InfoNCE Loss Provably Learns Cluster-Preserving Representations

Advait Parulekar ADVAITP@UTEXAS.EDU
University of Texas at Austin

Liam Collins LIAMC@UTEXAS.EDU
University of Texas at Austin

Karthikeyan Shanmugam KARTHIKEYANVS@GOOGLE.COM
Google Research India

Aryan Mokhtari MOKHTARI@AUSTIN.UTEXAS.EDU
University of Texas at Austin

Sanjay Shakkottai SANJAY.SHAKKOTTAI@UTEXAS.EDU
University of Texas at Austin

Editors: Gergely Neu and Lorenzo Rosasco

Abstract

The goal of contrasting learning is to learn a representation that preserves underlying clusters by keeping samples with similar content, e.g. the "dogness" of a dog, close to each other in the space generated by the representation. A common and successful approach for tackling this unsupervised learning problem is minimizing the InfoNCE loss associated with the training samples, where each sample is associated with their augmentations (positive samples such as rotation, crop) and a batch of negative samples (unrelated samples). To the best of our knowledge, it was unanswered if the representation learned by minimizing the InfoNCE loss preserves the underlying data clusters, as it only promotes learning a representation that is faithful to augmentations, i.e., an image and its augmentations have the same representation. Our main result is to show that the representation learned by InfoNCE with a finite number of negative samples is also consistent with respect to *clusters* in the data, under the condition that the augmentation sets within clusters may be non-overlapping but are close and intertwined, relative to the complexity of the learning function class.

Keywords: Contrastive learning, Representation learning, Self-supervised learning

1. Introduction

Representations pretrained on partially or completely unlabeled data are becoming ubiquitous in machine learning applications (Peters et al., 2018; Radford et al., 2021), in large part due to the availability of large unlabeled datasets and significant computing power offline, and the effectiveness of self-supervised representation learning algorithms, especially contrastive learning (CL). CL aims to learn representations that treat natural images similarly to their augmentations, while maximizing the average distance between random pairs of images. In recent years CL has demonstrated numerous successes in pretraining representations with unlabeled data that learn meaningful relationships between data points that generalize well to downstream tasks in computer vision (Hjelm et al., 2018; Oord et al., 2018; Bachman et al., 2019; Caron et al., 2020; Chen et al., 2020a,b; He et al., 2020; Henaff, 2020; Li et al., 2020; Misra and Maaten, 2020; Tian et al., 2020a,b) and natural language processing (Brown et al., 2020; Gao et al., 2021; Su et al., 2021; Radford et al., 2019).

Despite its empirical success, it is not well-understood how CL learns meaningful relationships between data points. Since data are unlabeled, the only immediate structure in datasets leveraged

by CL are the sets of images and their augmentations. Without further assumptions, this structure is insufficient to learn relationships between images across augmentation sets. To circumvent this issue there are two approaches. The first is to assume that augmentation sets of semantically similar natural images overlap, i.e. for two images of cats, some of the augmentations of each image are equivalent (Arora et al., 2019; HaoChen et al., 2021, 2022; Shen et al., 2022; Wang et al., 2022). However, this assumption is unlikely to hold in practice, as pointed out by recent work (Saunshi et al., 2022). The second approach is to consider inductive biases of the representation function class and/or optimization algorithm and use these to argue that only certain types of representations (that capture semantic relationships) can be learned.

Prior works have initiated the study of how inductive biases of the representation class can lead to meaningful representations in CL (Saunshi et al., 2022; HaoChen and Ma, 2022), but their analysis is for the spectral contrastive loss, which is not used in practice. Instead, the vast majority of CL approaches, including the widely popular SimCLR framework (Chen et al., 2020a), optimize a loss function based on InfoNCE (Gutmann and Hyvärinen, 2010; Oord et al., 2018). A variety of works have studied properties of the InfoNCE loss, but due to its unwieldy log-sum structure have made restrictive assumptions, such as having infinite (Wang and Isola, 2020; Robinson et al., 2020; Von Kügelgen et al., 2021) or only a single (Tosh et al., 2021; Huang et al., 2021) negative sample(s).

Main Contributions. We are given a collection of clusters of natural images, with each image associated with augmentations (positive samples such as 'rotation') and a finite set of negative samples (unrelated images). Using the InfoNCE loss, our goal is to learn a d dimensional representation $g \in G$, where $g = (f_1, f_2, \ldots, f_d)$ and $f \in \{f_i\}$ are binary functions mapping images to $\{-1, 1\}$ (thus g maps images on the hypercube $H_d = \{-1, 1\}^d$). Our setting is one where the function class has bounded expressivity with respect to the augmentation sets, meaning that the augmentation sets within clusters are intertwined, and hard to separate from the rest of the cluster using functions in F.

(Realizable Setting) Suppose there exists a representation $g \in G$ that is: *(a)* cluster preserving, and *(b)* different clusters of images are uniformly mapped over distinct vertices on the hypercube (qualitatively, class-balance in the image dataset). We show that with *any finite number of negative samples*, the representation learned by the InfoNCE loss is cluster-preserving and uniform. Furthermore, this learned representation when composed with a two-layer ReLU head, achieves zero downstream error on any cluster-preserving binary classification task. Our proof hinges on a novel Markov Chain construction showing that the InfoNCE loss of any non-uniform representation can be improved by "blurring" the representation through the Markov Chain transitions. Conversely, we show that solutions to the InfoNCE loss optimized over an arbitrarily powerful representation class $G_?$ cannot have meaningful downstream performance guarantees on such tasks.

(Agnostic Setting) In the agnostic (non-realizable) case, through sensitivity analysis, we show that for any close-to-uniform and non-cluster-preserving representation, there exists a representation that preserves one additional cluster and thus improves the InfoNCE loss. Our proof uses a novel partitioning of the image space that is of independent interest for future analysis of the InfoNCE loss.

1.1. Related Work

Several works have aimed to explain the success of contrastive learning in recent years. Wang and Liu (2021) and Wang and Isola (2020) showed empirically that CL encourages aligned and uniform representations, and improving alignment and uniformity improves downstream performance. The work in Chen et al. (2021) generalizes the InfoNCE loss to a larger family of losses with alignment

and uniformity terms weighted according to a hyperparameter. Early theoretical studies attributed the success of CL to its proclivity to maximize the mutual information between augmentations of the same image (Bachman et al., 2019), but later work cast doubt on this viewpoint by showing that optimizing a tighter bound on the mutual information leads to worse performance (McAllester and Stratos, 2020; Tschannen et al., 2019). The work in (Wang and Isola, 2020) further showed that solutions to the InfoNCE loss are aligned and uniform in the limit of infinite negative samples per batch.

A variety of works have studied CL's ability to recover meaningful clusters or latent variables in the data (Arora et al., 2019; Tosh et al., 2021; Zimmermann et al., 2021; Ash et al., 2021; Nozawa and Sato, 2021; HaoChen et al., 2021; Shen et al., 2022; HaoChen et al., 2022; HaoChen and Ma, 2022; Wang et al., 2022; Awasthi et al., 2022; Bao et al., 2022). However, the majority of these works consider arbitrary function classes, which requires strong assumptions on the connectedness of augmentation sets within each cluster, such as assuming positive pairs are conditionally independent given their cluster identity, in order to give downstream guarantees (Saunshi et al., 2022). The work by HaoChen and Ma (2022) is the most related work to ours, as they study function classes that induce a similar bias towards preserving clusters as ours without any assumption on the connectedness of augmentation sets. However, their study is focused on minimizing a spectral contrastive loss which serves as a surrogate for the more practically used InfoNCE loss. While studying spectral contrastive loss is enlightening and provides some intuition, it cannot be extended to the InfoNCE loss because of two major reasons: First, the loss function fails to highlight the role of finite batches of negative samples, which is a well-studied and key component of the InfoNCE loss (Awasthi et al., 2022; Bao et al., 2022; Ash et al., 2021; Nozawa and Sato, 2021). Second, their analysis does not translate to our setting because the key difficulty in our proof is to show that negative samples promote uniformity; this aspect directly follows with the spectral loss due to the covariance regularizer.

Additional theoretical works have studied the feature learning process of CL with (stochastic) gradient descent on linear (Tian, 2022a; Ji et al., 2021) and two-layer ReLU neural networks (Wen and Li, 2021; Tian, 2022b), properties augmentations must satisfy in order for CL to be successful (Tian et al., 2020b), the role of the projection head in CL (Wen and Li, 2022; Gupta et al., 2022), and the behavior of contrastive losses in (semi-)supervised settings (Khosla et al., 2020; Zheng et al., 2021; Chen et al., 2022). Several other works analyze non-contrastive self-supervised learning methods (Wei et al., 2020; Balestriero and LeCun, 2022; Garrido et al., 2022; Lee et al., 2021).

2. Problem Formulation

Our learning task consists of *(i)* a pretraining phase – wherein we are not provided supervised labels but rather only *associations* between images and *(ii)* a supervised learning phase in which we are provided (a few) labeled data points, labeled according to some specific downstream task. During the pre-training phase, we do not know what the downstream task is. However, we are provided *augmentations* of the raw data points that the learner knows should be classified the same way as the raw data for *any* downstream task. In a sense, the augmentations can be seen as modifying the data in a way that leaves the information contained in the data invariant with respect to the downstream tasks. Ideally, we aim to learn a representation that is invariant to such augmentations so that downstream learning can be statistically efficient. For interpretability, we will work in the setting of "images".

Images and augmentations. The images consist of features that are either important for classification or which function only as irrelevant details. Inspired by (Von Kügelgen et al., 2021),

we consider an image generation model that consists of *(i)* content variables denoted by $\mathbf{c}$ which capture innate qualities of the images (e.g., the 'catness' of a cat), and *(ii)* style variables denoted by $\mathbf{s}$ which capture the appearance of the image (e.g., 'rotation' and 'crop' for creating augmentations to an image; 'long tail' and 'furry' for different natural images of dogs). More precisely, each image x is generated according to $x = I(\mathbf{c}; \mathbf{s})$, where $I(\cdot;\cdot)$ is a mapping from the space of content and style variables to the space of images. We assume that the natural images are generated such that their content variables $\mathbf{c}$ belong to the set C and their style variables $\mathbf{s}$ belong to the set S .

We further consider that there is a set of augmentations , which is a set of functions mapping natural images to augmented images. An augmented image of an image x is denoted $A(x)$, where A . We assume that the augmented image preserves the content of the original image, while its style may differ from the original image. More precisely, if the original image is given by $x = I(\mathbf{c}; \mathbf{s})$, then its augmented image $A(x)$ satisfies the following property: $A(x) = A(I(\mathbf{c}; \mathbf{s})) = I(\mathbf{c}; \mathbf{s}^+)$ for some $\mathbf{s}^+$, where the set S contains S . So the augmented images have possibly different style variables *but the same content variables* as the natural images. Further, the set of augmented images of the image $x = I(\mathbf{c}; \mathbf{s})$ is called its *augmentation set* and is defined as $A(x) := A(I(\mathbf{c}; \mathbf{s})) := \{A(I(\mathbf{c}; \mathbf{s}))| A \}$, with all images having equal-sized augmentation sets for simplicity. We typically refer to an image $I(\mathbf{c}; \mathbf{s})$ as x and its augmentation $I(\mathbf{c}; \mathbf{s}^+)$ $A(x)$ as x^+, where, for all sets of images B, B denotes a random sample drawn uniformly from the set B. We let D denote the set of all images and their augmentations and D D denote the set of all natural images.

Representations and heads. We consider a function class F of binary functions, $f(x') \in \{-1, +1\}$, where x' is either an image x or its augmentation x^+. This is a function class with bounded expressivity (e.g., a class of functions that can be expressed as the thresholded output of a neuron from a neural network with bounded width and depth). The function class F and the augmentations define a set of *clean* functions F_c F that separate the data in a way that respects the augmentations, $F_c = \{f : f(x) = f(A(x)) \ \forall x \in D, A \in \}$. In other words, the binary function f is clean if it does not separate any image and its augmentations from each other.

We search over d-dimensional representations, denoted by G, such that each coordinate of the representation is an element of F, i.e., $g = (f_1, f_2, \ldots, f_d)$,. Thus a representation $g \in G := F^d$ is simply a concatenation of d binary classifiers, mapping an image x to the vertex of the Rademacher hypercube[1] $H_d = \{-1, 1\}^d$. Note that each $g \in G$ denotes only the representation (e.g., the body of a neural network). For downstream tasks, a full classifier is formed by composing g with a *head* $h \in J$ for some class J of heads (e.g., the final classification layer of a neural network).

Clusters. We consider an equivalence class on content variables such that

$$\mathbf{c} \quad \mathbf{c}' \quad ( ) \quad f(I(\mathbf{c}; \mathbf{s})) = f(I(\mathbf{c}'; \mathbf{s}')) \ \forall f \in F_c :$$

We refer to the images in the equivalence classes as clusters $_{\mathbf{c}}$, so

$$_{\mathbf{c}} = \{x : \exists \mathbf{c}' \quad \mathbf{c}, \mathbf{s} \text{ such that } x = I(\mathbf{c}', \mathbf{s})\}:$$

As an example, suppose that the content $\mathbf{c}$ captures the 'dogness' of an image. Then, different images of dogs would have the same content, but have different style variables (e.g., furry, skinny, long ears).

1. Representations in CL often map to the unit hypersphere (Wang and Isola, 2020). Here, we consider a discretized version of this output space for two reasons: (1) it allows us to construct a naturally restricted representation function class by extending natural properties of binary classifiers, and (2) it provides a tractable setting for us to show the first results that InfoNCE prefers cluster-preserving and uniform representations with finite samples, as it is still an open problem to determine uniform arrangements of finite points on the unit hypersphere (Thomson, 1904).

Recall that the augmentations of an image also share the same content, but the style might be chosen from a different set (e.g., rotation, cropping, blur).

Our partition of the space of images into clusters highlights the interplay between the richness of the function class F and the diversity in the augmentations . A more diverse set of augmentations generally results in a smaller set F_c, and hence generally increases the size of the clusters. Meanwhile, a richer class F generally results in a larger F_c, and smaller clusters.

In this work we show conditions under which for a fixed choice of F and , contrastive pretraining learns the equivalence class. This is useful for solving downstream classification tasks, as described below.

Goal of pretraining. Ultimately, we aim to find a representation that allows for easily solving tasks from a set of possible downstream binary classification tasks $h \in T$, where each task h maps an image to a binary label $f \; 1; 1g$. These tasks are assumed to be faithful to the clusters, meaning that for any pair of images $x; x^0$ belonging to the same cluster, $h(x) = h(x^0)$.

Note that during pretraining, the learner does not have any knowledge about which task will be assigned among the solvable ones. After pretraining, the learner fixes the representation but can learn a task-specific head when it encounters a downstream task. We define the error a representation g on the downstream task $h \in T$ with respect to the class $J \quad f ! : \mathbb{R}^d \; ! \; \mathbb{R}g$ of allowed heads as

$$L_{h;J}(g) := \inf_{! \; J} \mathrm{P}_{x \;\; D}[! \quad g(x) \not\in h(x)]: \tag{1}$$

The error of g on a family of downstream tasks $T \quad F_c$ is the worst case error among tasks in T:

$$L_{T;J}(g) := \sup_{h \in T} L_{h;J}(g): \tag{2}$$

To summarize, for a task that is realizable *with supervision* using function class F, we would like to learn a representation entirely from unlabelled data such that the task on the embedded images is still realizable for J. The overall motivation is that learning $! \in J$ can generally require fewer *labeled* samples than learning the joint model $! \quad g$.

2.1. InfoNCE loss

We denote $\mathrm{E}_{x;x^+} := \mathrm{E}_{x \;\; D \;;x^+ \;\; A(x)}$ and $\mathrm{E}_{x;x^+;\{x_i\}} := \mathrm{E}_{x \;\; D \;;x^+ \;\; A(x);\{x_i\} \;\; D}$ for simplicity. The InfoNCE loss we consider is given by[2]

$$L(g) = \underbrace{\mathrm{E}_{x;x^+}[g(x)^\top g(x^+)]}_{\text{alignment}} + \mathrm{E}_{x;x^+;\{x_i\}} \left[\log \left(e^{g(x)^\top g(x^+)} + \sum_{i=1} e^{g(x)^\top g(x_i)} \right) \right] \tag{3}$$

Following Wang and Isola (2020), we refer to the first term as the *alignment* term, or the *positive* term, and we refer to the second term as the *uniformity* term or the *negative* term. By minimizing the first term, we are maximizing the alignment between the representation of an image and its augmentation, and by minimizing the second term we are enforcing the representation of different images to be as different as possible.

2. For ease of exposition we consider the case wherein negative samples are drawn from the set of natural images, as in (Wen and Li, 2021). Although this may not hold in practice, it greatly simplifies the presentation of our results.

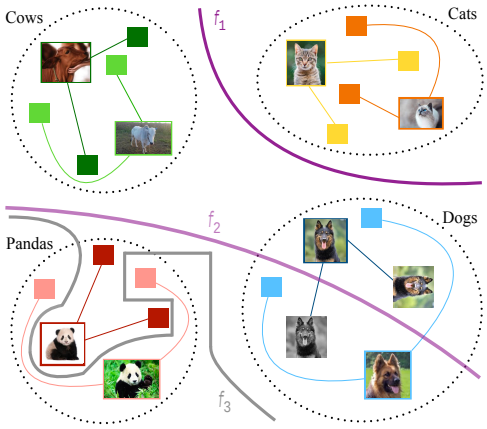

Figure 1: We illustrate a setting with four clusters, and two augmentation sets within each cluster indicated by linked rectangles in distinct colors. The clean function f_1 does not split any augmentation or cluster (meaning, it maps all images from the same augmentation set alike, and likewise for clusters). f_2 splits the cluster of dogs, and in accordance with Assumption 2, also splits augmentation sets within that cluster (and is therefore non-clean). Finally, f_3 violates Assumption 2, because it does not split any augmentation of pandas, yet it splits the pandas cluster. Images are from the Animals V2 dataset: DeepNets (2022).

The above formulation suggests that the representation learned by minimizing the above loss forces images and their augmentations to have a similar representation. What we show in the following sections is a stronger result which guarantees by minimizing the InfoNCE loss, all images that belong to the same cluster (share the same content) will have a similar representation.

3. Bounded Function Class

The goal of contrastive learning is to learn a representation from unlabeled samples that is useful for downstream tasks. Recall that the representations we consider map images to vertices on the Rademacher hypercube H_d. A "good" representation should map images from the same cluster to the same vertex, and images from distinct clusters to distinct vertices.

Intuitively, this seems possible if images having the same content (i.e., from the same cluster) along with their set of augmentations are "close and intertwined" (see Figure 1), such that any function $f \in F$ cannot split the cluster without also splitting an image from its augmentation. Note that we do *not* need connected clusters with overlapping augmentations (meaning two images have the same augmentation, which is an unrealistic assumption); merely that the cluster has a complex geometry relative to the function class.

3.1. Complexity of F Relative to Augmentations

We formalize the notion of bounded expressivity of F relative to the geometry of clusters. We use this assumption to show in Section 4 that solutions to the InfoNCE loss optimized over G satisfy useful uniformity and alignment properties that lead to downstream performance guarantees on tasks that adhere to the clusters. Formally, the function class F and the augmentations define a set of *clean* functions F_c F that separate the data in a way that respects the augmentations.

Definition 1 (Clean Function) $f \in F_c$ is clean $( )$ $f(x) = f(A(x))) \; 8x \in D \; ; A \in$.

In other words, the binary function f is clean if it does not separate any image and its augmentations from each other. Our main assumption is that if a classifier in F splits a cluster, then it is not clean.

Assumption 2 (Intertwined Augmentations) *For all* $f \in F$, *if* $f(x) \neq f(x')$ *for some* $x; x' \in$ $_c$, *then* $f(x'') \neq f(A(x''))$ *for some* $x'' \in$ $_c \setminus D$, *where* $A(x'') \in A(x'')$.

Note that if a classifier does not split any cluster, then it must be clean, since augmentation sets are contained within clusters. Thus, Assumption 2 implies that $f \in F_c$ if and only if f labels all images with the same content (belonging to the same cluster) alike, in other words it is *cluster-preserving*. This assumption holds if the augmentation sets within clusters are close and intertwined (they cannot be easily split from the rest of the cluster), relative to the complexity of F. Importantly, the augmentation sets need not overlap, meaning a single image need not be an augmentation to multiple natural images, consistent with practice (Saunshi et al., 2022). As prior works have pointed out (Saunshi et al., 2022; HaoChen and Ma, 2022), Assumption 2 or variants on the bounded complexity of the function class are necessary for the success of CL in the realistic setting in which the augmentation sets do not overlap.

However, while some condition like Assumption 2 is necessary, it is not clear if this suffices to show that CL learns useful representations. Consider the example in Figure 1. It may be the case, for instance, that CL on F does not learn the cluster-preserving classifiers, as in addition to trying to maximize the similarity between images and their augmentations, CL also tries to minimize the similarity between negative pairs of images. Thus, it may choose a non-cluster-preserving classifier such as f_2 in an effort to minimize similarity of negative pairs. This would lead to poor downstream generalization on tasks involving classifying dogs, since f_2 separates images of dogs. It thus becomes critical to quantify the extent to which non-cluster-preserving classifiers must intersect augmentation sets such that CL will not learn them, as we do in Section 5. Before this, we must show that even if CL learns a representation consisting of cluster-preserving classifiers, this representation generalizes well, which may not happen if it maps two or more clusters to the same vertex. For instance, if CL simply learned d copies of the cluster-preserving classifier f_1 in Figure 1, this representation would not be able to distinguish cows from pandas from dogs on downstream tasks. We thus desire representations to be *both* cluster-preserving and uniform such that their mapping is a bijection from clusters to vertices. Next, we show that when a cluster-preserving and uniform representation is realizable, CL with the InfoNCE loss learns it, even with finite negative samples per batch.

4. Results for the Realizable Setting

Our first result shows that when the dataset D and representation class G allow for mapping the data *uniformly* on the hypercube in a *cluster-preserving* manner, then the representation learned by minimizing the InfoNCE loss over G results in such a mapping. We first formally define the terms *uniform* and *cluster-preserving* below.

Definition 3 (Cluster-Preserving) *A* cluster-preserving *representation $g \in G$ is one that for all $c \in C$ and all $x, x' \in A_c$, $g(x) = g(x')$.*

Definition 4 (Uniform) *A* uniform *representation $g \in G$ satisfies $\Pr_{x \sim D}[g(x) = v] = 2^{-d}$ for all $v \in H_d$.*

Next, our results in this section assume a cluster-preserving and uniform representation exists in G.

Assumption 5 (Realizability) *There exists a $g \in G$ that is both cluster-preserving and uniform.*

In order for there to exist a representation that is both cluster-preserving and uniform, there must be an integral multiple of 2^d clusters in the dataset and they must be balanced. Before stating our main result, we must prove a key lemma that shows that among all "clean" representations, those

that minimize the InfoNCE loss are uniform. We define $G_c \subset G$ as the set of clean representations in G consisting of d concatenated clean classifiers from F_c.

Lemma 6 *If Assumptions 2 and 5 hold, $N > c \log d$ for an absolute constant c, $d > 3$, and $\ell \geq 1$, then $g \in \arg\min_{g \in G_c} L(g)$ if and only if g is uniform.*

Proof [Proof sketch] Since the optimization problem is over representations composed of clean functions, we know that for all $g \in G_c$, the term $g(x)^\top g(x^+)$ in the InfoNCE loss is exactly equal to d. Hence, by regrouping the terms in (3), the optimization problem simplifies to:

$$\min_{g \in G_c} L(g) = \min_{g \in G_c} \hat{L}(g) := \mathop{\mathbb{E}}_{x, x^+, \{x_i\}_{i=1}^{N\ell}} \log\left(1 + \sum_{i=1}^{N\ell} e^{g(x)^\top g(x_i) - d}\right) \tag{4}$$

By Assumption 5, at least one uniform representation belongs to the set G_c. We show that it minimizes the loss $\hat{L}(g)$. To do so, we observe that we can think of minimizing $\hat{L}(g)$ as an optimization with respect to distributions over the hypercube induced by g. To better understand this connection, consider the random variable $g(x)$ for $x \sim D$. Further, denote the corresponding induced distribution over $g(x)$ as D_g, i.e., D_g is a distribution over the vertices of the hypercube H_d. Letting $y = g(x)$, the objective above can now be rewritten in terms of these distributions:

$$\min_{\{D_g : g \in G_c\} g} L(D_g) := \mathop{\mathbb{E}}_{y, \{y_i\}_{i=1}^{N\ell} \sim D_g} \log\left(1 + \sum_{i=1}^{N\ell} e^{y^\top y_i - d}\right) \tag{5}$$

Suppose the the minimizing distribution was not uniform over the hypercube, i.e. for $D_g \in \arg\min_{\{D_g : g \in G_c\} g} L(D_g)$, $D_g \neq U$, where U is the uniform distribution over the hypercube H_d. For any sample $y, \{y_i\} \sim D_g$, consider a random walk that starts from this sample and evolves over time. For this random walk, denote the variables at time t by $y^t, \{y_i^t\}$ where y^t (and similarly y_i^t for all i), with $y^0 = y$ (correspondingly $y_i^0 = y_i$). The random walk evolves from y^{t-1} to y^t by flipping a uniformly random bit of y^{t-1} with probability $\frac{1}{2}$, and with probability $\frac{1}{2}$, not changing anything; this construction is independent across all samples. We now observe that this construction induces an irreducible, aperiodic Markov chain with uniform stationary distribution over the hypercube.

With this construction, the critical step in our proof is a surprising "monotonicity" property over time: we show in Appendix A that each transition over time decreases the function value as long as D_g is not uniform. Intuitively, *"blurring" the distribution D_g decreases the objective.*

This result implies that g is a minimizer of the loss $\hat{L}(g)$ if and only if g is a uniform representation. Consequently, we obtain that among all the representations in G_c, the ones that are uniform minimize the loss in (5) and the statement of Lemma 6 follows. See Appendix A.1 for details. ■

Using Lemma 6, we show our main result that all minimizers of the InfoNCE loss are uniform and cluster-preserving. To the best of our knowledge, this is the first result characterizing the minimizers of the InfoNCE loss with a finite batch of negative samples. The proof is provided in Appendix A.2.

Theorem 7 *If Assumptions 2 and 5 hold, and we have $d > 3$, $\ell \geq 1$, and $N > c \log d$ for an absolute constant c, then a representation $g \in G$ is a global minimizer of the loss $L(g)$ optimized over G if and only if it is uniform and cluster-preserving.*

4.1. Downstream Guarantees

Next we translate the aforementioned representation learning results for G into downstream performance guarantees. We consider the class of heads consisting of single-layer ReLU neural networks with m neurons. Formally,

$$J_{\text{ReLU}} := \{ f : f_{a,\mathbf{W},b} : \mathbb{R}^d \to \mathbb{R} \text{ s.t. } f_{a,\mathbf{W},b}(g(x)) = a^\top \text{ReLU}(\mathbf{W} g(x) - b),\ a, b \in \mathbb{R}^m,\ \mathbf{W} \in \mathbb{R}^{m \times d} \},$$

where $\text{ReLU}(h) = \max(h, 0)$.

Theorem 8 *Suppose there are exactly 2^d clusters, and the representation $g \in \arg\min_{g \in G} L(g)$ under Assumptions 2 and 5, $> c \log d$ for an absolute constant c and $d > 3$. Then for any set of cluster-preserving downstream tasks T, $L_{T, J_{ReLU}}(g) = 0$.*

Theorem 8 shows that any representation learned by minimizing the InfoNCE loss achieves zero downstream error on any task from F_c with a sufficiently wide two-layer ReLU head.

Next, we show that controlling the expressivity of G is necessary to achieve meaningful downstream performance guarantees. Suppose that instead of optimizing the InfoNCE loss over G, we instead optimized it over a representation class $G_? := F_?^d$ where $F_? := \{ f : D \to \{1, \dots, 1\} \}$ consists of all classifiers mapping from images to binary labels.

Theorem 9 *Let $> c \log d$ for an absolute constant c and $d > 3$. There exists a dataset D that satisfies Assumptions 2 and 5 for G, representation $g \in \arg\min_{g' \in G_?} L(g')$, and a downstream task $h \in F_c$ such that $L_{h, J_?}(g) \geq 0.5$, where $J_? = \{ f : H_d \to \{1, \dots, 1\} \}$ is the set of all mappings from $H_d \to \{1, \dots, 1\}$.*

5. Results for the Agnostic Setting

In this section, we consider the setting in which there may not exist any cluster-preserving and uniform representation (that is, Assumption 5 is violated). We show that even in this setting, the InfoNCE loss prioritizes cluster-preserving representations. Specifically, we show that if an optimal solution of the InfoNCE loss on G is close to uniform, then it must also be cluster-preserving. This result requires two new assumptions that we describe below.

First, the function class F must be closed under operations that make classifiers cluster-preserving, in the sense that if $f \in F$ and f does not preserve the cluster $\mathbf{c}$, then the two perturbations of f that preserve $\mathbf{c}$ (by assigning 1 to all images within it) and do not change f otherwise are also in F.

Assumption 10 (Expressivity of F) *For any cluster $\mathbf{c}$, if any $f \in F$ is such that $f(x) \neq f(x')$ for some $x, x' \in \mathbf{c}$, then $f' \in F$ and $f'' \in F$, where $f'(x) = f''(x) = f(x)\ \forall x \notin \mathbf{c}$, and $f'(x) = 1, f''(x) = -1\ \forall x \in \mathbf{c}$.*

Remark 11 *This assumption is used in the analysis to perturb candidate representations that are not cluster-preserving towards improved representations that are cluster-preserving but have a lower loss. If F is expressive enough to isolate each cluster $\mathbf{c}$ (that is, if $\{x \in \mathbf{c}\} \in F$ for all $\mathbf{c}$), then for Assumption 10 to be satisfied, it is sufficient for F to be closed under negations and the 'OR' operations, that is, if $f \in F \Rightarrow -f \in F$ and $f_1, f_2 \in F \Rightarrow f_1 \wedge f_2 \in F$.*

Next we define a regularity condition of a function class and augmentation set that captures the extent to which non-cluster-preserving classifiers classify images in positive pairs differently within clusters that they intersect. So far, we have only assumed that non-cluster-preserving classifiers misclassify at least *one* positive pair differently within any cluster they intersect (Assumption 2). However, for regular classes of binary classifiers and intertwined augmentation sets within clusters, we can expect that the number of positive pairs split in a cluster that are split by any binary classifier scales with the number of negative pairs in the same cluster that are split by the classifier. For a set of images $B \quad D$, we employ the notations $kBk := P_{x \quad D}[x \; 2 \; B]$ and $kBk := P_{x \quad DnD}[x \; 2 \; B]$.

Definition 12 (-Regularity) *For any $f \; 2 \; F$, let $\quad_f := fc \; 2 \; C : 9x; x^0 \; 2 \quad_c$ s.t. $f(x) \; 6= \; f(x^0)g$ be the set of content variables corresponding to clusters split by f. For all $c \; 2 \quad_f$ and $\quad 2 \; f \quad 1; 1g$, define*

$$f^{(c; \quad)}(x) := \begin{cases} f(x) & x \; 2 \quad_c \\ & x \; 2 \quad_c \end{cases}$$

as the classifier that outputs the same label as f on all images not in $\quad_c$ and on $\quad_c$. Further define

$$\quad_{f;c} := \min_{2 \; f \quad 1;1g} kfx \; 2 \quad_{c;} \; j \; f(x) \; 6= \; f^{(c; \quad)}(x)gk$$

as the minimum measure of the set on which $f^{c;}$ and f differ among all possible choices of $\quad 2 \; f \quad 1; 1g$. Then $(F; \quad)$ is regular if for all $c \; 2 \quad_f$,

$$kfA(x) : A \; 2 \quad ; x \; 2 \quad_{c;} \; ; \; f(x) \; 6= \; f(x^+)gk \qquad \quad_{f;c}:$$

Next, we state our regularity assumption and the result for the agnostic case.

Assumption 13 (-Regularity of $(F; \quad)$) *The pair $(F; \quad)$ is regular with $0{:}4$.*

Remark 14 *[Discussion of Assumption 13] Assumption 13 can be interpreted as a relationship between mislabelling of data and generalization error in the supervised learning problem associated with F. Specifically, suppose $f^{(c; \quad)}(x) := f^0(x)$ is a perfect classifier. Consider an entity that mislabels images before giving them to a supervised learner using class F. For some f, the quantity $kfx \; 2 \quad_{c;} \; j \; f(x) \; 6= \; f^0(x)gk$ for some cluster can be interpreted as a "mislabelling budget". The classifier f can be thought of solving a dataset in which the $fx \; 2 \quad_{c;} \; j \; f(x) \; 6= \; f^0(x)g$ has been mislabelled. Further, suppose generalization loss is given by the measure of augmented images (recall that these are not provided during training in the supervised problem considered in this remark) that are misclassified by f. This is precisely the quantity $kfA(x) : A \; 2 \quad ; x \; 2 \quad_{c;} \; ; \; f(x) \; 6= \; f(x^+)gk$. To summarize, this assumption says that training on the "true" labels, results in good generalization, but classifying a mislabelled dataset results in a generalization error that scales in the size of the mislabelling.*

Note that this is related to Assumption 2, which simply states that $\quad_f > 0$ whenever $\quad_{f;c} > 0$. Assumption 2 is sufficient in the realizable setting.

Theorem 15 *Suppose Assumptions 10 and 13 hold and $g = [f_1; :::; f_d]$ is* ***not*** *cluster-preserving with $\min_{j \; 2[d]} \min_{c2 \quad_{f_j}} P_{x;x^0 \quad D}[x; x^0 \; 2 \quad_{c;} f_j(x) \; 6= \; f_j(x^0)] \quad > 0$. Let $` \quad \frac{c}{} d2^d, \quad c \log(\frac{c}{})2^d$ for*

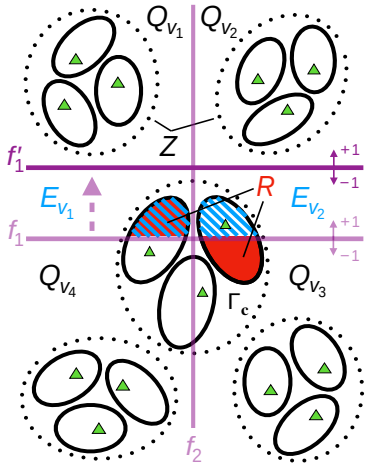

Figure 2: Example partitioning of images with $d = 2$ and D $\mathbb{R}^2$. Green triangles denote natural images and solid black ellipses denote their corresponding augmentation sets (here we have drawn the augmentation sets as compact convex sets for ease of presentation, but in reality they may be non-simply connected and non-smooth). Clusters are indicated by dotted black ellipses. The non-cluster-preserving representation $g = (f_1; f_2)$, and we construct $g^0 = (f_1^0; f_2)$ by making f_1 preserve the cluster $_{\mathbf{c}}$. The region R consisting of augmentations in $_{\mathbf{c}}$ misclassified by f_1 is shaded red, and the set E of images which are classified differently by f_1 and f_1^0 is indicated by blue diagonal lines. By Assumption 13, kRk kEk, where $kBk := \mathsf{P}_{x\ D}[x\ 2\ B]$ and $kBk := \mathsf{P}_{x\ D}[x\ 2\ B]$ for any set of images B D.

a sufficiently large constant c. *Moreover, suppose* g *is close to a uniform representation in the sense that*[3] $\mathsf{P}_{x\ D}[g(x) = v]$ $\frac{10}{cd2^d}$ *or* $\mathsf{P}_{x\ D}[g(x) = v]$ $\frac{}{100cd2^{2d}}$ *for all* $v\ 2\ H_d$. *Then* g *is* ***not*** *a minimizer of the InfoNCE loss.*

Proof [Proof sketch of Theorem 15] For a non-cluster-preserving representation g that is "close" to a uniform representation, we construct a nearby representation g^0 by changing one coordinate of g such that it preserves one additional cluster, and show that the resulting g^0 achieves smaller InfoNCE loss than g. Suppose WLOG that f_1 does not preserve the cluster $_{\mathbf{c}}$. Further, let $f_1^{(\mathbf{c};\ )}$ be the smallest perturbation of f_1 that preserves $_{\mathbf{c}}$, as defined in Definition 12. Denote $f_1^0 = f_1^{(\mathbf{c};\ )}$. By Assumption 10, $f_1^0\ 2\ F$. Construct $g^0 = [f_1^0; f_2; : : : ; f_d]\ 2\ G$. Note that g^0 is equivalent to g on all but one coordinate, and the one differing coordinate differs only on one cluster.

To characterize the variation in the InfoNCE loss when moving from g to g^0, we first consider a specific partition of the space of images defined based on the representations g and g^0. In particular, for a given vertex $v\ 2\ H_d$, consider the set $Q_v := fx\ 2\ D : g(x) = v; g^0(x) = vg$ which denotes the set of images that both g and g^0 map to vertex v, and the set $E_v := fx\ 2\ D : g(x) = v; g^0(x)$ ~~$=$~~ vg which denotes the set of images that g maps to v and g^0 maps to another vertex. Considering these definitions, the set $Q := [_{v2H_d} Q_v$ corresponds to the set of all images that g and g^0 map to the same vertex, while $E := [_{v2H_d} E_v$ denotes the set of all images which g and g^0 map to different vertices. Based on this construction, it is not hard to observe that for any v ~~$6=$~~ v^0 the sets Q_v, Q_{v^0}, E_v, and E_{v^0} are disjoint, and each image belongs to either some Q_v or E_v. Hence, the concatenation of these sets partitions the space of images. Figure 2 illustrates this partition for a special case with $d = 2$. The above partition is critical as we divide our sensitivity analysis into multiple cases based on the location of the positive and negative images in this partition.

Let us define L^+ and L as the alignment and uniformity losses in (3), respectively. We refer to L^+ as the positive part of the loss as it deals with positive samples (augmented images), and we refer to L as the negative part of the loss as it contains negative samples. To prove that moving from g to g^0 decreases the loss, i.e., $L(g)$ $L(g^0) > 0$, we show that the amount that

3. Note that this near-uniformity condition allows for representations that for each vertex put mass at least a constant factor of $\frac{1}{d}$ times 2^{-d}, or essentially treat the vertex as inactive, which allows for the case wherein the number of clusters is less than 2^d and some vertices are inactive for cluster-preserving representations.

the positive part of the loss decreases is more than the amount the negative part might increase: $L^+(g) \quad L^+(g') > L \quad (g') \quad L \quad (g)$. To do so, first, note that the variation in the positive part is

$$L^+(g) \quad L^+(g') = 2 \quad \mathrm{P}\ x \ 2\ Q; x^+ \ 2\ E \quad + \mathrm{P}\ x \ 2\ Q; x^+ \ 2\ E \quad : \tag{6}$$

This holds as $g(x)^{>} g(x^+) = g(x)^{0>} g'(x^+)$ except for the cases that $x \ 2\ Q; x^+ \ 2\ E$ or $x \ 2\ Q; x^+ \ 2\ E$. In these two cases, they differ by 2 . Note that the augmentations that belong to either of these two cases lie in the area shaded red in Figure 2. We refer to the set of augmentations in this region as R, in other words, R is the set of augmentations in $_{\mathbf{c}}$ that are classified differently than their natural image by f_1. Thus, we can write $L^+(g) \quad L^+(g') = 2 \ kRk$.

Next, we consider the difference in negative parts of the loss. To bound this difference, we leverage the partitioning of the space of images defined above to decompose the variation of the losses based on the set that image x belongs to. In particular, if we define the function

$$L_B(g) := \mathrm{E}_{x;x^+;\{x_i\}} \quad (B) \log \ e^{\,g(x)^{>} g(x^+)} + \sum_{i=1}^{\infty} e^{\,g(x)^{>} g(x_i)}$$

for any event B, where (B) is the indicator random variable for the event B, then using the fact that each image x either belongs to one of the Q_v's or E_v's we can write

$$L \quad (g') \quad L \quad (g) = \sum_{v 2 H_d} L_{\{x 2 Q_v\} g}(g') \quad L_{\{x 2 Q_v\} g}(g) \ + \ L_{\{x 2 E_v\} g}(g') \quad L_{\{x 2 E_v\} g}(g) \ ; \tag{7}$$

Since the cases with $x \ 2\ E_v$ utilize similar analysis for those with $x \ 2\ Q_v$, we focus on the $x \ 2\ Q_v$ cases here and defer the $x \ 2\ E_v$ cases to Appendix B.

To analyze $L_{\{x 2 Q_v\} g}(g') \quad L_{\{x 2 Q_v\} g}(g)$, we first observe that this difference is non-positive for a subset of the Q_v's. Note in Fig. 2 that if x belongs to Q_{v_1} or Q_{v_2}, then moving from f_1 to f_1' decreases the representation similarity for some pairs of negative samples (those with $x_i \ 2\ E$) while keeping the rest the same. So, the negative part of the loss cannot increase going from g to g' if x lies in either Q_{v_1} or Q_{v_2}. We formally define this set of Q_v's as $Z := \{Q_v : f_1'(x) \neq f_1'(x \ ) \ 8x \ 2\ Q_v; x \ \ 2\ E_v g\}$. At a high level, the reason this definition implies the negative part of the loss does not increase if $x \ 2\ Q_v \ 2\ Z$ is because f_1 and f_1' must agree on $Q_v \ 2\ Z$ and disagree on E, so since f_1' differs on $Q_v \ 2\ Z$ and E, f_1 must agree on these sets. Thus, the similarity between negative pairs consisting of $x \ 2\ Q_v \ 2\ Z$ and $x_i \ 2\ E$ diminishes when moving from g to g'. Thus, we have

$$\sum_{v 2 H_d} L_{\{x 2 Q_v\} g}(g') \quad L_{\{x 2 Q_v\} g}(g) \qquad \sum_{v 2 H_d} L_{\{x 2 Q_v \ 2 Z\} g}(g') \quad L_{\{x 2 Q_v \ 2 Z\} g}(g) \tag{8}$$

Now, for each event $\{x \ 2\ Q_v \ 2 \ Z\}$, we consider two cases depending on the number of negative samples in Q_v. (1) If there is at least one negative sample $x_i \ 2\ Q_v$, then $g'(x) = g'(x_i) = g(x) = g(x_i)$, so both the log-sums in $L_{\{x 2 Q_v \ 2 Z\} g}(g')$ and $L_{\{x 2 Q_v \ 2 Z\} g}(g)$ are dominated by $e^{\ d}$ terms and the losses do not significantly differ (using that log-sum is approximately a max operation). (2) If no negative samples lie in Q_v, then the dominant terms in the log-sum for $L_{\{x 2 Q_v \ 2 Z\} g}(g')$ may be a factor of e^2 larger than the dominant terms for $L_{\{x 2 Q_v \ 2 Z\} g}(g)$, requiring a sharp analysis to control the probability these events occur. Letting $n_{1;v}$ denote the number of negative samples in Q_v, we define

these two cases above as $B_{v;1} := \{x \in Q_v \cap Z; n_{1;v} > 0\}$ and $B_{v;2} := \{x \in Q_v \cap Z; n_{1;v} = 0\}$, respectively. Note that they form a partition of $\{x \in Q_v \cap Z\}$, so we have

$$L_{\{x \in Q_v \cap Z\}}(g') \quad L_{\{x \in Q_v \cap Z\}}(g) = \sum_{j=1}^{2} L_{B_{v;j}}(g') \quad L_{B_{v;j}}(g):$$

We detail each case below, where n_2 is the number of negative samples in E.

Case 1: $B_{1;v} := \{x \in Q_v \cap Z; n_{1;v} > 0\}$. In this case the dominant terms in the log sums for $L_{B_{1;v}}(g')$ and $L_{B_{1;v}}(g)$ are both $e^{\ d}$, although the losses may differ in the number of such terms, which can be, in the worst case, $n_{1;v} + n_2 + 1$ for g' and $n_{1;v}$ for g. This is because g and g' can disagree on at most n_2 negative samples, and they can also disagree on the positive sample. Thus,

$$L_{B_{1;v}}(g') \quad L_{B_{1;v}}(g) \quad \mathsf{E} \quad (B_{1;v}) \log \frac{n_{1;v}+n_2+1}{n_{1;v}} \qquad \mathsf{E}[\ (B_{1;v}) \frac{2(n_2+1)}{n_{1;v}+1}];$$

where the last inequality follows using $\log(1+x) \quad x$. We bound $\mathsf{E}[\ (B_{1;v}) \frac{2(n_2+1)}{n_{1;v}+1}]$ by writing the trinomial expansion of the expectation (note that the joint distribution of $(n_{1;v}; n_2)$ is trinomial with parameters $(kQ_vk\ ; kEk\ )$), and further simplifying to result in an upper bound of $kEk\ $. Importantly, this bound is $O(kRk)$ by Assumption 13 and independent of $\ $, so we control it by making $\ $ large enough.

Case 2: $B_{2;v} := \{x \in Q_v \cap Z; n_{1;v} = 0\}$. Since here there is no shared dominant $e^{\ d}$ term in the log-sums for $L_{B_{2;v}}(g')$ and $L_{B_{2;v}}(g)$, the dominant terms for g' may involve strictly larger similarities than those for g, corresponding to $x_i \in E$ and $x^+ \in E$ (the only samples on which g' and g can disagree). These events are bounded depending on whether $n_2 = 0$. If $n_2 = 0$, the loss of g' exceeds that of g iff $g'(x)^\top g'(x^+) = g(x)^\top g(x^+) + 2$, which occurs iff $x^+ \in E$. If $n_2 > 0$, the loss can increase by $2\ $ regardless of the value of x^+. Combining these sub-cases yields

$$\begin{aligned} L_{B_{2;v}}(g') \quad L_{B_{2;v}}(g) \quad & 2\ \ \mathsf{P}(x \in Q_v \cap Z; x^+ \in E)\mathsf{P}(n_{1;v} = 0; n_2 = 0) \\ & + 2\ \ \mathsf{P}(x \in Q_v \cap Z)\mathsf{P}(n_{1;v} = 0; n_2 > 0): \end{aligned}$$

For each term above, we need to show that the coefficient of $2\ $ is $o(kRk)$ even after it is summed over v. Note that both terms scale with the probability that $x \in Q_v \cap Z$ *and* no negative samples are in Q_v. To control this probability we leverage that the distribution induced by g is close to uniform in the sense that every "active" vertex v has mass $\mathsf{P}[g(x) = v] = \tilde{}(\frac{1}{d2^d})$. We use this fact to bound $\mathsf{P}[x \in Q_v]$. Note that the set of images that g maps to v is $Q_v \cup E_v$, yet for all $Q_z \cap Z$, $E_v = \ ;$ since, at a high level, f_1 must separate these Q_v from E. So, $\mathsf{P}[g(x) = v] = \mathsf{P}[x \in Q_v]$ for all $Q_v \cap Z$. Therefore, we can show that with large $\grave{}$ it is highly unlikely that $x \in Q_v \cap Z$ and none of the negative samples are in Q_v. To complete the bounds, we leverage the facts that $\mathsf{P}[x \in Q_v \cap Z; x^+ \in E]$ scales with kRk for the first term, and $\mathsf{P}[n_2 > 0]$ scales with $kEk\ $ for the second term, where $kEk\ = O(kRk)$ by Assumption 13.

After performing a similar analysis for $\{x \in E_v\}$ and summing the resulting bounds over $\{v \in H_d\}$, as in (6), we obtain $L\ \ (g') \quad L\ \ (g) < 2\ \ kRk = L^+(g) \quad L^+(g')$.

■

Theorem 15 shows that for large $\grave{}$ and $\ $, all minimizers of the InfoNCE loss that are near-uniform must be cluster-preserving regardless of the sizes of each cluster or the number of clusters. However,

it does not rule out that there could be a highly non-uniform and non-cluster-preserving optimal representation. In Appendix B.1, we show that if we re-weight the alignment and uniformity losses in the InfoNCE loss, we can ensure that *all* minimizers of the InfoNCE loss are cluster-preserving.

6. Conclusion

We study properties of minimizers of the InfoNCE loss optimized over function classes with restricted complexity relative to the complexity of augmentations in the dataset, in realistic settings with disjoint augmentation sets and finite negative samples. Our results show that such representations are uniform and cluster-preserving in the realizable setting, and must be cluster-preserving if they are close to uniform in the agnostic setting. We believe that our novel analytical tools, namely our stochastic argument for the optimality of representations and our inverse partitioning of the space of images, may be of use for future studies of the InfoNCE loss.

Acknowledgments

This research is supported in part by NSF Grants 2019844, 2107037 and 2112471, ONR Grant N00014-19-1-2566, the Machine Learning Lab (MLL) at UT Austin, and the Wireless Networking and Communications Group (WNCG) Industrial Affiliates Program.

References

Sanjeev Arora, Hrishikesh Khandeparkar, Mikhail Khodak, Orestis Plevrakis, and Nikunj Saunshi. A theoretical analysis of contrastive unsupervised representation learning. In *36th International Conference on Machine Learning, ICML 2019*, pages 9904–9923. International Machine Learning Society (IMLS), 2019.

Jordan T Ash, Surbhi Goel, Akshay Krishnamurthy, and Dipendra Misra. Investigating the role of negatives in contrastive representation learning. *arXiv preprint arXiv:2106.09943*, 2021.

Pranjal Awasthi, Nishanth Dikkala, and Pritish Kamath. Do more negative samples necessarily hurt in contrastive learning? In *International Conference on Machine Learning*, pages 1101–1116. PMLR, 2022.

Philip Bachman, R Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. *Advances in neural information processing systems*, 32, 2019.

Randall Balestriero and Yann LeCun. Contrastive and non-contrastive self-supervised learning recover global and local spectral embedding methods. *arXiv preprint arXiv:2205.11508*, 2022.

Han Bao, Yoshihiro Nagano, and Kento Nozawa. On the surrogate gap between contrastive and supervised losses. In *International Conference on Machine Learning*, pages 1585–1606. PMLR, 2022.

Pierre Bremaud. *Markov chains: Gibbs fields, Monte Carlo simulation, and queues; 1st ed.* Texts in applied mathematics. Springer, Berlin, 2001.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in Neural Information Processing Systems*, 33:9912–9924, 2020.

Mayee Chen, Daniel Y Fu, Avanika Narayan, Michael Zhang, Zhao Song, Kayvon Fatahalian, and Christopher Ré. Perfectly balanced: Improving transfer and robustness of supervised contrastive learning. In *International Conference on Machine Learning*, pages 3090–3122. PMLR, 2022.

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, 2020a.

Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. Big self-supervised models are strong semi-supervised learners. *Advances in neural information processing systems*, 33:22243–22255, 2020b.

Ting Chen, Calvin Luo, and Lala Li. Intriguing properties of contrastive losses. *Advances in Neural Information Processing Systems*, 34:11834–11845, 2021.

DeepNets. Animals - v2: Image classification dataset, Nov 2022. URL https://www.kaggle.com/datasets/utkarshsaxenadn/animal-image-classification-dataset.

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

Quentin Garrido, Yubei Chen, Adrien Bardes, Laurent Najman, and Yann Lecun. On the duality between contrastive and non-contrastive self-supervised learning. *arXiv preprint arXiv:2206.02574*, 2022.

Kartik Gupta, Thalaiyasingam Ajanthan, Anton van den Hengel, and Stephen Gould. Understanding and improving the role of projection head in self-supervised learning. *arXiv preprint arXiv:2212.11491*, 2022.

Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 297–304. JMLR Workshop and Conference Proceedings, 2010.

Jeff Z HaoChen and Tengyu Ma. A theoretical study of inductive biases in contrastive learning. *arXiv preprint arXiv:2211.14699*, 2022.

Jeff Z HaoChen, Colin Wei, Adrien Gaidon, and Tengyu Ma. Provable guarantees for self-supervised deep learning with spectral contrastive loss. *Advances in Neural Information Processing Systems*, 34:5000–5011, 2021.

Jeff Z HaoChen, Colin Wei, Ananya Kumar, and Tengyu Ma. Beyond separability: Analyzing the linear transferability of contrastive representations to related subpopulations. *arXiv preprint arXiv:2204.02683*, 2022.

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.

Olivier Henaff. Data-efficient image recognition with contrastive predictive coding. In *International conference on machine learning*, pages 4182–4192. PMLR, 2020.

R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*, 2018.

Weiran Huang, Mingyang Yi, and Xuyang Zhao. Towards the generalization of contrastive self-supervised learning. *arXiv preprint arXiv:2111.00743*, 2021.

Wenlong Ji, Zhun Deng, Ryumei Nakada, James Zou, and Linjun Zhang. The power of contrast for feature learning: A theoretical analysis. *arXiv preprint arXiv:2110.02473*, 2021.

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*, 33:18661–18673, 2020.

Jason D Lee, Qi Lei, Nikunj Saunshi, and Jiacheng Zhuo. Predicting what you already know helps: Provable self-supervised learning. *Advances in Neural Information Processing Systems*, 34: 309–323, 2021.

Junnan Li, Pan Zhou, Caiming Xiong, and Steven CH Hoi. Prototypical contrastive learning of unsupervised representations. *arXiv preprint arXiv:2005.04966*, 2020.

David McAllester and Karl Stratos. Formal limitations on the measurement of mutual information. In *International Conference on Artificial Intelligence and Statistics*, pages 875–884. PMLR, 2020.

Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6707–6717, 2020.

Kento Nozawa and Issei Sato. Understanding negative samples in instance discriminative self-supervised representation learning. *Advances in Neural Information Processing Systems*, 34: 5784–5797, 2021.

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

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. *CoRR*, abs/1802.05365, 2018. URL http://arxiv.org/abs/1802.05365.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

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.

Joshua Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. Contrastive learning with hard negative samples. *arXiv preprint arXiv:2010.04592*, 2020.

Nikunj Saunshi, Jordan Ash, Surbhi Goel, Dipendra Misra, Cyril Zhang, Sanjeev Arora, Sham Kakade, and Akshay Krishnamurthy. Understanding contrastive learning requires incorporating inductive biases. *arXiv preprint arXiv:2202.14037*, 2022.

Kendrick Shen, Robbie M Jones, Ananya Kumar, Sang Michael Xie, Jeff Z HaoChen, Tengyu Ma, and Percy Liang. Connect, not collapse: Explaining contrastive learning for unsupervised domain adaptation. In *International Conference on Machine Learning*, pages 19847–19878. PMLR, 2022.

Yixuan Su, Fangyu Liu, Zaiqiao Meng, Tian Lan, Lei Shu, Ehsan Shareghi, and Nigel Collier. Tacl: Improving bert pre-training with token-aware contrastive learning. *arXiv preprint arXiv:2111.04198*, 2021.

Joseph John Thomson. Xxiv. on the structure of the atom: an investigation of the stability and periods of oscillation of a number of corpuscles arranged at equal intervals around the circumference of a circle; with application of the results to the theory of atomic structure. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 7(39):237–265, 1904.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *European conference on computer vision*, pages 776–794. Springer, 2020a.

Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning? *Advances in Neural Information Processing Systems*, 33:6827–6839, 2020b.

Yuandong Tian. Deep contrastive learning is provably (almost) principal component analysis. *arXiv preprint arXiv:2201.12680*, 2022a.

Yuandong Tian. Understanding the role of nonlinearity in training dynamics of contrastive learning. *arXiv preprint arXiv:2206.01342*, 2022b.

Christopher Tosh, Akshay Krishnamurthy, and Daniel Hsu. Contrastive learning, multi-view redundancy, and linear models. In *Algorithmic Learning Theory*, pages 1179–1206. PMLR, 2021.

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

Julius Von Kügelgen, Yash Sharma, Luigi Gresele, Wieland Brendel, Bernhard Schölkopf, Michel Besserve, and Francesco Locatello. Self-supervised learning with data augmentations provably isolates content from style. *Advances in neural information processing systems*, 34:16451–16467, 2021.

Feng Wang and Huaping Liu. Understanding the behaviour of contrastive loss. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2495–2504, 2021.

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

Yifei Wang, Qi Zhang, Yisen Wang, Jiansheng Yang, and Zhouchen Lin. Chaos is a ladder: A new theoretical understanding of contrastive learning via augmentation overlap. *arXiv preprint arXiv:2203.13457*, 2022.

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

Zixin Wen and Yuanzhi Li. Toward understanding the feature learning process of self-supervised contrastive learning. In *International Conference on Machine Learning*, pages 11112–11122. PMLR, 2021.

Zixin Wen and Yuanzhi Li. The mechanism of prediction head in non-contrastive self-supervised learning. *arXiv preprint arXiv:2205.06226*, 2022.

Mingkai Zheng, Fei Wang, Shan You, Chen Qian, Changshui Zhang, Xiaogang Wang, and Chang Xu. Weakly supervised contrastive learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10042–10051, 2021.

Roland S Zimmermann, Yash Sharma, Steffen Schneider, Matthias Bethge, and Wieland Brendel. Contrastive learning inverts the data generating process. In *International Conference on Machine Learning*, pages 12979–12990. PMLR, 2021.

Appendix A. Proof of Theorem 7

We prove Theorem 7. To do so, we first prove Lemma 6, and use this to prove our main result.

A.1. Proof of Lemma 6

Lemma 16 (Lemma 6 Restated) *If Assumptions 2 and 5 hold, $\quad > c \log d$ for an absolute constant c, $d > 3$, and $\grave{} \quad 1$, then $g \quad 2 \arg\min_{g2G_c} L(g)$ if and only if g is uniform.*

Proof To prove this claim we first note that since we are optimizing over G_c and $g(x) = g(x^+)$ for all $x \; 2 \; D \; ; x^+ \; 2 \; A(x)$ and $g \; 2 \; G_c$, the optimization problem $\min_{g2G_c} L(g)$ is equivalent to

$$\min_{g2G_c} \hat{L}(g) = \mathsf{E}_{x;fx_i\,g\cdot} \quad \log \quad 1 + \sum_{i=1}^{\dot{}} e^{\; g(x)^{>} g(x_i) \quad d} \quad : \tag{9}$$

Note that we can think of this as an optimization over distributions over the hypercube induced by g. That is, consider the random variable Y supported on the hypercube that is given by $Y = g(X)$ for $X \quad D$, and denote its distribution as D_g. Using this notation, the optimization above can be rewritten in terms of distributions

$$\min_{D2fD_g:g2Gg} L(D) = \mathsf{E}_{y;fy_ig\cdot \quad D} \quad \log(1 + \sum_{i=1}^{\dot{}} e^{\; y^{>} y_i \quad d}) \quad :$$

where y_i corresponds to the representation of the i-th negative sample, and here we overload notation by using $\quad D$ to denote an i.i.d. draw from the distribution D.

Next, we define a Markov chain as follows. We begin with a fresh set of samples denoted by $y^0; y_1^0; : : : ; y^0$ that are drawn i.i.d. from the distribution $D^0 = D$. At each step, for each sample, we either with probability $\frac{1}{2}$ flip one bit uniformly at random, or with probability $\frac{1}{2}$ we do not change it. Concretely, we take for all i a random variable $j_{i;t} \; 2 \; [d]$ (both uniformly random and independent of each other and every other such sample) and set $(y_i^t)_{j_{i;t}} = \quad (y_i^{t \quad 1})_{j_{i;t}}$. After this operation, each y_i^t (and y) can be considered to be an i.i.d. drawn from D^t, where D^t is another distribution over H_d. We show that $L(D^{t \quad 1}) > L(D^t)$ if $D^{t \quad 1}$ is not uniform. Since $fD^t g_t$ converges to the uniform distribution by Lemma 19, these two arguments together imply the claim of Lemma 6.

To show $L(D^{t \quad 1}) > L(D^t)$ for the case that $D^{t \quad 1}$ is not the uniform distribution, considering the definition of L we need to study the variation in the inner products between the vectors $(y^t; y_i^t)$ when we move from one distribution to another. Note that as these vectors are binary vectors, their inner product can be written as a function of their Hamming distances. More precisely, for any pair $(y; y^0)$ we have $y^{>} y^0 = d \quad 2h(y; y^0)$, where the Hamming distance between them is defined as $h(y; y^0) := \sum_{j=1}^{\grave{}} \quad fy_j \; \neq \; y_j^0 g$ or the number of bits that are different in the two points $y; y^0$ (note that $\quad fU g$ is the indicator variable for the event U).

For ease of notation we let $h_i^t := h(y^t; y_i^t)$ for all $i; t$. Due to the fact that each of the y_i^t's are independent and identically distributed and are evolving according to a Markov chain, the h_i^t's also evolve according to a Markov chain. In particular, for every distribution D^t over H_d that describes the distribution of each y_i^t, there is induced a distribution D_h^t over $[d]$ that specifies the distribution for h_i^t. By direct computation, one can check that h_i^t has the following transition kernels which differ for different values of h_i:

For $2 \quad h_i^{t\ 1} \quad d \quad 2$:

$$h_i^t \ ! \ \begin{cases} h_i^{t\ 1} \quad 2 & \text{w.p. } \frac{(h_i^{t\ 1})(h_i^{t\ 1}\ 1)}{4d^2} \\ h_i^{t\ 1} \quad 1 & \text{w.p. } \frac{h_i^{t\ 1}}{2d} \\ h_i^{t\ 1} & \text{w.p. } \frac{1}{4} + \frac{(h_i^{t\ 1})(d \quad h_i^{t\ 1}+1)+(d \quad h_i^{t\ 1})(h_i^{t\ 1}+1)}{4d^2} \\ h_i^{t\ 1} + 1 & \text{w.p. } \frac{d \quad h_i^{t\ 1}}{2d} \\ h_i^{t\ 1} + 2 & \text{w.p. } \frac{(d \quad h_i^{t\ 1})(d \quad h_i^{t\ 1}+1)}{4d^2} \end{cases}$$

For $h_i^{t\ 1} = 1$:

$$1 \ ! \ \begin{cases} 0 & \text{w.p. } \frac{1}{2d} \\ 1 & \text{w.p. } \frac{1}{4} + \frac{1}{4d} + \frac{2(d\ 1)}{4d^2} \\ 2 & \text{w.p. } \frac{d\ 1}{2d} \\ 3 & \text{w.p. } \frac{(d\ 1)(d\ 2)}{4d^2} \end{cases} \tag{10}$$

For $h_i^{t\ 1} = d \quad 1$:

$$1 \ ! \ \begin{cases} d & \text{w.p. } \frac{1}{2d} \\ d \quad 1 & \text{w.p. } \frac{1}{4} + \frac{1}{4d} + \frac{2(d\ 1)}{4d^2} \\ d \quad 2 & \text{w.p. } \frac{d\ 1}{2d} \\ d \quad 3 & \text{w.p. } \frac{(d\ 1)(d\ 2)}{4d^2} \end{cases}$$

For $h_i^{t\ 1} = 0$:

$$0 \ ! \ \begin{cases} 0 & \text{w.p.} \frac{1}{4} + \frac{1}{4d} \\ 1 & \text{w.p.} \frac{1}{2} \\ 2 & \text{w.p.} \frac{d\ 1}{4d} \end{cases} \tag{11}$$

For $h_i^{t\ 1} = d$:

$$d \ ! \ \begin{cases} d & \text{w.p.} \frac{1}{4} + \frac{1}{4d} \\ d \quad 1 & \text{w.p.} \frac{1}{2} \\ d \quad 2 & \text{w.p.} \frac{d\ 1}{4d} \end{cases}$$

For ease of notation we drop the t superscripts and refer to quantities at time $t \quad 1$ without any superscript, and quantities at times t with a $^{\prime}$ superscript, e.g. $D^{t\ 1}$ as D and D^t as $D^{\prime}$, for the remainder of the proof.

Next, let $\mathbf{h} = [h_1; : : : ; h_{\grave{}}]$ denote the vector of concatenated Hamming distances between y and $fy_i g_{\grave{}\ i=1}$. Using the above definitions, and the fact that $d \quad y^{>} y_i$ is twice of the hamming distance between y and y_i the loss $\mathcal{L}$ can be written as

$$\mathcal{L}(D) = \mathsf{E}_{y; fy_i g_{\grave{}}\ D} \left[\log \left(1 + \sum_{i=1}^{\grave{}} e^{(y^{>} y_i \quad d)} \right) \right] = \mathsf{E}_{\mathbf{h}\ D_h} \left[\log \left(1 + \sum_{i=1}^{\grave{}} e^{\ 2\ h_i} \right) \right]$$

Now to characterize the difference between $\mathcal{L}(D)$ and $\mathcal{L}(D^{\prime})$ we need to study the evolution of the distribution of the Hamming distance $\mathbf{h}$ from D_h to $D_h^{\prime}$, i.e.,

$$\mathcal{L}(D^{\prime}) \quad \mathcal{L}(D) = \mathsf{E}_{\mathbf{h}^{\prime}\ D_h^{\prime}} \left[\log \left(1 + \sum_{i=1}^{\grave{}} e^{\ 2\ h_i^{\prime}} \right) \right] \quad \mathsf{E}_{\mathbf{h}\ D_h} \left[\log \left(1 + \sum_{i=1}^{\grave{}} e^{\ 2\ h_i} \right) \right]$$

For each i, consider the random variable $s_i = h_i^0 \quad h_i$ that indicates which of the transitions is undertaken by h_i, and let $\mathbf{s} := [s_1; \dots; s_\ell]$ be its concatenation. Note that each s_i takes values in $f \quad 2; \quad 1; 0; 1; 2g$ and its distribution depends on the value of h_i, defined according to the transition kernel of h_i defined above. Given this, we can express $L(D^0)$ as

$$L(D^0) = \mathrm{E}_{\mathbf{h} \quad D_h;\mathbf{s}} \left[\log 1 + \sum_{i=1}^{\ell} e^{\quad 2 \ (h_i + s_i))} \right]$$

Now we consider the difference $L(D^0) \quad L(D)$. According to the above definitions, this difference can be written as

$$\begin{aligned}
& L(D^0) \quad L(D) \\
&= \mathrm{E}_{\mathbf{h} \quad D_h;\mathbf{s}} \left[\log \quad 1 + \sum_{i=1}^{\ell} e^{\quad 2 \ (h_i + s_i)} \right] \quad \log \quad 1 + \sum_{i=1}^{\ell} e^{\quad 2 \ h_i} \Bigg] \\
&= \sum_{i=1}^{\ell} \mathrm{E}_{\mathbf{h} \quad D_h;\mathbf{s}} \left[\log \quad 1 + \sum_{j=1}^{i} e^{\quad (h_j + s_j)} + \sum_{j=i+1}^{\ell} e^{\quad 2 \ h_j} \right. \\
&\qquad \left. \log \quad 1 + \sum_{j=1}^{i-1} e^{\quad 2 \ (h_j + s_j)} + \sum_{j=i}^{\ell} e^{\quad 2 \ h_j} \right] \qquad (12)
\end{aligned}$$

where (12) follows from telescoping over each negative sample indexed by i. Now if we take out the i-th term of each of the above two expressions the difference can be written as

$$\begin{aligned}
& f(D^0) \quad f(D) \\
&= \sum_{i=1}^{\ell} \mathrm{E}_{\mathbf{h} \quad D_h;\mathbf{s}} \left[\log \quad 1 + \sum_{j=1}^{i-1} e^{\quad 2 \ (h_j + s_j)} + \sum_{j=i+1}^{\ell} e^{\quad 2 \ h_j} + e^{\quad 2 \ (h_i + s_i)} \right. \\
&\qquad \left. \log \quad 1 + \sum_{j=1}^{i-1} e^{\quad 2 \ h_j} + \sum_{j=i+1}^{\ell} e^{\quad 2 \ h_j} + e^{\quad 2 \ h_i} \right] \\
&= \sum_{i=1}^{\ell} \mathrm{E}_{h \quad i;s_i} \mathrm{E}_{h_i \quad D_h;s_i} \left[\log \quad 1 + \sum_{j=1}^{i-1} e^{\quad 2 \ (h_j + s_j)} + \sum_{j=i+1}^{\ell} e^{\quad 2 \ h_j} + e^{\quad 2 \ (h_i + \mathbf{s}_i)} \right. \\
&\qquad \left. \log \quad 1 + \sum_{j=1}^{i-1} e^{\quad 2 \ (h_j + s_j)} + \sum_{j=i+1}^{\ell} e^{\quad 2 \ h_j} + e^{\quad 2 \ h_i} \right]
\end{aligned}$$

To show that the RHS above is strictly less than 0 when D is not uniform, it is sufficient to show that each term of the sum is strictly less than zero. To do this, we show that the inner expectation is strictly negative. In other words, all that remains to prove is that for all instances of $C \ _i := 1 + \sum_{j=1}^{i \ 1} e^{\quad 2 \ (h_j + s_j)} + \sum_{j=i+1}^{\ell} e^{\quad 2 \ h_j}$, each of the terms satisfies

$$\mathrm{E}_{h_i \quad D_h;s_i} \left[\log \quad C \ _i + e^{\quad 2 \ (h_i + \mathbf{s}_i)} \quad \log \quad C \ _i + e^{\quad 2 \ h_i} \right] < 0 \qquad (13)$$

when D is not uniform. Once we have this result, the claim that $f(D^{\prime}) \quad f(D) < 0$ holds.

From now on, for ease of notation, we replace $C_{\ i}$ by C. To prove the claim in (13), we first introduce the function $\ _C(h)$ defined as

$$\begin{aligned} {}_C(h) &:= \mathrm{E}_{s_i} \quad \log \ C + e^{\ 2\ (h+s_i)} \quad \log \ C + e^{\ 2\ h} \\ &= \mathrm{E}_{s_i} \quad \log \ 1 + \frac{e^{\ 2\ (h+s_i)}}{C} \quad \log \ 1 + \frac{e^{\ 2\ h}}{C} \ : \end{aligned}$$

Considering this definition the claim in (13) can be translated into

$$\sum_{k=0}^{d} \mathrm{P}_{h \ \ D_h}[h = k] \ \ {}_C(h) < 0. \tag{14}$$

To show this, it is easiest to compare this expression with the analogous expression in the case that D is uniform. That is $y; \{y_i\}_{i=1}$ are drawn from the uniform distribution U. We let σ_h denote the distribution on $[d]$ of Hamming distances induced by U. By the stationarity of the uniform distribution (Lemma 19), the distribution of $h_i^{\prime} = h_i + s_i$ is identical to that of h_i if h_i is drawn from a uniform distribution. Thus we have the following result:

$$\begin{aligned} &\mathrm{E}_{h_i \ \ \sigma_h; s_i} \quad \log \ \ C_{\ i} + e^{\ 2\ (h_i+s_i)} \quad \log \ \ C_{\ i} + e^{\ 2\ h_i} \\ &= \mathrm{E}_{h_i^{\prime} \ \ \sigma_h^{\prime}} \quad \log \ \ C_{\ i} + e^{\ 2\ (h_i^{\prime})} \quad \mathrm{E}_{h_i \ \ \sigma_h} \quad \log \ \ C_{\ i} + e^{\ 2\ h_i} \\ &= 0; \end{aligned} \tag{15}$$

Next, we show that $\sum_{k=0}^{d} \mathrm{P}_{h \ \ D_h}[h = k] \ \ {}_C(h) \quad \sum_{k=0}^{d} \mathrm{P}_{h \ \ \sigma_h}[h = k] \ \ {}_C(h) < 0$, which by (15) immediately implies $\sum_{k=0}^{d} \mathrm{P}_{h \ \ D_h}[h = k] \ \ {}_C(h) < 0$. To achieve this we invoke Lemmas 17 and 18, which describe the behavior of $\mathrm{P}_{h \ \ D_h}[h = k] \ \ {}_C(h) \quad \sum_{k=0}^{d} \mathrm{P}_{h \ \ \sigma_h}[h = k] \ \ {}_C(h)$ and $\ _C(k)$, respectively. Using these lemmas we obtain:

$$\begin{aligned} &\sum_{k=0}^{d} \mathrm{P}_{h \ \ D_h}[h = k] \ \ {}_C(k) \quad \sum_{k=0}^{d} \mathrm{P}_{h \ \ \sigma_h}[h = k] \ \ {}_C(k) \\ &\quad \mathrm{P}_{h \ \ D_h}[h = 0] \quad \mathrm{P}_{h \ \ \sigma_h}[h = 0] \quad {}_C(0) \\ &\quad + \sum_{k>0} \binom{d}{k} \mathrm{P}_{h \ \ D_h}[h = 0] \quad \mathrm{P}_{h \ \ \sigma_h}[h = 0] \quad {}_C(k) \qquad (16) \\ &= \mathrm{P}_{h \ \ D_h}[h = 0] \quad \mathrm{P}_{h \ \ \sigma_h}[h = 0] \sum_{k \ 0} \binom{d}{k} \ {}_C(k) \qquad (17) \\ &< 0 \qquad (18) \end{aligned}$$

where (16) holds by Lemma 17-3, (17) follows by simply combining terms, and (18) holds by Lemma 17-2, which states that $\mathrm{P}_{h \ \ D_h}[h = 0] \quad \mathrm{P}_{h \ \ \sigma_h}[h = 0] > 0$, and Lemma 18, which states that $\sum_{k \ 0} \binom{d}{k} \ {}_C(k) < 0$. This completes the proof. ■

Lemma 17 *For any the distribution D_h on $[d]$ induced by any non-uniform distribution D on H_d, the following are true:*

1. $\mathrm{P}_{h \sim U_h}[h = k] = \binom{d}{k} \mathrm{P}_{h \sim U_h}[h = 0]$
2. $\mathrm{P}_{h \sim D_h}[h = 0] - \mathrm{P}_{h \sim U_h}[h = 0] > 0$
3. $\mathrm{P}_{h \sim D_h}[h = k] - \mathrm{P}_{Y \sim U}[h = k] < \binom{d}{k} \mathrm{P}_{h \sim D_h}[h = 0] - \mathrm{P}_{Y \sim U}[h(y; Y) = k]$

Proof For any vertex $v \in H_d$, we let $p_v := \mathrm{P}_{y \sim D}[y = v]$ for ease of notation.

1. Note that for any $k \in [d]$,

$$\mathrm{P}_{h \sim U_h}[h = k] = \mathrm{P}_{y;y' \sim U}[h(y; y') = k] = \mathrm{P}_{y \sim U}[h(w; y) = k] \tag{19}$$

for any fixed vertex $w \in H_d$ by the symmetry of the uniform distribution. Then

$$\begin{aligned}
\mathrm{P}_{h \sim U_h}[h = k] &= \mathrm{P}_{y \sim U}[h(w; y) = k] \\
&= \sum_{v \in H_d : h(w;v) = k} \mathrm{P}_{y \sim U}[y = v] \\
&= \sum_{v \in H_d : h(w;v) = k} \frac{1}{2^d} \\
&= \binom{d}{k} \frac{1}{2^d},
\end{aligned}$$

which implies that $\mathrm{P}_{h \sim U_h}[h = 0] = \frac{1}{2^d}$.

2. Using the above observation that $\mathrm{P}_{h \sim U_h}[h = 0] = \frac{1}{2^d}$, we have

$$\begin{aligned}
\mathrm{P}_{h \sim D_h}[h = 0] - \mathrm{P}_{h \sim U_h}[h = 0] &= \sum_{v \in H_d} p_v^2 - \frac{1}{2^{2d}} \\
&= \sum_{v \in H_d} p_v^2 - \frac{2p_v}{2^d} + \frac{1}{2^{2d}} + \sum_{v \in H_d} \frac{2}{2^{2d}} + \frac{2p_v}{2^d} \\
&= \sum_{v \in H_d} p_v - \frac{1}{2^d}^2 \\
&> 0
\end{aligned}$$

where the strict inequality holds since D is not uniform.

3. We argue similarly as in the proofs of the previous two statements. We have

$$\begin{aligned}
&2 \cdot \mathrm{P}_{h \sim D_h}[h = k] - \mathrm{P}_{h \sim U_h}[h = k] \\
&= 2 \sum_{v;u \in H_d; h(v;u) = k} p_v p_u - \frac{1}{2^{2d}}
\end{aligned} \tag{20}$$

$$= \sum_{v;u2H_d;h(v;u)=k} p_v^2 + 2p_v p_u \quad p_u^2 + \sum_{v;u2H_d;h(v;u)=k} p_v^2 + p_u^2 \quad \frac{2}{2^{2d}} \tag{21}$$

$$= \sum_{v;u2H_d;h(v;u)=k} p_v \quad p_u \ ^2 + \sum_{v;u2H_d;h(v;u)=k} p_v^2 \quad \frac{1}{2^{2d}} + p_u^2 \quad \frac{1}{2^{2d}}$$

$$= \sum_{v;u2H_d;h(v;u)=k} p_v \quad p_u \ ^2 + 2 \binom{d}{k} \sum_{v2H_d} p_v^2 \quad \frac{1}{2^{2d}} \tag{22}$$

$$= \sum_{v;u2H_d;h(v;u)=k} p_v \quad p_u \ ^2 + 2 \binom{d}{k} \quad P_{h \ D_h}[h = 0] \quad P_{h \ U_h}[h = 0]$$

$$< 2 \binom{d}{k} \quad P_{h \ D_h}[h = 0] \quad P_{Y \ U}[h(y; Y) = 0] \tag{23}$$

where (21) is obtained by adding and subtracting $p_v^2 + p_u^2$, (22) follows by the symmetry of the hypercube and the fact that for every $v \ 2 \ H_d$, there are $\binom{d}{k}$ vertices $u \ 2 \ H_d$ satisfying $h(v; u) = k$, and (23) follows by the fact that $p_v \ \not\in \ p_u$ for some $v; u \ 2 \ H_d$ for all non-uniform distributions D.

■

Lemma 18 *If* $\quad > c \log d$ *for an absolute constant* c, $d > 3$ *and* $C > 1$, *then* $\sum_{k \ 0} \binom{d}{k} \quad _C(k) < 0.$

Proof According to the transition matrix of h for the case that $h = 0$ we know that s could be either 0, 1, or 2, with probabilities denoted in (11). Hence, we can simplify the expression for $\quad _C(0)$ as

$$_C(0) = \log \quad 1 + \frac{1}{C} \quad \frac{d+1}{4d} + \log \quad 1 + \frac{e \ ^2}{C} \quad \frac{1}{2} + \log \quad 1 + \frac{e \ ^4}{C} \quad \frac{d \quad 1}{4d} \quad \log \quad 1 + \frac{1}{C}$$

$$= \quad \frac{3d \quad 1}{4d} \log \quad 1 + \frac{1}{C} \quad + \log \quad 1 + \frac{e \ ^2}{C} \quad \frac{1}{2} + \log \quad 1 + \frac{e \ ^4}{C} \quad \frac{d \quad 1}{4d}$$

We can similarly compute $\quad _C(1), \quad _C(2); : : :$ to obtain:

$$\sum_{k \ 0} \binom{d}{k} \quad _C(k)$$

$$= \log \quad 1 + \frac{1}{C} \quad \frac{3d+1}{4d} + \log \quad 1 + \frac{e \ ^2}{C} \quad \frac{1}{2} + \log \quad 1 + \frac{e \ ^4}{C} \quad \frac{d \quad 1}{4d}$$

$$+ \binom{d}{1} \quad \log(1 + \frac{1}{C}) \frac{1}{2d} + \log \quad 1 + \frac{e \ ^2}{C} \quad \frac{3d^2 + 3d \quad 2}{4d^2}$$

$$+ \log \quad 1 + \frac{e \ ^4}{C} \quad \frac{d \quad 1}{2d} + \log(1 + \frac{e \ ^6}{C}) \frac{(d \quad 1)(d \quad 2)}{4d^2}$$

$$+ \binom{d}{2} \quad \log \quad 1 + \frac{1}{C} \quad \frac{1}{2d^2}$$

$$
\begin{aligned}
&+ \log\left(1 + \frac{e^{-2}}{C}\right)\left(\frac{1}{d} + \log\left(1 + \frac{e^{-4}}{C}\right)\right)\left(\frac{3}{4} + \frac{2(d-1) + 3(d-2)}{4d^2}\right) \\
&+ \log\left(1 + \frac{e^{-6}}{C}\right)\left(\frac{(d-2)}{2d} + \log\left(1 + \frac{e^{-8}}{C}\right)\right)\left(\frac{(d-2)(d-3)}{4d^2}\right) \\
&+ \sum_{k>2}^{d} \binom{d}{k} \mathcal{C}(k) \\
&= \log\left(1 + \frac{1}{C}\right)\left(\frac{3d+1}{4d} + \frac{\binom{d}{1}}{2d} + \frac{\binom{d}{2}}{2d^2}\right) + \log\left(1 + \frac{e^{-2}}{C}\right)\left(\frac{1}{2} + \binom{d}{1}\frac{3d^2+3d-2}{4d^2} + \frac{\binom{d}{2}}{d}\right) \\
&+ \log\left(1 + \frac{e^{-4}}{C}\right)\left(\binom{d-1}{1}\frac{1}{4d} + \binom{d}{1}\binom{d-1}{1}\frac{1}{2d} + \binom{d}{2}\left(\frac{3}{4} + \frac{2(d-1) + 3(d-2)}{4d^2}\right)\right) \\
&+ \log\left(1 + \frac{e^{-6}}{C}\right)\left(\binom{d}{1}\frac{(d-1)(d-2)}{4d^2} + \binom{d}{2}\frac{(d-2)}{2d}\right) \\
&+ \log\left(1 + \frac{e^{-8}}{C}\right)\left(\binom{d}{2}\frac{(d-2)(d-3)}{4d^2} + \sum_{k>2}^{d}\binom{d}{k}\mathcal{C}(k)\right) \\
&= \log\left(1 + \frac{e^{-2}}{C}\right)\left(\frac{d^2}{4d}\right) \\
&+ \log\left(1 + \frac{e^{-4}}{C}\right)\left(\binom{d-1}{1}\frac{1}{4d} + \binom{d}{1}\binom{d-1}{1}\frac{1}{2d} + \binom{d}{2}\left(\frac{3}{4} + \frac{2(d-1) + 3(d-2)}{4d^2}\right)\right) \\
&+ \log\left(1 + \frac{e^{-6}}{C}\right)\left(\binom{d}{1}\frac{(d-1)(d-2)}{4d^2} + \binom{d}{2}\frac{(d-2)}{2d}\right) \\
&+ \log\left(1 + \frac{e^{-8}}{C}\right)\left(\binom{d}{2}\frac{(d-2)(d-3)}{4d^2} + \sum_{k>2}^{d}\binom{d}{k}\mathcal{C}(k)\right) \\
&\frac{e^{-2}}{C}\frac{d^2}{8d}\left(3d+2\right) + \frac{c'}{C}d^2 e^{-4} + \sum_{k>2}^{d}\binom{d}{k}\mathcal{C}(k) \qquad (24)
\end{aligned}
$$

where in (24) we have used the numerical inequalities $\log(1+x) \geq \frac{x}{2}$ for $x \in [0;1]$ and $\log(1+x) \leq x$, and $> c\log d$, and c' is a sufficiently large constant.

For $k > 2$, we again use $\log(1+x) \leq x$ to obtain

$$
\begin{aligned}
\mathcal{C}(k) &\leq \frac{e^{-2(k-2)}}{C}\frac{(k-1)k}{4d^2} + \frac{e^{-2(k-1)}}{C}\frac{k}{2d} \quad \frac{1}{2}\log\left(1 + \frac{e^{-2k}}{C}\right) + \frac{e^{-2(k+1)}}{2C} + \frac{e^{-2(k+2)}}{4C} \\
&\leq \frac{e^{-2(k-2)}}{C}\frac{(k-1)k}{4d^2} + \frac{c''}{C}e^{-2(k-1)} \qquad (25)
\end{aligned}
$$

for an absolute constant c''. Combining this bound with (24) yields

$$
\begin{aligned}
\sum_{k=0}^{d}\binom{d}{k}\mathcal{C}(k) &\leq \frac{e^{-2}}{C}\frac{d^2}{8d}\left(3d+2\right) + \frac{c'}{C}d^2 e^{-4} + \sum_{k=3}^{d}\frac{e^{-2(k-2)}}{C}\frac{(k-1)k}{4d^2} + \frac{c''}{C}e^{-2(k-1)} \\
&\quad \frac{e^{-2}}{C}\left(\frac{d^3}{8d^2}\left(3d^2+2d\right) + \frac{12}{8d^2}\right) + \frac{c'''}{C}d^2 e^{-4} \qquad (26) \\
&< 0 \qquad (27)
\end{aligned}
$$

where (26) holds for an absolute constant c^{000}, and (27) holds for a sufficiently large constant c and $d > 3$, where throughout we have used $> c \log d$. ■

Lemma 19 *$fD_t g_t$ converges to U_d.*

Proof The transition kernel of D_t is aperiodic and irreducible over a finite state space, and has a symmetric transition kernel, so it must converge to the uniform distribution (Bremaud, 2001). ■

A.2. Proof of Theorem 7

Now using the above results, we prove the main claim of Theorem 7.

Proof Note that the InfoNCE loss can be written as

$$L(g) = \underset{x;x^+;fx_i\ g`}{\mathsf{E}} \log \Big(1 + \sum_{i=1}^{\dot{X}} e^{\ g(x)^{>}(g(x_i)\ \ g(x^+))}\Big) \tag{28}$$

Considering that we search over representations composed of clean functions, we know that for all $g \ 2 \ G_c$, the term $g(x)^{>}g(x^+)$ is exactly equal to d. Hence, the optimizing $L(g)$ over G_c simplifies to minimizing

$$\hat{L}(g) := \underset{x;x^+;fx_i\ g`}{\mathsf{E}} \log \Big(1 + \sum_{i=1}^{\dot{X}} e^{\ g(x)^{>}g(x_i)\ \ \ d}\Big) \tag{29}$$

Below, we use the term 'clean representation' to indicate that the the representation is composed of clean functions, and a non-clean representation if at least one of the functions in the representation is not clean. Recall the definitions of the functions $L(g)$ in (28) and $\hat{L}(g)$ in (29). Note that since we always have $g(x)^{>}g(x^+) \quad d$, we can argue that for any g we have $L(g) \quad \hat{L}(g)$. Indeed, the equality holds when g is a clean representation. Moreover, for any non-clean representation g, we know that there exists at least one image x for which its representation $g(x)$ is not exactly aligned with the representation of one of its augmented images x^+. Therefore, for that pair $(x; x^+)$, $g(x)^{>}g(x^+) > \ \ d$. Hence, for some sample $x; x^+; fx_i\ g`$ with positive mass, we have: $g(x)^{>}(g(x_i)\ \ g(x^+)) > \ g(x)^{>}g(x_i) \quad d$. Therefore, for non-clean g we have $L(g) > \hat{L}(g)$ from (28) and (29).

Moreover, according to the result of Lemma 6, we know that the minimizer of the loss function $\hat{L}$ is a uniform representation, thus for any non-uniform representation g^0 and uniform representation g^{00} we have $\hat{L}(g^0) > \hat{L}(g^{00})$.

Considering these two observations, we show that a uniform representation composed of clean functions, denoted by $g^{\ }$, is an optimal solution of the loss L. We consider the following two cases:

Case 1: If the representation g^0 is not composed of clean functions, then we have

$$L(g^0) \overset{(a)}{>} \hat{L}(g^0) \overset{(b)}{} \hat{L}(g^{\ }) \overset{(c)}{=} L(g^{\ })$$

where (a) holds with strict inequality since g^0 is not clean (discussion above), (b) holds as $g^{\ }$ with a uniform distribution is an optimal solution of $\hat{L}$ (Lemma 6), and (c) holds because $g^{\ }$ is composed of clean functions.

Case 2: If the representation g^0 is composed of clean functions, but is not uniform, then we have

$$L(g^0) \overset{(a)}{} \hat{L}(g^0) \overset{(b)}{>} \hat{L}(g\ ) \overset{(c)}{=} L(g\ )$$

where (a) holds based on the definitions of L and $\hat{L}$, (b) holds since g^0 is not uniform and g is uniform (discussion above), and (c) holds because g is composed of clean functions.

Combining these two cases, we obtain that the representation g minimizes $L(\ )$ if and only if g is composed of clean functions and uniform. Furthermore, by Assumption 2, g being composed of clean functions implies that is it cluster-preserving. ■

Appendix B. Agnostic Case

In this section we prove Theorem 15.

Theorem 20 (Theorem 15 Restated) *Suppose Assumptions 10 and 13 hold and $g = [f_1; \dots; f_d]$ is **not** cluster-preserving with $\min_{j\in[d]} \min_{\mathbf{c}} \mathrm{P}_{x,x^0 \sim D}[x, x^0 \in \mathbf{c}; f_j(x) \neq f_j(x^0)] > 0$. Let $\ell \geq cd2^d$, $c \log(\frac{c}{\ell})2^d$ for a sufficiently large constant $c > 1$. Moreover, suppose g is close to a uniform representation in the sense that $\mathrm{P}_{x\sim D}[g(x) = v] \leq \frac{10}{cd2^d}$ or $\mathrm{P}_{x\sim D}[g(x) = v] \geq \frac{1}{100cd2^{2d}}$ for all $v \in H_d$. Then g is **not** a minimizer of the InfoNCE loss.*

Proof First we recall notations: for a set of images $B \subseteq D$, we employ the notations $\|B\| := \mathrm{P}_{x\sim D}[x \in B]$ and $\|B|D\| := \mathrm{P}_{x\sim D|D}[x \in B]$. Also, we let $\mathbf{c}_{\setminus} := \mathbf{c} \setminus D$.

As discussed in the proof sketch, we construct a representation g^0 that is close to g by changing one coordinate of g such that it preserves one additional cluster, and show that the resulting g^0 achieves smaller InfoNCE loss than g.

Suppose WLOG that f_1 does not preserve the cluster $\mathbf{c}$ for some $\mathbf{c} \in C$. That is, $\exists x, x^0 \in \mathbf{c}$ such that $f_1(x) \neq f_1(x^0)$. Further, let $f_1^{(\mathbf{c};\ )}$ be the smallest perturbation of f_1 that preserves $\mathbf{c}$. Specifically, $f^{(\mathbf{c};\ )}(x) := \begin{cases} f(x) & x \notin \mathbf{c} \\ x & x \in \mathbf{c} \end{cases}$, where $\in \arg\min_{\emptyset \neq f \subseteq [1;1g} \|f x \in \mathbf{c};\ \forall j \in f,\ f_1(x) \neq f_1^{(\mathbf{c};\ \emptyset)}(x)g\|$.

Denote $f_1^{\emptyset} = f_1^{(\mathbf{c};\ )}$. By Assumption 10, $f_1^{\emptyset} \in F$. Construct $g^0 = [f_1^{\emptyset}; f_2; \dots; f_d] \in G$, and note that g^0 is equivalent to g on all but one coordinate, and the one differing coordinate differs only on one cluster, for which this coordinate preserves the cluster in g^0 but does not preserve it in g.

We show that $L(g) \quad L(g^0) > 0$, where

$$L(g) \quad L(g^0) = L^+(g) \quad L^+(g^0) + L\ \ (g) \quad L\ \ (g^0): \tag{30}$$

where

$$L^+(g) := \quad \mathrm{E}_{x,x^+}\left[g(x)^\top g(x^+) \right]$$

$$L\ \ (g) := \mathrm{E}_{x,x^+,\{x_i\}}\ \log\ \ e^{\,g(x)^\top g(x^+)} + \sum_{i=1} e^{\,g(x)^\top g(x_i)}$$

where L^+ and L respectively correspond to the alignment and uniformity losses in 3. We will refer to these losses as the positive and negative losses, respectively. We show $L(g) \quad L(g^0) > 0$ by showing

$$L^+(g) \quad L^+(g^0) > L \quad (g^0) \quad L \quad (g) \tag{31}$$

by first computing the LHS explicitly, then upper bounding the RHS. To do this, we define a partitioning of the set of images D as follows. For all $v \ 2 \ H_d$, define the sets $Q_v := fx \ 2 \ D : g(x) = v; g^0(x) = vg$ and $E_v := fx \ 2 \ D : g(x) = v; g^0(x) \ \cancel{6}= vg$. In other words $Q_v \ [\ E_v$ is the set of images that g maps to v, and Q_v is the subset of this set which g^0 also maps to v, while E_v is the subset which g^0 does not map to v. Let $Q := \ [_{v2H_d} Q_v$ and $E := \ [_{v2H_d} E_v$. Observe that $Q \ [\ E = D$, and each of the Q_v and E_v's are disjoint, so they form a partition of D. Further note that $kEk \ = \ _{f_1;\mathbf{c}} := \min_{^0 2 f \ 1;1g} kfx \ 2 \ _{\mathbf{c};} \ j \ f_1(x) \ \cancel{6}= \ f_1^{(\mathbf{c};\ ^0)}(x)gk$, as defined in Definition 12.

Now we consider the difference in the positive losses. For all augmentations $x^+ \ 2 \ D \ n \ D$, define $A^{-1}(x^+)$ as the natural image from which the augmentation was derived, i.e. $A^{-1}(x^+) = x \quad ( ) \quad A(x) = x^+$ for some $A \ 2 \quad$. Moreover, define $R := fx^+ \ 2 \ E \ n \ D : A^{-1}(x^+) \ 2 \ Qg \ [\ fx^+ \ 2 \ Q \ n \ D \ : A^{-1}(x^+) \ 2 \ Eg = fx^+ \ 2 \ _{\mathbf{c}} : f_1(A^{-1}(x^+)) \ \cancel{6}= \ f_1(x^+)g$ as the set of augmentations in $_{\mathbf{c}}$ that f_1 classifies incorrectly. For any event B, let B denote the indicator random variable for B, i.e. $B = 1$ if B occurs and $B = 0$ otherwise. Using this notation and the construction of g^0 we can write the difference in positive losses as:

$$\begin{aligned}
L^+(g) \quad L^+(g^0) &= \mathrm{E}_{x;x^+}\left[g^0(x) > g^0(x^+) \quad g(x) > g(x^+) \right] \\
&= \mathrm{E}_{x;x^+} \quad f_1^0(x) f_1^0(x^+) \quad f_1(x) f_1(x^+) \\
&= \sum_{v2H_d} \mathrm{E}_{x;x^+} \quad fx \ 2 \ Q_v g(f_1^0(x) f_1^0(x^+) \quad f_1(x) f_1(x^+)) \\
&\qquad + \quad fx \ 2 \ E_v g(f_1^0(x) f_1^0(x^+) \quad f_1(x) f_1(x^+)) \\
&= 2 \sum_{v2H_d} \mathrm{E}_{x;x^+} \quad fx \ 2 \ Q_v g \ \ fx^+ \ 2 \ Eg + \quad fx \ 2 \ E_v g \ \ fx^+ \ 2 \ Qg \\
&= 2 \sum_{v2H_d} \mathrm{P}_{x \ \ D \ ;x^+ \ \ A(x)} \ x \ 2 \ Q_v; x^+ \ 2 \ E \\
&\qquad + \mathrm{P}_{x \ \ D \ ;x^+ \ \ A(x)} \ x \ 2 \ E_v; x^+ \ 2 \ Q \\
&= 2 \ kRk
\end{aligned}$$

where in the last equality we have used that all augmentation sets are of equal size.

Now we consider the negative losses. We first decompose the negative loss as

$$\begin{aligned}
L \quad (g) &= \mathrm{E}_{x;x^+;fx_i g^{\cdot}} \quad \log \quad e^{\ g(x)^{\top} g(x^+)} + \sum_{i=1}^{\cdot} e^{\ g(x)^{\top} g(x_i)} \\
&= \sum_{v2H_d} \mathrm{E}_{x;x^+;fx_i g^{\cdot}} \quad fx \ 2 \ Q_v g \log \quad e^{\ g(x)^{\top} g(x^+)} + \sum_{i=1}^{\cdot} e^{\ g(x)^{\top} g(x_i)} \\
&\quad + \sum_{v2H_d} \mathrm{E}_{x;x^+;fx_i g^{\cdot}} \quad fx \ 2 \ E_v g \log \quad e^{\ g(x)^{\top} g(x^+)} + \sum_{i=1}^{\cdot} e^{\ g(x)^{\top} g(x_i)}
\end{aligned}$$

$$= \sum_{v \in H_d} L_{Q_v}(g) + L_{E_v}(g)$$

where $L_{Q_v}(g) := \mathrm{E}_{x,x^+,\{x_i\}_{i=1}^{\ell}}\left[\mathbb{1}\{x \in Q_v\} \log\left(e^{g(x)^\top g(x^+)} + \sum_{i=1}^{\ell} e^{g(x)^\top g(x_i)}\right)\right]$ and $L_{E_v}(g) := \mathrm{E}_{x,x^+,\{x_i\}_{i=1}^{\ell}}\left[\mathbb{1}\{x \in E_v\} \log\left(e^{g(x)^\top g(x^+)} + \sum_{i=1}^{\ell} e^{g(x)^\top g(x_i)}\right)\right]$. Note that we need to upper bound

$$L(g^0) - L(g) = \sum_{v \in H_d} L_{Q_v}(g^0) - L_{Q_v}(g) + L_{E_v}(g^0) - L_{E_v}(g). \tag{32}$$

We analyze $L_{Q_v}(g^0) - L_{Q_v}(g)$ and $L_{E_v}(g^0) - L_{E_v}(g)$ separately for every $v \in H_d$. To do so, we define additional notations. For a batch of negative samples $\{x_i\}_{i=1}^{\ell}$ and a vertex $v \in H_d$, let $n_{1;v} := \sum_{i=1}^{\ell} \mathbb{1}\{x_i \in Q_v\}$, $n_2 := \sum_{i=1}^{\ell} \mathbb{1}\{x_i \in E\}$, $n_{3;v} := \sum_{i=1}^{\ell} \mathbb{1}\{x_i \in E_v\}$ and $n_4 := \sum_{i=1}^{\ell} \mathbb{1}\{x_i \in Q\}$. Using the fact that $g(x)^\top g(x_i) = d$ for all $x, x_i \in Q_v$, we have

$$L_{Q_v}(g) = \mathrm{E}_{x,x^+,\{x_i\}_{i=1}^{\ell}}\left[\mathbb{1}\{x \in Q_v\} \log\left(e^{g(x)^\top g(x^+)} + n_{1;v} e^{d} + \sum_{x_i \notin Q_v} e^{g(x)^\top g(x_i)}\right)\right]$$

$$= \mathrm{E}_{x,x^+,\{x_i\}_{i=1}^{\ell}}\left[\mathbb{1}\{x \in Q_v\} \log\left(e^{g(x)^\top g(x^+)} + n_{1;v} e^{d} + \sum_{x_i \in E} e^{g(x)^\top g(x_i)} + \sum_{x_i \notin Q_v \cup E} e^{g(x)^\top g(x_i)}\right)\right]$$

Next, using the fact that $g^0(x) = g(x)$ for all $x \notin E$, we have

$$L_{Q_v}(g^0) = \mathrm{E}_{x,x^+,\{x_i\}_{i=1}^{\ell}}\left[\mathbb{1}\{x \in Q_v\} \log\left(e^{g^0(x)^\top g^0(x^+)} + n_{1;v} e^{d} + \sum_{x_i \notin Q_v} e^{g^0(x)^\top g^0(x_i)}\right)\right]$$

$$= \mathrm{E}_{x,x^+,\{x_i\}_{i=1}^{\ell}}\left[\mathbb{1}\{x \in Q_v\} \log\left(e^{g^0(x)^\top g^0(x^+)} + n_{1;v} e^{d} + \sum_{x_i \in E} e^{g^0(x)^\top g^0(x_i)} + \sum_{x_i \notin Q_v \cup E} e^{g(x)^\top g(x_i)}\right)\right]$$

Before analyzing L_{E_v}, we first prove the following claims.

Claim B.1 *For all $x, x' \in E_v$, $g^0(x) = g^0(x')$.*

Proof By construction of g^0, $g^0(x)$ agrees with $g(x)$ on all but the first coordinate. Thus, $g(x) \neq g^0(x) \Rightarrow f_1(x) = -f_1^0(x)$. Consider any $x, x' \in E_v$. By definition of E_v, $g(x) = g(x') = v$, and $g^0(x) \neq g(x)$. Let $v = [v_1, v_2, \ldots, v_d]$, then we have $g^0(x) = g^0(x') = [-v_1, v_2, \ldots, v_d]$. ■

Claim B.2 *For all $v \in H_d$ and all $x \in E_v$, $x' \notin E$, $g^0(x)^\top g^0(x') = g(x)^\top g(x')$.*

Proof Consider any $x \in E_v, x' \in E$. As above, observe that the j-th coordinates of $g(x)$ and $g'(x)$ are the same for all $j > 1$ (and likewise for $g(x')$ and $g'(x')$) by construction of g. Moreover, $f_1(x) = f_1'(x)$ and $f_1(x') = f_1'(x')$ by definition of E. Thus $g'(x)^\top g'(x') - g(x)^\top g(x') = f_1'(x)f_1'(x') - (f_1'(x))(f_1'(x')) = 0$. ■

Claim B.3 *For all $v \in H_d$ and all $x \in E_v$, $x' \in Q$, $g'(x)^\top g'(x') \not\preceq g(x)^\top g(x')$.*

Proof By definition of E_v, $g'(x) \not\preceq g(x)$ for all $x \in E_v$, and by definition of Q, $g'(x') = g(x')$ for all $x' \in Q$. Thus $g'(x)^\top g'(x') \not\preceq g(x)^\top g(x')$. ■

Next we can decompose $L_{E_v}(g)$ as follows, using the fact that $g(x)^\top g(x_i) = d$ for all $x, x_i \in E_v$.

$$L_{E_v}(g) = \mathbb{E}_{x,x^+,\{x_i\}}\left[\mathbb{1}\{x \in E_v\} \log \frac{e^{g(x)^\top g(x^+)} + n_{3,v}e^{d}}{e^{g(x)^\top g(x^+)} + n_{3,v}e^{d} + \sum_{x_i \neq x} e^{g(x)^\top g(x_i)}}\right]$$

$$= \mathbb{E}_{x,x^+,\{x_i\}}\left[\mathbb{1}\{x \in E_v\} \log \frac{e^{g(x)^\top g(x^+)} + n_{3,v}e^{d}}{e^{g(x)^\top g(x^+)} + n_{3,v}e^{d} + \sum_{x_i \in Q} e^{g(x)^\top g(x_i)} + \sum_{x_i \neq x, x_i \notin E_v \cup Q} e^{g(x)^\top g(x_i)}}\right]$$

Next we use Claims B.1, B.2 and B.3 to obtain

$$L_{E_v}(g') = \mathbb{E}_{x,x^+,\{x_i\}}\left[\mathbb{1}\{x \in E_v\} \log \frac{e^{g'(x)^\top g'(x^+)} + n_{3,v}e^{d}}{e^{g'(x)^\top g'(x^+)} + n_{3,v}e^{d} + \sum_{x_i \neq x} e^{g'(x)^\top g'(x_i)}}\right]$$

$$= \mathbb{E}_{x,x^+,\{x_i\}}\left[\mathbb{1}\{x \in E_v\} \log \frac{e^{g'(x)^\top g'(x^+)} + n_{3,v}e^{d}}{e^{g'(x)^\top g'(x^+)} + n_{3,v}e^{d} + \sum_{x_i \in Q} e^{g'(x)^\top g'(x_i)} + \sum_{x_i \neq x, x_i \notin E_v \cup Q} e^{g(x)^\top g(x_i)}}\right]$$

Next, define $Z := \{x \in Q : f_1'(x) \not\preceq f_1'(x'); \forall x' \in E\} \subseteq Q$. as the set of images that f_1' labels differently than it does the samples in E (note that $f_1'(x) = f_1'(x')$ for all $x, x' \in E$, and $f_1'(x) = f_1'(x')$ for all $x, x' \in Q_v$, so the $\forall$ condition in the definition of Z could be replaced with 'for some'). Also note that the definition of Z here differs slightly from the one in the proof sketch in Section 5 for ease of notation.

Next we prove two claims regarding properties of Z.

Claim B.4 *For all $v \in H_d$ exactly one of the following holds: (i) $Q_v \setminus Z = Q_v$ or (ii) $Q_v \cap Z = Q_v$.*

Proof Suppose $x \in Q_v \setminus Z$. Then, for all $x' \in Q_v$, $f_1'(x') = f_1'(x)$ by definition of Q_v. Thus $f_1'(x') = f_1'(x) \not\preceq f_1'(x'')$ for any $x'' \in E$ since $x \in Z$. This implies $x' \in Z$, therefore $Q_v \setminus Z = Q_v$.

Likewise, suppose $x \in Q_v \cap Z$. Then, for all $x' \in Q_v$, $f_1'(x') = f_1'(x)$ by definition of Q_v. Thus $f_1'(x') = f_1'(x) = f_1'(x'')$ for any $x'' \in E$ since $x \notin Z$. This implies $x' \notin Z$, therefore $Q_v \cap Z = Q_v$. ■

Claim B.5 *For all* $v \in H_d$, $x \in Q_v \setminus Z$, $x' \in E$, $g^0(x)^\top g^0(x') = g(x)^\top g(x')$.

Proof Suppose $x \in Q_v \in Z$. For any $x' \in E$, then $f_1(x_i) = f_1^0(x_i)$ by definition of E, $f_1(x) = f_1^0(x)$ by definition of Q_v and $f_1^0(x) = f_1^0(x_i)$ by definition of Z. Therefore, $f_1(x) = f_1(x_i)$ and $g^0(x)^\top g^0(x') = g(x)^\top g(x')$, noting that g and g^0 agree on all but the first coordinate. ■

Claim B.6 *For all* $v \in H_d$, $Q_v \setminus Z = \emptyset$ *and* $Q_v \not\subseteq \emptyset \implies E_v = \emptyset$.

Proof From Claim B.4, $Q_v \setminus Z = \emptyset \implies Q_v \cap Z = Q_v$. Suppose $x \in Q_v \cap Z$ and $x^0 \in E_v$. Then $f_1(x) = f_1(x^0)$ by definition of Q_v and E_v. Also, $f_1^0(x) = f_1^0(x^0)$ since $x \in Z$ and $x^0 \in E$, and $f_1(x^0) \not= f_1^0(x^0)$ by definition of E. Therefore $f_1(x) \not= f_1^0(x)$, but this contradicts the definition of Q_v. ■

We use Claim B.6 later in the proof. For now we use Claims B.4 and B.5 to bound $L_{Q_v}(g^0) - L_{Q_v}(g)$ for all v such that $Q_v \not= \emptyset$:

$$
\begin{aligned}
& L_{Q_v}(g^0) - L_{Q_v}(g) \\
&= \mathbb{E}_{x,x^+,\{x_i\}} \Big[\mathbb{1}\{x \in Q_v\} \\
&\quad \log \frac{e^{g^0(x)^\top g^0(x^+)} + n_{1,v} e^{-d} + \sum_{x_i \in E} e^{g^0(x)^\top g^0(x_i)} + \sum_{x_i \notin Q_v \cup E} e^{g(x)^\top g(x_i)}}{e^{g(x)^\top g(x^+)} + n_{1,v} e^{-d} + \sum_{x_i \in E} e^{g(x)^\top g(x_i)} + \sum_{x_i \notin Q_v \cup E} e^{g(x)^\top g(x_i)}} \Big] \\
&= \mathbb{E}_{x,x^+,\{x_i\}} \Big[\mathbb{1}\{x \in Q_v\} \\
&\quad \log \frac{e^{g(x)^\top g(x^+)} + n_{1,v} e^{-d} + e^{-2} \sum_{x_i \in E} e^{g(x)^\top g(x_i)} + \sum_{x_i \notin Q_v \cup E} e^{g(x)^\top g(x_i)}}{e^{g(x)^\top g(x^+)} + n_{1,v} e^{-d} + \sum_{x_i \in E} e^{g(x)^\top g(x_i)} + \sum_{x_i \notin Q_v \cup E} e^{g(x)^\top g(x_i)}} \Big] \qquad (33) \\
&\le 0 \qquad (34)
\end{aligned}
$$

Thus, we have

$$
\begin{aligned}
L(g^0) - L(g) = & \sum_{v \in H_d} L_{Q_v}(g^0) - L_{Q_v}(g) + L_{E_v}(g^0) - L_{E_v}(g) \\
& \sum_{v \in H_d : Q_v \cap Z = Q_v} L_{Q_v}(g^0) - L_{Q_v}(g) + \sum_{v \in H_d} L_{E_v}(g^0) - L_{E_v}(g) \qquad (35)
\end{aligned}
$$

For each $v \in H_d : Q_v \cap Z = Q_v$, we consider three cases: (1) $n_2 = 0$, (2) $n_2 > 0; n_{1;v=0}$, and (3) $n_2 > 0; n_{1;v} > 0$. In particular we decompose $L_{Q_v}(g^0) - L_{Q_v}(g)$ as follows:

$$L_{Q_v}(g^0) - L_{Q_v}(g)$$

$$
\begin{aligned}
&= \mathrm{E}_{x;x^+;\{x_i\}}\big[\ \{x \in Q_v \cap Z\}\ \{n_2 = 0\}\big] \\
&\quad \log\Big(e^{g'(x)\cdot g'(x^+)} + n_{1;v}e^{\delta} + \sum_{x_i \in E} e^{g'(x)\cdot g'(x_i)} + \sum_{x_i \in Q_v \setminus E} e^{g(x)\cdot g(x_i)}\Big) \\
&\quad \log\Big(e^{g(x)\cdot g(x^+)} + n_{1;v}e^{\delta} + \sum_{x_i \in E} e^{g(x)\cdot g(x_i)} + \sum_{x_i \in Q_v \setminus E} e^{g(x)\cdot g(x_i)}\Big) \\
&+ \mathrm{E}_{x;x^+;\{x_i\}}\big[\ \{x \in Q_v \cap Z\}\ \{n_2 > 0; n_{1;v} = 0\}\big] \\
&\quad \log\Big(e^{g'(x)\cdot g'(x^+)} + n_{1;v}e^{\delta} + \sum_{x_i \in E} e^{g'(x)\cdot g'(x_i)} + \sum_{x_i \in Q_v \setminus E} e^{g(x)\cdot g(x_i)}\Big) \\
&\quad \log\Big(e^{g(x)\cdot g(x^+)} + n_{1;v}e^{\delta} + \sum_{x_i \in E} e^{g(x)\cdot g(x_i)} + \sum_{x_i \in Q_v \setminus E} e^{g(x)\cdot g(x_i)}\Big) \\
&+ \mathrm{E}_{x;x^+;\{x_i\}}\big[\ \{x \in Q_v \cap Z\}\ \{n_2 > 0; n_{1;v} > 0\}\big] \\
&\quad \log\Big(e^{g'(x)\cdot g'(x^+)} + n_{1;v}e^{\delta} + \sum_{x_i \in E} e^{g'(x)\cdot g'(x_i)} + \sum_{x_i \in Q_v \setminus E} e^{g(x)\cdot g(x_i)}\Big) \\
&\quad \log\Big(e^{g(x)\cdot g(x^+)} + n_{1;v}e^{\delta} + \sum_{x_i \in E} e^{g(x)\cdot g(x_i)} + \sum_{x_i \in Q_v \setminus E} e^{g(x)\cdot g(x_i)}\Big)
\end{aligned}
$$

Likewise, for each $v \in H_d$, we decompose $L_{E_v}(g') - L_{E_v}(g)$ as:

$$
\begin{aligned}
&L_{E_v}(g') - L_{E_v}(g) \\
&= \mathrm{E}_{x;x^+;\{x_i\}}\big[\ \{x \in E_v\}\ \{n_4 = 0\}\big] \\
&\quad \log\Big(e^{g'(x)\cdot g'(x^+)} + n_{3;v}e^{\delta} + \sum_{x_i \in Q} e^{g'(x)\cdot g'(x_i)} + \sum_{x_i \in E_v \setminus Q} e^{g(x)\cdot g(x_i)}\Big) \\
&\quad \log\Big(e^{g(x)\cdot g(x^+)} + n_{3;v}e^{\delta} + \sum_{x_i \in Q} e^{g(x)\cdot g(x_i)} + \sum_{x_i \in E_v \setminus Q} e^{g(x)\cdot g(x_i)}\Big) \\
&+ \mathrm{E}_{x;x^+;\{x_i\}}\big[\ \{x \in E_v\}\ \{n_4 > 0; n_{3;v} = 0\}\big] \\
&\quad \log\Big(e^{g'(x)\cdot g'(x^+)} + n_{3;v}e^{\delta} + \sum_{x_i \in Q} e^{g'(x)\cdot g'(x_i)} + \sum_{x_i \in E_v \setminus Q} e^{g(x)\cdot g(x_i)}\Big) \\
&\quad \log\Big(e^{g(x)\cdot g(x^+)} + n_{3;v}e^{\delta} + \sum_{x_i \in Q} e^{g(x)\cdot g(x_i)} + \sum_{x_i \in E_v \setminus Q} e^{g(x)\cdot g(x_i)}\Big) \\
&+ \mathrm{E}_{x;x^+;\{x_i\}}\big[\ \{x \in E_v\}\ \{n_4 > 0; n_{3;v} > 0\}\big]
\end{aligned}
$$

$$\log \quad e^{\,g^0(x)^\top g^0(x^+)} + n_{3;v} e^{\,d} + \sum_{x_i \in Q} e^{\,g^0(x)^\top g^0(x_i)} + \sum_{x_i \notin E_v[Q} e^{\,g(x)^\top g(x_i)}$$

$$\log \quad e^{\,g(x)^\top g(x^+)} + n_{3;v} e^{\,d} + \sum_{x_i \in Q} e^{\,g(x)^\top g(x_i)} + \sum_{x_i \notin E_v[Q} e^{\,g(x)^\top g(x_i)}$$

Thus, for each v in (32), we need to upper bound a total of six terms. We consider each of these six terms individually, starting with the three terms with $\{x \in Q_v \cap Z\}$ factors.

1. $x \in Q_v \cap Z; n_2 = 0$.

   In this case, we have

$$\begin{aligned}
((1)) &:= \mathbb{E}_{x;x^+;\{x_i\}} \quad \{x \in Q_v \cap Z\} \quad \{n_2 = 0\} \\
&\log \quad e^{\,g^0(x)^\top g^0(x^+)} + n_{1;v} e^{\,d} + \sum_{x_i \notin Q_v[E} e^{\,g(x)^\top g(x_i)} \\
&\log \quad e^{\,g(x)^\top g(x^+)} + n_{1;v} e^{\,d} + \sum_{x_i \notin Q_v[E} e^{\,g(x)^\top g(x_i)} \qquad (36) \\
&\mathbb{E}_{x;x^+;\{x_i\}} \quad \{x \in Q_v \cap Z\} \quad \{n_2 = 0\} \\
&\log \quad e^{\,g^0(x)^\top g^0(x^+)} + n_{1;v} e^{\,d} \qquad \log \quad e^{\,g(x)^\top g(x^+)} + n_{1;v} e^{\,d} \qquad (37) \\
&= \mathbb{E}_{x;x^+;\{x_i\}} \quad \{x \in Q_v \cap Z; x^+ \in E; n_2 = 0\} \log \quad \frac{e^2 \; e^{\,g(x)^\top g(x^+)} + n_{1;v} e^{\,d}}{e^{\,g(x)^\top g(x^+)} + n_{1;v} e^{\,d}} \qquad (38) \\
&\mathbb{E}_{x;x^+;\{x_i\}} \quad \{x \in Q_v \cap Z; x^+ \in E; n_2 = 0\} \log \quad \frac{e^{\,d} + n_{1;v} e^{\,d}}{e^{\,(d\ 2)} + n_{1;v} e^{\,d}} \qquad (39) \\
&= \mathbb{E}_{x;x^+;\{x_i\}} \quad \{x \in Q_v \cap Z; x^+ \in E; n_2 = 0\} \log \quad \frac{1 + n_{1;v}}{e^{\ 2} + n_{1;v}} \\
&= \mathbb{E}_{x;x^+;\{x_i\}} \quad \{x \in Q_v \cap Z; x^+ \in E; n_2 = 0; n_{1;v} = 0\} \log \quad \frac{1 + n_{1;v}}{e^{\ 2} + n_{1;v}} \\
&\quad + \mathbb{E}_{x;x^+;\{x_i\}} \quad \{x \in Q_v \cap Z; x^+ \in E; n_2 = 0; n_{1;v} > 0\} \log \quad \frac{1 + n_{1;v}}{e^{\ 2} + n_{1;v}} \\
&\quad 2 \quad \mathbb{E}_{x;x^+;\{x_i\}} \quad \{x \in Q_v \cap Z; x^+ \in E; n_2 = 0; n_{1;v} = 0\} \\
&\quad + \log(2) \mathbb{E}_{x;x^+;\{x_i\}} \quad \{x \in Q_v \cap Z; x^+ \in E; n_2 = 0; n_{1;v} > 0\} \qquad (40)
\end{aligned}$$

   where (37) follows by the submodularity of the $\log(\,)$ function and the fact that $g^0(x)^\top g^0(x^+) \le g(x)^\top g(x^+)$ by construction of g^0, (54) follows by the fact that if $x \in Q_v$, then $g^0(x)^\top g^0(x^+) = g(x)^\top g(x^+)$ for all $x^+ \notin B$, and (39) follows since $h(x) := \frac{ax+c}{x+c}$ is monotonically increasing

for $a > 1$. Next, by the independence of x_i from x and x^+,

$$\begin{aligned}
&\mathsf{E}_{x;x^+;\{x_i\}}\big[\mathbb{1}\{x \in Q_v \cap Z;\, x^+ \in E;\, n_2 = 0;\, n_{1;v} = 0\}\big] \\
&\quad = \mathsf{P}(x \in Q_v \cap Z \mid x^+ \in B)\mathsf{P}(n_{1;v} = 0;\, n_2 = 0) \\
&\quad = \mathsf{P}(x \in Q_v \cap Z \mid x^+ \in B)(1 - |Q_v \cap Z| - |E|)^{\ell} \qquad (41)
\end{aligned}$$

Similarly, for the second term in (40), we have

$$\begin{aligned}
&\mathsf{E}_{x;x^+;\{x_i\}}\big[\mathbb{1}\{x \in Q_v \cap Z;\, x^+ \in E;\, n_2 = 0;\, n_{1;v} > 0\}\big] \\
&\quad = \mathsf{P}(x \in Q_v \cap Z \mid x^+ \in E)\mathsf{P}(n_{1;v} > 0 | n_2 = 0)\mathsf{P}(n_2 = 0) \qquad (42) \\
&\quad \mathsf{P}(x \in Q_v \cap Z \mid x^+ \in E)\min\Big(1;\, \frac{\ell |Q_v \cap Z|}{1 - |E|}\Big)(1 - |E|)^{\ell} \qquad (43)
\end{aligned}$$

By combining (43), (41), and (40), we obtain the following upper bound on (36):

$$\begin{aligned}
&((1)) \quad \mathsf{P}(x \in Q_v \cap Z \mid x^+ \in E) \\
&\qquad 2(1 - |Q_v \cap Z| - |E|)^{\ell} + \log(2)\min\Big(1;\, \frac{\ell |Q_v \cap Z|}{1 - |E|}\Big)(1 - |E|)^{\ell} : \qquad (44)
\end{aligned}$$

2. $x \in Q_v \cap Z; n_{1;v} = 0; n_2 > 0$

In this case we have:

$$\begin{aligned}
&((2)) := \mathsf{E}_{x;x^+;\{x_i\}}\big[\mathbb{1}\{x \in Q_v \cap Z; n_{1;v} = 0; n_2 > 0\}\big] \\
&\quad \log\Big(e^{g^0(x)\cdot g^0(x^+)} + n_{1;v}e^{d} + \sum_{x_i \in E} e^{g^0(x)\cdot g^0(x_i)} + \sum_{x_i \in Q_v \setminus E} e^{g(x)\cdot g(x_i)}\Big) \\
&\quad \log\Big(e^{g(x)\cdot g(x^+)} + n_{1;v}e^{d} + \sum_{x_i \in E} e^{g(x)\cdot g(x_i)} + \sum_{x_i \in Q_v \setminus E} e^{g(x)\cdot g(x_i)}\Big) \qquad (45)
\end{aligned}$$

$$\begin{aligned}
&= \mathsf{E}_{x;x^+;\{x_i\}}\big[\mathbb{1}\{x \in Q_v \cap Z; n_{1;v} = 0; n_2 > 0\}\big] \\
&\quad \log\Big(e^{g^0(x)\cdot g^0(x^+)} + \sum_{x_i \in E} e^{g^0(x)\cdot g^0(x_i)} + \sum_{x_i \in Q_v \setminus E} e^{g(x)\cdot g(x_i)}\Big) \\
&\quad \log\Big(e^{g(x)\cdot g(x^+)} + \sum_{x_i \in E} e^{g(x)\cdot g(x_i)} + \sum_{x_i \in Q_v \setminus E} e^{g(x)\cdot g(x_i)}\Big) \qquad (46)
\end{aligned}$$

$$\begin{aligned}
&\mathsf{E}_{x;x^+;\{x_i\}}\big[\mathbb{1}\{x \in Q_v \cap Z; n_{1;v} = 0; n_2 > 0\}\big] \\
&\quad \log\Big(e^{g^0(x)\cdot g^0(x^+)} + \sum_{x_i \in E} e^{g^0(x)\cdot g^0(x_i)}
\end{aligned}$$

$$\log\left(e^{g(x)^\top g(x^+)} + \sum_{x_i \in E} e^{g(x)^\top g(x_i)}\right) \tag{47}$$

where (47) follows by the submodularity of the $\log()$ function and the facts that $g^0(x)^\top g^0(x^+) \geq g(x)^\top g(x^+)$ and $g^0(x)^\top g^0(x_i) \geq g(x)^\top g(x_i)$ for all $x \in Q_v \cap Z; x_i \notin E$ by definition of g^0, C and E. Next,

$$\begin{aligned}
((2)) &\leq 2\,\mathrm{P}(x \in Q_v \cap Z)\mathrm{E}_{\{x_i\}\sim g}[\,\{n_{1,v} = 0\} \cap \{n_2 > 0\}] && (48)\\
&= 2\,\mathrm{P}(x \in Q_v \cap Z)\mathrm{E}_{\{x_i\}\sim g}[\,\{\cap_i \{x_i \notin Q_v \cap Z\}\} \cap \{\cup_i \{x_i \in E\}\}]\\
&= 2\,\mathrm{P}(x \in Q_v \cap Z)\mathrm{P}(\cup_i \{x_i \in E\} \mid \cap_i \{x_i \notin Q_v \cap Z\})\mathrm{P}(\cap_i \{x_i \notin Q_v \cap Z\})\\
&\leq 2\,\mathrm{P}(x \in Q_v \cap Z)\mathrm{P}(\cap_i \{x_i \notin Q_v \cap Z\}) \min\left(1; \sum_{i=1}^{\infty} \mathrm{P}(x_i \in E \mid x_i \notin Q_v \cap Z)\right) && (49)\\
&= 2\,|Q_v \cap Z|\,(1 - |Q_v \cap Z|)^{\infty} \min\left(1; \frac{\infty |E|}{1 - |Q_v \cap Z|}\right)
\end{aligned}$$

where (48) follows since $g^0(x)^\top g^0(x^+) \geq g(x)^\top g(x^+) \geq 2$ and $g^0(x)^\top g^0(x_i) \geq g(x)^\top g(x_i) \geq 2$ for all x, x_i, and (49) follows by a union bound.

3. $x \in Q_v \cap Z; n_1 > 0; n_2 > 0$.

In this case, we have

$$\begin{aligned}
((3)) &:= \mathrm{E}\{x \in Q_v \cap Z; n_{1,v} > 0; n_2 > 0\}\\
&\log\left(e^{g^0(x)^\top g^0(x^+)} + n_{1,v} e^{-\tau^{-1}} + \sum_{x_i \in E} e^{g^0(x)^\top g^0(x_i)} + \sum_{x_i \notin Q_v \cup E} e^{g(x)^\top g(x_i)}\right)\\
&\log\left(e^{g(x)^\top g(x^+)} + n_{1,v} e^{-\tau^{-1}} + \sum_{x_i \in E} e^{g(x)^\top g(x_i)} + \sum_{x_i \notin Q_v \cup E} e^{g(x)^\top g(x_i)}\right)
\end{aligned} \tag{50}$$

$$\begin{aligned}
&\mathrm{E}\{x \in Q_v \cap Z; n_{1,v} > 0; n_2 > 0\}\\
&\log\left(e^{g^0(x)^\top g^0(x^+)} + n_{1,v} e^{-\tau^{-1}} + e^2 \sum_{x_i \in E} e^{g(x)^\top g(x_i)}\right)\\
&\log\left(e^{g(x)^\top g(x^+)} + n_{1,v} e^{-\tau^{-1}} + \sum_{x_i \in E} e^{g(x)^\top g(x_i)}\right)
\end{aligned} \tag{51}$$

where (51) follows by the submodularity of $\log()$ and the facts that $g^0(x)^\top g^0(x^+) \geq g(x)^\top g(x^+)$ and $g^0(x)^\top g^0(x_i) = g(x)^\top g(x_i) + 2$ for all $x \in Q_v \cap Z; x_i \notin E$ by definition of g^0, Z and E. Continuing, we obtain

$$((3)) \quad \mathrm{E}\{x \in Q_v \cap Z; n_{1,v} > 0; n_2 > 0\}$$

$$\log\ e^{g^0(x)^\top g^0(x^+)} + (n_{1;v} + n_2)e^{d}$$

$$\log\ e^{g(x)^\top g(x^+)} + n_{1;v}e^{d} + n_2 e^{(d\ 2)} \tag{52}$$

$$= \mathrm{E}\quad fx\ 2\ Q_v\ n\ Z; n_{1;v} > 0; n_2 > 0; x^+\ \not\in\ Eg$$

$$\log\ e^{g^0(x)^\top g^0(x^+)} + (n_{1;v} + n_2)e^{d}$$

$$\log\ e^{g(x)^\top g(x^+)} + n_{1;v}e^{d} + n_2 e^{(d\ 2)}$$

$$+ \mathrm{E}\quad fx\ 2\ Q_v\ n\ Z; n_{1;v} > 0; n_2 > 0; x^+\ 2\ Eg$$

$$\log\ e^{g^0(x)^\top g^0(x^+)} + (n_{1;v} + n_2)e^{d}$$

$$\log\ e^{g(x)^\top g(x^+)} + n_{1;v}e^{d} + n_2 e^{(d\ 2)} \tag{53}$$

where (52) follows since $h(x) := \frac{a+cx}{b+x}$ is an increasing function of x for $x > 0; c > \frac{a}{b}$ (here, $\frac{a}{b}$ $\frac{2}{1+e^{\ 2}}$ 2 and $c = e^2 > 2$). Note that $x\ 2\ Q_v; x^+\ 2\ E \ =)\ \ g^0(x)^\top g^0(x^+) = g(x)^\top g(x^+) + 2$, and $x\ 2\ Q_v; x^+\ \not\in\ E \ =)\ \ g^0(x)^\top g^0(x^+) = g(x)^\top g(x^+)$. Using this we find

$$\begin{aligned}
((3)) \quad & \mathrm{E}\quad fx\ 2\ Q_v\ n\ Z; n_{1;v} > 0; n_2 > 0; x^+\ \not\in\ Eg \log\ \frac{n_{1;v} + n_2}{n_{1;v}} \\
& + \mathrm{E}\quad fx\ 2\ Q_v\ n\ Z; n_{1;v} > 0; n_2 > 0; x^+\ 2\ Bg \log\ \frac{n_{1;v} + n_2 + 1}{n_{1;v}} \\
& \mathrm{E}\quad fx\ 2\ Q_v\ n\ Z; n_{1;v} > 0; n_2 > 0; x^+\ \not\in\ Bg \frac{2n_2}{n_{1;v} + 1} \\
& + \mathrm{E}\quad fx\ 2\ Q_v\ n\ Z; n_{1;v} > 0; n_2 > 0; x^+\ 2\ Eg \frac{2n_2 + 2}{n_{1;v} + 1} \qquad (54) \\
= & \mathrm{E}\quad fx\ 2\ Q_v\ n\ Z; n_{1;v} > 0; n_2 > 0g \frac{2n_2}{n_{1;v} + 1} \\
& + \mathrm{E}\quad fx\ 2\ Q_v\ n\ Z; n_{1;v} > 0; n_2 > 0; x^+\ 2\ Bg \frac{2}{n_{1;v} + 1} \\
= & \mathrm{E}\quad fx\ 2\ Q_v\ n\ Z; n_{1;v} > 0; n_2 > 0g \frac{2n_2}{n_{1;v} + 1} \\
& + \mathrm{E}\quad fx\ 2\ Q_v\ n\ Z; n_{1;v} > 0; n_2 > 0; x^+\ 2\ Eg \frac{2}{n_{1;v} + 1} \qquad (55)
\end{aligned}$$

where (54) follows using the inequality $\log(1 + x)$ x. Thus we are left with two terms in (55). For the first term we have (ignoring notation overload, as after the first line, $n_{1;v}$ and n_2

change from random variables to dummy variables):

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{1}\{x \in Q_v \cap Z\} \mid n_{1,v} > 0, n_2 > 0, Q_v \setminus E = \emptyset\right] \frac{2n_2}{n_{1,v}+1} \\
&= \mathbb{P}(x \in Q_v \cap Z) \sum_{n_{1,v}, n_2 : n_{1,v}+n_2 = g} \mathbb{1}\{n_{1,v} > 0, n_2 > 0\} \binom{g}{n_{1,v}, n_2, g - n_{1,v} - n_2} \\
&\qquad |Q_v \cap Z|^{n_{1,v}} |E|^{n_2} (1 - |Q_v \cap Z| - |E|)^{g - n_{1,v} - n_2} \frac{2n_2}{n_{1,v}+1} \qquad (56) \\
&= 2|Q_v \cap Z| \sum_{n_{1,v}, n_2 : n_{1,v}, n_2 > 0, n_{1,v}+n_2 = g} \binom{g-1}{n_{1,v}, n_2 - 1, g - n_{1,v} - n_2} \\
&\qquad |Q_v \cap Z|^{n_{1,v}} |E|^{n_2} (1 - |Q_v \cap Z| - |E|)^{g - n_{1,v} - n_2} \frac{n_2}{n_{1,v}+1} \qquad (57) \\
&= 2|Q_v \cap Z| \frac{|E|}{|Q_v \cap Z|} \sum_{n_{1,v}, n_2 : n_{1,v}, n_2 > 0, n_{1,v}+n_2 = g} \binom{g-1}{n_{1,v}+1, n_2 - 1, g - n_{1,v} - n_2} \\
&\qquad |Q_v \cap Z|^{n_{1,v}+1} |E|^{n_2 - 1} (1 - |Q_v \cap Z| - |E|)^{g - n_{1,v} - n_2} \qquad (58) \\
&\quad 2|Q_v \cap Z| \frac{|E|}{|Q_v \cap Z|} \\
&\qquad \sum_{n_{1,v}+1, n_2 - 1 : n_{1,v}+1, n_2 - 1 > 0, n_{1,v}+1+n_2 - 1 = g} \binom{g-1}{n_{1,v}+1, n_2 - 1, g - (n_{1,v}+1) - (n_2 - 1)} \\
&\qquad |Q_v \cap Z|^{n_{1,v}+1} |E|^{n_2 - 1} (1 - |Q_v \cap Z| - |E|)^{(n_{1,v}+1) - (n_2 - 1)} \qquad (59) \\
&= 2|E| \left(|Q_v \cap Z| + |E| + 1 - |Q_v \cap Z| - |E|\right)^{g-1} \\
&= 2|E|
\end{aligned}$$

where in (59) we have added terms to the sum to complete the trinomial expansion, and the last equality follows since Q_v and E are disjoint.

Now we need to consider the last term in (55), which corresponds to the case wherein the positive inner products are not equal for g and g'. For this term, we simply have

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{1}\{x \in Q_v \cap Z, n_{1,v} > 0, n_2 > 0, x^+ \in E\} \frac{2}{n_{1,v}+1}\right] \\
&\quad \mathbb{P}\left[x \in Q_v \cap Z, x^+ \in E\right] \qquad (60)
\end{aligned}$$

In total, for the case $n_{1,v} > 0, n_2 > 0$ and $x \in Q_v \cap Z$, we have

$$((3)) \quad 2|E| + \mathbb{P}\left[x \in Q_v \cap Z, x^+ \in E\right] \qquad (61)$$

4. $x \in E_v, n_4 = 0$.

This case is symmetric to Case 1, so we argue similarly.

$$((4)) := \mathbb{E}_{x, x^+, \{x_i\}_{i=1}^{g-1}}\left[\mathbb{1}\{x \in E_v\} \mathbb{1}\{n_4 = 0\}\right]$$

$$\log\left(e^{g^0(x)^\top g^0(x^+)} + n_{3,v}e^{d} + \sum_{x_i \in Q} e^{g^0(x)^\top g^0(x_i)} + \sum_{x_i \in E_v \setminus Q} e^{g(x)^\top g(x_i)}\right)$$

$$\log\left(e^{g(x)^\top g(x^+)} + n_{3,v}e^{d} + \sum_{x_i \in Q} e^{g(x)^\top g(x_i)} + \sum_{x_i \in E_v \setminus Q} e^{g(x)^\top g(x_i)}\right)$$

$$= \mathbb{E}_{x,x^+,\{x_i\}}\left[\{x \in E_v\} \cap n_4 = 0\right]$$

$$\log\left(e^{g^0(x)^\top g^0(x^+)} + n_{3,v}e^{d} + \sum_{x_i \in E_v \setminus Q} e^{g(x)^\top g(x_i)}\right)$$

$$\log\left(e^{g(x)^\top g(x^+)} + n_{3,v}e^{d} + \sum_{x_i \in E_v \setminus Q} e^{g(x)^\top g(x_i)}\right)$$

$$\mathbb{E}_{x,x^+,\{x_i\}}\left[\{x \in E_v\} \cap n_4 = 0\right] \log \frac{e^{g^0(x)^\top g^0(x^+)} + n_{3,v}e^{d}}{e^{g(x)^\top g(x^+)} + n_{3,v}e^{d}} \tag{62}$$

where (62) follows since $g^0(x)^\top g^0(x^+) \geq g(x)^\top g(x^+)$ and $\log()$ is submodular. Next we intersect with the events $\{x^+ \in Q\}$ and $\{x^+ \notin Q\}$, obtaining

$$\begin{aligned}
((4)) \quad & \mathbb{E}_{x,x^+,\{x_i\}}\left[\{x \in E_v, n_4 = 0, x^+ \in Q\}\right] \log \frac{e^{g^0(x)^\top g^0(x^+)} + n_{3,v}e^{d}}{e^{g(x)^\top g(x^+)} + n_{3,v}e^{d}} \\
& + \mathbb{E}_{x,x^+,\{x_i\}}\left[\{x \in E_v, n_4 = 0, x^+ \notin Q\}\right] \log \frac{e^{g^0(x)^\top g^0(x^+)} + n_{3,v}e^{d}}{e^{g(x)^\top g(x^+)} + n_{3,v}e^{d}} \\
= & \mathbb{E}_{x,x^+,\{x_i\}}\left[\{x \in E_v, n_4 = 0, x^+ \in Q\}\right] \log \frac{e^2\, e^{g(x)^\top g(x^+)} + n_{3,v}e^{d}}{e^{g(x)^\top g(x^+)} + n_{3,v}e^{d}} \\
& \mathbb{E}_{x,x^+,\{x_i\}}\left[\{x \in E_v, n_4 = 0, x^+ \in Q\}\right] \log \frac{e^{d} + n_{3,v}e^{d}}{e^{(d-2)} + n_{3,v}e^{d}} \\
= & \mathbb{E}_{x,x^+,\{x_i\}}\left[\{x \in E_v, n_4 = 0, x^+ \in Q\}\right] \log \frac{1 + n_{3,v}}{e^{-2} + n_{3,v}} \\
& 2\ \mathbb{E}_{x,x^+,\{x_i\}}\left[\{x \in E_v, n_4 = 0, x^+ \in Q, n_{3,v} = 0\}\right] \\
& + \log(2)\mathbb{E}_{x,x^+,\{x_i\}}\left[\{x \in E_v, n_4 = 0, x^+ \in Q, n_{3,v} > 0\}\right]
\end{aligned} \tag{63}$$

where

$$\begin{aligned}
& \mathbb{E}_{x,x^+,\{x_i\}}\left[\{x \in E_v, n_4 = 0, x^+ \in Q, n_{3,v} = 0\}\right] \\
& = \mathbb{P}(x \in E_v \setminus x^+ \in Q)\mathbb{P}(n_{3,v} = 0, n_4 = 0) && (64) \\
& = \mathbb{P}(x \in E_v \setminus x^+ \in Q)(1 - \|E_v\| - \|Q\|)^{\top} && (65)
\end{aligned}$$

$$\begin{aligned}
& \mathbb{E}_{x,x^+,\{x_i\}}\left[\{x \in E_v, n_4 = 0, x^+ \in Q, n_{3,v} > 0\}\right] \\
& = \mathbb{P}(x \in E_v \setminus x^+ \in Q)\mathbb{P}(n_{3,v} > 0, n_4 = 0) \\
& \mathbb{P}(x \in E_v \setminus x^+ \in Q)\min\left(1, \frac{(\|E_v\|)^{\top}}{1 - \|Q\|}\right)(1 - \|Q\|)^{\top} && (66)
\end{aligned}$$

so in total for this case we have

$$((4)) \quad P(x \in E_v \mid x^+ \in Q) \leq 2\,(1 - \|E_v\| - \|Q\|)^{\ell} + \log(2)\min\left(1, \frac{\ell \|E_v\|}{1 - \|Q\|}\right)(1 - \|Q\|)^{\ell}$$

5. $x \in E_v, n_4 > 0, n_{3,v} = 0$.
Define $n_5 := \sum_{i=1}^{\ell} \{x_i \in Q \cap Z\}$. We have

$$((5)) = \mathbb{E}_{x,x^+,\{x_i\}_{\ell}} \left[\{x \in E_v, n_4 > 0, n_{3,v} = 0\} \log \frac{e^{g'(x)^\top g'(x^+)} + \sum_{x_i \in Q} e^{g'(x)^\top g'(x_i)} + \sum_{x_i \notin E_v \cup Q} e^{g(x)^\top g(x_i)}}{e^{g(x)^\top g(x^+)} + \sum_{x_i \in Q} e^{g(x)^\top g(x_i)} + \sum_{x_i \notin E_v \cup Q} e^{g(x)^\top g(x_i)}} \right]$$

since we are intersecting with the event $\{n_{3,v} = 0\}$. Next we split the negative samples in Q into those in $Q \cap Z$ and those in Z, noting that $g'(x)^\top g'(x_i) = g(x)^\top g(x_i) - 2$ for $x_i \in Z$ and $g'(x)^\top g'(x_i) = g(x)^\top g(x_i) + 2$ for $x_i \in Q \cap Z$.

$$((5)) = \mathbb{E}_{x,x^+,\{x_i\}_{\ell}} \left[\{x \in E_v, n_4 > 0, n_{3,v} = 0\} \log \frac{e^{g'(x)^\top g'(x^+)} + e^{2} \sum_{x_i \in Q \cap Z} e^{g(x)^\top g(x_i)} + e^{-2} \sum_{x_i \in C} e^{g(x)^\top g(x_i)} + \sum_{x_i \notin E_v \cup Q} e^{g(x)^\top g(x_i)}}{e^{g(x)^\top g(x^+)} + \sum_{x_i \in Q \cap Z} e^{g(x)^\top g(x_i)} + \sum_{x_i \in C} e^{g(x)^\top g(x_i)} + \sum_{x_i \notin E_v \cup Q} e^{g(x)^\top g(x_i)}} \right]$$

$$\mathbb{E}_{x,x^+,\{x_i\}_{\ell}} \left[\{x \in E_v, n_4 > 0, n_{3,v} = 0\} \log \frac{e^{g'(x)^\top g'(x^+)} + e^{2} \sum_{x_i \in Q \cap Z} e^{g(x)^\top g(x_i)} + \sum_{x_i \notin E_v \cup Q} e^{g(x)^\top g(x_i)}}{e^{g(x)^\top g(x^+)} + \sum_{x_i \in Q \cap Z} e^{g(x)^\top g(x_i)} + \sum_{x_i \notin E_v \cup Q} e^{g(x)^\top g(x_i)}} \right]$$

$$\mathbb{E}_{x,x^+,\{x_i\}_{\ell}} \left[\{x \in E_v, n_4 > 0, n_{3,v} = 0\} \right]$$

$$\log\left(e^{g^0(x)^\top g^0(x^+)} + n_5 e^{d} + \sum_{x_i \in E_v[Q} e^{g(x)^\top g(x_i)}\right)$$
$$\log\left(e^{g(x)^\top g(x^+)} + n_5 e^{(d-2)} + \sum_{x_i \in E_v[Q} e^{g(x)^\top g(x_i)}\right) \tag{67}$$

where (67) follows since $h(x) := \frac{a+e^2 x}{b+x}$ is an increasing function of x for $a \geq e^2 b$. Next we intersect with $\{x^+ \in Q\}$ and $\{x^+ \notin Q\}$ to obtain

$$\begin{aligned}
((5)) \quad & \mathbb{E}_{\{x,x^+,\{x_i\}\}}\left[\mathbb{1}\{x \in E_v, n_4 > 0, n_{3,v} = 0, x^+ \in Q\}\right. \\
& \log\left((n_5+1)e^{d} + \sum_{x_i \in E_v[Q} e^{g(x)^\top g(x_i)}\right) \\
& \left.\log\left((n_5+1)e^{(d-2)} + \sum_{x_i \in E_v[Q} e^{g(x)^\top g(x_i)}\right)\right] \\
+ & \mathbb{E}_{\{x,x^+,\{x_i\}\}}\left[\mathbb{1}\{x \in E_v, n_4 > 0, n_{3,v} = 0, x^+ \notin Q\}\right. \\
& \log\left(n_5 e^{d} + \sum_{x_i \in E_v[Q} e^{g(x)^\top g(x_i)}\right) \\
& \left.\log\left(n_5 e^{(d-2)} + \sum_{x_i \in E_v[Q} e^{g(x)^\top g(x_i)}\right)\right]
\end{aligned} \tag{68}$$

We have two terms above. For the first term,

$$\begin{aligned}
& \mathbb{E}_{\{x,x^+,\{x_i\}\}}\left[\mathbb{1}\{x \in E_v, n_4 > 0, n_{3,v} = 0, x^+ \in Q\}\right. \\
& \log\left((n_5+1)e^{d} + \sum_{x_i \in E_v[Q} e^{g(x)^\top g(x_i)}\right) \\
& \left.\log\left((n_5+1)e^{(d-2)} + \sum_{x_i \in E_v[Q} e^{g(x)^\top g(x_i)}\right)\right] \\
& \mathbb{E}_{\{x,x^+,\{x_i\}\}}\left[\mathbb{1}\{x \in E_v, n_4 > 0, n_{3,v} = 0, x^+ \in Q\} \log\frac{(n_5+1)e^{d}}{(n_5+1)e^{(d-2)}}\right] \\
& 2\,\mathbb{E}_{\{x,x^+,\{x_i\}\}}\left[\mathbb{1}\{x \in E_v, n_4 > 0, n_{3,v} = 0, x^+ \in Q\}\right]
\end{aligned} \tag{69}$$

Similarly, for the second term in (68), we have

$$\begin{aligned}
& \mathbb{E}_{\{x,x^+,\{x_i\}\}}\left[\mathbb{1}\{x \in E_v, n_4 > 0, n_{3,v} = 0, x^+ \notin Q\}\right. \\
& \left.\log\left(n_5 e^{d} + \sum_{x_i \in E_v[Q} e^{g(x)^\top g(x_i)}\right) \quad \log\left(n_5 e^{(d-2)} + \sum_{x_i \in E_v[Q} e^{g(x)^\top g(x_i)}\right)\right]
\end{aligned}$$

$$\mathbb{E}_{x,x^+,\{x_i\}}\left[\, x \in \mathcal{E}_v, n_5 > 0, n_{3,v} = 0, x^+ \notin Q \right] \log \frac{n_5 e^{d}}{n_5 e^{(d-2)}}$$
$$= 2\ \mathbb{E}_{x,x^+,\{x_i\}}\left[\, x \in \mathcal{E}_v, n_5 > 0, n_{3,v} = 0, x^+ \notin Q \right]$$

By summing the upper bounds on the two terms in (68), we obtain

$$\begin{aligned}((5)) \quad & 2\ \mathbb{E}_{x,x^+,\{x_i\}}\left[\, x \in \mathcal{E}_v, n_4 > 0, n_{3,v} = 0, x^+ \in Q \right] \\ & + 2\ \mathbb{E}_{x,x^+,\{x_i\}}\left[\, x \in \mathcal{E}_v, n_5 > 0, n_{3,v} = 0, x^+ \notin Q \right] \\ & 2\ \mathbb{P}(x \in \mathcal{E}_v)\mathbb{P}(n_4 > 0, n_{3,v} = 0) \\ & = 2\ \|\mathcal{E}_v\|\ (1 - \|\mathcal{E}_v\|)\ \min\left(1, \frac{1-\|Q\|}{1-\|\mathcal{E}_v\|}\right). \end{aligned} \tag{70}$$

6. $x \in \mathcal{E}_v, n_4 > 0, n_{3,v} > 0$.

In this case, we argue similarly as in Case 3 to obtain

$$\begin{aligned}((6)) &= \mathbb{E}_{x,x^+,\{x_i\}}\left[\, x \in \mathcal{E}_v, n_4 > 0, n_{3,v} > 0 \right] \\ & \log\left(e^{g'(x)^\top g'(x^+)} + n_{3,v} e^{d} + \sum_{x_i \in Q} e^{g'(x)^\top g'(x_i)} + \sum_{x_i \notin \mathcal{E}_v \cup Q} e^{g(x)^\top g(x_i)} \right) \\ & \log\left(e^{g(x)^\top g(x^+)} + n_{3,v} e^{d} + \sum_{x_i \in Q} e^{g(x)^\top g(x_i)} + \sum_{x_i \notin \mathcal{E}_v \cup Q} e^{g(x)^\top g(x_i)} \right) \\ & \mathbb{E}_{x,x^+,\{x_i\}}\left[\, x \in \mathcal{E}_v, n_4 > 0, n_{3,v} > 0 \right] \\ & \log\left(e^{g'(x)^\top g'(x^+)} + n_{3,v} e^{d} + e^{2} \sum_{x_i \in Q \cap Z} e^{g(x)^\top g(x_i)} + \sum_{x_i \notin \mathcal{E}_v \cup Q} e^{g(x)^\top g(x_i)} \right) \\ & \log\left(e^{g(x)^\top g(x^+)} + n_{3,v} e^{d} + \sum_{x_i \in Q \cap Z} e^{g(x)^\top g(x_i)} + \sum_{x_i \notin \mathcal{E}_v \cup Q} e^{g(x)^\top g(x_i)} \right) \\ & \mathbb{E}_{x,x^+,\{x_i\}}\left[\, x \in \mathcal{E}_v, n_4 > 0, n_{3,v} > 0 \right] \\ & \log\left(e^{g'(x)^\top g'(x^+)} + (n_{3,v} + n_5) e^{d} \right) \\ & \log\left(e^{g(x)^\top g(x^+)} + n_{3,v} e^{d} + n_5 e^{(d-2)} \right) \end{aligned} \tag{71}$$

where (71) follows by the analogous argument as in (52). Next we intersect with $x^+ \in Q$ and $x^+ \notin Q$. We have

$$\begin{aligned}((6)) \quad & \mathbb{E}_{x,x^+,\{x_i\}}\left[\, x \in \mathcal{E}_v, n_4 > 0, n_{3,v} > 0 \right] \\ & \log\left(e^{g'(x)^\top g'(x^+)} + (n_{3,v} + n_5) e^{d} \right) \end{aligned}$$

$$\log\ e^{g(x)^{\top}g(x^+)} + n_{3;v}e^{\ d} + n_5 e^{\ (d\ \ 2)}$$

$$\begin{aligned}
&\mathrm{E}_{x;x^+;fx_i\ g^{\backprime}}\ \ fx\ 2\ E_v; n_4 > 0; n_{3;v} > 0; x^+\ 2\ Qg \log\ \frac{n_{3;v}+n_5+1}{n_{3;v}}\\
&+\ \mathrm{E}_{x;x^+;fx_i\ g^{\backprime}}\ \ fx\ 2\ E_v; n_4 > 0; n_{3;v} > 0; x^+\ \not{2}\ Qg \log\ \frac{n_{3;v}+n_5}{n_{3;v}} \qquad (72)\\
&\mathrm{E}_{x;x^+;fx_i\ g^{\backprime}}\ \ fx\ 2\ E_v; n_{3;v} > 0; x^+\ 2\ Qg \log(n_5+2)\\
&+\ \mathrm{E}_{x;x^+;fx_i\ g^{\backprime}}\ \ fx\ 2\ E_v; n_{3;v} > 0; x^+\ \not{2}\ Qg \log(n_5+1)\\
&=\ \mathrm{P}(x\ 2\ E_v; x^+\ 2\ Q)\mathrm{E}_{x;x^+;fx_i\ g^{\backprime}}\ \ fn_{3;v} > 0g \log(n_5+2)\\
&+\ \mathrm{P}(x\ 2\ E_v; x^+\ \not{2}\ Q)\mathrm{E}_{x;x^+;fx_i\ g^{\backprime}}\ \ fn_{3;v} > 0g \log(n_5+1)\\
&\mathrm{P}(x\ 2\ E_v; x^+\ 2\ Q)\log(\mathrm{E}[n_5+2]) + \mathrm{P}(x\ 2\ E_v; x^+\ \not{2}\ Q)\log(\mathrm{E}[n_5+1]) \qquad (73)\\
&\mathrm{P}(x\ 2\ E_v; x^+\ 2\ Q)\log(\grave{}kQ\ n\ Zk\ +2) + kE_vk\ \log(\grave{}kQ\ n\ Zk\ +1)
\end{aligned}$$

where (72) follows since if $x\ 2\ E_v$, then $x^+\ 2\ Q\ \ (\ )\ \ g^{\emptyset}(x)^{\top}g^{\emptyset}(x^+) = g(x)^{\top}g(x^+) + 2$ and by the submodularity of $\log()$, and (73) follows by Jensen's Inequality and upper bounding $fn_{3;v} > 0g\ \ 1$.

Now we combine all six cases and sum over $v\ 2\ H_d$. We obtain

$$\begin{aligned}
&L\ \ (g^{\emptyset})\ \ \ L\ \ (g)\\
&\sum_{v2H_d}\ \mathrm{P}(x\ 2\ Q_v\ n\ Z\ \backslash\ x^+\ 2\ E)\\
&\qquad 2\ \ (1\ \ \ kQ_v\ n\ Zk\ \ \ kEk\ )\grave{}\ + \log(2)\min\ \ 1; \tfrac{\grave{}kQ_v nZk}{1\ \ kEk}\ \ (1\ \ \ kEk\ )\grave{}\\
&+\ 2\ \ kQ_v\ n\ Zk\ \ (1\ \ \ kQ_v\ n\ Zk\ )\grave{}\ \min\ \ 1; \tfrac{\grave{}kEk}{1\ \ kQ_v nZk}\\
&+\ 2kEk\ \ +\ \mathrm{P}\ \ x\ 2\ Q_v\ n\ Z; x^+\ 2\ E\\
&+\ \mathrm{P}(x\ 2\ E_v\ \backslash\ x^+\ 2\ Q)\ \ 2\ \ (1\ \ \ kE_vk\ \ \ kQ\ n\ Zk\ )\grave{}\\
&\qquad\qquad + \log(2)\min\ \ 1; \tfrac{\grave{}kE_vk}{1\ \ kQk}\ \ \ (1\ \ \ kQ\ n\ Zk\ )\grave{}\\
&+\ 2\ \ kE_vk\ \ (1\ \ \ kE_vk\ )\grave{}\ \min\ \ 1; \tfrac{\grave{}kQk}{1\ \ kE_vk}\\
&+\ \mathrm{P}(x\ 2\ E_v; x^+\ 2\ Q)\log(\grave{}kQ\ n\ Zk\ +2) + kE_vk\ \log(\grave{}kQ\ n\ Zk\ +1) \qquad (74)\\
&kRk\log(\grave{}\ +2) + kEk\ \ (2^{d+1} + \log(\grave{}\ +1))\\
&+\ \sum_{v2H_d}\ \mathrm{P}(x\ 2\ Q_v\ \backslash\ x^+\ 2\ E)\ \ 2\ \ (1\ \ \ kQ_vk\ \ \ kEk\ )\grave{}\ + \log(2)(1\ \ \ kEk\ )\grave{}\\
&+\ 2\ \ \grave{}kEk\ kQ_v\ n\ Zk\ \ (1\ \ \ kQ_v\ n\ Zk\ )\grave{}\ \ \ 1\\
&+\ \mathrm{P}(x\ 2\ E_v\ \backslash\ x^+\ 2\ Q)\ \ 2\ \ (1\ \ \ kE_vk\ \ \ kQk\ )\grave{}\ + \log(2)(1\ \ \ kQk\ )\grave{}\\
&+\ 2\ \ kE_vk\ \ (1\ \ \ kE_vk\ )\grave{} \qquad (75)
\end{aligned}$$

$$
\begin{aligned}
&kRk(\log(\grave{} + 2) + \log(2)) + kEk\ (2^{d+1} + \log(\grave{} + 1)) \\
&+ 2 \sum_{v2H_d} P(x\ 2\ Q_v \setminus x^+\ 2\ E)(kQk \quad kQ_vk\ )\grave{} + P(x\ 2\ E_v \setminus x^+\ 2\ Q)(kEk \quad kE_vk\ )\grave{} \\
&+ \grave{}kEk\ kQ_v\ n\ Zk\ (1 \quad kQ_v\ n\ Zk\ )\grave{}\ {}^1 + kE_vk\ (1 \quad kE_vk\ )\grave{} \\
&kRk(\log(\grave{} + 2) + \log(2)) + kEk\ (2^{d+1} + \log(\grave{} + 1)) + 2\ {}^{(\grave{}\ 1)}\ kRk \\
&+ 2 \sum_{v2H_d} P(x\ 2\ Q_v \setminus x^+\ 2\ E)(kQk \quad kQ_vk\ )\grave{} \\
&+ \grave{}kEk\ kQ_v\ n\ Zk\ (1 \quad kQ_v\ n\ Zk\ )\grave{}\ {}^1 + kE_vk\ (1 \quad kE_vk\ )\grave{} \qquad (76)
\end{aligned}
$$

where (75) follows since $\min(x; y) \quad x$, $\sum_{v2H_d} kE_vk = kEk$, and $kRk = \sum_{v2H_d} P(x\ 2\ Q_v \setminus x^+\ 2\ E_v) + P(x\ 2\ E_v \setminus x^+\ 2\ Q)$, and (76) follows since $kEk \quad \frac{1}{2}$ by construction of $g^{\emptyset}$ (since for $f_1^{\emptyset} = f_1^{(\mathbf{c};\ )}$, is chosen such that the induced kEk cannot be larger than $\frac{1}{2}$). It remains to bound the three terms in the sum in (76).

To do so, we first define $\bar{E}_v := fx^+\ 2\ B : \bar{A}\ {}^1(x^+)\ 2\ Q_vg$ as the set of partial augmentation sets that are in B, corresponding to sets whose natural image is in Q_v (where $\bar{A}\ {}^1(x^+)$ is the natural image from which the augmented image x^+ was generated, i.e. $\bar{A}\ {}^1(x^+) = x \quad (\ ) \quad A(x) = x^+$ for some $A\ 2 \quad$). We have

$$
\sum_{v2H_d} P(x\ 2\ Q_v \setminus x^+\ 2\ E)(kQk \quad kQ_vk\ )\grave{} = \sum_{v2H_d} k\bar{E}_vk(kQk \quad kQ_vk\ )\grave{} \qquad (77)
$$

$$
\sum_{v2H_d} kE_vk(1 \quad kQ_vk\ )\grave{} \sum_{v2H_d} k\bar{E}_vk \quad 1 \quad k\bar{E}_vk\ \grave{} \qquad (78)
$$

where (77) and (78) follow since all augmentation sets are equal size.

Note $\sum_v k\bar{E}_vk = kRk$, and $h(x_v) := x_v(1 \quad x_v)\grave{}$ is maximized on $x_v\ 2\ [0; 1]$ at $x_v = \frac{1}{\grave{}+1}$. Thus, $\sum_v kE_vk \quad 1 \quad k\bar{E}_vk\ \grave{}$ is upper bounded by setting $k\bar{E}_vk = \frac{1}{\grave{}+1}$ for all v. Thus we obtain

$$
\begin{aligned}
\sum_{v2H_d} P(x\ 2\ Q_v \setminus x^+\ 2\ B)(kQk \quad kQ_vk\ )\grave{} &\sum_{v2H_d} \frac{1}{\grave{}+1} \quad 1 \quad \frac{1}{\grave{}+1}\ \grave{} \\
&= \frac{2^d}{\grave{}+1} \quad 1 \quad \frac{1}{\grave{}+1}\ \grave{} \\
&\overline{cd}e^{\frac{\grave{}}{\grave{}+1}}
\end{aligned}
$$

where we have used $\grave{} \quad cd2^d$ for a constant c in the last line. Next we consider $\sum_{p2H_d} kE_vk\ (1 \quad kE_vk\ )\grave{}$, and use a tighter method of bounding this sum than above. Note that $\sum_{v2H_d} kE_vk = kEk$. If $\frac{kEk}{2^d} \quad \frac{1}{\grave{}+1}$, then by the concavity of $h(x_v) := x_v(1 \quad x_v)\grave{}$ on the interval $x_v\ 2\ [0; \frac{1}{\grave{}+1}]$,

the sum is maximized by setting $kE_vk = \frac{kEk}{2^d}$ for all v. Otherwise, the sum is upper bounded by setting $kE_vk = \frac{1}{\ell+1}$ for all v. Thus we have

$$\begin{aligned}
\sum_{v \in H_d} kE_vk\,(1 - kE_vk)^\ell &\le kEk \cdot \frac{2^d}{\ell+1} \cdot kEk \cdot 1 - \frac{kEk}{2^d}\Big)^\ell \\
&+ \Big(1 - kEk\Big) \cdot \frac{2^d}{\ell+1} \cdot \frac{2^d}{\ell+1} \cdot 1 - \frac{1}{\ell+1}\Big)^\ell \\
&\le kEk \cdot \frac{2^d}{\ell+1} \cdot kEk \cdot e^{-\frac{kEk\,\ell}{2^d}} \\
&+ \Big(1 - kEk\Big) \cdot \frac{2^d}{\ell+1} \cdot kEk \cdot e^{-\frac{\ell}{\ell+1}}
\end{aligned} \tag{79}$$

Finally, note that $D_g(v) := \mathsf{P}_{x \sim D}[g(x) = v] = kQ_vk + kE_vk$. We have that for all v, $Q_v \cap Z \neq \emptyset \implies Q_v \cap Z = Q_v; E_v = \emptyset$ by Claims B.4 and B.6. Thus for all $v : Q_v \cap Z \neq \emptyset$, $D_g(v) = kQ_vk = kQ_v \cap Zk$. This allows us to use that g is near uniform, i.e. $D_g(v) > \frac{1}{c_1 d 2^d}$ or $D_g(v) < \frac{1}{c_2 d 2^{2d}}$ for all $v \in H_d$ for some constants c_1, c_2. We have $\ell \ge c d 2^d$ and choose $c_1 < c$, such that $\frac{1}{c_1 d 2^d} > \frac{1}{\ell+1}$. Since $h(x_v) := x_v(1 - x_v)^\ell$ is a decreasing function of x_v for $x_v \ge \frac{1}{\ell+1}$, we can bound the last sum in (76) as

$$\begin{aligned}
\ell \sum_{v \in H_d} kQ_v \cap Zk\,(1 - kQ_v \cap Zk)^\ell &\le 1 + \ell 2^d \max\Big(\frac{1}{c_2 d 2^{2d}}, \frac{1}{c_1 d 2^d}\Big(1 - \frac{1}{c_1 d 2^d}\Big)^\ell\Big) \\
&\le \max\Big(\frac{1}{c_2 d 2^d}, \frac{1}{c_1 d} e^{-\frac{\ell}{c_1 d 2^d}}\Big) \\
&\le \max\Big(\frac{c}{c_2}, \frac{c}{c_1} e^{-(\frac{c}{c_1} - 1)}\Big)
\end{aligned} \tag{80}$$

Before we combining these bounds with (76), we first show that $kRk = \Omega(\ell)$.

Claim B.7 *Let* $\min_{j \in [d]} \min_{\mathbf{c}^0 \in \{-1,+1\}^{\ell_j}} \mathsf{P}_{x,x^0 \sim D}[x, x^0 \in \mathbf{c}^0; f_j(x) \neq f_j(x^0)]$ *as defined in the statement of Theorem 15. Then* $\Omega(\ell) kRk$.

Proof From our choice of f_1^0, we have

$$\begin{aligned}
&\min_{j \in [d]} \min_{\mathbf{c}^0 \in \{-1,+1\}^{\ell_j}} \mathsf{P}_{x,x^0 \sim D}[x, x^0 \in \mathbf{c}^0; f_j(x) \neq f_j(x^0)] \\
&\quad 2\mathsf{P}_{x,x^0 \sim D}[x \in \mathbf{c} \cap E; x^0 \in E] \\
&= 2\mathsf{P}_{x \sim D}[x \in \mathbf{c} \cap E]\mathsf{P}_{x \sim D}[x \in E] \\
&\quad 2\mathsf{P}_{x \sim D}[x \in E] \\
&\quad 2(\mathsf{P}_{x \sim D}[x \in E] + \mathsf{P}_{x^+ \sim DnD}[x^+ \in E]) \\
&= 2(kEk + kEk).
\end{aligned} \tag{81}$$

Note that $kEk \ge \frac{1}{2} kRk$ by Assumption 13. Observe that

$$kEk = \mathsf{P}_{x^+ \sim DnD}[x^+ \in E]$$

$$= \mathsf{P}_{x^+ \quad DnD}\,[x^+ \;2\; E; A^{\;1}(x^+) \;2\; E] + \mathsf{P}_{x^+ \quad DnD}\,[x^+ \;2\; E; A^{\;1}(x^+) \;\not\in\; E]$$
$$\mathsf{P}_{x \;\; D \;; x^+ \;\; A(x)}[x \;2\; E; x^+ \;2\; E] + kRk$$
$$kEk \;\; + kRk \tag{82}$$

Note that $kEk \quad \frac{1}{-}kRk$ by Assumption 13. Combining this with (81) and (82) yields

$$2(2kEk \;\; + kRk) \quad \frac{6}{-}kRk$$

■

Finally, using Claim B.7 with $\quad \frac{c^0}{-}\log \;\; \frac{c}{-} \;\; 2^d$, we obtain from (76) that

$$\begin{aligned}
&L \;\; (g^0) \quad L \;\; (g) \\
&\quad kRk(\log(` + 2) + \log(2)) + \frac{kRk}{\quad}(2^{d+1} + \log(` + 1)) + 2^{\;(` \;\; 1)} \;\; kRk \\
&\quad + \frac{2}{cd}e^{\;\frac{`}{` + 1}} + 2 \;\; \max\;\; \frac{c}{c_2}; \frac{c}{c_1}e^{\;(\frac{c}{c_1} \;\; 1)} \\
&\quad + 2 \quad kEk \quad \frac{2^d}{` + 1} \;\; kEk\; e^{\;\frac{kEk\;`}{2^d}} + 2 \quad kEk \;\; > \frac{2^d}{` + 1} \;\; kEk\; e^{\;\frac{`}{` + 1}} \\
&\quad\quad kRk \;\; \frac{\log(2cd2^{d=} \;\; + 4) + \frac{1}{-}\log(cd2^{d=} \;\; + 1)}{c^0 \log(\frac{c}{-})2^{d=}} + \frac{2}{c^0 \log(\frac{c}{-})2^d} + \frac{2}{2^{cd2^{d=}}} \\
&\quad\quad\quad + \frac{12}{cd}e^{\;\frac{`}{` + 1}} + 2\max\;\; \frac{c}{c_2}; \frac{c}{c_1}e^{\;(\frac{c}{c_1} \;\; 1)} \\
&\quad + 2 \;\; kEk \quad \overset{\mathsf{n}}{kEk} \quad \overset{\mathsf{o}}{\frac{}{cd}}\, e^{\;\frac{kEk\;`}{2^d}} + \quad kEk \;\; > \frac{2^d}{` + 1} \quad e^{\;\frac{`}{` + 1}} \\
&\quad\quad kRk \;\; \frac{\log(4c^{=} \;\; + 4) + \frac{1}{-}\log(2c^{=} \;\; + 1)}{2c^0 \log(\frac{c}{-})^{=}} + \frac{1}{c^0 \log(\frac{c}{-})} + \frac{1}{2^{c^{=}}} \\
&\quad\quad\quad + \frac{12}{cd}e^{\;\frac{`}{` + 1}} + 2\max\;\; \frac{c}{c_2}; \frac{c}{c_1}e^{\;(\frac{c}{c_1} \;\; 1)} \\
&\quad + \quad kRk \;\; \frac{12}{cd} \quad kEk \quad \frac{6kRk}{cd} \quad + \frac{1}{-} \quad kEk \;\; > \frac{2^d}{` + 1} \quad e^{\;\frac{`}{` + 1}} \\
&\quad 2 \;\; kRk \;\; \frac{2c}{c_2} + \frac{1}{100} \quad + 2 \;\; kRk\frac{1{:}01}{e} \qquad (83) \\
&< 2 \;\; kRk = L^+(g) \quad L^+(g^0) \qquad (84)
\end{aligned}$$

where (83) follows for a sufficiently large constants c^0 and $c > c_1$. Use $\quad 0{:}4$ and set $c^0 = c$, $c_1 = \frac{c}{10}$, $c_2 = 100c$ and make c sufficiently large to obtain

$$L \;\; (g^0) \quad L \;\; (g) < 2 \;\; kRk = L^+(g) \quad L^+(g^0){:}$$

■

B.1. Remark: Modified version of Theorem 15

Theorem 15 shows that if a minimizer of the InfoNCE loss is close to uniform, then it must be clean. Here we show that the InfoNCE loss can be interpreted as the Lagrangian which, under appropriate choice of hyperparameters, we can formally show to be minimized only by clean representations.

Weighted InfoNCE loss. Consider the following constrained optimization problem that tries to maximize uniformity while preserving alignment:

$$\min_{g \in G} \mathsf{E}_{x,x^+,\{x_i\}_{i=1}^{\ell}} \log \left(e^{g(x)^\top g(x^+)} + \sum_{i=1}^{\ell} e^{g(x)^\top g(x_i)} \right)$$
$$\text{s.t.} \quad \mathsf{E}_{x,x^+}[g(x)^\top g(x^+)] = d \tag{85}$$

The unconstrained penalized version of this problem is

$$\min_{g \in G} \; -\mathsf{E}_{x,x^+}\left[g(x)^\top g(x^+) \right] + \mathsf{E}_{x,x^+,\{x_i\}_{i=1}^{\ell}} \left[\log \left(e^{g(x)^\top g(x^+)} + \sum_{i=1}^{\ell} e^{g(x)^\top g(x_i)} \right) \right] \tag{86}$$

Note that the above objective with penalty coefficient $=$ is equal to the InfoNCE objective. This formulation motivates alternate choices of depending on how strictly we would like to enforce alignment. We refer to the loss

$$L_{\;}(g) = -\mathsf{E}_{x,x^+}\left[g(x)^\top g(x^+) \right] + \mathsf{E}_{x,x^+,\{x_i\}_{i=1}^{\ell}} \left[\log \left(e^{g(x)^\top g(x^+)} + \sum_{i=1}^{\ell} e^{g(x)^\top g(x_i)} \right) \right] \tag{87}$$

as the Weighted InfoNCE loss which is equivalent to the penalized version of (85) with $=$. We note that this loss is similar to the generalized InfoNCE loss proposed by Chen et al. (2021).

Corollary 21 *Consider the same setting as Theorem 15 but with any* > 0 *and with the Weighted InfoNCE loss with* $2^{d+1} + \frac{1}{}$. *Then for sufficiently large* ℓ *and* , *all solutions of the Weighted InfoNCE objective are cluster-preserving.*

Proof The result follows from the analysis in the proof of Theorem 15. The analysis for the difference in positive terms is identical except that they are scaled by , so we have

$$L^+(g) \quad L^+(g^0) = 2 \quad kRk \tag{88}$$

For the difference in negative terms, the analysis is again identical except that we can no longer use that g is close to uniform. The only place we have used this is to bound $\ell \sum_{v \in H_d} kQ_v n Z k \; (1 \; kQ_v n Z k \;)^\ell$ in (80). Here, we bound this term using the fact that $h(x) := x(1 \quad x)^\ell$ is maximized on the interval $x \in [0, 1]$ at $x = \frac{1}{\ell+1}$.

$$\ell \sum_{v \in H_d} kQ_v n Z k \; (1 \quad kQ_v n Z k \;)^\ell \quad \ell \sum_{v \in H_d} \frac{1}{\ell+1} \quad 1 \quad \frac{1}{\ell+1}^\ell$$
$$= 2^d \quad \frac{\ell}{\ell+1} \quad \frac{\ell}{\ell+1}^{\ell+1}$$

$$\frac{2^d}{e} \tag{89}$$

Replacing this bound and executing the same analysis as in (83) yields

$$\begin{aligned} L\ (g^{\prime})\quad L\ (g) & \\ 2\ kRk\ \ 2^d + \frac{1}{100}\quad + 2\ kRk\frac{1{:}01}{e} & \\ < 2\quad kRk & \end{aligned} \tag{90}$$

completing the proof. ■

Appendix C. Proofs of Downstream Guarantees

C.1. Proof of Theorem 8

Proof Because there are exactly 2^d clusters, from Theorem 7, we know that since g is uniform, it maps each cluster to a unique vertex on the d-dimensional hypercube. Let $_f := f\ _{\mathbf{c}} : f(x) = 1\ 8x\ 2\ \ _{\mathbf{c}}g$ be the set of clusters which f labels 1, and let $N = j\ _f j$. Similarly let $^{c}_{f} := f\ _{\mathbf{c}} : f(x) = \ \ 1\ 8x\ 2\ \ _{\mathbf{c}}g$. For all $j\ 2\ [N]$, let the j-th row of $\mathsf{W}\ 2\ \mathsf{R}^{m\ \ d}$ equal the vertex corresponding to the mapping of the j-th cluster in $_f$ by g . For all $j\ 2\ [m\ \ N]$, let the $j + N$-th row of W equal the vertex corresponding to the mapping of the j-th cluster in $^{c}_{f}$ by g Then for any x such that $f(x) = 1$, $\mathsf{W}g\ (x)$ has exactly one element with value d among the first N elements, and all other elements are at most $d\ \ 2$. On the other hand, for any x such that $f(x) = \ \ 1$, $\mathsf{W}g\ (x)$ has exactly one element with value d among the last $m\ \ N$ elements, and all other elements are at most $d\ \ 2$. Set $b = (d\ \ 2)\ \ 1_m$, that is, $d\ \ 2$ times the m-dimensional vector of ones, and $a = [1^{>}_{N};\ \ 1^{>}_{m\ \ N}]^{>}$, that is, the m-dimensional vector whose first m elements are 1 and whose last $m\ \ N$ elements are $\ \ 1$. Then $a^{>}\mathrm{ReLU}(\mathsf{W}g(x)\ \ b) = f(x)$ for all x. ■

C.2. Proof of Theorem 9

Proof Since we are in the realizable setting, D has $m := 2^d$, equal-size clusters, where $d > 3$. Moreover, since $F_?$ is arbitrarily powerful and the augmentation sets are disjoint, for every pair of augmentation sets $(A(x); A(x))$, there exists an $f\ 2\ F_?$ such that $f(x^+) \neq f(x^+)$ for all $x^+\ 2\ A(x)$ and $x\ 2\ A(x)$, and f does not intersect any other augmentation set. Further, $G_?$ can map augmentation sets to arbitrarily different vertices, even if these sets lie in the same cluster. In other words, there are clean representations in $G_?$ (meaning they are faithful to all augmentation sets) that split clusters by augmentation sets.

Suppose the number of augmentation sets in each cluster is $k\ \ m\ \ 2^d$ for some $k\ 2\ \mathsf{N}^{+}$, and all augmentation sets are of equal size M. Then there exists a clean and uniform representation $g\ 2\ G_?$ such that for each vertex $v\ 2\ H_d$, kM of the images in the set $fx\ 2\ D\ \ : g(x) = vg$ are in each cluster. In other words, for all $x\ 2\ D\ \ $ and $x^+\ \ \ A(x)$, $g(x) = g(x^+)$. Thus, we can apply Theorem 7 to obtain $g\ 2\ \arg\min_{g^{\prime}2G_?} L(g^{\prime})$ (note that in Theorem 7, cluster-preserving is equivalent to clean since we are optimizing over the restricted class G, and the same proof can be applied exactly as is,

with the word "cluster-preserving" replaced by "clean", to show that $g \in \arg\min_{g' \in G_?} L(g')$ if and only if g is clean and uniform).

Since any head $! \in J$ composed with g must yield the same prediction for all images mapped to the same vertex on H_d, and all vertices have the same number of images from each cluster mapped to them, the number of images with predicted label 1 must be the same for all clusters, and likewise for 1. Thus, for any downstream binary classification task h that satisfies $h(x) = h(x')$ for all $x, x' \in$ c for all c $\in C$ and $P_{x \sim D}[h(x) = 1] = 0.5$, any $!$ must have $L_f(! \circ g) \geq 0.5$. ■