D}(f \circ \phi, \mu)$$

$$= \mathbb{E}_{S_1 \sim \mu_1^m} \ldots \mathbb{E}_{S_K \sim \mu_K^m} \left[\sup_{f \in \mathcal{F}} \left(\sum_{c=1}^{K} \pi_c (\mathbb{E}_{\mu_c}[L_\gamma(\rho_f(\phi(x), c)))] - \hat{\mathbb{E}}_{S_c \sim \mu_c^m}[L_\gamma(\rho_f(\phi(x), c))]) \right) \right]$$

Note that the sup here is taken only on the classifier function class and not on the classifier and the feature map together. For a given class c and feature map ϕ define:

$$\mathcal{G}_c = \left\{ h | h(z) = L_\rho \circ \rho_f(z, c) : f \in \mathcal{F}, z \in \mathcal{Z} \right\}.$$

Using the fact that $\sup(a + b) \leq \sup a + \sup b$, we have:

$$\mathbb{D}(f \circ \phi, \mu) \leq \sum_{c=1}^{K} \pi_c \mathbb{E}_{S_c \sim \mu_c} \sup_{f \in \mathcal{F}} \left(\mathbb{E}_{\mu_c}[L_\gamma(\rho_f(\phi(x), c)))] - \hat{\mathbb{E}}_{S_c \sim \mu_c^m}[L_\gamma(\rho_f(\phi(x), c))] \right)$$

$$= \sum_{c=1}^{K} \pi_c \mathbb{E}_{S_c \sim \mu_c} \left[\sup_{h \in \mathcal{G}_c} \left(\mathbb{E}_{\mu_c}[h(\phi(x))] - \hat{\mathbb{E}}_{S \sim \mu_c^m}[h(\phi(x))] \right) \right], \tag{8.6}$$

where the last equality follows from the definition of the function class $\mathcal{G}_c$

We are left now with bouding each class dependent deviation. We drop the index c from S_c in what follows in order to avoid cumbersome notations. Considering an

independent sample of same size $\tilde{S}$ from μ_c we have:

$$\mathbb{E}_{S\sim\mu_c^m}\left[\sup_{h\in\mathcal{G}_c}\left(\mathbb{E}_{\mu_c}[h(\phi(x))]-\hat{\mathbb{E}}_{S\sim\mu_c^m}[h(\phi(x))]\right)\right]\leq\mathbb{E}_{S,\tilde{S}\sim\mu_c^m}\left[\sup_{h\in\mathcal{G}_c}\hat{\mathbb{E}}_S[h(\phi(x)]-\hat{\mathbb{E}}_{\tilde{S}}[h(\phi(x))]\right]. \tag{8.7}$$

Note that $h(z)=L_\gamma(\rho_f(z,c))$ is lipchitz with lipchitz constant $\frac{1}{\gamma}\text{Lip}(\rho_f(.,c))$, since L_γ is lipchitz with lipchitz constant $\frac{1}{\gamma}$and by assumption the margin $\rho_f(z,c)$ is lipchitz in its first argument. By the dual of the Wasserstein 1 distance we have:

$$\mathcal{W}_1(\phi_\#p_S,\phi_\#p_{\tilde{S}})=\sup_{h,\text{Lip}(h)\leq 1}\hat{\mathbb{E}}_S[h(\phi(x)]-\hat{\mathbb{E}}_{\tilde{S}}[h(\phi(x))]$$

Since $\mathcal{G}_c$ are subset of lipchitz of functions with lipchitz constant $\frac{\text{Lip}(\rho_f(.,c))}{\gamma}$, it follows that:

$$\sup_{h\in\mathcal{G}_c}\hat{\mathbb{E}}_S[h(\phi(x)]-\hat{\mathbb{E}}_{\tilde{S}}[h(\phi(x))\leq\frac{\text{Lip}(\rho_f(.,c))}{\gamma}\mathcal{W}_1(\phi_\#p_S,\phi_\#p_{\tilde{S}}) \tag{8.8}$$

It follows from equation 8.7 and equation 8.8, that:

$$\begin{aligned}\mathbb{E}_{S\sim\mu_c^m}\left[\sup_{h\in\mathcal{G}_c}\left(\mathbb{E}_{\mu_c}[h(\phi(x))]-\hat{\mathbb{E}}_{S\sim\mu_c^m}[h(\phi(x))]\right)\right]&\leq\frac{\text{Lip}(\rho_f(.,c))}{\gamma}\mathbb{E}_{S,\tilde{S}\sim\mu_c^m}\mathcal{W}_1(\phi_\#p_S,\phi_\#p_{\tilde{S}})\\&=\frac{\text{Lip}(\rho_f(.,c))}{\gamma}\text{Var}_{m_c}(\phi_\#\mu_c).\end{aligned} \tag{8.9}$$

Finally Plugging equation 8.9 in equation 8.6 we obtain finally:

$$\mathbb{D}(f\circ\phi,\mu)\leq\frac{\sum_{c=1}^K\pi_c\text{Lip}(\rho_f(.,c))\text{Var}_{m_c}(\phi_\#\mu_c)}{\gamma} \tag{8.10}$$

Using equation 8.10 and noting that,

$$L_\gamma(\rho_f(\phi(x),c))\leq\mathbb{1}_{\rho_f(\phi(x),c)\leq\gamma}$$

we finally have by equation 8.5, the following generalization bound, that holds with

probability $1-\delta$:

$$R_\mu(f\circ\phi) \le \sum_{c=1}^{K}\pi_c\hat{\mathbb{E}}_{S\sim\mu_c^m}[\mathbb{1}_{\rho_f(\phi(x),c)\le\gamma}] + \frac{1}{\gamma}\sum_{c=1}^{K}\pi_c \mathrm{Lip}(\rho_f(.,c))\mathrm{Var}_{m_c}(\phi_\#\mu_c) + \sqrt{\frac{\log(1/\delta)}{2m}}$$
$$= \hat{R}_\gamma(f\circ\phi) + \mathbb{E}_{c\sim p_y}\left[\frac{\mathrm{Lip}(\rho_{f(.)}(\cdot,c))}{\gamma}\mathrm{Var}_{m_c}(\phi_\#\mu_c)\right] + \sqrt{\frac{\log(1/\delta)}{2m}}$$

□

This connection suggests that k-variance is a natural alternative to Rademacher complexity if the margin is Lipschitz. The following lemma shows that this holds when the functions f_j are Lipschitz:

Lemma 8.3. *The margin $\rho_f(.,y)$ is Lipschitz in its first argument if each of the f_j is Lipschitz.*

Proof. Assume $f_c(z) = \max_{y'\neq y} f_{y'}(z)$ and $f_{c'}(z') = \max_{y'\neq y} f_{y'}(z')$. Ties are broken by taking the largest index among the ones achieving the max.

$$\begin{aligned}
\rho_f(z,y) - \rho_f(z',y) &= f_y(z) - \max_{y'\neq y} f_{y'}(z) - (f_y(z') - \max_{y'\neq y} f_{y'}(z')) \\
&= f_y(z) - f_y(z') + f_{c'}(z') - f_c(z) \\
&\le L||z-z'|| + f_{c'}(z') - f_{c'}(z) \\
&\le L||z-z'|| + L||z-z'|| \\
&= 2L||z-z'||
\end{aligned}$$

where we used that all f_y are lipchitz and the fact that $f_c(z) \ge f_{c'}(z)$. Note that,

$$\begin{aligned}
\rho_f(z,y) - \rho_f(z',y) &= f_y(z) - f_y(z') + f_{c'}(z') - f_c(z) \\
&\ge -L||z-z'|| + f_c(z') - f_c(z) \\
&\ge -L||z-z'|| - L||z-z'|| \\
&= -2L||z-z'||.
\end{aligned}$$

where we used that all f_y are lipchitz and the fact that $f_{c'}(z') \geq f_c(z')$. Combining this two inequalities give the result. □

The bound in Theorem 8.2 is minimized when (a) the k-variance of features within each class is small, (b) the classifier has large margin, and (c) the Lipschitz constant of f is small. In particular, (a) and (b) express the idea of *concentration* and *separation* of the feature distribution, which we will further discuss in Section 8.5.

Compared to the margin bound with Rademacher complexity [111], Theorem 8.2 studies a fixed encoder, allowing the bound to capture the structure of the feature distribution. Although the Rademacher-based bound is also data-dependent, it only depends on the distribution over inputs and therefore can neither capture the effect of label corruption nor explain how the structure of the feature distribution $\phi_{\#}(\mu)$ affects generalization. Importantly, it is also non-trivial to estimate the Rademacher complexity empirically, which makes it hard to apply the bound in practice.

8.3.2 Gradient Normalized (GN) Margin Bounds with k-Variance

We next extend our theorem to use the *gradient-normalized margin*, a variation of the margin equation 8.1 that empirically improves generalization and adversarial robustness [53, 90]. Elsayed et al. [53] proposed it to approximate the minimum distance to a decision boundary, and Jiang et al. [90] simplified it to

$$\tilde{\rho}_f(\phi(x), y) := \rho_f(\phi(x), y)/(\|\nabla_\phi \rho_f(\phi(x), y)\|_2 + \epsilon),$$

where ϵ is a small value (10^{-6} in practice) that prevents the margin from going to infinity. The gradient here is $\nabla_\phi \rho_f(\phi(x), y) := \nabla_\phi f_y(\phi(x)) - \nabla_\phi f_{y_{\max}}(\phi(x))$, where ties among the $y_{\max}$ are broken arbitrarily as in [53, 90] (ignoring subgradients). The gradient-normalized margin $\tilde{\rho}_f(x, y)$ can be interpreted as the first order Taylor approximation of the minimum distance of the input x to the decision boundary for the class pair (y, y') [53]. In particular, the distance is defined as the norm of the minimal perturbation in the input or feature space to make the prediction change. See also Lemma B.9 in the supplement for an interpretation of this margin in terms of

robust feature separation. Defining the margin loss $\hat{R}^{\nabla}_{\gamma,m}(f) = \mathbb{E}_{(x,y)\sim S}[\mathbb{1}_{\tilde{\rho}_f(\phi(x),y)\leq\gamma}]$, we extend Theorem 8.2 to the gradient-normalized margin.

Theorem 8.4 (Gradient-Normalized Margin Bound)**.** *Let $f = [f_1, \cdots, f_k] \in \mathcal{F} = [\mathcal{F}_1, \cdots, \mathcal{F}_k]$ where $\mathcal{F}_i : \mathcal{X} \to \mathbb{R}$. Fix $\gamma > 0$. Then, for any $\delta > 0$, the following bound holds for all $f \in \mathcal{F}$ with probability at least $1 - \delta > 0$:*

$$R_\mu(f \circ \phi) \leq \hat{R}^{\nabla}_{\gamma,m}(f \circ \phi) + \mathbb{E}_{c\sim p}\left[\frac{\mathrm{Lip}(\tilde{\rho}_{f(.)}(\cdot, c))}{\gamma}\mathrm{Var}_{m_c}(\phi_{\#}\mu_c)\right] + \sqrt{\frac{\log(1/\delta)}{2m}},$$

where $\mathrm{Lip}(\tilde{\rho}_{f(.)}(\cdot, c)) = \sup_{x,x'\in\mathcal{X}} \frac{|\tilde{\rho}_f(\phi(x),c) - \tilde{\rho}_f(\phi(x'),c)|}{||\phi(x)-\phi(x')||}$ *is the Lipschitz constant defined w.r.t ϕ.*

Proof. It is enough to show that:

$$R_\mu(f \circ \phi) \leq \mathbb{E}_{(x,y)}[L_\gamma(\tilde{\rho}_f(\phi(x), y))],$$

and the rest of the proof is the same as in Theorem 2. For any $\xi(x, y) > 0$, and $\gamma > 0$

$$\begin{aligned}
R_\mu(f \circ \phi) &= \mathbb{P}_{(x,y)}(\arg\max_c f_c(\phi(x)) \neq y) \\
&\leq \mathbb{P}(f_y(\phi(x)) - \max_{y'\neq y} f_{y'}(\phi(x)) \leq 0) \\
&= \mathbb{P}\left(\frac{f_y(\phi(x)) - \max_{y'\neq y} f_{y'}(\phi(x))}{\xi(x)} \leq 0\right) \\
&\leq \mathbb{E}\left[\mathbb{1}_{\frac{f_y(\phi(x)) - \max_{y'\neq y} f_{y'}(\phi(x))}{\xi(x,y)} \leq 0}\right] \\
&\leq \mathbb{E}\left[L_\gamma\left(\frac{f_y(\phi(x)) - \max_{y'\neq y} f_{y'}(\phi(x))}{\xi(x, y)}\right)\right].
\end{aligned}$$

Setting $\xi(x, y) = \|\nabla_\phi \rho_f(\phi(x), y)\|_2 + \epsilon$, gives the result. □

8.3.3 Estimation Error of k-Variance

So far, our bounds have used the k-variance, which is an expectation. Lemma 8.5 bounds the estimation error when estimating the k-variance from data. This may

be viewed as a generalization error for the learned features in terms of k-variance, motivated by the connections of k-variance to test error.

Lemma 8.5 (Estimation Error of k-Variance / "Generalization Error" of the Encoder)**.** *Given a distribution μ and n empirical samples $\{S^j, \tilde{S}^j\}_{j=1}^n$ where each $S^j, \tilde{S}^j \sim \mu^k$, define the empirical Wasserstein-1 k-variance: $\widehat{\mathrm{Var}}_{k,n}(\phi_\#\mu) = \frac{1}{n}\sum_{j=1}^n \mathcal{W}_1(\phi_\#\mu_{S^j}, \phi_\#\mu_{\tilde{S}^j})$. Suppose the encoder satisfies $\sup_{x,x'} \|\phi(x) - \phi(x')\| \le B$, then with probability at least $1-\delta > 0$, we have*

$$\mathrm{Var}_k(\phi_\#\mu) \le \widehat{\mathrm{Var}}_{k,n}(\phi_\#\mu) + \sqrt{\frac{2B^2\log(1/\delta)}{nk}}.$$

Proof. We would like to estimation to the k-variance with $\widehat{\mathrm{Var}}_k(\phi_\#\mu) = \frac{1}{n}\sum_{j=1}^n \mathcal{W}_1(\phi_\# p_{S^j}, \phi_\# p_{\tilde{S}^j})$ as a function of the nk independent samples from which it is computed, each sample being a pair $(x_i, \tilde{x}_i)$. To apply the McDiarmid's Inequality, we have to examine the stability of the empirical k-variance. The Kantorovich–Rubinstein duality gives us the general formula of $\mathcal{W}1$ distance:

$$\mathcal{W}_1(P, Q) = \sup_{\mathrm{Lip}(f)\le 1} \mathbb{E}_P[f] - \mathbb{E}_Q[f]$$

In our case, separately for each j, we can write

$$\mathcal{W}_1(\phi_\# p_{S^j}, \phi_\# p_{\tilde{S}^j}) = \sup_{\mathrm{Lip}(f)\le 1} \frac{1}{k}\sum_{\ell=1}^k (f(\phi(x_\ell^j)) - f(\phi(\tilde{x}_\ell^j))).$$

Recall that the $(x_\ell, \tilde{x}_\ell)$ are independent across ℓ and j. Consider replacing one of the elements $(x_i^j, \tilde{x}_i^j)$ with some $(x'^j_i, \tilde{x}'^j_i)$, forming $p_{\bar{S}^j}$ and $p_{\bar{\tilde{S}}^j}$. Since the $(x_\ell^j, \tilde{x}_\ell^j)$ are identically distributed, by symmetry we can set $i = 1$. We then bound

$$\begin{aligned}
&\mathcal{W}_1(\phi_\# p_{S^j}, \phi_\# p_{\tilde{S}^j}) - \mathcal{W}_1(\phi_\# p_{\bar{S}^j}, \phi_\# p_{\bar{\tilde{S}}^j}) \\
&= \sup_{\mathrm{Lip}(f)\le 1} \frac{1}{k}\left((f(\phi(x_1^j)) - f(\phi(\tilde{x}_1^j))) + \sum_{\ell=2}^k (f(\phi(x_\ell^j)) - f(\phi(\tilde{x}_\ell^j))) \right)
\end{aligned}$$

$$
\begin{aligned}
&\quad - \sup_{\mathrm{Lip}(f)\leq 1} \frac{1}{k}\left((f(\phi(x_1'^j)) - f(\phi(\tilde{x}_1'^j))) + \sum_{\ell=2}^{k} (f(\phi(x_\ell^j)) - f(\phi(\tilde{x}_\ell^j))) \right) \\
&\leq \frac{1}{k} \sup_{\mathrm{Lip}(f)\leq 1} \Big(f(\phi(x_1^j)) - f(\phi(\tilde{x}_1^j)) + f(\phi(x_1'^j)) - f(\phi(\tilde{x}_1'^j)) \Big) \\
&\leq \frac{1}{k} \sup_{\mathrm{Lip}(f)\leq 1} \Big(f(\phi(x_1^j)) - f(\phi(\tilde{x}_1^j)) \Big) + \frac{1}{k} \sup_{\mathrm{Lip}(f)\leq 1} \Big(f(\phi(x_1'^j)) - f(\phi(\tilde{x}_1'^j)) \Big) \\
&\leq \frac{\|\phi(x_1^j) - \phi(x_1'^j)\| + \|\phi(x_1'^j) - \phi(\tilde{x}_1'^j)\|}{k}, \\
&\leq \frac{2B}{k}
\end{aligned}
$$

where we have used in the third inequality the fact that the sup is a contraction ($\sup_h A(h) - \sup_h B(h) \leq \sup_h (A(h) - B(h))$), and the definition of the Lipschitzity in the fourth inequality. By symmetry and scaling the right hand side with $\frac{1}{n}$, we have :

$$
\left| \frac{1}{n} \sum_{j=1}^{n} \mathcal{W}_1(\phi_{\#} p_{S^j}, \phi_{\#} p_{\tilde{S}^j}) - \frac{1}{n} \sum_{j=1}^{n} \mathcal{W}_1(\phi_{\#} p_{\bar{S}^j}, \phi_{\#} p_{\bar{\tilde{S}}^j}) \right| \leq \frac{2B}{kn}.
$$

We are now ready to apply the McDiarmid Inequality with nk samples, which yields:

$$
\mathbb{P}(\mathrm{Var}_k(\phi_{\#}\mu) - \widehat{\mathrm{Var}}_{k,n}(\phi_{\#}\mu) \geq t) \leq \exp\left(\frac{-t^2 nk}{2B^2} \right).
$$

Setting the probability to be less than δ and solving for t, we can see that this probability is less than δ if and only if $t \geq \sqrt{\frac{2B^2 \log(1/\delta)}{nk}}$. Therefore, with probability at least $1 - \delta$,

$$
\mathbb{E}_{S,\tilde{S}}[\mathcal{W}_1(\phi_{\#} p_S, \phi_{\#} p_{\tilde{S}})] \leq \frac{1}{n} \sum_{j=1}^{n} \mathcal{W}_1(\phi_{\#} p_{S^j}, \phi_{\#} p_{\tilde{S}^j}) + \sqrt{\frac{2B^2 \log(1/\delta)}{nk}}.
$$

□

We can then combine Lemma 8.5 with our margin bounds to obtain full generalization bounds. The following corollary states the empirical version of Theorem 8.2:

Corollary 8.6. *Given the setting in Theorem 8.2 and Lemma 8.5, with probability*

at least $1-\delta$*, for* $m = \sum_{c=1}^{K} \lfloor \frac{m_c}{2n} \rfloor$*,* $R_\mu(f \circ \phi)$ *is upper bounded by*

$$\hat{R}_{\gamma,m}(f \circ \phi) + \mathbb{E}_{c \sim p_y} \left[\frac{\mathrm{Lip}(\rho_f(\cdot, c))}{\gamma} \left(\widehat{\mathrm{Var}}_{\lfloor \frac{m_c}{2n} \rfloor, n}(\phi_\# \mu_c) + 2B \sqrt{\frac{\log(2K/\delta)}{n \lfloor \frac{m_c}{2n} \rfloor}} \right) \right] + \sqrt{\frac{\log(\frac{2}{\delta})}{2m}}.$$

Note that the same result holds for the gradient normalized margin $\tilde{\rho}_f$. The proof of this corollary is a simple application of a union bound on the concentration of the k-variance for each class and the concentration of the empirical risk. We end this section by bounding the variance of the empirical k-variance. While Solomon et al. [175] proved a high-probability concentration result using McDiarmid's inequality, we here use the Efron-Stein inequality to directly bound the variance.

Theorem 8.7 (Empirical variance)**.** *Given a distribution* μ *and an encoder* ϕ*, we have*

$$\mathrm{Var}\left[\widehat{\mathrm{Var}}_{k,n}(\phi_\# \mu) \right] \leq \frac{4 \mathrm{Var}_\mu(\phi(X))}{nk},$$

where $\mathrm{Var}_\mu(\phi(X)) = \mathbb{E}_{x \sim \mu}[||\phi(x) - \mathbb{E}_{x \sim \mu} \phi(x)||^2]$ *is the variance of* $\phi_\# \mu$*.*

Proof. We use the Efron Stein inequality:

Lemma 8.8 (Efron Stein Inequality)**.** *Let* $X := (X_1, \ldots, X_m)$ *be an* m*-tuple of* $\mathcal{X}$*-valued independent random variables, and let* X_i' *be independent copies of* X_i *with the same distribution. Suppose* $g : \mathcal{X}^m \to \mathbb{R}$ *is a map, and define* $X^{(i)} = (X_1, \ldots, X_{i-1}, X_i', X_{i+1} \ldots X_m)$*. Then*

$$\mathrm{Var}(g(X)) \leq \frac{1}{2} \sum_{i=1}^{m} \mathbb{E}\left[(g(X) - g(X^{(i)}))^2 \right]. \tag{8.11}$$

Consider $\frac{1}{n} \sum_{j=1}^{n} \mathcal{W}_1(\hat{\mu}_k^j, \hat{\mu}_k'^j)$ as a function of the nk independent samples from which it is computed, each sample being a pair (x_i^j, y_i^j). Using Kantorovich–Rubinstein

duality, we have the general formula:

$$\mathcal{W}_1(P,Q) = \sup_{\|f\|_{\text{Lip}} \leq 1} \mathbb{E}_P[f] - \mathbb{E}_Q[f]$$

where $\|\cdot\|_{\text{Lip}}$ is the Lipschitz norm. In our case, separately for each j, we can write

$$\mathcal{W}_1(\hat{\mu}_k^j, \hat{\mu}_k'^j) = \mathcal{W}_1\left(\frac{1}{k}\sum_{\ell=1}^k \delta_{x_\ell^j}, \frac{1}{k}\sum_{\ell=1}^k \delta_{y_\ell^j}\right) = \sup_{\|f\|_{\text{Lip}} \leq 1} \frac{1}{k}\sum_{\ell=1}^k (f(x_\ell^j) - f(y_\ell^j)).$$

Recall that the (x_ℓ^j, y_ℓ^j) are independent across ℓ and j. Consider replacing one of the elements (x_i^j, y_i^j) with some $(x_i'^j, y_i'^j)$, forming $\bar{\mu}_k^j$ and $\bar{\mu}_k'^j$. Since the (x_ℓ^j, y_ℓ^j) are identically distributed, by symmetry we can set $i = 1$. We then bound

$$\begin{aligned}
\mathcal{W}_1(\hat{\mu}_k^j, \hat{\mu}_k'^j) - \mathcal{W}_1(\bar{\mu}_k^j, \bar{\mu}_k'^j) &= \sup_{\|f\|_{\text{Lip}} \leq 1} \frac{1}{k}\left((f(x_1^j) - f(y_1^j)) + \sum_{\ell=2}^k (f(x_\ell^j) - f(y_\ell^j))\right) \\
&\quad - \sup_{\|f\|_{\text{Lip}} \leq 1} \frac{1}{k}\left((f(x_1'^j) - f(y_1'^j)) + \sum_{\ell=2}^k (f(x_\ell^j) - f(y_\ell^j))\right) \\
&\leq \frac{1}{k} \sup_{\|f\|_{\text{Lip}} \leq 1} (f(x_1^j) - f(x_1'^j)) + (f(y_1'^j) - f(y_1^j)) \\
&\leq \frac{\|x_1^j - x_1'^j\| + \|y_1^j - y_1'^j\|}{k},
\end{aligned}$$

where we have used the definition of the Lipschitz norm. By symmetry, this yields (scaling by $\frac{1}{n}$ as in the expression in the theorem)

$$\left|\frac{1}{n}\mathcal{W}_1(\hat{\mu}_k^j, \hat{\mu}_k'^j) - \frac{1}{n}\mathcal{W}_1(\bar{\mu}_k^j, \bar{\mu}_k'^j)\right| \leq \frac{\|x_1^j - x_1'^j\| + \|y_1^j - y_1'^j\|}{kn}$$

It follows that:

$$\begin{aligned}
\mathbb{E}\left[\left(\frac{1}{n}\mathcal{W}_1(\hat{\mu}_k^j, \hat{\mu}_k'^j) - \frac{1}{n}\mathcal{W}_1(\bar{\mu}_k^j, \bar{\mu}_k'^j)\right)^2\right] &\leq \frac{\mathbb{E}\left[(\|x_1^j - x_1'^j\| + \|y_1^j - y_1'^j\|)^2\right]}{k^2 n^2} \\
&= \frac{2(\mathbb{E}_{x,x' \sim \mu}\|x - x'\|^2 + (\mathbb{E}_{x,x' \sim \mu}\|x - x'\|)^2)}{k^2 n^2}
\end{aligned}$$

for each of the independent nk random variables (x_i^j, y_i^j), where we have used the fact that x_i^j and $y_i'^j$ are i.i.d. We can substitute this into the Efron Stein inequality above to obtain

$$\begin{aligned}
\mathrm{Var}\left[\widehat{\mathrm{Var}}_{k,n}(\mu)\right] &\leq \frac{\mathbb{E}_{x,x'\sim\mu}\|x-x'\|^2 + (\mathbb{E}_{x,x'\sim\mu}\|x-x'\|)^2}{kn} \\
&= \frac{2\mathrm{Var}_\mu(X) + (\mathbb{E}_{x,x'\sim\mu}\|x-x'\|)^2}{kn} \\
&\leq \frac{2\mathrm{Var}_\mu(X) + \mathbb{E}_{x,x'\sim\mu}\|x-x'\|^2}{kn} \\
&\leq \frac{4\mathrm{Var}_\mu(X)}{kn}
\end{aligned}$$

where used that the variance $\mathrm{Var}_\mu(X) = \frac{1}{2}\mathbb{E}_{x,x'\sim\mu}\|x-x'\|^2$, and Jensen inequality.

□

Theorem 8.7 implies that if the feature distribution $\phi_\#\mu$ has bounded variance, the variance of the empirical k-variance decreases as k and n increase. The values of k we used in practice were large enough that the empirical variance of k-variance was small even when we set $n = 1$.

8.4 Measuring Generalization with Normalized Margins

We now empirically compare the generalization behavior of neural networks to the predictions of our margin bounds. To provide a unified view of the bound, we set the second term in the right hand side of the bound to a constant. For instance, for Theorem 8.2, we choose $\gamma = \gamma_0 \cdot \mathbb{E}_{c\sim p}\left[\mathrm{Lip}(\rho_f(\cdot, c)) \cdot \mathrm{Var}_{m_c}(\phi_\#\mu_c)\right]$, yielding $R_\mu(f\circ\phi) \leq \hat{R}_{\gamma,m}(f\circ\phi) + 1/\gamma_0 + \mathcal{O}(m^{-1/2})$, where

$$\hat{R}_{\gamma,m}(f\circ\phi) = \hat{\mathbb{E}}_{(x,y)\sim S}\left[\mathbb{1}\left(\rho_f(\phi(x), y)/\mathbb{E}_{c\sim p}\left[\mathrm{Lip}(\rho_f(\cdot, c)) \cdot \mathrm{Var}_{m_c}(\phi_\#\mu_c)\right] \leq \gamma_0\right)\right].$$

and $\mathbb{1}(\cdot)$ is the indicator function. This implies the model generalizes better if the normalized margin is larger. We therefore consider the distribution of the k-variance

normalized margin, where each data point is transformed into a single scalar via

$$\frac{\rho_f(\phi(x), y)}{\mathbb{E}_{c\sim p}\left[\widehat{\mathrm{Var}}_{\lfloor \frac{m_c}{2} \rfloor,1}(\phi_{\#}\mu_c) \cdot \widehat{\mathrm{Lip}}(\rho_f(\cdot, c))\right]} \quad \text{and} \quad \frac{\tilde{\rho}_f(\phi(x), y)}{\mathbb{E}_{c\sim p}\left[\widehat{\mathrm{Var}}_{\lfloor \frac{m_c}{2} \rfloor,1}(\phi_{\#}\mu_c) \cdot \widehat{\mathrm{Lip}}(\tilde{\rho}_f(\cdot, c))\right]},$$

respectively. For simplicity, we set k and n as $k = \lfloor m_c/2 \rfloor$ and $n = 1$. We refer to these normalized margins as k-Variance normalized Margin (kV-Margin) and k-Variance Gradient Normalized Margin (kV-GN-Margin), respectively.

It is NP-hard to compute the exact Lipschitz constant of ReLU networks [96, 167]. Various approaches have been proposed to estimate the Lipschitz constant for ReLU networks [55, 96], however they remain computationally expensive. As we show in Appendix A.5.4, a naive spectral upper bound on the Lipschitz constant leads to poor results in predicting generalization. On the other hand, as observed by [96], a simple lower bound can be obtained for the Lipschitz constant of ReLU networks by taking the supremum of the norm of the Jacobian on the training set.[1] Letting $y^* = \arg\max_{y\neq c} f_y(\phi(x))$, the Lipschitz constant of the margin can therefore be empirically approximated as

$$\widehat{\mathrm{Lip}}(\rho_f(\cdot, c)) := \max_{x\in S_c} \|\nabla_x f_c(\phi(x)) - \nabla_x f_{y^*}(\phi(x))\|,$$

where $S_c = \{(x_i, y_i) \in S \mid y_i = c\}$ is the set of empirical samples for class c (as noted in [96], although this does not lead to correct computation of Jacobians for ReLU networks, it empirically performs well). In practice, we take the maximum over samples in the training set. We refer the reader to [138] and [190] for an analysis of the estimation error of the Lipschitz constant from finite subsets. In the supplement (App. B.3.1), we show that for piecewise linear hypotheses such as ReLU networks, the norm of the Jacobian of the gradient-normalized margin is very close to 1 almost everywhere. We thus simply set the Lipschitz constant to 1 for the gradient-normalized margin.

[1] In general, the Lipschitz constant of smooth, scalar valued functions is equal to the supremum of the norm of the input Jacobian in the domain [56, 96, 167].

	CIFAR VGG	**SVHN** NiN	**CINIC** FCN bn	**CINIC** FCN	**Flowers** NiN	**Pets** NiN	**Fashion** VGG	**CIFAR** NiN
Margin†	13.59	16.32	2.03	2.99	0.33	1.24	0.45	5.45
SN-Margin† [13]	5.28	3.11	0.24	2.89	0.10	1.00	0.49	6.15
GN-Margin 1st [90]	3.53	35.42	26.69	6.78	4.43	1.61	1.04	13.49
GN-Margin 8th [90]	0.39	31.81	7.17	1.70	0.17	0.79	2.12	1.16
TV-GN-Margin 1st [90]	19.22	36.90	31.70	16.56	4.67	4.20	0.16	**25.06**
TV-GN-Margin 8th [90]	38.18	41.52	6.59	16.70	0.43	5.65	**2.35**	10.11
kV-Margin† 1st	5.34	26.78	**37.00**	**16.93**	**6.26**	2.07	1.82	15.75
kV-Margin† 8th	30.42	26.75	6.05	15.19	0.78	1.76	0.33	2.26
kV-GN-Margin† 1st	17.95	44.57	30.61	16.02	4.48	3.92	0.61	21.20
kV-GN-Margin† 8th	**40.92**	**45.61**	6.54	15.80	1.13	**5.92**	0.29	8.07

Table 8.1: Mutual information scores on PGDL tasks. We compare different margins across tasks in PGDL. The first and second rows indicate the datasets and the architecture types used by tasks. The methods that are supported with theoretical bounds are marked with †. Our k-variance normalized margins outperform the baselines in 6 out of 8 tasks in PGDL dataset.

8.4.1 Experiment: Predicting Generalization in Deep Learning

We evaluate our margin bounds on the Predicting Generalization in Deep Learning (PGDL) dataset [92]. The dataset consists of 8 tasks, each task contains a collection of models trained with different hyperparameters. The models in the same task share the same dataset and model type, but can have different depths and hidden sizes. The goal is to find a *complexity measure* of networks that correlates with their generalization error. In particular, the complexity measure maps the model and training dataset to a real number, where the output should rank the models in the same order as the generalization error. The performance is then measured by the Conditional Mutual Information (CMI). Intuitively, CMI measures the *minimal* mutual information between complexity measure and generalization error conditioned on different sets of hyperparameters. To achieve high CMI, the measure must be robust to all possible settings including different architectures, learning rates, batch sizes, etc. Please refer to [92] for details.

Experimental Setup. We compare our k-variance normalized margins (kV-Margin and kV-GN-Margin) with Spectrally-Normalized Margin (SN-Margin) [13], Gradient-Normalized Margin (GN-Margin) [53], and total-variance-normalized GN-Margin (TV-GN-Margin) [90]. Note that the (TV-GN-Margin) of [90] corresponds to $\frac{\tilde{\rho}_f(\phi(x),y)}{\sqrt{\mathrm{Var}_{x\sim\mu}(||\phi(x)||^2)}}$. Comparing to our kV-GN-Margin, our normalization is theoretically motivated and involves the Lipschitz constants of f as well as the generalized notion of k-variance. As some of the margins are defined with respect to different layers of networks, we present the results with respect to the shallow layer (first layer) and the deep layer (8th layer if the number of convolutional layers is greater than 8, otherwise the deepest convolutional layer). To produce a scalar measurement, we use the median to summarize the margin distribution, which can be interpreted as finding the margin γ that makes the margin loss ≈ 0.5. We found that using expectation or other quantiles leads to similar results. The Wasserstein-1 distance in k-variance is computed exactly, with the linear program in the POT library [57]. All of our experiments are run on 6 TITAN X (Pascal) GPUs.

To ease the computational cost, all margins and k-variances are estimated with random subsets of size $\min(200\times\#\text{classes}, \text{data\_size})$ sampled from the training data. The average results over 4 subsets are shown in Table 8.1. Standard deviations are given in App. A.5.2, as well as the effect of varying the size of the subset in App. A.5.3 . Our k-variance normalized margins outperform the baselines in 6 out of 8 tasks. Notably, our margins are the only ones achieving good empirical performance while being supported with theoretical bounds.

Margin Visualization. To provide a qualitative comparison, we select four models from the first task of PGDL (CIFAR10/VGG), which have generalization error 24.9%, 26.2%, 28.6%, and 31.8%, respectively. We visualize margin distributions for each in Figure 8-1. Without proper normalization, Margin and GN-Margin struggle to discriminate these models. Similar to the observation in [91], SN-Margin even negatively correlates with generalization. Among all the apporaches, kV-GN-Margin is the only measure that correctly orders and distinguishes between all four models. This is consistent with Table 8.1, where kV-GN-Margin achieves the highest score.

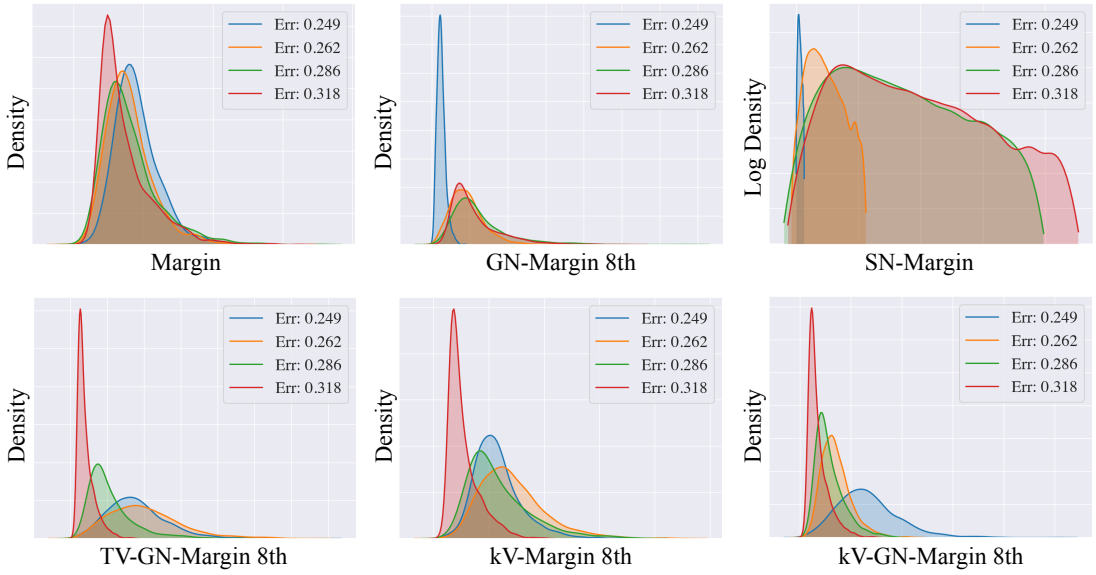


Figure 8-1: Margin Visualization of PGDL Models. From left to right, the correct order of the margin distributions should be red, green, orange, and blue. kV-GN-Margin is the only measure that behaves consistently with the generalization error.

Next, we compare our approach against the winning solution of the PGDL competition: Mixup*DBI [141]. Mixup*DBI uses the geometric mean of the Mixup accuracy [223] and Davies Bouldin Index (DBI) to predict generalization. In particular, they use DBI to measure the clustering quality of intermediate representations of neural networks. For fair comparison, we calculate the geometric mean of the Mixup accuracy and the median of the k-variance normalized margins and show the results in Table 8.2. Following [141], all approaches use the representations from the first layer. Our Mixup*kV-GN-Margin outperforms the state-of-the-art [141] in 5 out of the 8 tasks.

	CIFAR VGG	**SVHN** NiN	**CINIC** FCN bn	**CINIC** FCN	**Flowers** NiN	**Pets** NiN	**Fashion** VGG	**CIFAR** NiN
Mixup*DBI [141]	0.00	42.31	31.79	15.92	**43.99**	**12.59**	**9.24**	25.86
Mixup*kV-Margin	7.37	27.76	**39.77**	20.87	9.14	4.83	1.32	22.30
Mixup*kV-GN-Margin	**20.73**	**48.99**	36.27	**22.15**	4.91	11.56	0.51	**25.88**

Table 8.2: Mutual information scores on PGDL tasks with Mixup. We compare with the winner (Mixup*DBI) of the PGDL competition [92]. Scores of Mixup*DBI from [141].

8.4.2 Experiment: Label Corruption

A sanity check proposed in [222] is to examine whether the generalization measures are able to capture the effect of label corruption. Following the experiment setup in [222], we train two Wide-ResNets [218], one with true labels (generalization error = 12.9%) and one with random labels (generalization error = 89.7%) on CIFAR-10 [108]. Both models achieve 100% training accuracy. We select the feature from the second residual block to compute all the margins that involve intermediate features and show the results in Figure 8-2. Without k-variance normalization, margin and GN-Margin can hardly distinguish these two cases, while k-variance normalized margins correctly discriminate them.

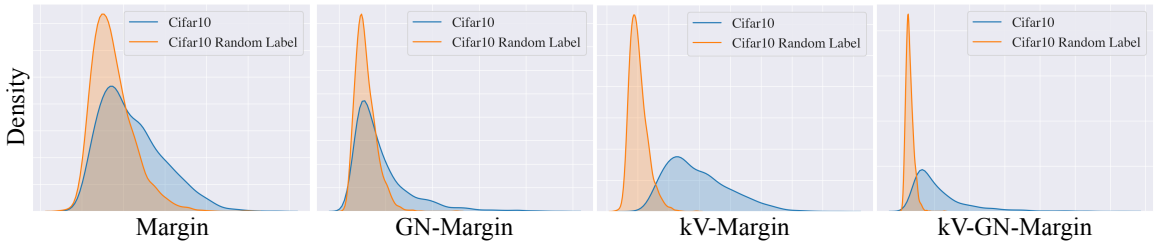


Figure 8-2: Margin distributions with clean or random labels. Without k-variance normalization, Margin and GN-Margin struggle to distinguish the models trained with clean labels or random labels.

8.4.3 Experiment: Task Hardness

Next, we demonstrate our margins are able to measure the "hardness" of learning tasks. We say that a learning task is hard if the generalization error appears large for well-trained models. Different from the PGDL benchmark, where only models trained on the same dataset are compared, we visualize the margin distributions of Wide-ResNets trained on CIFAR-10 [108], SVHN [142], and MNIST [113], which have generalization error 12.9%, 5.3%, and 0.7%, respectively. The margins are measured on the respective datasets. In Figure 8-3, we again see that k-variance normalized margins reflect the hardness better than the baselines. For instance, CIFAR-10 and SVHN are indicated to be harder than MNIST as the margins are smaller.

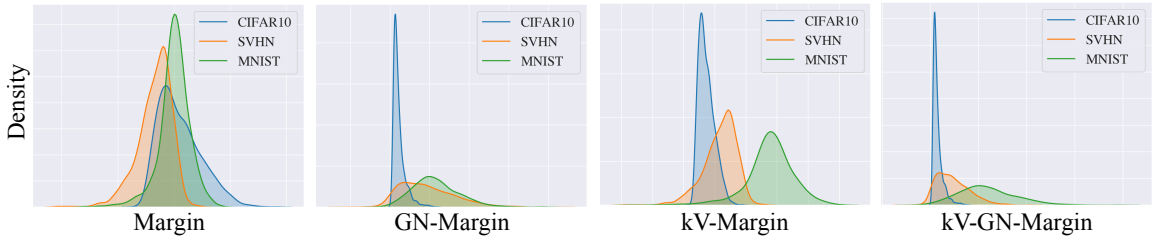


Figure 8-3: CIFAR-10, SVHN, and MNIST have different hardness. Although models achieve 100% training accuracy on each task, the test accuracy differs. With k-variance normalization, the margin distributions of the models are able to recognize the hardness of the tasks.

8.5 Concentration and Separation of Representations

8.5.1 Concentration of Representations

In this section, we study how the structural properties of the feature distributions enable fast learning. Following the Wasserstein-2 k-variance analysis in [175], we apply bounds by Weed and Bach [203] to demonstrate the fast convergence rate of Wasserstein-1 k-variance when (1) the distribution has low intrinsic dimension or (2) the support is clusterable.

Proposition 8.9. (Low-dimensional Measures, Informal) *For $\phi_{\#}\mu \in \mathrm{Prob}(\mathbb{R}^d)$, we have $\mathrm{Var}_m(\phi_{\#}\mu) \leq \mathcal{O}(m^{-1/d})$ for $d > 2$. If $\phi_{\#}\mu$ is supported on an approximately d'-dimensional set where $d' < d$, we obtain a better rate: $\mathrm{Var}_m(\phi_{\#}\mu) \leq \mathcal{O}(m^{-1/d'})$.*

We defer the complete statement to the supplement. Without any assumption, the rate gets significantly worse as the feature dimension d increases. Nevertheless, for an intrinsically d'-dimensional measure, the variance decreases with a faster rate. For clustered features, we can obtain an even stronger rate:

Proposition 8.10. (Clusterable Measures) *A distribution μ is (n, Δ)-clusterable if $\mathrm{supp}(\mu)$ lies in the union of n balls of radius at most Δ. If $\phi_{\#}\mu$ is (n, Δ)-clusterable, then for all $m \leq n(2\Delta)^{-2}$, $\mathrm{Var}_m(\phi_{\#}\mu) \leq 24\sqrt{\frac{n}{m}}$.*

We arrive at the parametric rate $\mathcal{O}(m^{-1/2})$ if the cluster radius Δ is sufficiently small. In particular, the fast rate holds for large m when the clusters are well concen-

trated. Different from conventional studies that focus on the complexity of a complete function class (such as Rademacher complexity [14]), our k-variance bounds capture the concentration of the feature distribution.

8.5.2 Separation of Representations

We showed in the previous section that the concentration of the representations is captured by the k-variance and that this notion translates the properties of the underlying probability measures into generalization bounds. Next, we show that maximizing the margin sheds light on the separation of the underlying representations in terms of Wasserstein distance.

Lemma 8.11. *(***Large Margin and Feature Separation***) Assume there exist $f_y, y = 1 \dots K$ that are L-Lipschitz and satisfy the max margin constraint $\rho_f(\phi(x), y) \geq \gamma$ for all $(x, y) \sim D$, i.e: $f_y(\phi(x)) \geq f_{y'}(\phi(x)) + \gamma,\ \forall\ y' \neq y, \forall x \in supp(\mu_y)$. Then $\forall y \neq y'$, $\mathcal{W}_1(\phi_\#(\mu_y), \phi_\#(\mu_{y'})) \geq \frac{\gamma}{L}$.*

Proof. Since f_y and $f_{y'}$ are L lipchitz, it follows that $g(z) = f_y(z) - f_{y'}(z)$ is $2L$ Lipchitz, and hence $\frac{g}{2L}$ is Lip_1.

$$
\begin{aligned}
\mathcal{W}_1(\phi_\#(\mu_y), \phi_\#(\mu_{y'})) &= \sup_{f \in Lip_1} \mathbb{E}_{x \sim p_y} f(\phi(x)) - \mathbb{E}_{x \sim \mu_{y'}} f(\phi(x)) \\
&\geq \frac{1}{2L} \Big(\mathbb{E}_{x \sim p_y} g(\phi(x)) - \mathbb{E}_{x \sim \mu_{y'}} g(\phi(x)) \Big) \\
&= \frac{1}{2L} \Big(\mathbb{E}_{x \sim p_y} [f_y(\phi(x)) - f_{y'}(\phi(x))] + \mathbb{E}_{x \sim \mu_{y'}} [f_{y'}(\phi(x)) - f_y(\phi(x))] \Big) \\
&\geq \frac{1}{2L} (2\gamma) \ \text{(By Assumption on } f) \\
&= \frac{\gamma}{L}
\end{aligned}
$$

□

Lemma 8.11 states that large margins imply Wasserstein separation of the representations of each class. It also sheds light on the Lipschitz constant of the downstream classifier $\mathcal{F}$: $L \geq \gamma / \min_{y,y'} \mathcal{W}_1(\phi_\#(\mu_y), \phi_\#(\mu_{y'}))$. One would need more complex classifiers, i.e, those with a larger Lipschitz constant, to correctly classify classes

that are close to other classes, by Wasserstein distance, in feature space. We further relate the margin loss to Wasserstein separation:

Lemma 8.12. *Define the pairwise margin loss $R_\gamma^{y,y'}$ for $y, y' \in \mathcal{Y}$ as*

$$R_\gamma^{y,y'}(f \circ \phi) = \frac{1}{2}\left(\mathbb{E}_{x\sim\mu_y}[\gamma - f_y(\phi(x)) + f_{y'}(\phi(x))]_+ + \mathbb{E}_{x\sim\mu_{y'}}[\gamma - f_{y'}(\phi(x)) + f_y(\phi(x))]_+\right).$$

Assume f_c is L-Lipschitz for all $c \in \mathcal{Y}$. Given a margin $\gamma > 0$, for all $y \neq y'$, we have:

$$\mathcal{W}_1(\phi_\#(\mu_y), \phi_\#(\mu_{y'})) \geq \frac{1}{L}\left(\gamma - R_\gamma^{y,y'}(f \circ \phi)\right).$$

Proof of Lemma 8.12. We follow the same notation of the proof above but we don't make any assumption on $f_y, f_{y'}$ except that they are Lip_L:

$$\begin{aligned}
&\mathcal{W}_1(\phi_\#(\mu_y), \phi_\#(\mu_{y'})) = \sup_{f\in Lip_1} \mathbb{E}_{x\sim p_y} f(\phi(x)) - \mathbb{E}_{x\sim\mu_{y'}} f(\phi(x)) \\
&\geq \frac{1}{2L}\left(\mathbb{E}_{x\sim p_y} g(\phi(x)) - \mathbb{E}_{x\sim\mu_{y'}} g(\phi(x))\right) \\
&= \frac{1}{2L}\left(\int [f_y(z) - f_{y'}(z))] d\phi_\#(\mu_y)(z) + \int [f_{y'}(z) - f_y(z)] d\phi_\#(\mu_{y'})(z)\right) \\
&= \frac{1}{2L}\left(\gamma - \int [\gamma - (f_y(z) - f_{y'}(z))] d\phi_\#(\mu_y)(z) + \gamma - \int [\gamma - (f_{y'}(z) - f_y(z))] d\phi_\#(\mu_{y'})(z)\right) \\
&\geq \frac{1}{2L}\left(2\gamma - \int [\gamma - (f_y(z) - f_{y'}(z))]_+ d\phi_\#(\mu_y)(z) - \int [\gamma - (f_{y'}(z) - f_y(z))]_+ d\phi_\#(\mu_{y'})(z)\right)
\end{aligned}$$

where the last inequality follows from the fact that for $t \in \mathbb{R}$, we have $t \leq [t]_+ = \max(t, 0)$ Hence we have:

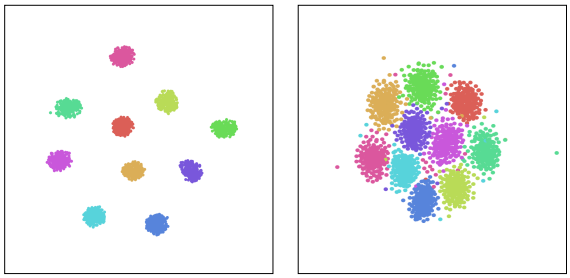
$$\begin{aligned}
&\mathcal{W}_1(\phi_\#(\mu_y), \phi_\#(\mu_{y'})) \\
&\geq \frac{1}{L}\left(\gamma - \frac{1}{2}\left(\mathbb{E}_{\mu_y}[\gamma - f_y(\phi(x)) + f_{y'}(\phi(x))]_+ + \mathbb{E}_{\mu_{y'}}[\gamma - f_{y'}(\phi(x)) + f_y(\phi(x))]_+\right)\right).
\end{aligned}$$

□

We show a similar relation for the gradient-normalized margin [53] in the supplement (Lemma B.9): gradient normalization results in a robust Wasserstein separation

of the representations, making the feature separation between classes robust to adversarial perturbations.

Example: Clean vs. Random Labels. Finally, we provide an illustrative example on how concentration and separation are associated with generalization. For the label corruption setting from Section 8.4.2, Figure 8-4 shows t-SNE visualizations [192] of the representations learned with true or random labels on CIFAR-10. Training with clean labels leads to well-clustered representations. Although the model trained with random labels has 100% training accuracy, the resulting feature distribution is less concentrated and separated, implying worse generalization.



(a) Clean Labels (b) Random Labels

Figure 8-4: t-SNE visualization of representations. Classes are indicated by colors.

Discussion The notions of concentration and separation are tightly related to self-supervised learning with contrastive loss [24, 36, 81, 99]. The key idea of contrastive learning is to minimize the variance of features within the class, while maximize the distance for features between classes [200]. Recently, Khosla et al. [99] show that contrastive loss outperforms the standard cross entropy loss in several classification datasets. We hypothesize the empirical success of contrastive learning is a result of concentration and separation in the feature space.

For example, in the preceding chapters, both Figure 3-4 and 5-4 demonstrate that the features derived from contrastive learning exhibit high degrees of concentration and separation. Furthermore, upon eliminating the false negatives and false positives, this trend of distinct concentration and separation becomes even more pronounced.

8.6 Conclusion

In this chapter, we present k-variance normalized margin bounds, a new data-dependent generalization bound based on optimal transport. The proposed bounds predict the generalization error well on the large scale PGDL dataset [92]. We use our theoretical bounds to shed light on the role of the feature distribution in generalization.

Chapter 9

Epilogue

Since the rise of deep learning, neural networks have been greatly dominant, marking a significant paradigm shift in the field. The landscape has further evolved with the advent of self-supervised models, a change that is gradually redefining the rules of the game. These models, trained from large-scale datasets, have become an indispensable backbone for numerous applications. The evolution has also given birth to new foundation models like Segment Anything [103], Stable Diffusion [164], and more, effectively ushering the field of artificial intelligence into the next stage of its development.

Concurrently, the scale of data and complexity has grown exponentially, moving towards multimodal learning that encompasses vision, text, audio, motion, etc [66]. As a result, the datasets have become noisier and the computational expense of training these models has increased substantially, but this is where our works shine! For instance, a recent works by Radenovic et al. [158], demonstrate that our algorithms continue to improve performance, even in the era of foundation models.

Looking ahead, it is almost certain that the size of datasets and models will continue to increase exponentially. This growth, however, raises important questions: Will the benefits derived from our algorithms diminish as we scale up datasets into the hundreds or thousands of billions? Can we still control errors and biases at such a scale and construct a reliable and responsible AI system? The mere modification of algorithmic components, such as adjusting the loss function, may prove insufficient

in this context. Specifically, there could be a compelling need to curate specialized datasets for fine-tuning, ensuring that model behaviors are both corrected and aligned with intended outcomes.

Beyond the practical ramifications of scale, there are also implications for theoretical analysis. The pertinence of theoretical analysis might be called into question when confronted with overwhelmingly large sample sizes. Notably, as the sample size tends towards infinity, the generalization error correspondingly approaches a minimum. Consequently, a shift in focus might be necessitated: Instead of merely discussing generalization within an i.i.d. setting, it may become more pertinent to explore extrapolation. This includes deliberations on the types of distributions to which models can extrapolate, or the types of regularization that empower models to learn such extrapolations.

While there is still a great deal of uncertainty regarding the future of AI, it is my hope that this dissertation can serve as a stepping stone toward the development of a more reliable and robust AI system. The future of artificial intelligence remains promising, albeit riddled with challenges that we, as a community, must tackle head-on.

Appendix A

Additional Experiments and Details

A.1 False Negative Pairs: DCL

A.1.1 Experiment Details

CIFAR10 and STL10 We adopt PyTorch to implement SimCLR [24] with ResNet-50 [79] as the encoder architecture and use the Adam optimizer [100] with learning rate 0.001 and weight decay $1e-6$. We set the temperature t as 0.5 and the dimension of the latent vector as 128. All the models are trained for 400 epochs. The data augmentation uses the following PyTorch code:

The models are evaluated by training a linear classifier with cross entropy loss after fixing the learned embedding. We again use the Adam optimizer with learning rate 0.001 and weight decay $1e-6$.

Imagenet-100 We adopt the official code[1] of contrastive multiview coding (CMC) [186]. To implement the debiased objective, we only modify the "NCE/NCECriterion.py" file and left the rest of the code unchanged. The temperature of CMC is set to 0.07, which often makes the estimator $\frac{1}{\tau^-}\left(\frac{1}{N}\sum_{i=1}^{N} e^{f(x)^T f(u_i)} - \tau^+ \frac{1}{M}\sum_{i=1}^{M} e^{f(x)^T f(v_i)}\right)$ less than $e^{-1/t}$. To retain the learning signal, when the estimator is less than $e^{-1/t}$, we will optimize the biased loss instead. This improves the convergence and stability

[1] https://github.com/HobbitLong/CMC/

of our method.

Sentence Embedding We adopt the official code[2] of quick-thought (QT) vectors [119]. To implement the debiased objective, we only modify the "src/s2v-model.py" file and left the rest of the code unchanged. Since the official BookCorpus [104] dataset is missing, we use the unofficial version [3] for the experiments. The feature vector of QT is not normalized, therefore, we simply constrain the estimator described in equation (7) to be greater than zero.

Reinforcement Learning We adopt the official code[4] of Contrastive unsupervised representations for reinforcement learning (CURL) [178]. To implement the debiased objective, we only modify the "curl-sac.py" file and left the rest of the code unchanged. We again constrain the estimator described in equation (7) to be greater than zero since the feature vector of CURL is not normalized.

A.2 Hard Negative Pairs: HCL

A.2.1 Graph Representation Learning

We describe in detail the hard sampling method for graphs whose results are reported in Section 4.2.3. Before getting that point, in the interests of completeness we cover some required background details on the InfoGraph method of [181]. For further information see the original paper [181].

Background on Graph Representations

We observe a set of graphs $\mathbf{G} = \{G_j \in \mathbb{G}\}_{j=1}^{n}$ sampled according to a distribution p over an ambient graph space $\mathbb{G}$. Each node u in a graph G is assumed to have features $h_u^{(0)}$ living in some Euclidean space. We consider a K-layer graph neural network,

[2]https://github.com/lajanugen/S2V
[3]https://github.com/soskek/bookcorpus
[4]https://github.com/MishaLaskin/curl

whose k-th layer iteratively computes updated embeddings for each node $v \in G$ in the following way,

$$h_v^{(k)} = \text{COMBINE}^{(k)}\left(h_v^{(k-1)}, \text{AGGREGATE}^{(k)}\left(\left\{\left(h_v^{(k-1)}, h_u^{(k-1)}, e_{uv}\right) : u \in \mathcal{N}(v)\right\}\right)\right)$$

where $\text{COMBINE}^{(k)}$ and $\text{AGGREGATE}^{(k)}$ are parameterized learnable functions and $\mathcal{N}(v)$ denotes the set of neighboring nodes of v. The K embeddings for a node u are collected together to obtain a single final summary embedding for u. As recommended by [213] we use concatenation, $h^u = h^u(G) = \text{CONCAT}\left(\{h_u^{(k)}\}_{k=1}^K\right)$ to obtain an embedding in $\mathbb{R}^d$. Finally, the node representations are combined together into a length d graph level embedding using a readout function,

$$H(G) = \text{READOUT}\left(\{h^u\}_{u \in G}\right)$$

which is typically taken to be a simple permutation invariant function such as the sum or mean. The InfoGraph method aims to maximize the mutual information between the graph level embedding $H(G)$ and patch-level embeddings $h^u(G)$ using the following objective,

$$\max_h \mathbb{E}_{G \sim p} \frac{1}{|G|} \sum_{u \in G} I\left(h^u(G); H(G)\right)$$

In practice the population distribution p is replaced by its empirical counterpart, and the mutual information I is replaced by a variational approximation I_T. In line with [181] we use the Jensen-Shannon mutual information estimator as formulated by [**?**]. It is defined using a neural network discriminator $T : \mathbb{R}^{2d} \to \mathbb{R}$ as,

$$I_T\left(h^u(G); H(G)\right) = \mathbb{E}_{G \sim p}\left[-\text{sp}(-T\left(h^u(G), H(G)\right))\right] - \mathbb{E}_{(G,G') \sim p \times p}\left[\text{sp}(T\left(h^u(G), H(G')\right))\right]$$

where $\text{sp}(z) = \log(1 + e^z)$ denotes the softplus function. The finial objective is the

joint maximization over h and T,

$$\max_{\theta,\psi} \mathbb{E}_{G\sim p} \frac{1}{|G|} \sum_{u\in G} I_T\left(h^u(G); H(G)\right)$$

Hard Negative Sampling for Learning Graph Representations

In order to derive a simple modification of the NCE hard sampling technique that is appropriate for use with InfoGraph, we first provide a mildly generalized view of hard sampling. Recall that the NCE contrastive objective can be decomposed into two constituent pieces,

$$\mathcal{L}(f, q) = \mathcal{L}_{\text{align}}(f) + \mathcal{L}_{\text{unif}}(f, q)$$

where q is in fact a family of distributions $q(x^-; x)$ over x^- that is indexed by the possible values of the anchor x. $\mathcal{L}_{\text{align}}$ performs the role of "aligning" positive pairs (embedding near to one-another), while $\mathcal{L}_{\text{unif}}$ repels negative pairs. The hard sampling framework aims to solve,

$$\inf_f \sup_q \mathcal{L}(f, q).$$

In the case of NCE loss we take,

$$\begin{aligned}
\mathcal{L}_{\text{align}}(f) &= -\mathbb{E}_{\substack{x\sim p \\ x^+\sim p_x^+}} f(x)^T f(x^+), \\
\mathcal{L}_{\text{unif}}(f, q) &= \mathbb{E}_{\substack{x\sim p \\ x^+\sim p_x^+}} \log\left\{e^{f(x)^T f(x^+)} + Q\mathbb{E}_{x^-\sim q}[e^{f(x)^T f(x^-)}]\right\}.
\end{aligned}$$

View this view, we can easily adapt to the InfoGraph framework, taking

$$\begin{aligned}
\mathcal{L}_{\text{align}}(h, T) &= -\mathbb{E}_{G\sim p} \frac{1}{|G|} \sum_{u\in G} \text{sp}(-T\left(h^u(G), H(G)\right)), \\
\mathcal{L}_{\text{unif}}(h, T, q) &= -\mathbb{E}_{G\sim p} \frac{1}{|G|} \sum_{u\in G} \mathbb{E}_{G'\sim q} \text{sp}(T\left(h^u(G), H(G')\right))
\end{aligned}$$

Denote by $\hat{p}$ the distribution over nodes $u \in \mathbb{R}^s$ defined by first sampling $G \sim p$, then sampling $u \in G$ uniformly over all nodes of G. Then these two terms can be simplified to

$$\mathcal{L}_{\text{align}}(h, T) = -\mathbb{E}_{u\sim\hat{p}}\text{sp}(-T\left(h^u(G), H(G)\right)),$$
$$\mathcal{L}_{\text{unif}}(h, T, q) = -\mathbb{E}_{(u,G')\sim\hat{p}\times q}\text{sp}(T\left(h^u(G), H(G')\right))$$

At this point it becomes clear that, just as with NCE, a distribution $q^* \in \arg\max_q \mathcal{L}(f, q)$ in the InfoGraph framework if it is supported on $\arg\max_{G'\in\mathbb{G}} \text{sp}(T\left(h^u(G), H(G')\right))$. Although this is still hard to compute exactly, it can be approximated by,

$$q_u^\beta(G') \propto \exp\left(\beta T(h^u(G), H(G))\right) \cdot p(G').$$

A.2.2 Experimental Details

Visual Representations We implement SimCLR in PyTorch. We use a ResNet-50 [79] as the backbone with embedding dimension 2048 (the representation used for linear readout), and projection head into the lower 128-dimensional space (the embedding used in the contrastive objective). We use the Adam optimizer [100] with learning rate 0.001 and weight decay 10^{-6}. Since we adopt the SimCLR framework, the number of negative samples $N = 2(\text{batch size} - 1)$. Since we always take the batch size to be a power of 2 $(16, 32, 64, 128, 256)$ the negative batch sizes are $30, 62, 126, 254, 510$ respectively. Unless otherwise stated, all models are trained for 400 epochs.

Annealing β Method: We detail the annealing method whose results are given in Figure A-1. The idea is to reduce the concentration parameter down to zero as training progresses. Specifically, suppose we have e number of total training epochs. We also specify a number ℓ of "changes" to the concentration parameter we shall make. We initialize the concentration parameter $\beta_1 = \beta$ (where this β is the number

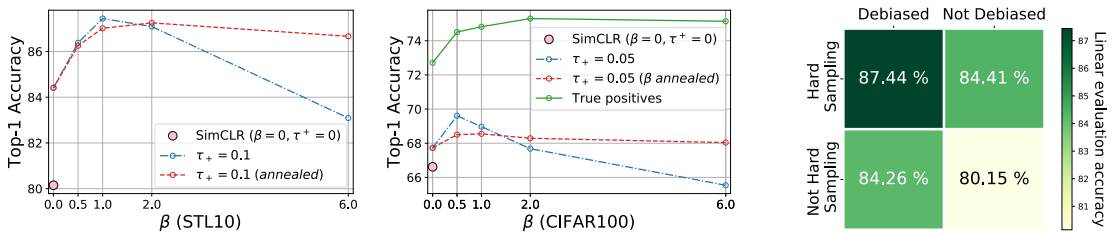


Figure A-1: Left: the effect of varying concentration parameter β on linear readout accuracy. Middle: linear readout accuracy as concentration parameter β varyies, in the case of contrastive learning (fully unsupervised), using true positive samples (uses label information), and an annealing method that improves robustness to the choice of β (see Appendix A.2.2 for details). Right: STL10 linear readout accuracy for hard sampling with and without debiasing, and non-hard sampling ($\beta = 0$) with and without debiasing. Best results come from using both simultaneously.

reported in Figure A-1), then once every e/ℓ epochs we reduce β_i by β/ℓ. In other words, if we are currently on β_i, then $\beta_{i+1} = \beta_i - \beta/\ell$, and we switch from β_i to β_{i+1} in epoch number $i \cdot e/\ell$.

The idea of this method is to select particularly difficult negative samples early on order to obtain useful gradient information early on, but later (once the embedding is already quite good) we reduce the "hardness" level so as to reduce the harmful effect of only approximately correcting for false negatives (negatives with the same labels as the anchor). We also found the annealing in the opposite direction ("down") achieved similar performance.

A.3 False Positive Pairs: RINCE

A.3.1 Exact Number of CIFAR-10 and ACAV100M Experiments

We first provide the exact numbers for CIFAR-10 and ACAV100M experiments in Table A.1, A.2, and A.3.

η	InfoNCE	$q = 0.01$	$q = 0.1$	$q = 0.5$	$q = 1.0$
0.0	93.4±0.2	93.4±0.2	93.2±0.1	93.3±0.1	93.0±0.2
0.2	93.1±0.1	93.3±0.3	93.0±0.1	93.2±0.2	92.9±0.3
0.4	90.7±0.2	93.0±0.2	92.0±0.9	93.1±0.1	92.8±0.1
0.6	88.2±0.4	90.8±0.2	90.6±0.3	92.9±0.2	92.4±0.2
0.8	87.1±0.5	89.1±0.2	89.3±0.1	89.9±0.3	91.6±0.3
1.0	87.1±1.0	88.7±0.1	89.3±0.4	89.3±0.6	88.2±0.3

Table A.1: CIFAR-10 Label Noise

η	InfoNCE	$q = 0.01$	$q = 0.1$	$q = 0.5$	$q = 1.0$
0.0	91.1±0.1	91.6±0.1	91.5±0.1	91.8±0.2	90.7±0.1
0.2	89.3±0.1	89.8±0.2	89.7±0.1	90.4±0.1	90.9±0.1
0.4	87.3±0.4	87.7±0.5	87.5±0.2	88.8±0.1	89.0±0.1
0.6	84.5±0.2	85.4±0.2	85.3±0.2	86.6±0.1	86.3±0.2
0.8	80.6±0.1	81.2±0.2	80.3±0.2	82.5±0.2	82.8±0.3
1.0	71.0±0.5	71.2±0.6	71.8±0.4	71.5±0.3	72.7±0.2

Table A.2: CIFAR-10 Augmentation Noise

model	20K	50K	100K	200K	500K
InfoNCE (100 epoch)	72.482	75.205	77.161	79.937	82.717
InfoNCE (150 epoch)	72.429	76.13	78.8	80.095	83.082
InfoNCE (200 epoch)	72.429	76.183	78.641	79.94	83.388
RINCE (100 epoch)	73.635	76.685	78.694	81.153	83.505
RINCE (150 epoch)	74.632	77.505	79.064	82.263	83.399
RINCE (200 epoch)	74.253	78.086	79.355	82.368	83.769

Table A.3: Top1 accuracy on UCF101 of models trained on ACAV100M.

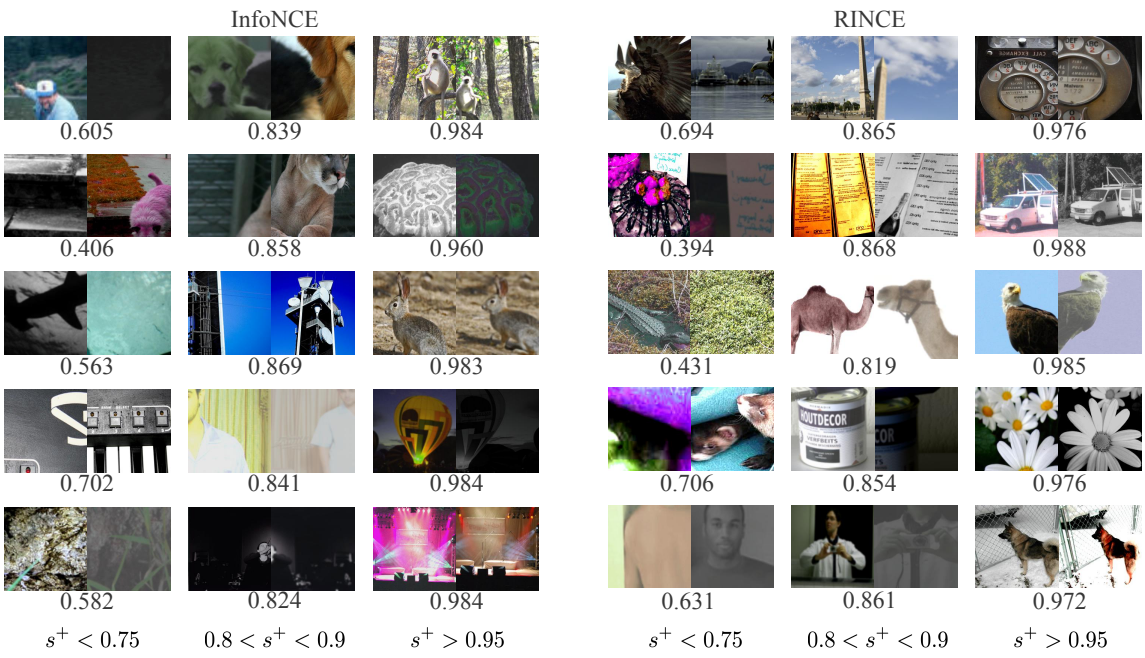
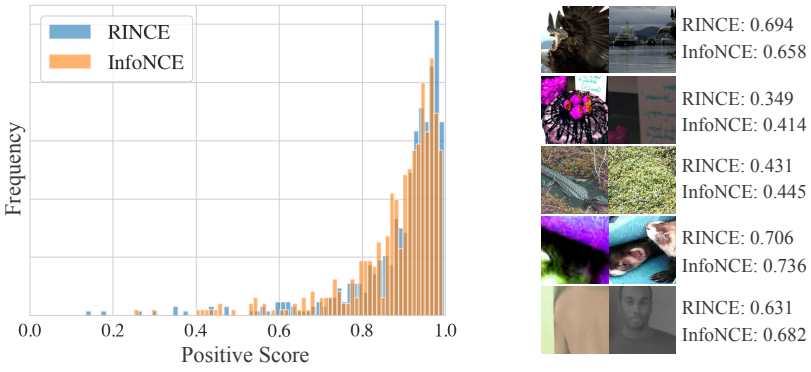


Figure A-2: Positive pairs and their scores. The corresponding positive scores are shown below the image pairs. The positive scores $s^+ \in [-1, 1]$ are output by the trained InfoNCE and RINCE model (temperature $= 1$). Pairs that have lower scores are visually noisy, while informative pairs often have higher scores.

A.3.2 Positive Scores and Views, Continue

We extend our analysis of Figure 5-5 to InfoNCE baseline and discuss the impact of implicit weighting. We can see that the positive scores in both InfoNCE and RINCE models are correlated to the noisiness of positive pairs.

We then study the distribution of positive scores and compare the positive scores output by InfoNCE and RINCE on noisy views. As Figure A-3 (a) shows, the positive scores of clean pairs output by RINCE is slightly higher, making the density of RINCE around score 1.0 larger than InfoNCE. Figure A-3 (b) gives a closer look on scores versus noisy views. We can see that InfoNCE tends to output higher scores for noisy views than RINCE, corroborating our analysis: InfoNCE tends to maximize the positive score of hard (noisy) pairs. This inherently makes the positive scores of clean pairs lower for InfoNCE, explaining the discrepancy between InfoNCE and RINCE in (a).



(a) Distribution of Positive Scores (b) Noisy Views and Scores

Figure A-3: Comparison between RINCE and InfoNCE. (a) Distribution of Positive Scores for RINCE and InfoNCE; (b) InfoNCE outputs higher scores for noisy pairs.

A.3.3 Ablation Study on λ

Finally, we provide an ablation study on how λ affect the performance of RINCE with CIFAR-10 augmentation noise experiments. We can see that in both clean and noise setting, RINCE is not sensitive to the choice of λ as long as it is not too large. Therefore, we simply set $\lambda = 0.01$ for all vision experiments and $\lambda = 0.025$ for graph experiments.

Noise Rate	0.0	0.4
RINCE ($\lambda = 0.01$)	91.54	89.65
RINCE ($\lambda = 0.05$)	**91.81**	**89.81**
RINCE ($\lambda = 0.1$)	91.32	89.9
RINCE ($\lambda = 0.2$)	90.55	89.69
RINCE ($\lambda = 0.4$)	90.89	89.39

Table A.4: CIFAR-10 Augmentation Noise

A.3.4 Experiment Details

CIFAR-10 We follow the experiment setup in [36], where the SimCLR [24] models are trained with Adam optimizer for 500 epochs with learning rate 0.001 and weight decay 1e−6. The encoder is ResNet-50 and the dimension of the latent vector is 128. The temperature is set to $t = 0.5$. The models are then evaluated by training a linear classifier for 100 epochs with learning rate 0.001 and weight decay 1e−6.

ImageNet: SimCLR We adopt the SimCLR implementation[5] from PyTorch Lightning [54]. In addition, we spot a bug and fix the implementation of negative masking of PyTorch Lightning and achieve 68.9 top-1 accuracy on ImageNet (the one reported in the PyTorch Lightning's website is 68.4). To implement RINCE, we only modify the lines that calculates loss.

ImageNet: Mocov3 We adopt the official code[6] from Mocov3 [28]. To implement RINCE, we only modify the lines that calculates loss in `moco/builder.py`.

Kinetics-400 We adopt the official implementation[7] from [132]. Similarly, we only modify the loss function in the `criterions` directory. In particular, we use the SimCLR style implementation for both InfoNCE and RINCE loss. We also adopt the same hyperparameters described in the git repository for training. We set the learning rate to $1e-3$ to finetune the models on downstream classification tasks such as UCF101 and HMDB51 with the provided evaluation code.

ACAV100M We again modify the official implementation of [132] for the ACAV100M experiments, where we modify the data loader to adopt it to ACAV100M. Different from Kinetics-400 experiments, the input size is set to 8×224^2 during the finetuning process for computational efficiency. We again use the exact same set of hyperparameters from [132] for both training and testing.

TU-Dataset We adopt the official implementation[8] from [215]. To implement RINCE, we only modify the loss in `gsimclr.py` file.

[5] `https://github.com/PyTorchLightning/lightning-bolts/tree/master/pl_bolts/models/self_supervised/simclr`

[6] `https://github.com/facebookresearch/moco-v3`

[7] `https://github.com/facebookresearch/AVID-CMA`

[8] `https://github.com/Shen-Lab/GraphCL/tree/master/unsupervised_TU`

A.4 Debiasing Foundation Models

A.4.1 Prompts

In this section, we provide the exact prompt we use for all the experiments in the paper in Table A.5, A.6, A.7, A.8.

	Class Prompt
$y = 0$	This is a picture of a landbird.
$y = 1$	This is a picture of a waterbird.
	Spurious Prompt
	This is a land background. This is a picture of a forest.
	This is a picture of a moutain. This is a picture of a wood.
	This is a water background. This is a picture of an ocean.
	This is a picture of a beach. This is a picture of a port.
	Positive Pairs
enumerate of	This is a picture of a landbird with land background.
	This is a picture of a landbird with water background.
	This is a picture of a landbird in the ocean
	This is a picture of a landbird in the water.
	This is a picture of a landbird in the forest.
enumerate of	This is a picture of a waterbird with land background.
	This is a picture of a waterbird with water background.
	This is a picture of a waterbird in the ocean
	This is a picture of a waterbird in the water.
	This is a picture of a waterbird in the forest.

Table A.5: Prompts for WaterBird Dataset. The spurious prompts are sentences that describe the spurious features, that is, the background of the images. In addition to keywords such as land/water background, we further include terms that describe similar concepts, such as forest or ocean. The positive pairs consist of sentences that describe the same type of bird (landbird/waterbird), while phrases that describe the background are appended afterward.

A.4.2 Importance of Class Name in Positive Pairs

In this section, we study the importance of class names in the positive pairs. In particular, instead of using "a photo of a [class name] with [spurious attribute]", we

	Class Prompt
$y = 0$	A photo of a celebrity with dark hair.
$y = 1$	A photo of a celebrity with blond hair.
	Spurious Prompt
	A photo of a male. A photo of a male celebrity. A photo of a man. A photo of a female. A photo of a female celebrity. A photo of a woman.
	Positive Pairs
	(A photo of a male celebrity with dark hair., A photo of a female celebrity with dark hair.) (A photo of a male celebrity with blond hair., A photo of a female celebrity with blond hair.)

Table A.6: Prompts for CelebA Dataset. We found two positive pairs are sufficient to mitigate the biases for CelebA dataset.

Prompt:	A photo of a [CONCEPT] person.
CONCEPT:	good, evil, smart, dumb, attractive, unattractive, lawful, criminal, friendly, unfriendly

Table A.7: Prompts for text-image retrieval on FairFace Dataset. We adopt the 10 training concepts from [16] to construct the prompts for FairFace. These concepts are irrelevant to gender, race, or age, which makes them suitable for evaluating the model biases.

instead use "a photo of a [spurious attribute]" to estimate the calibration matrix. The results are shown in Table A.9. We can see that the performance significantly drops after removing the class name from the prompt, emphasizing the importance of class-conditioned prompts.

A.4.3 Human Evaluation

We generate 100 images for each profession for evaluation. Therefore, there are 1500 images for each model in total. The random seed is fixed for the original and debiased Stable Diffusion models. In particular, automatic evaluation and human evaluation adopt the same set of images for fair comparison. The interface for human evaluation is shown in Figure A-4. Note that some generated images might be corrupted, or do not even contain humans. In this case, the annotators can click 3 or 6 to indicate

GENDER:	A photo of a male [profession]. A photo of a female [profession].
RACE:	A photo of a white [profession]. A photo of a black [profession]. A photo of an Asian [profession]. A photo of an Indian [profession]. A photo of a Latino [profession].

Table A.8: Prompts for debiasing generative models. We simply use prompts that describe the gender and race attribute as prompts, where we avoid using ambiguous terms such as white and black as the model could wrongly interpret it as the color of the photo.

	Waterbird			**CelebA**		
	WG	Avg	Gap	WG	Avg	Gap
Class-Agnostic	57.5	81.4	23.9	52.8	85.2	32.4
Class-Conditioned	74.0	78.7	4.7	82.2	84.4	2.2

Table A.9: Importance of target classes in the positive pairs. Removing [class name] in the positive pairs significantly degenerate the robustness of the zero-shot models.

that the current image is not identifiable. We remove these images while calculating the discrepancy.

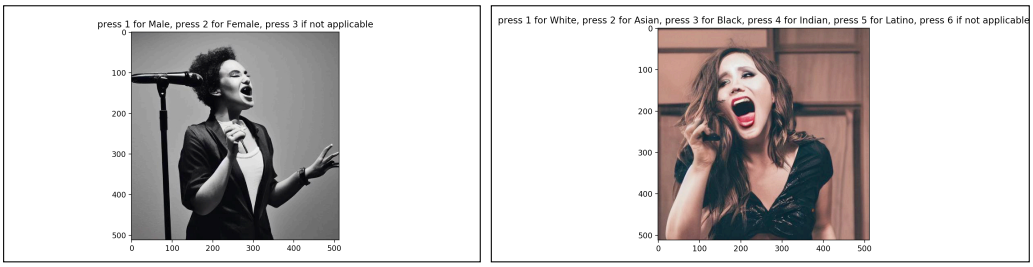


Figure A-4: Interface for human evaluation. We construct a simple labeling script for the annotator to easily label the sensitive attributes. Once they select the label, the program will ask whether they are sure about the answer. One can reselect the answer or press "enter" to switch to the next image.

A.4.4 More Samples from Biased and Debiased Generative Models

In this section, we show more generated images from Stable Diffusion 2.1 to provide a qualitative experiment. We can see that the proposed debiasing approach significantly

improves the diverisity across training and testing professions as as Figure A-5, A-6, A-8 and A-9 show. Nevertheless, there are also failure cases, where both our approach and the original model fail. For instance, biased and debiased models fail to generate females for many engineer-related professions such as carpenter as Figure A-7 shows.

Figure A-5: Generation against Gender Bias on Training Set. After debiasing, we can see that the gender diversity of Stable Diffusion greatly improves.

A.5 Generalization Theory of Representation Learning

A.5.1 Summary of Margins

A.5.2 Variance of Empirical Estimation

In Table 8.1, we show the average scores over 4 random sampled subsets. We now show the standard deviation in Table A.12. Overall, the standard deviation of the estimation is fairly small, consistent to the observation in Theorem 8.7.

Figure A-6: Generation against Gender Bias on Testing Set. We can see that by applying the calibration matrix, the gender distributions are more balanced.

Figure A-7: Failure Case for Generation against Gender Bias. There are also failure cases, where both our approach and the original model fail. For instance, biased and debiased models fail to generate females for professions such as carpenter and builders.

A.5.3 The effect of k in k-Variance

We next show the ablation study with respect to m (data size) in Table A.13. In particular, we draw $\overline{m}_c \times$ #classes samples where $\overline{m}_c = 50, 100$, and 200. Note that if the class distribution p is not uniform, m_c could be different for each class. The scores are computed with one subset for computational efficiency. Since the sample size per class of Flowers and Pets datasets are smaller than 50, the ablation study is not applicable.

Figure A-8: Generation against Race Bias on Training Set. We can observe a clear difference before and after debiasing, where the diversity is improved after debiasing.

A.5.4 Spectral Approximation to Lipschitz Constant

In section 8.4, we use the supermum of the norm of the jacobian on the training set as an approximation to Lipschitz constant, which is a simple lower bound of Lipschitz constant for ReLU networks [96]. It is well known that the spectral complexity, the multiplication of spectral norm of weights, is an upper bound on the Lipschitz constant of ReLU networks [130]. We replace the $\widehat{\text{Lip}}$ in kV-Margin with the spectral complexity of the network and show the results in Table A.14. The norm of the jacobian yields much better results than spectral complexity, which aligns with the observations in [51, 91].

A.5.5 Experiment Details

PGDL Dataset The models and datasets are accessible with Keras API [31] (integrated with TensorFlow [1]): `https://github.com/google-research/google-research/tree/master/pgdl` (Apache 2.0 License). We use the official evaluation code of PGDL competition [92]. All the scores can be computed with one TITAN X (Pascal) GPUs. The intuition behind the sample size $\min(200 \times \#\text{classes}, \text{data\_size})$ is that we want the average sample size for each class is 200. Note that if the class distribution p is not

Figure A-9: Generation against Race Bias on Testing Set. We can again see that the diversity is improved after debising even for unseen classes.

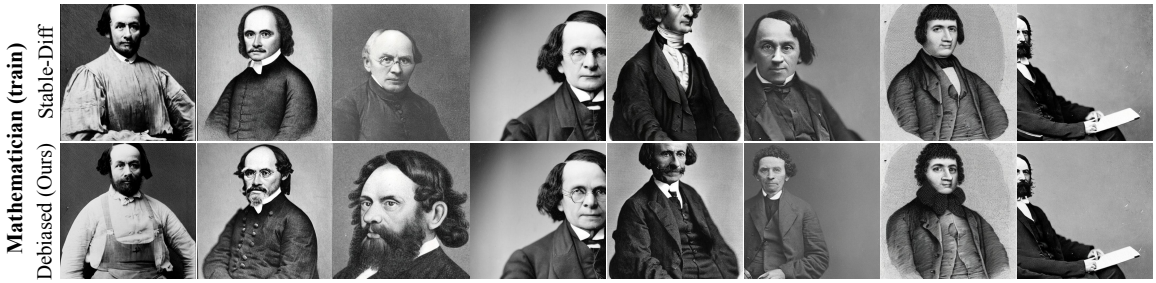


Figure A-10: Failure Case for Generation against Racial Bias. For certain classes that are highly correlated with historical figures such as mathematicians, our approach does not improve the diversity a lot due to the strong biases from the data.

uniform, the sample size for each class could be different. However, the sample size per class of Flowers and Pets datasets are smaller than $200 \times$ #classes, we constrain the sample size to be dataset size at most. We follow the setting in [141] to calculate the mixup accuracy with label-wise mixup.

Other Experiments The experiments in section 8.4.2 are run with the code from [222]: `https://github.com/pluskid/fitting-random-labels` (MIT License). We trained the models with the exact same code and visualize the margins with our own implementation via PyTorch [153]. For the experiments in section 8.4.3, we only change the data loader part of the code. The models of MNIST and SVHN are trained

Train	Actor, Architect, Audiologist, Author, Baker, Barber, Blacksmith Bricklayer, Bus Driver, Butcher, Chef, Chemist, Cleaner, Coach Comedian, Computer Programmer, Construction Worker, Consultant, Counselor, Dancer, Dentist, Designer, Dietitian, DJ, Doctor, Driver, Economist, Electrician, Engineer, Entrepreneur, Farmer, Florist Graphic Designer, Hairdresser, Historian, Journalist, Judge, Lawyer Librarian, Magician, Makeup Artist, Mathematician, Marine Biologist Mechanic, Model, Musician, Nanny, Nurse, Optician, Painter Pastry Chef, Pediatrician, Photographer, Plumber, Police Officer, Politician Professor, Psychologist, Real Estate Agent, Receptionist, Recruiter, Researcher Sailor, Salesperson, Surveyor, Singer, Social Worker, Software Developer Statistician, Surgeon, Teacher, Technician, Therapist, Tour Guide Translator, Vet, Videographer, Waiter, Writer, Zoologist
Test	Accountant, Astronaut, Biologist, Carpenter, Civil Engineer, Clerk, Detective Editor, Firefighter, Interpreter, Manager, Nutritionist, Paramedic, Pharmacist Physicist, Pilot, Reporter, Security Guard, Scientist, Web Developer

Table A.10: 100 Training and Testing Professions. We use GPT-4 to list 100 job titles to form our training and testing set. The similar approach can also be extended to other debiasing tasks by querying large language models.

	Definition
Margin	$\rho_f(\phi(x), y)$
SN-Margin [13]	$\rho_f(\phi(x), y)/SC(f \circ \phi)$
GN-Margin [90]	$\tilde{\rho}_f(\phi(x), y) =$ $\rho_f(\phi(x), y)/(\|\nabla_\phi \rho_f(\phi(x), y)\|_2 + \epsilon)$
TV-GN-Margin [90]	$\tilde{\rho}_f(\phi(x), y)/\sqrt{\mathrm{Var}_{x\sim\mu}(\|\|\phi(x)\|\|^2)}$
kV-Margin (Ours)	$\rho_f(\phi(x), y)/\mathbb{E}_{c\sim p}[\mathrm{Var}_{m_c}(\phi_{\#}\mu_c) \cdot \mathrm{Lip}(\rho_f(\cdot, c))]$
kV-GN-Margin (Ours)	$\tilde{\rho}_f(\phi(x), y)/\mathbb{E}_{c\sim p}[\mathrm{Var}_{m_c}(\phi_{\#}\mu_c) \cdot \mathrm{Lip}(\tilde{\rho}_f(\cdot, c))]$

Table A.11: Definitions of margins. The SC stands for the spectral complexity defined in [13]. We use the empirical estimation of k-variance and Lipschitz constant defined in section 8.4 to calculate kV-Margin and kV-GN-Margin.

for 10 and 20 epochs, respectively. To visualize the t-SNE in section 8.5, we use the default parameter in scikit-learn [156] (sklearn.manifold.TSNE) with the output from the 4th residual block of the network.

	CIFAR VGG	SVHN NiN	CINIC FCN bn	CINIC FCN	Flowers NiN	Pets NiN	Fashion VGG	CIFAR NiN
Margin†	0.25	0.84	0.16	0.13	0.01	0.04	0.06	0.59
SN-Margin† [13]	0.07	0.06	0.01	0.03	0.00	0.01	0.01	0.00
GN-Margin 1st [90]	0.18	0.17	0.27	0.15	0.06	0.02	0.10	0.52
GN-Margin 8th [90]	0.03	1.44	0.09	0.04	0.01	0.00	0.05	0.14
TV-GN-Margin 1st [90]	0.26	0.78	0.49	0.62	0.03	0.05	0.03	1.29
TV-GN-Margin 8th [90]	0.31	0.35	0.18	0.19	0.01	0.14	0.09	0.73
kV-Margin† 1st	0.40	1.57	0.55	0.45	0.07	0.03	0.23	2.78
kV-Margin† 8th	0.64	0.89	0.24	0.21	0.02	0.03	0.07	0.84
kV-GN-Margin† 1st	0.15	0.56	0.47	0.72	0.02	0.04	0.06	1.70
kV-GN-Margin† 8th	0.81	0.93	0.16	0.33	0.03	0.01	0.04	0.44

Table A.12: Standard deviation of CMI score on PGDL tasks.

	CIFAR VGG	SVHN NiN	CINIC FCN bn	CINIC FCN	Fashion VGG	CIFAR NiN
kV-Margin 1st (50)	7.23	30.21	37.21	17.65	1.74	14.39
kV-Margin 1st (100)	5.83	29.11	36.45	17.51	1.89	13.89
kV-Margin 1st (200)	4.81	29.79	36.23	17.01	2.37	12.63
kV-Margin 8th (50)	31.66	28.10	5.82	15.13	0.36	1.54
kV-Margin 8th (100)	29.72	27.20	6.01	15.10	0.37	1.43
kV-Margin 8th (200)	28.14	27.72	5.84	15.27	0.19	3.11
kV-GN-Margin 1st (50)	19.58	45.42	31.29	15.39	0.55	23.59
kV-GN-Margin 1st (100)	18.17	45.24	30.78	15.66	0.56	21.85
kV-GN-Margin 1st (200)	17.81	44.93	30.30	15.64	0.78	20.80
kV-GN-Margin 8th (50)	40.75	44.71	6.83	15.64	0.36	9.36
kV-GN-Margin 8th (100)	41.09	46.28	6.71	15.99	0.31	8.14
kV-GN-Margin 8th (200)	41.05	47.57	6.63	15.96	0.25	8.66

Table A.13: The role of data size in estimating k-variance The number between brackets denotes the average class size $\overline{m}_c$.

	CIFAR VGG	SVHN NiN	CINIC FCN bn	CINIC FCN	Flowers NiN	Pets NiN	Fashion VGG	CIFAR NiN
Spectral 1st	3.20	1.19	0.31	2.68	0.24	2.43	0.58	7.06
Spectral 8th	1.08	2.26	0.69	0.91	0.08	0.99	1.99	4.72
Jacobian Norm 1st	5.34	26.78	37.00	16.93	6.26	2.11	1.82	15.75
Jacobian Norm 8th	30.42	26.75	6.05	15.19	0.78	1.60	0.33	2.26

Table A.14: k-vairance normalized margins with spectral complexity. We show the score of kV-Margin with different approximations to Lipschitz constant. Empirically, gradient norm of data points yields better results.

Appendix B

Additional Theory and Proof

B.1 Theory of False Negative Samples

We provide the proof of Theorem 3.4 here.

Proof of Theorem 3.4. I first decompose the probability as follows,

$$
\begin{aligned}
&\mathbb{P}\bigg(\bigg| -\log\frac{h(x,x^+)}{h(x,x^+)+Qg(x,\{u_i\}_{i=1}^N,\{v_i\}_{i=1}^M)}+\log\frac{h(x,x^+)}{h(x,x^+)+Q\mathbb{E}_{x^-\sim p_x^-}h(x,x^-)}\bigg|\geq\varepsilon\bigg)\\
&=\mathbb{P}\bigg(\bigg|\log\big\{h(x,x^+)+Qg(x,\{u_i\}_{i=1}^N,\{v_i\}_{i=1}^M)\big\}-\log\big\{h(x,x^+)+Q\mathbb{E}_{x^-\sim p_x^-}h(x,x^-)\big\}\bigg|\geq\varepsilon\bigg)\\
&=\mathbb{P}\bigg(\log\big\{h(x,x^+)+Qg(x,\{u_i\}_{i=1}^N,\{v_i\}_{i=1}^M)\big\}-\log\big\{h(x,x^+)+Q\mathbb{E}_{x^-\sim p_x^-}h(x,x^-)\big\}\geq\varepsilon\bigg)\\
&\quad+\mathbb{P}\bigg(-\log\big\{h(x,x^+)+Qg(x,\{u_i\}_{i=1}^N,\{v_i\}_{i=1}^M)\big\}+\log\big\{h(x,x^+)+Q\mathbb{E}_{x^-\sim p_x^-}h(x,x^-)\big\}\geq\varepsilon\bigg)
\end{aligned}
$$

where the final equality holds simply because $|X|\geq\varepsilon$ if and only if $X\geq\varepsilon$ or $-X\geq\varepsilon$.

Consider the first term; it can be bounded as follows,

$$\mathbb{P}\Big(\log\big\{h(x,x^+) + Qg(x,\{u_i\}_{i=1}^N,\{v_i\}_{i=1}^M)\big\} - \log\big\{h(x,x^+) + Q\mathbb{E}_{x^-\sim p_x^-}h(x,x^-)\big\} \geq \varepsilon\Big)$$

$$\mathbb{P}\Big(\log \frac{h(x,x^+) + Qg(x,\{u_i\}_{i=1}^N,\{v_i\}_{i=1}^M)}{h(x,x^+) + Q\mathbb{E}_{x^-\sim p_x^-}h(x,x^-)} \geq \varepsilon\Big)$$

$$\leq \mathbb{P}\Big(\frac{Qg(x,\{u_i\}_{i=1}^N,\{v_i\}_{i=1}^M) - Q\mathbb{E}_{x^-\sim p_x^-}h(x,x^-)}{h(x,x^+) + Q\mathbb{E}_{x^-\sim p_x^-}h(x,x^-)} \geq \varepsilon\Big)$$

$$= \mathbb{P}\Big(g(x,\{u_i\}_{i=1}^N,\{v_i\}_{i=1}^M) - \mathbb{E}_{x^-\sim p_x^-}h(x,x^-) \geq \varepsilon\Big\{\frac{1}{Q}h(x,x^+) + \mathbb{E}_{x^-\sim p_x^-}h(x,x^-)\Big\}\Big)$$

$$\leq \mathbb{P}\Big(g(x,\{u_i\}_{i=1}^N,\{v_i\}_{i=1}^M) - \mathbb{E}_{x^-\sim p_x^-}h(x,x^-) \geq \varepsilon e^{-1}\Big). \tag{B.1}$$

The first inequality follows by applying the fact that $\log x \leq x-1$ for $x>0$. The second inequality holds since $\frac{1}{Q}h(x,x^+) + \mathbb{E}_{x^-\sim p_x^-}h(x,x^-) \geq 1/e$. Next we move on to bounding the second term, which proceeds similarly, using the same two bounds.

$$\mathbb{P}\Big\{ -\log\big(h(x,x^+) + Qg(x,\{u_i\}_{i=1}^N,\{v_i\}_{i=1}^M)\big\} + \log\big\{h(x,x^+) + Q\mathbb{E}_{x^-\sim p_x^-}h(x,x^-)\big\} \geq \varepsilon\Big)$$

$$= \mathbb{P}\Big(\log \frac{h(x,x^+) + Q\mathbb{E}_{x^-\sim p_x^-}h(x,x^-)}{h(x,x^+) + Qg(x,\{u_i\}_{i=1}^N,\{v_i\}_{i=1}^M)} \geq \varepsilon\Big)$$

$$\leq \mathbb{P}\Big(\frac{Q\mathbb{E}_{x^-\sim p_x^-}h(x,x^-) - Qg(x,\{u_i\}_{i=1}^N,\{v_i\}_{i=1}^M)}{h(x,x^+) + Qg(x,\{u_i\}_{i=1}^N,\{v_i\}_{i=1}^M)} \geq \varepsilon\Big)$$

$$= \mathbb{P}\Big(\mathbb{E}_{x^-\sim p_x^-}h(x,x^-) - g(x,\{u_i\}_{i=1}^N,\{v_i\}_{i=1}^M) \geq \varepsilon\Big\{\frac{1}{Q}h(x,x^+) + g(x,\{u_i\}_{i=1}^N,\{v_i\}_{i=1}^M)\Big\}\Big)$$

$$\leq \mathbb{P}\Big(\mathbb{E}_{x^-\sim p_x^-}h(x,x^-) - g(x,\{u_i\}_{i=1}^N,\{v_i\}_{i=1}^M) \geq \varepsilon e^{-1}\Big). \tag{B.2}$$

Combining equation equation B.1 and equation equation B.2, we have

$$\mathbb{P}(\Delta \geq \varepsilon) \leq \mathbb{P}\Big(\big|g(x,\{u_i\}_{i=1}^N,\{v_i\}_{i=1}^M) - \mathbb{E}_{x^-\sim p_x^-}h(x,x^-)\big| \geq \varepsilon e^{-1}\Big).$$

It therefore suffices to bound the right hand tail probability. We are bounding the tail of a difference of the form $|\max(a,b)-c|$ where $c \geq b$. Notice that $|\max(a,b)-c| \leq |a-c|$. If $a>b$ then this relation is obvious, while if $a \leq b$ we have $|\max(a,b)-c| = |b-c| = c-b \leq c-a \leq |a-c|$. Using this elementary observation we can decompose

the random variable whose tail we wish to control as follows,

$$\begin{aligned}\Big|g(x,\{u_i\}_{i=1}^N,\{v_i\}_{i=1}^M) - \mathbb{E}_{x^-\sim p_x^-}h(x,x^-)\Big| \\ &\leq \frac{1}{\tau^-}\left|\frac{1}{N}\sum_{i=1}^N \mathbb{E}_{x\sim p}h(x,u_i) - \mathbb{E}_{\substack{x^-\sim p\\ x\sim p}}h(x,x^-)\right| \\ &+ \frac{\tau^+}{\tau^-}\left|\frac{1}{M}\sum_{i=1}^M \mathbb{E}_{x\sim p}h(x,v_i) - \mathbb{E}_{\substack{x^-\sim p_x^+\\ x\sim p}}h(x,x^-)\right|\end{aligned}$$

Using this observation we find that

$$\begin{aligned}&\mathbb{P}\bigg(\Big|g(x,\{u_i\}_{i=1}^N,\{v_i\}_{i=1}^M) - \mathbb{E}_{x^-\sim p_x^-}h(x,x^-)\Big| \geq \varepsilon e^{-1}\bigg) \\ &\leq \mathbb{P}\bigg(\Big|\frac{1}{\tau^-}\left(\frac{1}{N}\sum_{i=1}^N e^{f(x)^T f(u_i)} - \tau^+\frac{1}{M}\sum_{i=1}^M e^{f(x)^T f(v_i)}\right) - \mathbb{E}_{x^-\sim p_x^-}h(x,x^-)\Big| \geq \varepsilon e^{-1}\bigg) \\ &\leq \mathrm{I}(\varepsilon) + \mathrm{II}(\varepsilon).\end{aligned}$$

where

$$\begin{aligned}\mathrm{I}(\varepsilon) &= \mathbb{P}\left(\frac{1}{\tau^-}\left|\frac{1}{N}\sum_{i=1}^N h(x,u_i) - \mathbb{E}_{x^-\sim p}h(x,x^-)\right| \geq \frac{\varepsilon e^{-1}}{2}\right) \\ \mathrm{II}(\varepsilon) &= \mathbb{P}\left(\frac{\tau^+}{\tau^-}\left|\frac{1}{M}\sum_{i=1}^M h(x,v_i) - \mathbb{E}_{x^-\sim p_x^+}h(x,x^-)\right| \geq \frac{\varepsilon e^{-1}}{2}\right).\end{aligned}$$

Hoeffding's inequality states that if $X, X_1, \ldots, X_N$ are i.i.d random variables bounded in the range $[a,b]$ then,

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^N X_i - \mathbb{E}X\right| \geq \varepsilon\right) \leq 2\exp\left(-\frac{2N\varepsilon^2}{b-a}\right)$$

In our particular case $e^{-1} \leq h(x,\bar{x}) \leq e$, yielding the following bound on the tails of both terms,

$$\mathrm{I}(\varepsilon) \leq 2\exp\left(-\frac{N\varepsilon^2(\tau^-)^2}{2e^3}\right) \quad \text{and} \quad \mathrm{II}(\varepsilon) \leq 2\exp\left(-\frac{M\varepsilon^2(\tau^-/\tau^+)^2}{2e^3}\right).$$

□

B.2 Theory of Hard Negative Samples

Here, we provide the proof for some lemmas we use in Chapter 4.

Lemma B.1. *Consider the same setting as introduced in Theorem 4.5. In particular define*

$$\xi' = \min_{c,c^-:c\neq c^-} \|\mathbf{v}_c - \mathbf{v}_{c-}\|, \qquad \xi = \min_{c,c^-:c\neq c^-} \left\|\mathbf{v}_c^* - \mathbf{v}_{c-}^*\right\|.$$

where $\{\mathbf{v}_c^* \in \mathbb{S}^{d-1}/t\}_{c\in\mathcal{C}}$ *is a solution to Problem 4.7, and* $\{\mathbf{v}_c \in \mathbb{S}^{d-1}/t\}_{c\in\mathcal{C}}$ *is defined via* $\mathbf{v}_c = f(x_c)$ *with* $x_c \in \arg\min_{\overline{\{x^+:h(x^+)=c\}}} \mathbb{E}_{x\sim p(\cdot|c)}\left\|f(x) - f(x^+)\right\|^2$ *for each* $c \in \mathcal{C}$*. Then we have,*

$$\xi' \geq \sqrt{\xi^2 - 2|\mathcal{C}|\,(1+1/t)\varepsilon^{1/2}}.$$

Proof. Define,

$$X = \min_{c^-:c^-\neq c} \|\mathbf{v}_c - \mathbf{v}_{c-}\|^2, \qquad X^* = \min_{c^-:c^-\neq c} \left\|\mathbf{v}_c^* - \mathbf{v}_{c-}^*\right\|^2.$$

X and X^* are random due to the randomness of $c \sim \rho$. We can split up the following expectation by conditioning on the event $\{X \leq X^*\}$ and its complement,

$$\mathbb{E}|X - X^*| = \mathbb{P}(X \geq X^*)\mathbb{E}[X - X^*] + \mathbb{P}(X \leq X^*)\mathbb{E}[X^* - X]. \tag{B.3}$$

Using $\mathcal{L}^*_\infty(f) - \inf_{\bar{f}\text{ measurable}} \mathcal{L}^*_\infty(\bar{f}) \leq \varepsilon$ and the notational re-writing of the objective $\mathcal{L}^*_\infty$ introduced before Theorem 4.4, we observe the following fact, whose proof we give in a separate lemma after the conclusion of this proof.

Fact (see lemma B.2): $\mathbb{E}X^* - 2(1 + 1/t)\sqrt{\varepsilon} \leq \mathbb{E}X \leq \mathbb{E}X^*$.

This fact implies in particular $\mathbb{E}[X - X^*] \leq 0$ and $\mathbb{E}[X^* - X] \leq 2(1 + 1/t)\sqrt{\varepsilon}$. Inserting both inequalities into Eqn. B.3 we find that $\mathbb{E}|X - X^*| \leq 2(1 + 1/t)\sqrt{\varepsilon}$. In other words, since ρ is uniform,

$$\frac{1}{|\mathcal{C}|}\sum_{c\in\mathcal{C}}\left|\min_{c^-:c^-\neq c}\|\mathbf{v}_c - \mathbf{v}_{c-}\|^2 - \min_{c^-:c^-\neq c}\left\|\mathbf{v}^*_c - \mathbf{v}^*_{c-}\right\|^2\right| \leq 2(1 + 1/t)\sqrt{\varepsilon}.$$

From which we can say that for any $c \in \mathcal{C}$,

$$\left|\min_{c^-:c^-\neq c}\|\mathbf{v}_c - \mathbf{v}_{c-}\|^2 - \min_{c^-:c^-\neq c}\left\|\mathbf{v}^*_c - \mathbf{v}^*_{c-}\right\|^2\right| \leq 2|\mathcal{C}|\,(1 + 1/t)\sqrt{\varepsilon}.$$

So we have

$$\min_{c^-:c^-\neq c}\|\mathbf{v}_c - \mathbf{v}_{c-}\| \geq \sqrt{\min_{c^-:c^-\neq c}\left\|\mathbf{v}^*_c - \mathbf{v}^*_{c-}\right\|^2 - 2|\mathcal{C}|\,(1 + 1/t)\varepsilon^{1/2}} \geq \sqrt{\xi^2 - 2|\mathcal{C}|\,(1 + 1/t)\varepsilon^{1/2}}.$$

Since this holds for any $c \in \mathcal{C}$, we conclude that $\xi' \geq \sqrt{\xi^2 - 2|\mathcal{C}|\,(1 + 1/t)\varepsilon^{1/2}}$. □

Lemma B.2. *Consider the same setting as introduced in Theorem 4.5. Define also,*

$$X = \min_{c^-:c^-\neq c}\|\mathbf{v}_c - \mathbf{v}_{c-}\|^2, \qquad X^* = \min_{c^-:c^-\neq c}\left\|\mathbf{v}^*_c - \mathbf{v}^*_{c-}\right\|^2,$$

where $\mathbf{v}_c = f(x_c)$ *with* $x_c \in \arg\min_{\overline{\{x^+:h(x^+)=c\}}} \mathbb{E}_{x\sim p(\cdot|c)}\left\|f(x) - f(x^+)\right\|^2$ *for each*

$c \in \mathcal{C}$. *We have,*

$$\mathbb{E}X^* - 2(1+1/t)\sqrt{\varepsilon} \leq \mathbb{E}X \leq \mathbb{E}X^*.$$

Proof. By Theorem 4.7 we know there is an f^* attaining the minimum $\inf_{\bar{f} \text{ measurable}} \mathcal{L}^*_\infty(\bar{f})$ and that this f^* attains $\mathcal{L}^*_{\text{align}}(f^*) = 0$, and also minimizes the uniformity term $\mathcal{L}^*_{\text{unif}}(f)$, taking the value $\mathcal{L}^*_{\text{unif}}(f^*) = \mathbb{E}_{c\sim\rho} \max_{c^-:c^-\neq c} \mathbf{v_c^*}^\top \mathbf{v}^*_{c^-}$. Because of this we find,

$$\begin{aligned}
\mathcal{L}^*_{\text{unif}}(f) &\leq \big(\mathcal{L}^*_\infty(f) - \mathcal{L}^*_\infty(f^*)\big) + \big(\mathcal{L}^*_{\text{align}}(f^*) - \mathcal{L}^*_{\text{align}}(f)\big) + \mathcal{L}^*_{\text{unif}}(f^*) \\
&\leq \big(\mathcal{L}^*_\infty(f) - \mathcal{L}^*_\infty(f^*)\big) + \mathcal{L}^*_{\text{unif}}(f^*) \\
&\leq \varepsilon + \mathcal{L}^*_{\text{unif}}(f^*) \\
&= \varepsilon + \mathbb{E}_{c\sim\rho} \max_{c^-:c^-\neq c} \mathbf{v_c^*}^\top \mathbf{v}^*_{c^-}.
\end{aligned}$$

Since we would like to bound $\mathbb{E}_{c\sim\rho} \max_{c^-:c^-\neq c} \mathbf{v_c}^\top \mathbf{v}_{c^-}$ in terms of $\mathbb{E}_{c\sim\rho} \max_{c^-:c^-\neq c} \mathbf{v_c^*}^\top \mathbf{v}^*_{c^-}$, this observation means that is suffices to bound $\mathbb{E}_{c\sim\rho} \max_{c^-:c^-\neq c} \mathbf{v_c}^\top \mathbf{v}_{c^-}$ in terms of $\mathcal{L}^*_{\text{unif}}(f)$. To this end, note that for a fixed c, and x such that $h(x) = c$ we have,

$$\begin{aligned}
\sup_{x^-\nsim x} f(x)^\top f(x^-) &= \sup_{x^-\nsim x} \left\{ \mathbf{v}_c{}^\top f(x^-) + (f(x) - \mathbf{v}_c)^\top f(x^-) \right\} \\
&= \sup_{x^-\nsim x} \mathbf{v}_c{}^\top f(x^-) - \left\| f(x) - \mathbf{v}_c \right\| / t \\
&\geq \max_{x^-\in\{x_c\}_{c\in\mathcal{C}}} \mathbf{v}_c{}^\top f(x^-) - \left\| f(x) - \mathbf{v}_c \right\| / t \\
&= \max_{c^-\neq c} \mathbf{v}_c{}^\top \mathbf{v}_{c^-} - \left\| f(x) - \mathbf{v}_c \right\| / t
\end{aligned}$$

where the inequality follows since $\{x_c\}_{c\in\mathcal{C}}$ is a subset of the closure of $\{x^- : x^- \nsim x\}$. Taking expectations over c, x,

$$\begin{aligned}\mathcal{L}^*_{\text{unif}}(f) &= \mathbb{E}_{x,c} \sup_{x^- \nsim x} f(x)^\top f(x^-) \\ &\geq \mathbb{E}_{c\sim\rho} \max_{c^- \neq c} \mathbf{v}_c{}^\top \mathbf{v}_{c^-} - \mathbb{E}_{x,c} \left\| f(x) - \mathbf{v}_c \right\| / t \\ &\geq \mathbb{E}_{c\sim\rho} \max_{c^- \neq c} \mathbf{v}_c{}^\top \mathbf{v}_{c^-} - \sqrt{\mathbb{E}_{x,c} \left\| f(x) - \mathbf{v}_c \right\|^2} / t \\ &\geq \mathbb{E}_{c\sim\rho} \max_{c^- \neq c} \mathbf{v}_c{}^\top \mathbf{v}_{c^-} - \sqrt{\varepsilon}/t.\end{aligned}$$

So since $\varepsilon \leq \sqrt{\varepsilon}$, we have found that

$$\mathbb{E}_{c\sim\rho} \max_{c^- \neq c} \mathbf{v}_c{}^\top \mathbf{v}_{c^-} \leq \sqrt{\varepsilon}/t + \varepsilon + \mathbb{E}_{c\sim\rho} \max_{c^-:c^-\neq c} \mathbf{v}_{\mathbf{c}}^{*\top} \mathbf{v}^*_{c^-} \leq (1+1/t)\sqrt{\varepsilon} + \mathbb{E}_{c\sim\rho} \max_{c^-:c^-\neq c} \mathbf{v}_{\mathbf{c}}^{*\top} \mathbf{v}^*_{c^-}.$$

Of course we also have,

$$\mathbb{E}_{c\sim\rho} \max_{c^-:c^-\neq c} \mathbf{v}_{\mathbf{c}}^{*\top} \mathbf{v}^*_{c^-} = \mathcal{L}^*_{\text{unif}}(f^*) \leq \mathbb{E}_{c\sim\rho} \max_{c^-:c^-\neq c} \mathbf{v}_{\mathbf{c}}{}^\top \mathbf{v}_{c^-}$$

since the embedding $f(x) = \mathbf{v}_c$ whenever $h(x) = c$ is also a feasible solution. Combining these two inequalities with the simple identity $\mathbf{x}^\top \mathbf{y} = 1/t^2 - \|\mathbf{x} - \mathbf{y}\|^2 / 2$ for all length $1/t$ vectors $\mathbf{x}, \mathbf{y}$, we find,

$$\begin{aligned}1/t^2 - \mathbb{E}_{c\sim\rho} \max_{c^-:c^-\neq c} \left\| \mathbf{v}^*_{\mathbf{c}} - \mathbf{v}^*_{c^-} \right\|^2 / 2 &\leq 1/t^2 - \mathbb{E}_{c\sim\rho} \max_{c^-:c^-\neq c} \left\| \mathbf{v}_{\mathbf{c}} - \mathbf{v}_{c^-} \right\|^2 / 2 \\ &\leq 1/t^2 - \mathbb{E}_{c\sim\rho} \max_{c^-:c^-\neq c} \left\| \mathbf{v}^*_{\mathbf{c}} - \mathbf{v}^*_{c^-} \right\|^2 / 2 + (1 + 1/t)\sqrt{\varepsilon}.\end{aligned}$$

Subtracting $1/t^2$ and multiplying by -2 yields the result.

□

B.3 Theory of Representation Learning

B.3.1 Estimating the Lipschitz Constant of the GN-Margin

$$\widehat{\text{Lip}}(\tilde{\rho}_f(\cdot, c)) = \max_{x \in S_c} \left\| \nabla_\phi \frac{\rho_f(\phi(x), c)}{\|\nabla_\phi \rho_f(\phi(x), y)\|_2 + \epsilon} \right\|_2 = \max_{x \in S_c} \frac{\|\nabla_\phi \rho_f(\phi(x), y)\|_2}{\|\nabla_\phi \rho_f(\phi(x), y)\|_2 + \epsilon} \approx 1 \tag{B.4}$$

Proof of equation B.4 We first expand the derivative as follows:

$$\begin{aligned}
\widehat{\text{Lip}}(\tilde{\rho}_f(\cdot, c)) &= \max_{x \in \mathcal{X}} \left\| \nabla_\phi \frac{\rho_f(\phi(x), c)}{\|\nabla_\phi \rho_f(\phi(x), y)\|_2 + \epsilon} \right\|_2 \\
&= \max_{x \in \mathcal{X}} \left\| \frac{\nabla_\phi \rho_f(\phi(x), c)(\|\nabla_\phi \rho_f(\phi(x), y)\|_2 + \epsilon) - \rho_f(\phi(x), c)\nabla_\phi \|\nabla_\phi \rho_f(\phi(x), y)\|_2}{(\|\nabla_\phi \rho_f(\phi(x), y)\|_2 + \epsilon)^2} \right\|_2 .
\end{aligned}$$

Note that ρ_f is piecewise linear as f is ReLU networks. For points where ρ_f is differentiable i.e that do not lie on the boundary between linear regions, the second order derivative is zero. In particular, we have $\nabla_\phi \|\nabla_\phi \rho_f(\phi(x), y)\|_2 = 0$. Therefore, excluding from $\mathcal{X}$ non differentiable points of ρ_f, the empirical Lipschitz estimation (lower bound) can be written as

$$\begin{aligned}
\widehat{\text{Lip}}(\tilde{\rho}_f(\cdot, c)) &= \max_{x \in \mathcal{X}} \left\| \frac{\nabla_\phi \rho_f(\phi(x), c)(\|\nabla_\phi \rho_f(\phi(x), y)\|_2 + \epsilon)}{(\|\nabla_\phi \rho_f(\phi(x), y)\|_2 + \epsilon)^2} \right\|_2 \\
&= \max_{x \in \mathcal{X}} \left\| \frac{\nabla_\phi \rho_f(\phi(x), c)}{\|\nabla_\phi \rho_f(\phi(x), y)\|_2 + \epsilon} \right\|_2 \\
&= \max_{x \in \mathcal{X}} \frac{\|\nabla_\phi \rho_f(\phi(x), c)\|_2}{\|\nabla_\phi \rho_f(\phi(x), y)\|_2 + \epsilon} \\
&\leq 1,
\end{aligned}$$

We can see that $\widehat{\text{Lip}}$ is tightly upper bounded by 1 when ϵ is a very small value.

Discussion of Lower and Upper Bounds on the Lipschitz constant Note that the approximation of the lipchitz constant will result in additional error in the

generalization bound as follows:

$$\hat{R}_{\gamma,m}(f\circ\phi)+\mathbb{E}_{c\sim p_y}\left[\frac{\widehat{\mathrm{Lip}}(\rho_f(\cdot,c))}{\gamma}\left(\widehat{\mathrm{Var}}_{\lfloor\frac{m_c}{2n}\rfloor,n}(\phi_{\#}\mu_c)+2B\sqrt{\frac{\log(2K/\delta)}{n\lfloor\frac{m_c}{2n}\rfloor}}\right)\right]$$
$$+\mathbb{E}_{c\sim p_y}\left[\frac{\mathrm{Lip}(\rho_f(\cdot,c))-\widehat{\mathrm{Lip}}(\rho_f(\cdot,c))}{\gamma}\left(\widehat{\mathrm{Var}}_{\lfloor\frac{m_c}{2n}\rfloor,n}(\phi_{\#}\mu_c)+2B\sqrt{\frac{\log(2K/\delta)}{n\lfloor\frac{m_c}{2n}\rfloor}}\right)\right]+\sqrt{\frac{\log(\frac{2}{\delta})}{2m}}.$$

While for an upper bound on the lipschitz constant the third term is negative and can be ignored in the generalization bound. For a lower bound this error term $\mathrm{Lip}(\rho_f(\cdot,c))-\widehat{\mathrm{Lip}}(\rho_f(\cdot,c))$ results in additional positive error term. Bounding this error term is beyond the scope of this work and we leave it for a future work.

B.3.2 Proof of Proposition 8.9

We will prove these two arguments separately with the following two propositions.

Proposition B.3. *For any $\phi_{\#}\mu\in\mathrm{Prob}(\mathbb{R}^d)$, we have $\mathrm{Var}_m(\phi_{\#}\mu)\leq\mathcal{O}(m^{-1/d})$ for $d>2$.*

Proof. The result is an application of Theorem 1 of [58].

Theorem B.4 ((Fournier and Guillin [58])). *Let $\mu\in\mathrm{Prob}(\mathbb{R}^d)$ and let $p>0$. Define $M_q(\mu)=\int_{\mathbb{R}^d}|x|^q\mu(dx)$ be the q-th moment for μ and assume $M_q(\mu)\leq\infty$ for some $q>p$. There exists a constant C depending only on p,d,q such that, for all $m\geq1$, $p\in(0,d/2)$ and $q\neq d/(d-p)$,*

$$\mathbb{E}_{S\sim\mu^m}[\mathcal{W}_p(\mu_S,\mu)]\leq CM_q^{p/q}(m^{-p/d}+m^{-(q-p)/q}).$$

By the triangle inequality and setting $p=1$, we have

$$\begin{aligned}\mathrm{Var}_m(\phi_{\#}\mu)=\mathbb{E}_{S,\tilde{S}\sim\mu^m}[\mathcal{W}_1(\phi_{\#}\mu_S,\phi_{\#}\mu_{\tilde{S}})]&\leq2\mathbb{E}_{S\sim\mu^m}[\mathcal{W}_1(\mu,\mu_S)]\\&\leq2CM_q^{1/q}(m^{-1/d}+m^{-(q-1)/q}).\end{aligned}$$

Note that the term $m^{-(q-1)/q}$ is small and can be removed. For instance, plugging $q = 2$, we can see that the first term dominates the second term which completes the proof for the first argument. □

We then demonstrate the case when the measure has low-dimensional structure.

Definition B.5. (Low-dimensional Measures) *Given a set $S \subseteq \mathcal{X}$, the ϵ-covering number of S, denoted as $\mathcal{N}_\epsilon(S)$, is the minimum n such that there exists n closed balls $B_1, \cdots, B_n$ of diameter ϵ such that $S \subseteq \bigcup_{1 \leq i \leq n} B_i$. For any $S \subseteq X$, the ϵ-fattening of S is $S_\epsilon := \{y : D(y, S) \leq \epsilon\}$, where D denotes the Euclidean distance.*

Proposition B.6. *Suppose* $\text{supp}(\phi_\# \mu) \subseteq S_\epsilon$ *for some* $\epsilon > 0$, *where* S *satisfies* $\mathcal{N}_{\epsilon'}(S) \leq (3\epsilon')^{-d}$ *for all* $\epsilon' \leq 1/27$ *and some* $d > 2$. *Then, for all* $m \leq (3\epsilon)^{-d}$, *we have* $\text{Var}_m(\phi_\# \mu) \leq 2C_1 m^{-1/d}$, *where* $C_1 = 54 + 27/(3^{\frac{d}{2}-1} - 1)$.

Proof. an application of Weed and Bach [203]'s Proposition 15 for $p = 1$.

Proposition B.7 ((Weed and Bach [203])). *Suppose* $\text{supp}(\mu) \subseteq S_\epsilon$ *for some* $\epsilon > 0$, *where* S *satisfies* $\mathcal{N}_{\epsilon'}(S) \leq (3\epsilon')^{-d}$ *for all* $\epsilon' \leq 1/27$ *and some* $d > 2p$. *Then, for all* $m \leq (3\epsilon)^{-d}$, *we have*

$$\mathbb{E}_{S \sim \mu^m}[\mathcal{W}_p^p(\mu, \mu_S)] \leq C_1 m^{-p/d},$$

where

$$C_1 = 27^p \left(2 + \frac{1}{3^{\frac{d}{2}-p} - 1}\right).$$

By the triangle inequality and setting $p = 1$, we have

$$\text{Var}_m(\phi_\# \mu) = \mathbb{E}_{S, \tilde{S} \sim \mu^m}[\mathcal{W}_1(\phi_\# \mu_S, \phi_\# \mu_{\tilde{S}})] \leq 2\mathbb{E}_{S \sim \mu^m}[\mathcal{W}_p^p(\mu, \mu_S)] \leq 2C_1 m^{-1/d},$$

where $C_1 = 54 + 27/(3^{\frac{d}{2}-1} - 1)$.

□

B.3.3 Proof of Proposition 8.10

Proof. The results is an application of Weed and Bach [203]'s Proposition 13 for $p = 1$.

Proposition B.8 (Weed and Bach [203])**.** *If μ is (n, Δ)-clusterable, then for all $m \leq n(2\Delta)^{-2p}$,*

$$\mathbb{E}_{S\sim\mu^m}[\mathcal{W}_p^p(\mu, \mu_S)] \leq (9^p + 3)\sqrt{\frac{n}{m}}.$$

Similarly, by the triangle inequality, we have

$$\mathrm{Var}_m(\phi_{\#}\mu) = \mathbb{E}_{S,\tilde{S}\sim\mu^m}[\mathcal{W}_1(\phi_{\#}\mu_S, \phi_{\#}\mu_{\tilde{S}})] \leq 2\mathbb{E}_{S\sim\mu^m}[\mathcal{W}_p^p(\mu, \mu_S)] \leq 24\sqrt{\frac{n}{m}}.$$

□

Lemma B.9 (Robust Feature Separation and Max-Gradient-Margin classifiers)**.** *Let $\mathcal{F}$ be function class satisfying assumption 1 and assumption 2 (ii) in [61] (piece-wise smoothness and growth and jump of the gradient) . Assume $f_y, f_{y'} \in Lip_L \cap \mathcal{F}$ and M bounded. Assume that f is such that for all y :*

$$f_y(\phi(x)) > f_{y'}(\phi(x)) + \gamma + \delta_n||\nabla_z f_y(\phi(x)) - \nabla_x f_{y'}(\phi(x))||_2, \forall x \in supp(\hat{\mu}_y), \forall y' \neq y$$

Then:

$$\sup_{\mu, \mathcal{W}_\infty(\mu,\hat{\mu})\leq\delta_n} \mathcal{W}_1(\phi_{\#}\mu_y, \phi_{\#}\mu_{y'}) \geq \frac{\gamma}{L} - \delta_n M - \varepsilon_n.$$

where $\varepsilon_n = O(1/\sqrt{n})$ $\hat{\mu}$ is defined as follows: $\hat{\mu}(x, c)$ be such that $\hat{\mu}(x|c = 1) = \hat{\mu}_y(x)$ and $\hat{\mu}(x|c = -1) = \hat{\mu}_{y'}(x)$, let $\hat{\mu}(c = 1) = \hat{\mu}(c = -1) = \frac{1}{2}$(similar definition holds for μ).

Proof. Without Loss of generality assume $\phi(x) = x$.

$$\begin{aligned}\mathcal{W}_1(\mu_y, \mu_{y'}) &= \sup_{f \in Lip_1} \mathbb{E}_{\mu_y} f(x) - \mathbb{E}_{\mu_{y'}} f(x) \\ &= \sup_{f \in Lip_1} \mathbb{E}_{(c,x)\sim\mu} 2cf(x) \\ &= -\inf_{f \in Lip_1} -2\mathbb{E}_{(c,x)\sim\mu} cf(x)\end{aligned}$$

This form of $\mathcal{W}_1$ suggests studying the following robust risk, for technical reason we will use another functional class $\mathcal{F} \subset Lip_1$ instead of Lip_1:

$$\inf_{f \in \mathcal{F}} \sup_{\mu, \mathcal{W}_\infty(\mu,\hat{\mu}) \leq \delta_n} -2\mathbb{E}_{(c,x)\sim\mu} cf(x)$$

Applying here theorem 1 (1) of Gao et al [61] see also example 12, for $\mathcal{F}$ of function satisfying assumption 1 and assumption 2 in Gao et al in addition to being lipchitz we have:

$$\sup_{\mu, \mathcal{W}_\infty(\mu,\hat{\mu}) \leq \delta_n} -2\mathbb{E}_{(c,x)\sim\mu} cf(x) \leq -2\mathbb{E}_{(c,x)\sim\mu} cf(x) + \delta_n 2\mathbb{E}_{\hat{\mu}} ||\nabla_{(c,x)} cf(x)|| + \varepsilon_n$$

Note that :

$$2\mathbb{E}_{(c,x)\sim\mu} cf(x) = \mathbb{E}_{\mu_y} f(x) - \mathbb{E}_{\mu_{y'}} f(x)$$

and

$$\nabla_{(c,x)} cf(x) = (f(x), c\nabla_x f(x))$$

and

$$||\nabla_{(c,x)} cf(x)|| = \sqrt{f(x)^2 + ||\nabla_x f(x)||^2} \leq |f(x)| + ||\nabla_x f(x)|| \leq M + ||\nabla_x f(x)||$$

where we used

$$\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}, \text{ and } |f(x)| \leq M$$

Hence we have:

$$\inf_{f\in\mathcal{F}} \sup_{\mu,\mathcal{W}_\infty(\mu,\hat{\mu})\leq\delta_n} -2\mathbb{E}_{(c,x)\sim\mu} cf(x) = \inf_{f\in\mathcal{F}} -2\mathbb{E}_{(c,x)\sim\mu} cf(x) + \delta_n 2\mathbb{E}_{\hat{\mu}} ||\nabla_{(c,x)} cf(x)|| + \varepsilon_n$$
$$\leq \inf_{f\in\mathcal{F}} -\mathbb{E}_{\mu_y} f + \mathbb{E}_{\mu_{y'}} f + \delta_n \mathbb{E}_{p_y}(||\nabla_x f(x)|| + |f(x)|) + \delta_n \mathbb{E}_{\mu_{y'}}(||\nabla_x f(x)|| + |f(x)|) + \varepsilon_n$$
$$\leq \inf_{f\in\mathcal{F}} -\mathbb{E}_{\mu_y}(f(x) - \delta_n ||\nabla_x f(x)||) + \mathbb{E}_{\mu_{y'}}(f(x) + \delta_n ||\nabla_x f(x)||) + \delta_n M + \varepsilon_n$$

Let $g(x) = \frac{f_y(x) - f_{y'}(x)}{2L}$ we have:

$$\begin{aligned}
&\inf_{f\in\mathcal{F}} -\mathbb{E}_{\mu_y}(f(x) - \delta_n ||\nabla_x f(x)||) + \mathbb{E}_{\mu_{y'}}(f(x) + \delta_n ||\nabla_x f(x)||) \\
&\leq -\mathbb{E}_{\mu_y}(g(x) - \delta_n ||\nabla_x g(x)||) + \mathbb{E}_{\mu_{y'}}(g(x) + \delta_n ||\nabla_x g(x)||) \\
&= -\mathbb{E}_{\mu_y}(g(x) - \delta_n ||\nabla_x g(x)||) - \mathbb{E}_{\mu_{y'}}(-g(x) - \delta_n ||\nabla_x g(x)||) \\
&\leq \frac{-2\gamma}{2L} = \frac{-\gamma}{L}.
\end{aligned}$$

It follows that there exists a robust classifier between the two classes y, y':

$$\inf_{f\in\mathcal{F}} \sup_{\mu,\mathcal{W}_\infty(\mu,\hat{\mu})\leq\delta_n} -2\mathbb{E}_{(c,x)\sim\mu} cf(x) \leq -\frac{\gamma}{L} + \delta_n M + \varepsilon_n$$

Note that:

$$-\inf_{f\in\mathcal{F}} \sup_{\mu,\mathcal{W}_\infty(\mu,\hat{\mu})\leq\delta_n} -2\mathbb{E}_{(c,x)\sim\mu} cf(x) = \sup_{f\in\mathcal{F}} \inf_{\mu,\mathcal{W}_\infty(\mu,\hat{\mu})\leq\delta_n} 2\mathbb{E}_{(c,x)\sim\mu} cf(x)$$

Hence:

$$\sup_{f\in\mathcal{F}} \inf_{\mu,\mathcal{W}_\infty(\mu,\hat{\mu})\leq\delta_n} 2\mathbb{E}_{(c,x)\sim\mu} cf(x) \geq \frac{\gamma}{L} - \delta_n M - \varepsilon_n.$$

On the other hand we have:

$$\sup_{f\in\mathcal{F}} \sup_{\mu,\mathcal{W}_\infty(\mu,\hat{\mu})\leq\delta_n} 2\mathbb{E}_{(c,x)\sim\mu} cf(x) \geq \sup_{f\in\mathcal{F}} \inf_{\mu,\mathcal{W}_\infty(\mu,\hat{\mu})\leq\delta_n} 2\mathbb{E}_{(c,x)\sim\mu} cf(x) \geq \frac{\gamma}{L} - \delta_n M - \varepsilon_n.$$

Note that $\mathcal{F} \subset Lip_1$

$$\sup_{f \in Lip_1} \sup_{\mu, \mathcal{W}_\infty(\mu,\hat{\mu}) \leq \delta_n} 2\mathbb{E}_{(c,x)\sim\mu} cf(x) \geq \sup_{f \in \mathcal{F}} \sup_{\mu, \mathcal{W}_\infty(\mu,\hat{\mu}) \leq \delta_n} 2\mathbb{E}_{(c,x)\sim\mu} cf(x) \geq \frac{\gamma}{L} - \delta_n M - \varepsilon_n.$$

We can now swap the two sups and obtain:

$$\begin{aligned} &\sup_{f \in Lip_1} \sup_{\mu, \mathcal{W}_\infty(\mu,\hat{\mu}) \leq \delta_n} 2\mathbb{E}_{(c,x)\sim\mu} cf(x) \\ = &\sup_{\mu, \mathcal{W}_\infty(\mu,\hat{\mu}) \leq \delta_n} \sup_{f \in Lip_1} 2\mathbb{E}_{(c,x)\sim\mu} cf(x) = \sup_{\mu, \mathcal{W}_\infty(\mu,\hat{\mu}) \leq \delta_n} W_1(\mu_y, \mu_{y'}) \end{aligned}$$

and finally we have:

$$\sup_{\mu, \mathcal{W}_\infty(\mu,\hat{\mu}) \leq \delta_n} \mathcal{W}_1(\mu_y, \mu_{y'}) \geq \frac{\gamma}{L} - \delta_n M - \varepsilon_n.$$

□

Bibliography

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. URL `https://www.tensorflow.org/`. Software available from tensorflow.org.

[2] Bijaya Adhikari, Yao Zhang, Naren Ramakrishnan, and B Aditya Prakash. Sub2vec: Feature learning for subgraphs. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 170–182. Springer, 2018.

[3] Sandhini Agarwal, Gretchen Krueger, Jack Clark, Alec Radford, Jong Wook Kim, and Miles Brundage. Evaluating clip: towards characterization of broader capabilities and downstream implications. *arXiv preprint arXiv:2108.02818*, 2021.

[4] Mohsan Alvi, Andrew Zisserman, and Christoffer Nellåker. Turning a blind eye: Explicit removal of biases and variation from deep neural network embeddings. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018.

[5] Humam Alwassel, Dhruv Mahajan, Bruno Korbar, Lorenzo Torresani, Bernard Ghanem, and Du Tran. Self-supervised learning by cross-modal audio-video clustering. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[6] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

[7] Sanjeev Arora, Rong Ge, Behnam Neyshabur, and Yi Zhang. Stronger generalization bounds for deep nets via a compression approach. In *International Conference on Machine Learning*, pages 254–263. PMLR, 2018.

[8] Sanjeev Arora, Hrishikesh Khandeparkar, Mikhail Khodak, Orestis Plevrakis, and Nikunj Saunshi. A theoretical analysis of contrastive unsupervised representation learning. *International Conference on Machine Learning*, 2019.

[9] Devansh Arpit, Stanislaw Jastrzebski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, et al. A closer look at memorization in deep networks. In *International Conference on Machine Learning*, pages 233–242. PMLR, 2017.

[10] Yuki M Asano, Mandela Patrick, Christian Rupprecht, and Andrea Vedaldi. Labelling unlabelled videos from scratch with multi-modal self-supervision. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[11] Philip Bachman, R Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. pages 15535–15545, 2019.

[12] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in neural information processing systems*, 2020.

[13] Peter Bartlett, Dylan J Foster, and Matus Telgarsky. Spectrally-normalized margin bounds for neural networks. In *Advances in Neural Information Processing Systems*, 2017.

[14] Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

[15] Peter L Bartlett, Olivier Bousquet, Shahar Mendelson, et al. Local rademacher complexities. *The Annals of Statistics*, 33(4):1497–1537, 2005.

[16] Hugo Berg, Siobhan Mackenzie Hall, Yash Bhalgat, Wonsuk Yang, Hannah Rose Kirk, Aleksandar Shtedritski, and Max Bain. A prompt array keeps the bias away: Debiasing vision-language models with adversarial learning. *arXiv preprint arXiv:2203.11933*, 2022.

[17] Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. Multimodal datasets: misogyny, pornography, and malignant stereotypes. *arXiv preprint arXiv:2110.01963*, 2021.

[18] Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. Language (technology) is power: A critical survey of" bias" in nlp. *arXiv preprint arXiv:2005.14050*, 2020.

[19] Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory*, pages 92–100, 1998.

[20] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29, 2016.

[21] Bernhard E Boser, Isabelle M Guyon, and Vladimir N Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152, 1992.

[22] Yuan Cao and Quanquan Gu. Generalization bounds of stochastic gradient descent for wide and deep neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[23] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[24] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[25] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in neural information processing systems*, pages 2172–2180, 2016.

[26] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

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

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

[29] Jaemin Cho, Abhay Zala, and Mohit Bansal. Dall-eval: Probing the reasoning skills and social biases of text-to-image generative transformers. *arXiv preprint arXiv:2202.04053*, 2022.

[30] Kristy Choi, Aditya Grover, Trisha Singh, Rui Shu, and Stefano Ermon. Fair generative modeling via weak supervision. In *International Conference on Machine Learning*, pages 1887–1898. PMLR, 2020.

[31] François Chollet. keras. `https://github.com/keras-team/keras`, 2015.

[32] Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, pages 539–546. IEEE, 2005.

[33] Marthinus Christoffel, Gang Niu, and Masashi Sugiyama. Class-prior estimation for learning from positive and unlabeled data. In *Asian Conference on Machine Learning*, pages 221–236, 2016.

[34] Ching-Yao Chuang and Stefanie Jegelka. Tree mover's distance: Bridging graph metrics and stability of graph neural networks. *Advances in Neural Information Processing Systems*, 35:2944–2957, 2022.

[35] Ching-Yao Chuang and Youssef Mroueh. Fair mixup: Fairness via interpolation. In *International Conference on Learning Representations*, 2021.

[36] Ching-Yao Chuang, Joshua Robinson, Yen-Chen Lin, Antonio Torralba, and Stefanie Jegelka. Debiased contrastive learning. *Advances in neural information processing systems*, 33:8765–8775, 2020.

[37] Ching-Yao Chuang, Antonio Torralba, and Stefanie Jegelka. Estimating generalization under distribution shifts via domain-invariant representations. In *International Conference on Machine Learning*. PMLR, 2020.

[38] Ching-Yao Chuang, Youssef Mroueh, Kristjan Greenewald, Antonio Torralba, and Stefanie Jegelka. Measuring generalization with optimal transport. *Advances in Neural Information Processing Systems*, 34:8294–8306, 2021.

[39] Ching-Yao Chuang, R Devon Hjelm, Xin Wang, Vibhav Vineet, Neel Joshi, Antonio Torralba, Stefanie Jegelka, and Yale Song. Robust contrastive learning against noisy views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16670–16681, 2022.

[40] Ching-Yao Chuang, Varun Jampani, Yuanzhen Li, Antonio Torralba, and Stefanie Jegelka. Debiasing vision-language models via biased prompts. *arXiv preprint arXiv:2302.00070*, 2023.

[41] Ching-Yao Chuang, Stefanie Jegelka, and David Alvarez-Melis. Infoot: Information maximizing optimal transport. In *International Conference on Machine Learning*, pages 6228–6242. PMLR, 2023.

[42] Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 215–223, 2011.

[43] Corinna Cortes and Vladimir Vapnik. Support-Vector Networks. *Machine learning*, 20(3):273–297, 1995.

[44] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[45] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, 2019.

[46] Bill Dolan, Chris Quirk, and Chris Brockett. Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In *Proceedings of the 20th international conference on Computational Linguistics*, page 350. Association for Computational Linguistics, 2004.

[47] Alexey Dosovitskiy, Jost Tobias Springenberg, Martin Riedmiller, and Thomas Brox. Discriminative unsupervised feature learning with convolutional neural networks. In *Advances in neural information processing systems*, pages 766–774, 2014.

[48] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[49] Marthinus C Du Plessis, Gang Niu, and Masashi Sugiyama. Analysis of learning from positive and unlabeled data. pages 703–711, 2014.

[50] Gintare Karolina Dziugaite and Daniel M Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. In *Proceedings of the International Conference on Uncertainty in Artificial Intelligence*, 2017.

[51] Gintare Karolina Dziugaite, Alexandre Drouin, Brady Neal, Nitarshan Rajkumar, Ethan Caballero, Linbo Wang, Ioannis Mitliagkas, and Daniel M Roy. In search of robust measures of generalization. *Advances in Neural Information Processing Systems*, 33, 2020.

[52] Charles Elkan and Keith Noto. Learning classifiers from only positive and unlabeled data. In *ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 213–220, 2008.

[53] Gamaleldin Elsayed, Dilip Krishnan, Hossein Mobahi, Kevin Regan, and Samy Bengio. Large margin deep networks for classification. *Advances in Neural Information Processing Systems*, 31:842–852, 2018.

[54] William Falcon and Kyunghyun Cho. A framework for contrastive self-supervised learning and designing a new approach. *arXiv preprint arXiv:2009.00104*, 2020.

[55] Mahyar Fazlyab, Alexander Robey, Hamed Hassani, Manfred Morari, and George J Pappas. Efficient and accurate estimation of lipschitz constants for deep neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[56] Herbert Federer. *Geometric measure theory*. Springer, 2014.

[57] Rémi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z Alaya, Aurélie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, et al. Pot: Python optimal transport. *Journal of Machine Learning Research*, 22(78):1–8, 2021.

[58] Nicolas Fournier and Arnaud Guillin. On the rate of convergence in wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162 (3):707–738, 2015.

[59] Felix Friedrich, Patrick Schramowski, Manuel Brack, Lukas Struppek, Dominik Hintersdorf, Sasha Luccioni, and Kristian Kersting. Fair diffusion: Instructing text-to-image generation models on fairness. *arXiv preprint arXiv:2302.10893*, 2023.

[60] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *arXiv preprint arXiv:2110.04544*, 2021.

[61] Rui Gao, Xi Chen, and Anton J Kleywegt. Wasserstein distributionally robust optimization and variation regularization. *arXiv preprint arXiv:1712.06050*, 2017.

[62] Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644, 2018.

[63] Sahin Cem Geyik, Stuart Ambler, and Krishnaram Kenthapadi. Fairness-aware ranking in search & recommendation systems with application to linkedin talent search. In *Proceedings of the 25th acm sigkdd international conference on knowledge discovery & data mining*, pages 2221–2231, 2019.

[64] Aritra Ghosh, Naresh Manwani, and PS Sastry. Making risk minimization tolerant to label noise. *Neurocomputing*, 160:93–107, 2015.

[65] Aritra Ghosh, Himanshu Kumar, and PS Sastry. Robust loss functions under label noise for deep neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

[66] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15180–15190, 2023.

[67] Noah Golowich, Alexander Rakhlin, and Ohad Shamir. Size-independent sample complexity of neural networks. In *Conference On Learning Theory*, pages 297–299. PMLR, 2018.

[68] Hila Gonen and Yoav Goldberg. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. *arXiv preprint arXiv:1903.03862*, 2019.

[69] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[70] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. *Advances in neural information processing systems*, 2020.

[71] Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 855–864, 2016.

[72] Aditya Grover, Jiaming Song, Ashish Kapoor, Kenneth Tran, Alekh Agarwal, Eric J Horvitz, and Stefano Ermon. Bias correction of learned generative models using likelihood-free importance weighting. *Advances in neural information processing systems*, 32, 2019.

[73] Michael Gutmann and Aapo Hyvärinen. Noise-Contrastive Estimation: A new estimation principle for unnormalized statistical models. pages 297–304, 2010.

[74] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742. IEEE, 2006.

[75] Melissa Hall, Laura Gustafson, Aaron Adcock, Ishan Misra, and Candace Ross. Vision-language models performing zero-shot tasks exhibit gender-based disparities. *arXiv preprint arXiv:2301.11100*, 2023.

[76] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. *Advances in neural information processing systems*, 2018.

[77] Tengda Han, Weidi Xie, and Andrew Zisserman. Video representation learning by dense predictive coding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[78] Tengda Han, Weidi Xie, and Andrew Zisserman. Self-supervised co-training for video representation learning. *Advances in neural information processing systems*, 2020.

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

[80] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. *arXiv preprint arXiv:1911.05722*, 2019.

[81] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

[82] Olivier J Hénaff, Aravind Srinivas, Jeffrey De Fauw, Ali Razavi, Carl Doersch, SM Eslami, and Aaron van den Oord. Data-efficient image recognition with contrastive predictive coding. pages 6661–6671, 2020.

[83] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*, 2018.

[84] F Hirzebruch N Hitchin L Hörmander, NJA Sloane B Totaro, and A Vershik M Waldschmidt. Grundlehren der mathematischen wissenschaften 332. 2006.

[85] Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177, 2004.

[86] Po-Sen Huang, Huan Zhang, Ray Jiang, Robert Stanforth, Johannes Welbl, Jack Rae, Vishal Maini, Dani Yogatama, and Pushmeet Kohli. Reducing sentiment bias in language models via counterfactual evaluation. *arXiv preprint arXiv:1911.03064*, 2019.

[87] Pavel Izmailov, Polina Kirichenko, Nate Gruver, and Andrew Gordon Wilson. On feature learning in the presence of spurious correlations. *arXiv preprint arXiv:2210.11369*, 2022.

[88] Shantanu Jain, Martha White, and Predrag Radivojac. Estimating the class prior and posterior from noisy positives and unlabeled data. In *Advances in neural information processing systems*, pages 2693–2701, 2016.

[89] Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *International Conference on Machine Learning*, 2018.

[90] Yiding Jiang, Dilip Krishnan, Hossein Mobahi, and Samy Bengio. Predicting the generalization gap in deep networks with margin distributions. In *International Conference on Learning Representations*, 2018.

[91] Yiding Jiang, Behnam Neyshabur, Hossein Mobahi, Dilip Krishnan, and Samy Bengio. Fantastic generalization measures and where to find them. In *International Conference on Learning Representations*, 2019.

[92] Yiding Jiang, Pierre Foret, Scott Yak, Daniel M Roy, Hossein Mobahi, Gintare Karolina Dziugaite, Samy Bengio, Suriya Gunasekar, Isabelle Guyon, and Behnam Neyshabur. Neurips 2020 competition: Predicting generalization in deep learning. *arXiv preprint arXiv:2012.07976*, 2020.

[93] Yiding Jiang, Parth Natekar, Manik Sharma, Sumukh K Aithal, Dhruva Kashyap, Natarajan Subramanyam, Carlos Lassance, Daniel M Roy, Gintare Karolina Dziugaite, Suriya Gunasekar, et al. Methods and analysis of the first competition in predicting generalization of deep learning. In *NeurIPS 2020 Competition and Demonstration Track*, pages 170–190. PMLR, 2021.

[94] Li Jing, Pascal Vincent, Yann LeCun, and Yuandong Tian. Understanding dimensional collapse in contrastive self-supervised learning. *arXiv preprint arXiv:2110.09348*, 2021.

[95] Longlong Jing and Yingli Tian. Self-supervised spatiotemporal feature learning by video geometric transformations. *arXiv preprint arXiv:1811.11387*, 2(7):8, 2018.

[96] Matt Jordan and Alexandros G Dimakis. Exactly computing the local lipschitz constant of relu networks. In *Advances in Neural Information Processing Systems*, volume 33, pages 7344–7353. Curran Associates, Inc., 2020.

[97] Kimmo Kärkkäinen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age. *arXiv preprint arXiv:1908.04913*, 2019.

[98] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

[99] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33, 2020.

[100] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 2015.

[101] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[102] Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Last layer re-training is sufficient for robustness to spurious correlations. *arXiv preprint arXiv:2204.02937*, 2022.

[103] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023.

[104] Ryan Kiros, Yukun Zhu, Russ R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Skip-thought vectors. In *Advances in neural information processing systems*, pages 3294–3302, 2015.

[105] Ryuichi Kiryo, Gang Niu, Marthinus C du Plessis, and Masashi Sugiyama. Positive-unlabeled learning with non-negative risk estimator. In *Advances in neural information processing systems*, pages 1675–1685, 2017.

[106] Vladimir Koltchinskii, Dmitry Panchenko, et al. Empirical margin distributions and bounding the generalization error of combined classifiers. *Annals of statistics*, 30(1):1–50, 2002.

[107] Bruno Korbar, Du Tran, and Lorenzo Torresani. Cooperative learning of audio and video models from self-supervised synchronization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[108] Alex Krizhevsky et al. Learning multiple layers of features from tiny images. 2009.

[109] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *2011 International conference on computer vision*, pages 2556–2563. IEEE, 2011.

[110] Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. *arXiv preprint arXiv:2202.10054*, 2022.

[111] Vitaly Kuznetsov, Mehryar Mohri, and U Syed. Rademacher complexity margin bounds for learning with a large number of classes. In *ICML Workshop on Extreme Classification: Learning with a Very Large Number of Labels*, 2015.

[112] Carlos Lassance, Louis Béthune, Myriam Bontonou, Mounia Hamidouche, and Vincent Gripon. Ranking deep learning generalization using label variation in latent geometry graphs. *arXiv preprint arXiv:2011.12737*, 2020.

[113] Yann LeCun and Corinna Cortes. MNIST handwritten digit database. 2010. URL `http://yann.lecun.com/exdb/mnist/`.

[114] Sangho Lee, Jiwan Chung, Youngjae Yu, Gunhee Kim, Thomas Breuel, Gal Chechik, and Yale Song. Acav100m: Automatic curation of large-scale datasets for audio-visual video representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10274–10284, 2021.

[115] Yuncheng Li, Jianchao Yang, Yale Song, Liangliang Cao, Jiebo Luo, and Li-Jia Li. Learning from noisy labels with distillation. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

[116] Paul Pu Liang, Irene Mengze Li, Emily Zheng, Yao Chong Lim, Ruslan Salakhutdinov, and Louis-Philippe Morency. Towards debiasing sentence representations. *arXiv preprint arXiv:2007.08100*, 2020.

[117] Tongliang Liu and Dacheng Tao. Classification with noisy labels by importance reweighting. *IEEE Transactions on pattern analysis and machine intelligence*, 38(3):447–461, 2015.

[118] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015.

[119] Lajanugen Logeswaran and Honglak Lee. An efficient framework for learning sentence representations. *arXiv preprint arXiv:1803.02893*, 2018.

[120] Shuang Ma, Zhaoyang Zeng, Daniel McDuff, and Yale Song. Active contrastive learning of audio-visual video representations. In *International Conference on Learning Representations*, 2020.

[121] David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning adversarially fair and transferable representations. *International Conference on Machine Learning*, 2018.

[122] Thomas Manzini, Yao Chong Lim, Yulia Tsvetkov, and Alan W Black. Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings. *arXiv preprint arXiv:1904.04047*, 2019.

[123] K. V. Mardia and P. Jupp. *Directional Statistics*. John Wiley and Sons Ltd., second edition, 2000.

[124] David McAllester and Karl Stratos. Formal limitations on the measurement of mutual information. In *International Conference on Artificial Intelligence and Statistics*, 2020.

[125] David A McAllester. Pac-bayesian model averaging. In *Proceedings of the twelfth annual conference on Computational learning theory*, pages 164–170, 1999.

[126] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35, 2021.

[127] Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-end learning of visual representations from uncurated instructional videos. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9879–9889, 2020.

[128] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[129] P Mishkin, L Ahmad, M Brundage, G Krueger, and G Sastry. Dall· e 2 preview-risks and limitations. *Noudettu*, 28:2022, 2022.

[130] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations*, 2018.

[131] Pedro Morgado, Ishan Misra, and Nuno Vasconcelos. Robust audio-visual instance discrimination. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12934–12945, 2021.

[132] Pedro Morgado, Nuno Vasconcelos, and Ishan Misra. Audio-visual instance discrimination with cross-modal agreement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12475–12486, 2021.

[133] Christopher Morris, Nils M Kriege, Franka Bause, Kristian Kersting, Petra Mutzel, and Marion Neumann. Tudataset: A collection of benchmark datasets for learning with graphs. *arXiv preprint arXiv:2007.08663*, 2020.

[134] Oleg R Musin and Alexey S Tarasov. The tammes problem for n= 14. *Experimental Mathematics*, 24(4):460–468, 2015.

[135] Moin Nadeem, Anna Bethke, and Siva Reddy. Stereoset: Measuring stereotypical bias in pretrained language models. *arXiv preprint arXiv:2004.09456*, 2020.

[136] Vaishnavh Nagarajan and J Zico Kolter. Generalization in deep networks: The role of distance from initialization. *arXiv preprint arXiv:1901.01672*, 2019.

[137] Vaishnavh Nagarajan and J Zico Kolter. Uniform convergence may be unable to explain generalization in deep learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[138] Assaf Naor and Yuval Rabani. On lipschitz extension from finite subsets. *Israel Journal of Mathematics*, 219(1):115–161, 2017.

[139] Annamalai Narayanan, Mahinthan Chandramohan, Rajasekar Venkatesan, Lihui Chen, Yang Liu, and Shantanu Jaiswal. graph2vec: Learning distributed representations of graphs. *arXiv preprint arXiv:1707.05005*, 2017.

[140] Nagarajan Natarajan, Inderjit S Dhillon, Pradeep K Ravikumar, and Ambuj Tewari. Learning with noisy labels. *Advances in neural information processing systems*, 26:1196–1204, 2013.

[141] Parth Natekar and Manik Sharma. Representation based complexity measures for predicting generalization in deep learning. *arXiv preprint arXiv:2012.02775*, 2020.

[142] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011.

[143] Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. Norm-based capacity control in neural networks. In *Conference on Learning Theory*, pages 1376–1401. PMLR, 2015.

[144] Behnam Neyshabur, Srinadh Bhojanapalli, and Nathan Srebro. A pac-bayesian approach to spectrally-normalized margin bounds for neural networks. In *International Conference on Learning Representations*, 2018.

[145] Behnam Neyshabur, Zhiyuan Li, Srinadh Bhojanapalli, Yann LeCun, and Nathan Srebro. Towards understanding the role of over-parametrization in generalization of neural networks. In *International Conference on Learning Representations*, 2019.

[146] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European Conference on Computer Vision*, pages 69–84. Springer, 2016.

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

[148] OpenAI. Gpt-4 technical report. *arXiv*, 2023.

[149] Sherjil Ozair, Corey Lynch, Yoshua Bengio, Aaron van den Oord, Sergey Levine, and Pierre Sermanet. Wasserstein dependency measure for representation learning. *arXiv preprint arXiv:1903.11780*, 2019.

[150] Bo Pang and Lillian Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd annual meeting on Association for Computational Linguistics*, page 271. Association for Computational Linguistics, 2004.

[151] Bo Pang and Lillian Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd annual meeting on association for computational linguistics*, pages 115–124. Association for Computational Linguistics, 2005.

[152] Otávio Parraga, Martin D More, Christian M Oliveira, Nathan S Gavenski, Lucas S Kupssinskü, Adilson Medronha, Luis V Moura, Gabriel S Simões, and Rodrigo C Barros. Debiasing methods for fairer neural models in vision and language research: A survey. *arXiv preprint arXiv:2211.05617*, 2022.

[153] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. URL `http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf`.

[154] Mandela Patrick, Yuki M Asano, Polina Kuznetsova, Ruth Fong, João F Henriques, Geoffrey Zweig, and Andrea Vedaldi. On compositions of transformations in contrastive self-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

[155] Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. Making deep neural networks robust to label noise: A loss correction approach. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1944–1952, 2017.

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

[157] Ben Poole, Sherjil Ozair, Aaron Van Den Oord, Alex Alemi, and George Tucker. On variational bounds of mutual information. In *International Conference on Machine Learning*, pages 5171–5180. PMLR, 2019.

[158] Filip Radenovic, Abhimanyu Dubey, Abhishek Kadian, Todor Mihaylov, Simon Vandenhende, Yash Patel, Yi Wen, Vignesh Ramanathan, and Dhruv Mahajan. Filtering, distillation, and hard negatives for vision-language pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6967–6977, 2023.

[159] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[160] Vikram V Ramaswamy, Sunnie SY Kim, and Olga Russakovsky. Fair attribute classification through latent space de-biasing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9301–9310, 2021.

[161] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

[162] Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. Learning to reweight examples for robust deep learning. In *International Conference on Machine Learning*, pages 4334–4343. PMLR, 2018.

[163] Joshua Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. Contrastive learning with hard negative samples. In *International Conference on Learning Representations*, 2021.

[164] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

[165] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

[166] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022.

[167] Kevin Scaman and Aladin Virmaux. Lipschitz regularity of deep neural networks: analysis and efficient estimation. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 3839–3848, 2018.

[168] Yair Schiff, Brian Quanz, Payel Das, and Pin-Yu Chen. Gi and pal scores: Deep neural network generalization statistics. *arXiv preprint arXiv:2104.03469*, 2021.

[169] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. pages 815–823, 2015.

[170] K Schütte and BL Van der Waerden. Auf welcher kugel haben 5, 6, 7, 8 oder 9 punkte mit mindestabstand eins platz? *Mathematische Annalen*, 123(1): 96–124, 1951.

[171] Pierre Sermanet, Corey Lynch, Yevgen Chebotar, Jasmine Hsu, Eric Jang, Stefan Schaal, Sergey Levine, and Google Brain. Time-contrastive networks:

Self-supervised learning from video. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1134–1141, 2018.

[172] Ashish Seth, Mayur Hemani, and Chirag Agarwal. Dear: Debiasing vision-language models with additive residuals. *arXiv preprint arXiv:2303.10431*, 2023.

[173] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in neural information processing systems*, pages 4077–4087, 2017.

[174] Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. *Advances in neural information processing systems*, 29, 2016.

[175] Justin Solomon, Kristjan Greenewald, and Haikady N Nagaraja. k-variance: A clustered notion of variance. *arXiv preprint arXiv:2012.06958*, 2020.

[176] Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. Deep metric learning via lifted structured feature embedding. pages 4004–4012, 2016.

[177] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

[178] Aravind Srinivas, Michael Laskin, and Pieter Abbeel. CURL: Contrastive unsupervised representations for reinforcement learning. pages 10360–10371, 2020.

[179] Sainbayar Sukhbaatar, Joan Bruna, Manohar Paluri, Lubomir Bourdev, and Rob Fergus. Training convolutional networks with noisy labels. In *3rd International Conference on Learning Representations, ICLR 2015*, 2015.

[180] Chen Sun, Fabien Baradel, Kevin Murphy, and Cordelia Schmid. Learning video representations using contrastive bidirectional transformer. *arXiv preprint arXiv:1906.05743*, 2019.

[181] Fan-Yun Sun, Jordan Hoffmann, Vikas Verma, and Jian Tang. InfoGraph: Unsupervised and semi-supervised graph-level representation learning via mutual information maximization. 2020.

[182] Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. Mitigating gender bias in natural language processing: Literature review. *arXiv preprint arXiv:1906.08976*, 2019.

[183] Pieter Merkus Lambertus Tammes. On the origin of number and arrangement of the places of exit on the surface of pollen-grains. *Recueil des travaux botaniques néerlandais*, 27(1):1–84, 1930.

[184] Yuval Tassa, Yotam Doron, Alistair Muldal, Tom Erez, Yazhe Li, Diego de Las Casas, David Budden, Abbas Abdolmaleki, Josh Merel, Andrew Lefrancq, et al. Deepmind control suite. *arXiv preprint arXiv:1801.00690*, 2018.

[185] Christopher TH Teo and Ngai-Man Cheung. Measuring fairness in generative models. *arXiv preprint arXiv:2107.07754*, 2021.

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

[187] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 776–794. Springer, 2020.

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

[189] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018.

[190] Adrien Vacher, Boris Muzellec, Alessandro Rudi, Francis Bach, and Francois-Xavier Vialard. A dimension-free computational upper-bound for smooth optimal transport estimation. *arXiv preprint arXiv:2101.05380*, 2021.

[191] Guillermo Valle-Perez, Chico Q Camargo, and Ard A Louis. Deep learning generalizes because the parameter-function map is biased towards simple functions. In *International Conference on Learning Representations*, 2019.

[192] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[193] Vladimir N Vapnik and A Ya Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. In *Measures of complexity*, pages 11–30. Springer, 2015.

[194] Andreas Veit, Neil Alldrin, Gal Chechik, Ivan Krasin, Abhinav Gupta, and Serge Belongie. Learning from noisy large-scale datasets with minimal supervision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017.

[195] Petar Veličković, William Fedus, William L Hamilton, Pietro Liò, Yoshua Bengio, and R Devon Hjelm. Deep graph infomax. *arXiv preprint arXiv:1809.10341*, 2018.

[196] Ellen M Voorhees. Overview of trec 2002.

[197] Jialu Wang, Yang Liu, and Xin Eric Wang. Are gender-neutral queries really gender-neutral? mitigating gender bias in image search. *arXiv preprint arXiv:2109.05433*, 2021.

[198] Mei Wang and Weihong Deng. Mitigating bias in face recognition using skewness-aware reinforcement learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9322–9331, 2020.

[199] Mei Wang, Weihong Deng, Jiani Hu, Xunqiang Tao, and Yaohai Huang. Racial faces in the wild: Reducing racial bias by information maximization adaptation network. In *Proceedings of the ieee/cvf international conference on computer vision*, pages 692–702, 2019.

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

[201] Yisen Wang, Xingjun Ma, Zaiyi Chen, Yuan Luo, Jinfeng Yi, and James Bailey. Symmetric cross entropy for robust learning with noisy labels. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 322–330, 2019.

[202] Zeyu Wang, Klint Qinami, Ioannis Christos Karakozis, Kyle Genova, Prem Nair, Kenji Hata, and Olga Russakovsky. Towards fairness in visual recognition: Effective strategies for bias mitigation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8919–8928, 2020.

[203] Jonathan Weed and Francis Bach. Sharp asymptotic and finite-sample rates of convergence of empirical measures in wasserstein distance. *Bernoulli*, 25(4A): 2620–2648, 2019.

[204] Colin Wei and Tengyu Ma. Improved sample complexities for deep networks and robust classification via an all-layer margin. In *International Conference on Learning Representations*, 2020.

[205] Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*, 2021.

[206] Janyce Wiebe, Theresa Wilson, and Claire Cardie. Annotating expressions of opinions and emotions in language. *Language resources and evaluation*, 39(2-3): 165–210, 2005.

[207] Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, et al. Robust fine-tuning of zero-shot models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7959–7971, 2022.

[208] Zhirong Wu, Yuanjun Xiong, Stella Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance-level discrimination. *arXiv preprint arXiv:1805.01978*, 2018.

[209] Tong Xiao, Tian Xia, Yi Yang, Chang Huang, and Xiaogang Wang. Learning from massive noisy labeled data for image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2691–2699, 2015.

[210] Chang Xu, Dacheng Tao, and Chao Xu. A survey on multi-view learning. *arXiv:1304.5634*, 2013.

[211] Dejing Xu, Jun Xiao, Zhou Zhao, Jian Shao, Di Xie, and Yueting Zhuang. Self-supervised spatiotemporal learning via video clip order prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10334–10343, 2019.

[212] Depeng Xu, Shuhan Yuan, Lu Zhang, and Xintao Wu. Fairgan: Fairness-aware generative adversarial networks. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 570–575. IEEE, 2018.

[213] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In *International Conference on Learning Representations*, 2018.

[214] Xueting Yan, Ishan Misra, Abhinav Gupta, Deepti Ghadiyaram, and Dhruv Mahajan. Clusterfit: Improving generalization of visual representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6509–6518, 2020.

[215] Yuning You, Tianlong Chen, Yongduo Sui, Ting Chen, Zhangyang Wang, and Yang Shen. Graph contrastive learning with augmentations. *Advances in Neural Information Processing Systems*, 33:5812–5823, 2020.

[216] Yuning You, Tianlong Chen, Yang Shen, and Zhangyang Wang. Graph contrastive learning automated. *arXiv preprint arXiv:2106.07594*, 2021.

[217] Xingrui Yu, Bo Han, Jiangchao Yao, Gang Niu, Ivor Tsang, and Masashi Sugiyama. How does disagreement help generalization against label corruption? In *International Conference on Machine Learning*, pages 7164–7173. PMLR, 2019.

[218] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *Proceedings of the British Machine Vision Conference (BMVC)*, pages 87.1–87.12. BMVA Press, September 2016.

[219] John Zarka, Florentin Guth, and Stéphane Mallat. Separation and concentration in deep networks. In *International Conference on Learning Representations*, 2021.

[220] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*. PMLR, 2021.

[221] Xiaohua Zhai, Avital Oliver, Alexander Kolesnikov, and Lucas Beyer. S4l: Self-supervised semi-supervised learning. In *Proceedings of the IEEE international conference on computer vision*, pages 1476–1485, 2019.

[222] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*, 2017.

[223] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018.

[224] Michael Zhang and Christopher Ré. Contrastive adapters for foundation model group robustness. *arXiv preprint arXiv:2207.07180*, 2022.

[225] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *European conference on computer vision*, pages 649–666. Springer, 2016.

[226] Zhilu Zhang and Mert R Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. *Advances in neural information processing systems*, 2018.

[227] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *arXiv preprint arXiv:1707.09457*, 2017.

[228] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. Gender bias in contextualized word embeddings. *arXiv preprint arXiv:1904.03310*, 2019.