A MESSAGE PASSING PERSPECTIVE ON LEARNING DYNAMICS OF CONTRASTIVE LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

In recent years, contrastive learning achieves impressive results on self-supervised visual representation learning, but there still lacks a rigorous understanding of its learning dynamics. In this paper, we show that if we cast a contrastive objective equivalently into the feature space, then its learning dynamics admits an interpretable form. Specifically, we show that its gradient descent corresponds to a specific message passing scheme on the corresponding augmentation graph. Based on this perspective, we theoretically characterize how contrastive learning gradually learns discriminative features with the alignment update and the uniformity update. Meanwhile, this perspective also establishes an intriguing connection between contrastive learning and Message Passing Graph Neural Networks (MP-GNNs). This connection not only provides a unified understanding of many techniques independently developed in each community, but also enables us to borrow techniques from MP-GNNs to design new contrastive learning variants, such as graph attention, graph rewiring, jumpy knowledge techniques, etc. We believe that our message passing perspective not only provides a new theoretical understanding of contrastive learning dynamics, but also bridges the two seemingly independent areas together, which could inspire more interleaving studies to benefit from each other.

1 INTRODUCTION

Contrastive Learning (CL) has become arguably the most effective approach to learning visual representations from unlabeled data (Chen et al., 2020b; He et al., 2020; Chen et al., 2020c;d; 2021; Caron et al., 2021). However, till now, we actually know little about how CL gradually learns meaningful features from unlabeled data. Recently, there has been a burst of interest in the theory of CL in order to explain its effectiveness. However, despite the remarkable progress that has been made, existing theories of CL are established for either an *arbitrary* function f in the function class $\mathcal{F}$ (Saunshi et al., 2019; Wang et al., 2022) or the optimal f^* with minimal contrastive loss (Wang & Isola, 2020; HaoChen et al., 2021; Wang et al., 2022). Instead, a theoretical characterization of the learning dynamics is largely overlooked, which is the focus of this work.

Perhaps surprisingly, from a feature space perspective, we find out that the optimization dynamics of contrastive learning corresponds to a specific message passing scheme among different samples. Specifically, based on a reformulation of the alignment and uniformity losses of the contrastive loss, we show that the derived alignment and uniformity updates actually correspond to message passing on two different graphs: the alignment update on the augmentation graph defined by data augmentations, and the uniformity update on the affinity graph defined by feature similarities. Therefore, the combined contrastive update is a competition between two message passing rules. Based on this perspective, we further show that the equilibrium of contrastive learning can be achieved when the two message rules are balanced, *i.e.,* when the learned distribution $\mathcal{P}_\theta$ matches the ground-truth data distribution $\mathcal{P}_d$, which provides a clear picture for understanding the dynamics of contrastive learning.

Meanwhile, as message passing is a general paradigm in many scenarios, the message passing perspective of contrastive learning above also allows us to establish some intriguing connections to these seemingly different areas. One particular example is in graph representation learning. Message Passing Graph Neural Networks (MP-GNNs) are the prevailing designs in modern Graph Neural Networks (GNNs), including numerous variants like GCN (Kipf & Welling, 2017), GAT

(Veličković et al., 2018), and even the Transformers (Vaswani et al., 2017). There is a vast literature studying its diffusion dynamics and representation power (Li et al., 2018; Oono & Suzuki, 2020; Wang et al., 2021; Dong et al., 2021; Xu et al., 2019; 2018). Therefore, establishing a connection between contrastive learning (CL) and MP-GNNs will hopefully bring new theoretical and empirical insights for understanding and designing contrastive learning methods. In this work, we illustrate this benefit from three aspects: 1) we establish formal connections between the basic message passing mechanisms in two domains; 2) based on this connection, we discover some close analogies among the representative techniques independently developed in each domain; and 3) borrowing techniques from MP-GNNs, we design two new contrastive learning variants, and demonstrate their effectiveness on benchmark datasets. We summarize our contributions as follows:

- **Learning Dynamics.** We reformulate of the contrastive learning into the feature space and develop a new decomposition of the alignment and uniformity loss. Based on this framework, we show that the alignment and uniformity updates correspond to two different message passing schemes, and characterize the equilibrium states under the combined update. This message perspective provides a new understanding of contrastive learning dynamics.
- **Connecting CL and MP-GNNs.** Through the message passing perspective of contrastive learning (CL), we show that we can establish an intriguing connection between CL and MP-GNNs. We not only formally establish the equivalence between alignment update and graph convolution, uniformity update and self-attention, but also point out the inherent analogies between important techniques independently developed in each domain.
- **New Designs Inspired by MP-GNNs.** We also demonstrate the empirical benefits of this connection by designing two new contrastive learning variants borrowing techniques from MP-GNNs: one is to avoid the feature collapse of alignment update by multi-stage aggregation, and one is to adaptively align different positive samples with by incorporating the attention mechanism. Empirically, we show that both techniques leads to clear benefits on benchmark datasets. In turn, their empirical successes also help verify the validness of our established connection between CL and MP-GNNs.

2 A Message Passing Perspective on Contrastive Learning

In this section, we develop a message passing perspective for understanding the dynamics of contrastive learning. We begin by reformulating the contrastive loss into the feature space with a new decomposition. We then study the update rules derived from the alignment and uniformity losses, and explain their behaviors from a message passing perspective. When combined together, we also characterize how the two updates strike a balance at the equilibrium states. And we finish this section with a proof-of-idea to illustrate the effectiveness of our derive message passing rules.

2.1 Background, Reformulation, and Decomposition

We begin our discussion by introducing the canonical formulation of contrastive learning methods in the parameter space, and present their equivalent formulation in the feature space.

Contrastive Learning (CL). Given two positive samples (x, x^+) generated by data augmentations, and an independently sampled negative sample x', we can learn an encoder $f_\theta : \mathbb{R}^d \to \mathbb{R}^m$ with the wide adopted InfoNCE loss (Oord et al., 2018):

$$\mathcal{L}_{\text{nce}}(\theta) = -\mathbb{E}_{x,x^+}[f_\theta(x)^\top f_\theta(x^+)] + \mathbb{E}_x \log \mathbb{E}_{x'}[\exp(f_\theta(x)^\top f_\theta(x'))], \tag{1}$$

where the form term pulls positive samples (x, x^+) together by encouraging their similarity, and the latter term pushes negative pairs (x, x') apart. In practice, we typically randomly draw M negative samples to approximate the second term.

In contrastive learning, the encoder is parameterized by deep neural networks, making it hardly amenable for formal analysis. Wen & Li (2021) resort to single-layer networks with strong assumptions on data distribution, but it is far from practice. Instead, in this work, we focus on the dynamics of the learned *data features* $f_\theta(x)$ in the feature space $\mathcal{R}^m$. As over-parameterized neural networks are very expressive and adaptive, we assume the feature matrix F_θ to be *unconstrained* (updated freely ignoring dependences on parameters and sample complexity) as in HaoChen et al. (2021). This assumption is also widely adopted in deep learning theory, such as, layer peeled models (Mixon et al.,

2022; Zhu et al., 2021). The unconstrained feature assumption enables us to focus on the relationship between augmented samples. As pointed out by Wang et al. (2022), the support overlap between augmented data is the key for contrastive learning to generalize to downstream tasks. However, we still do not have a clear view of how data augmentations affect feature learning, which is the focus of this work. For a formal exposition, we model data augmentations through the language of augmentation graph.

Following HaoChen et al. (2021), we assume a finite sample space $\mathcal{X}$ (can be exponentially large). Given a set of natural data as $\bar{\mathcal{X}} = \{\bar{x} \mid \bar{x} \in \mathbb{R}^d\}$, we first draw a natural example $\bar{x} \in \mathcal{P}_d(\bar{\mathcal{X}})$, and then draw a pair of samples x, x^+ from a random augmentation of $\bar{x}$ with distribution $\mathcal{A}(\cdot|\bar{x})$. We denote $\mathcal{X}$ as the collection of augmented "views", *i.e.,* $\mathcal{X} = \cup_{\bar{x}\in\bar{\mathcal{X}}} \operatorname{supp}(\mathcal{A}(\cdot|\bar{x}))$. Regarding the N samples in $\mathcal{X}$ as N nodes, we can construct an augmentation graph $\mathcal{G} = (\mathcal{X}, A)$ with an adjacency matrix A representing the mutual connectivity. The edge weight $A_{xx'} \geq 0$ between x and x' is defined as their joint probability $A_{xx'} = \mathcal{P}_d(x, x') = \mathbb{E}_{\bar{x}}\mathcal{A}(x|\bar{x})\mathcal{A}(x'|\bar{x})$. The normalized adjacency matrix is $\bar{A} = D^{-1/2}AD^{-1/2}$, where $D = \deg(A)$ denotes the diagonal degree matrix, *i.e.,* $D_{xx} = w_x = \sum_{x'} A_{xx'}$. The symmetrically normalized Laplacian matrix $\bar{L} = I - \bar{A}$ is known to be positive semi-definite (Chung, 1997). As natural data and data augmentations are uniformly sampled in practice (Chen et al., 2020b), we assume uniform distribution $w_x = 1/N$ for simplicity.

New Alignment and Uniformity Losses in Feature Space. Here, we reformulate the contrastive losses into the feature space, and decompose them into two parts, the alignment loss for positive pairs, and the uniformity loss for negative pairs ($\mathcal{L}_{\text{unif}}$ for InfoNCE loss, and $\mathcal{L}_{\text{unif}}^{(\text{sp})}$ for spectral loss):

$$\mathcal{L}_{\text{align}}(F) = \operatorname{Tr}(F^\top \bar{L} F) = \frac{1}{2}\mathbb{E}_{x,x^+}\|f(x) - f(x^+)\|^2, \tag{2a}$$

$$\mathcal{L}_{\text{unif}}(F) = \operatorname{LSE}(FF^\top) - \|F\|^2 = \mathbb{E}_x \log \mathbb{E}_{x'} \exp(f(x)^\top f(x')) - \mathbb{E}_x\|f(x)\|^2, \tag{2b}$$

where $\operatorname{LSE}(X) = \operatorname{Tr}(D\log(\deg(D\exp(D^{-1/2}XD^{-1/2}))))$ serve as a pseudo "norm".[1] The following proposition establishes the equivalence between original loss and our reformulated one.

Proposition 1. *When the x-th row of F is defined as $F_x = \sqrt{w_x} f_\theta(x)$, we have:*

$$\mathcal{L}_{\text{nce}}(\theta) = \mathcal{L}_{\text{align}}(F_\theta) + \mathcal{L}_{\text{unif}}(F_\theta). \tag{3}$$

Remark on Loss Decomposition. Notably, our decomposition of contrastive loss is different from the canonical decomposition proposed by Wang & Isola (2020). Taking the InfoNCE loss as an example, they directly take the two terms of Eq. 1 as the alignment and uniformity losses, which, however, leads to two improper implications without feature normalization. First, minimizing their alignment loss $\mathcal{L}_{\text{align}}(\theta) = -\mathbb{E}_{x,x^+} f_\theta(x)^\top f_\theta(x^+)$ alone is an ill-posed problem, as it approaches $-\infty$ when we simply scale the norm of F. In this case, the alignment loss gets smaller while the distance between positive pairs gets larger. Second, minimizing the uniformity loss alone will bridge the distance closer. Instead, with our decomposition, the alignment loss (Eq. 2a) is positive definite and will bring samples together with smaller loss, and a larger uniformity loss (Eq. 2b) will expand their distances (Figure 1). Thus, our decomposition in Eq. 2 seems more properly defined than theirs.

Based on the reformulation in Eq. 2, we will study the dynamics of each objective in Section 2.2.1 and give a unified analysis of their equilibrium in Section 2.3, both from a message passing perspective.

2.2 A Message Passing Perspective on Alignment and Uniformity

2.2.1 Alignment Update is Message Passing on Augmentation Graph

In fact, the reformulated alignment loss $\mathcal{L}_{\text{align}}(F) = \operatorname{Tr}(F^\top \bar{L} F)$ is widely known as the *Laplacian regularization* that has wide applications in various graph-related scenarios (Ng et al., 2001; Ando & Zhang, 2006). Its gradient descent with step size $\alpha > 0$ gives the following update rule:

$$\text{(global update)} \quad F^{(t+1)} = F^{(t)} - \alpha\nabla_{F^{(t)}}\mathcal{L}_{\text{align}}(F^{(t)}) = \left[(1-2\alpha)I + 2\alpha\bar{A}\right]F^{(t)}. \tag{4}$$

One familiar with graphs can immediately recognize that this update rule is a matrix-form message passing scheme of features $F^{(t)}$ with a propagation matrix $P = (1-2\alpha)I + 2\alpha\bar{A}$. For those less familiar with the graph literature, let us take a look at the *local* update of the feature F_x of each x:

$$\text{(local update)} \quad F_x^{(t+1)} = (1-2\alpha)F_x^{(t+1)} + 2\alpha\textstyle\sum_{x'\in\mathcal{N}_x} \bar{A}_{xx'}F_{x'}^{(t+1)}. \tag{5}$$

[1] Here $\log, \exp$ are element-wise operations.

We can see that the alignment loss updates each feature F_x by a weighted sum that aggregates features from its neighborhood $\mathcal{N}_x = \{x' \in \mathcal{X} | \bar{A}_{xx'} > 0\}$. In the context of contrastive learning, this reveals that the alignment loss is to propagate features among positive pairs, and in each step, more alike positive pairs x, x^+ (with higher edge weights A_{xx^+}) will exchange more information.

Next, let us turn our focus to how this message passing scheme affects downstream classification. As for contrastive learning, because positive samples come from the same class, this alignment loss will help cluster same-class samples together. However, it does not mean that any augmentations can attain a minimal intra-class variance. Below, we formally characterize the effect of data augmentations on the feature clustering of each class k. We measure the degree of clustering by the distance between all intra-class features to a one-dimensional subspace $\mathcal{M}$.

Proposition 2. *Denote the distance between a feature matrix X and a subspace Ω by $d_\Omega(X) := \inf\{\|X - Y\|_F \mid Y \in \Omega\}$. Assume that data augmentations are label-preserving. Then, for samples in each class k, denoting their adjacency matrix as $\bar{A}_k$ and their features at t-th step as $F_k^{(t)}$, the message passing in Eq. 4 leads to a smaller or equal distance to an one-dimensional subspace $\mathcal{M}$*

$$d_{\mathcal{M}}(F_k^{(t+1)}) \leq |1 - 2\alpha\lambda_{\mathcal{G}_k}| \cdot d_{\mathcal{M}}(F_k^{(t)}), \tag{6}$$

where $\lambda_{\mathcal{G}_k} \in [0, 2]$ denotes the intra-class algebraic connectivity of $\mathcal{G}$ (Chung, 1997), and $\mathcal{M} = \{e_1 \otimes y | y \in \mathbb{R}^c\}$, and e_1 is the eigenvector corresponding to the largest eigenvalue of $\bar{A}_k$.

An important message of Proposition 2 is that the alignment update does *not* necessarily bring better clustering, which depends on the algebraic connectivity $\lambda_{\mathcal{G}_k}$: 1) only a non-zero $\lambda_{\mathcal{G}_k}$ guarantees the strict decrease of intra-class distance, and 2) a larger $\lambda_{\mathcal{G}_k}$ indicates an even faster decrease. According to spectral graph theory, $\lambda_{\mathcal{G}_k}$ is the second-smallest eigenvalue of the Laplacian L, and its magnitude reflects how connected the graph is. In particular, $\lambda_{\mathcal{G}_k} > 0$ holds if and only if the intra-class graph is connected (Chung, 1997). Indeed, when the class is not connected, the alignment update can never bridge the disjoint components together. While in contrastive learning, the design of data augmentations determines $\mathcal{G}_k$ as well as $\lambda_{\mathcal{G}_k}$. Therefore, our message passing analysis *quantitatively* characterize how augmentations affect feature clustering in the contrastive learning dynamics.

2.2.2 Uniformity Update is Message Passing on the Affinity Graph

Contrary to the alignment loss, the uniformity loss instead penalizes the similarity between every sample pair $x, x' \in \mathcal{X}$ to encourage feature uniformity. Perhaps surprisingly, from a feature space perspective, its gradient descent rule, namely the uniformity update, also corresponds to a message passing scheme with the affinity graph $\mathcal{G}'$ with a "fake" adjacency matrix A':

$$\text{(global update)} \quad F^{(t+1)} = F^{(t)} - \alpha \nabla_{F^{(t)}} \mathcal{L}_{\text{unif}}(F^{(t)}) = [(1 + 2\alpha)I - 2\alpha \bar{A}'_{\text{sym}}]F^{(t)}, \tag{7}$$

where $\bar{A}'_{\text{sym}} = {D'_{\text{exp}}}^{-1} A'_{\text{exp}} + A'_{\text{exp}} {D'_{\text{exp}}}^{-1}$, and $A'_{\text{exp}} = \exp(D^{-1/2} F^{(t)} F^{(t)\top} D^{-1/2})$, $D'_{\text{exp}} = \deg(A'_{\text{exp}})$. When we further apply stop gradient to the target branch (as adopted in BYOL (Grill et al., 2020) and SimSiam (Chen & He, 2021)), we can further simplify it to

$$\text{(global update with stop gradient)} \quad F^{(t+1)} = [(1+\alpha)I - \alpha \bar{A}']F^{(t)}, \text{ where } \bar{A}' = {D'_{\text{exp}}}^{-1} A'_{\text{exp}}. \tag{8}$$

We can also interpret the uniformity update as a specific message passing scheme on the *affinity graph* $\mathcal{G}' = (\mathcal{X}, A')$ with node set $\mathcal{X}$ and adjacency matrix A'. In comparison to the augmentation graph $\mathcal{G}$ that defines the edge weights via data augmentations, the affinity graph is instead constructed through the estimated *feature similarity* between (x, x'). The contrastive learning with two graphs resembles the wake-sleep algorithm in classical unsupervised learning (Hinton et al., 1995).

Further, we notice that the affinity matrix $\bar{A}'$ is induced by the RBF kernel $k_{\text{RBF}}(u, v) = \exp(-\|u - v\|^2/2)$ when the features are normalized, *i.e.,* $\|f(x)\| = 1$ (often adopted in practice (Chen et al., 2020b)). Generally speaking, the RBF kernel is more preferable than the linear kernel as it is more non-Euclidean and non-parametric. Besides, because the message passing in the alignment update utilizes $\bar{A}$, which is also non-negative and normalized, the InfoNCE uniformity update with $\bar{A}'_{\text{nce}}$ can be more stable and require less hyperparameter tuning to cooperate with the alignment update. Thus, our message passing view not only provides a new understanding of the uniformity update, but also helps explain the popularity of the InfoNCE loss.

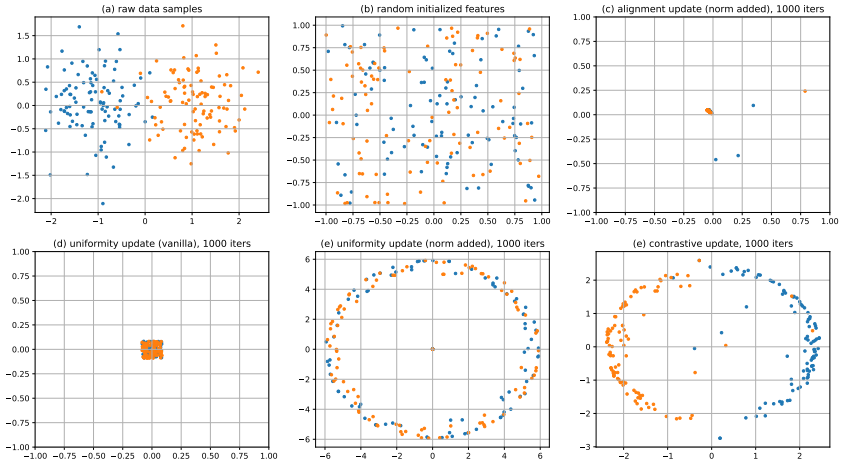


Figure 1: Mimicking contrastive learning with our message passing rules on synthetic data. The data samples $\mathcal{D}$ are generated following the isotropic Gaussian distribution with means $(-1, 0)$ and $(1, 0)$ (two classes, blue and yellow), and variance 0.7. We then construct an augmentation graph $\mathcal{G}$ by drawing edged $A_{ij} = 1$ for any data pair satisfying $\|x_i - x_j\|_2 \leq 0.4$. For mimicking contrastive learning, we firstly initialize all features $\{f_i\}$ uniformly in $[-1, 1]^2$, and then perform message passing of these features using $\mathcal{G}$ for 1,000 steps with step size 0.1. Best viewed in color.

2.3 The Equilibrium Between the Alignment and Uniformity Update

Combining the alignment update and the uniformity update discussed above, we arrive at our final update rule for contrastive learning with the (regularized) InfoNCE loss (with stop gradient):

$$F^{(t+1)} = F^{(t)} - \alpha \nabla_{F^{(t)}} \mathcal{L}_{\text{nce}}(F^{(t)}) = F^{(t)} + \alpha(\bar{A} - \bar{A}')F^{(t)}, \tag{9}$$

which is a special message passing scheme using the difference between the augmentation (data) and affinity (estimated) adjacency matrices $A_{\text{diff}} = \bar{A} - \bar{A}'$. Accordingly, we can reformulate the contrastive loss more intuitively as:

$$\mathcal{L}(F) = -\operatorname{Tr}(\operatorname{sg}(A_{\text{diff}})FF^\top) = -\operatorname{Tr}(\bar{A}FF^\top) + \operatorname{Tr}(\operatorname{sg}(A')FF^\top) = \mathcal{L}_{\text{align}}(F) + \mathcal{L}_{\text{unif}}(F), \tag{10}$$

where $\operatorname{sg}(\cdot)$ means the stop gradient operator.

Equilibrium Distribution. The reformulation above (Eq. 10) also reveals the stationary point of the learning dynamics, that is, $\bar{A}_{\text{diff}} = \bar{A} - A' = 0$, which implies the model equilibrium as follows.

Proposition 3. *The contrastive learning dynamics with the InfoNCE loss saturates when the real and the fake adjacency matrices agree, i.e., $\bar{A} = \bar{A}'_{\text{nce}}$, which means that element-wisely, we have:*

$$\forall\, x, x^+ \in \mathcal{X}, \;\; \mathcal{P}_d(x^+|x) = \mathcal{P}_\theta(x^+|x) \triangleq \frac{\exp(f_\theta(x)^\top f_\theta(x^+))}{\sum_{x'} \exp(f_\theta(x)^\top f_\theta(x'))}. \tag{11}$$

That is, the ground-truth data conditional distribution $\mathcal{P}_d(x^+|x)$ equals to the Boltzmann distribution $\mathcal{P}_\theta(x^+|x)$ estimated from the encoder features.

In this way, our analysis reveals that contrastive learning (with InfoNCE loss) implicitly learns a probabilistic model $\mathcal{P}_\theta(x^+|x)$ (Eq. 11), and its goal is to approximate the conditional data distribution $P_d(x^+|x)$ defined via the augmentation graph $\mathcal{G}$. This explains the effectiveness of contrastive learning on out-of-distribution detection tasks (Winkens et al., 2020) that rely on good density estimation. In practical implementations of InfoNCE, *e.g.,* SimCLR, a temperature scalar τ is often introduced to rescale the cosine similarities. With a flexible choice of τ, the model distribution $\mathcal{P}_\theta$ could better approximate sharp (like one-hot) data distribution $\mathcal{P}_d$. A detailed comparison to related work is included in Appendix C.

2.4 A Proof-of-idea Experiment

At last, we verify the derived three message passing rules (alignment, uniformity, contrastive) through a synthetic experiment using on a synthetic augmentation graph to mimic contrastive learning in

the feature space. From Figure 1, we can see that the results generally align well with our theory. First, the alignment update (Section 2.2.1) indeed bridges intra-class samples closer, but there is also a risk of feature collapse. Second, the uniformity update (Section 2.2.2) indeed keeps all samples uniformly distributed. Third, by combining the alignment and uniformity updates, the contrastive update (Section 2.3) can successfully cluster intra-class samples together while keeping inter-class samples apart, even after long iterations. In this way, we show that like the actual learning behaviors of contrastive learning, the derived contrastive message passing rules can indeed learn stable and discriminative features from unlabeled data.

3 Contrastive Learning (CL) and Message Passing Graph Neural Networks (MP-GNNs): Connections, Analogies, and New Designs

In the previous section, we have shown the contrastive learning dynamics corresponds nicely to a message passing scheme on two different graphs, $\bar{A}$ and $\bar{A}'$. In a wider context, message passing is a general technique with broad applications in graph theory, Bayesian inference, graph embedding, etc. As a result, our interpretation of contrastive learning from a message passing perspective enables us to establish some interesting connections between contrastive learning and these different areas.

A noticeable application of message passing is graph representation learning. In recent years, Message Passing Graph Neural Networks (MP-GNNs) become the *de facto* paradigm for GNNs. Various techniques have been explored in terms of different kinds of graph structures and learning tasks. By drawing a connection between contrastive learning and MP-GNNs, we could leverage these techniques to design better contrastive learning algorithms, as is the goal of this section.

To achieve this, we first establish the basic connections between the basic paradigms in Section 3.1. Built upon these connections, in Section 3.2, we point out the common ideas behind many representative techniques independently proposed by each community in different names. Lastly, we show that we could further borrow some new techniques from the MP-GNN literature for designing new contrastive learning methods in Section 3.3.

3.1 Connections between the Two Paradigms

Regarding the alignment update and the uniformity update in contrastive learning, we show that they do have a close connection to message passing schemes in graph neural networks.

Graph Convolution and Alignment Update. A graph data with N nodes typically consist of two inputs, the input feature $X \in \mathbb{R}^{N\times d}$, and the (normalized) adjacency matrix $\bar{A} \in \mathbb{R}^{n\times n}$. For a feature matrix $H^{(t)} \in \mathbb{R}^{n\times m}$ at the t-th layer, the canonical graph convolution scheme in GCN (Kipf & Welling, 2017) gives $H^{(t+1)} = \bar{A}H^{(t)}$. Later, DGC (Wang et al., 2021) generalizes it to $H^{(t+1)} = (1-\Delta t)I + \Delta t\bar{A}H^{(t)}$ with a flexible step size Δt for finite difference, which reduces the canonical form when $\Delta t = 1$. Comparing it to the alignment update rule in Eq. 4, we can easily notice their equivalence when we take the augmentation graph $\bar{A}$ as the input graph for a GCN (or DGC), and adopt a specific step size $\Delta t = 2\alpha$. In other words, the gradient descent of F using the alignment loss (an optimization step) is equivalent to the graph convolution of F based on the augmentation graph $\bar{A}$ (an architectural module). Their differences in names and forms are only superficial, since inherently, they share the same message passing mechanism.

Oversmoothing and Feature Collapse. The equivalence above is the key for us to unlock many intriguing connections between the two seemingly different areas. A particular example is the oversmoothing problem, one of the centric topics in MP-GNNs with tons of discussions, to name a few, Li et al. (2018), Oono & Suzuki (2020), Shi et al. (2022). Oversmoothing refers to the phenomenon that when we apply the graph convolution above for many times, the node features will become indistinguishable from each other and lose discriminative ability. Meanwhile, we have also observed the same phenomenon for the alignment update in Figure 1. Indeed, it is also well known in contrastive learning that with the alignment loss alone, the features will collapse to a single point, known as *complete feature collapse*. Hua et al. (2021) show that when equipped with BN at the output, the features will not fully collapse but still collapse to a low dimensional subspace, named *dimensional collapse*. In fact, both two feature collapse scenarios with the augmentation graph can be characterized using existing oversmoothing analysis of a general graph (Oono & Suzuki, 2020). Therefore, this connection also opens new paths for the theoretical analysis of contrastive learning.

Self-Attention and Uniformity Update. Besides the connection in the alignment update, we also notice an interesting connection between the uniformity update rule and the self-attention mechanism. Self-attention is the key component of Transformer (Vaswani et al., 2017), which recently obtains promising progresses in various scenarios like images, video, text, graph, molecules (Lin et al., 2021). For an input feature $H^{(t)} \in \mathbb{R}^{N \times m}$ at the t-th step, the self-attention module gives update:

$$H^{(t+1)} = \bar{A}' H^{(t)}, \text{ where } \bar{A}' = {D'_{\text{exp}}}^{-1} A'_{\text{exp}}, A'_{\text{exp}} = \exp(H^{(t)} H^{(t)\top}). \tag{12}$$

When comparing it to the uniformity update in Eq. 8, we can see that the two rules are equivalent when $\alpha = -1$. Because the step size is negative, self-attention is actually a *gradient ascent* step that will maximize the feature uniformity loss (Eq. 2b). Consequently, stacking self-attention will reduce the uniformity between features and lead to more collapsed representations, which explains the feature collapse behavior observed in Transformers (Shi et al., 2022). Different from previous work (Dong et al., 2021), our analysis above stems from the connection between contrastive learning and MP-GNNs, and it provides a simple explanation of this phenomenon from an optimization perspective.

3.2 Analogies in Existing Techniques

In the discussion above, we establish connections between the two paradigms. Besides, we know that many important techniques and variants have been built upon the basic paradigms in each community. Here, through a unified perspective of the two domains, we spot some interesting connections between techniques developed independently in each domain, while the inherent connections are never revealed. Below, we give two examples to illustrate benefits from these analogies.

NodeNorm / LayerNorm and ℓ_2 Normalization. In contrastive learning, since SimCLR (Chen et al., 2020b), a common practice is to apply ℓ_2 normalization before calculating the InfoNCE loss, *i.e.,* $f(x)/\|f(x)\|$. As a result, the alignment loss becomes equivalent to the *cosine similarity* between the features of positive pairs. Yet, little was known about the real benefits of ℓ_2 normalization. As another side of the coin, this node-wise feature normalization technique, in the name of NodeNorm (Zhou et al., 2021), has already been shown as an effective approach to mitigate the performance degradation of deep GCNs by controlling the feature variance. Zhou et al. (2021) also show that LayerNorm (Ba et al., 2016) with both centering and normalization can also obtain similar performance. These discussions in MP-GNNs also help us understand the effectiveness of ℓ_2 normalization.

PairNorm / BatchNorm and Feature Centering. Another common technique in contrastive learning is feature centering, *i.e.,* $f(x) - \mu$, where μ is the mean of all sample (node) features. In DINO (Caron et al., 2021), it is adopted to prevent feature collapse. Similarly, Hua et al. (2021) also show that BatchNorm (Ioffe & Szegedy, 2015) (with feature centering as a part) can alleviate feature collapse. On the other side, PairNorm (Zhao & Akoglu, 2020) combines feature centering and feature normalization (mentioned above) to alleviate oversmoothing. In particular, they show that PairNorm can preserve the total feature distance of the input (initial features). This, in turn, could be adopted as an explanation for how feature centering works in contrastive learning.

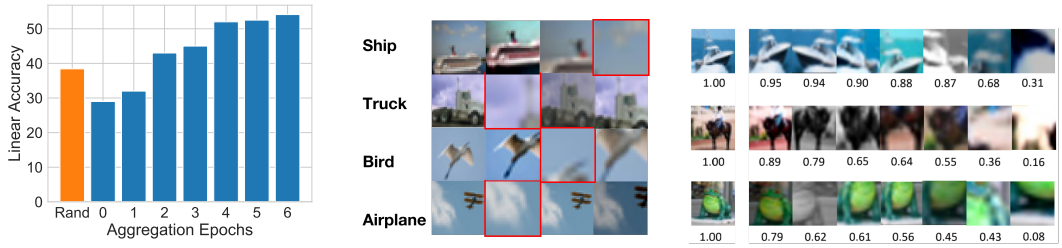
3.3 Inspired New Designs

Besides the analogies drawn above, we find that there are also many other valuable techniques from the MP-GNN literature that have not found their (exact) counterparts in contrastive learning. In this section, we explore two new connections from different aspects: multi-stage aggregation for avoiding feature collapse, and graph attention for better feature alignment.

3.3.1 New Connection I: Multi-stage Graph Aggregation

As discussed above, the alignment loss alone will lead to feature collapse in contrastive learning. Although existing works like BYOL (Grill et al., 2020) and SimSiam (Chen & He, 2021) can alleviate feature collapse via asymmetric architectural designs, we are still not fully clear how they work. Instead, the connection between feature collapse and oversmoothing (Section 3.1) inspires us to borrow techniques from the MP-GNNs to alleviate the feature collapse issue of contrastive learning.

In the MP-GNN literature, a principled solution to oversmoothing is the *jump knowledge* technique proposed in JK-Net (Xu et al., 2018). Instead of adopting features only from the last step $H^{(T)}$ (which might already oversmooth), JK-Net aggregates features from multiple stages, *e.g.,* $H^{(1)}, \ldots, H^{(T)}$,



(a) multi-stage aggregation (b) data augmentation examples (c) estimated attention coefficients

Figure 2: Experiments on CIFAR-10: (a) Performance of multi-stage aggregation (linear accuracy (%)) and aggregation epochs s. (b) natural (1st column) samples and their augmented views (2-4th columns) using SimCLR augmentations. (c) Estimated scores between the anchors (1st column) and the augmented views (2-8th columns) using a pretrained SimCLR model.

as the final feature for classification. In this way, the knowledge from previous non-over-smoothing steps could directly jump into the final output, which helps improve its discriminative ability. This idea of multi-stage aggregation is also widely adopted in many other MP-GNNs, such as APPNP (Klicpera et al., 2019), GCNII (Chen et al., 2020a), SIGN (Rossi et al., 2020), *etc.*

In this work, we transfer this idea to the contrastive learning scenario, that is, to alleviate feature collapse by combining features from multiple previous epochs. Specifically, we create a memory bank $\mathcal{M}$ to store the latest s-epoch features from the same *natural* sample $\bar{x}$. At the t-th epoch, for x generated from $\bar{x}$, we have stored old features of it as $z_{\bar{x}}^{(t-1)}, z_{\bar{x}}^{(t-2)}, \ldots, z_{\bar{x}}^{(t-r)}$ where $r = \min(s, t)$. Then we replace the original positive feature $f_\theta(x^+)$ with the aggregation of the old features in the memory bank, where we simply adopt the sum aggregation for simplicity, *i.e.,* $z_{\bar{x}} = \frac{1}{r}\sum_{i=1}^{r} z_{\bar{x}}^{(t-i)}$. Next, we optimize the encoder network f_θ with the following *multi-stage alignment* loss *alone*,

$$\mathcal{L}_{\mathrm{multi-align}}(\theta) = -\mathbb{E}_{\bar{x}}\mathbb{E}_{x|\bar{x}}(f_\theta(x)^\top z_{\bar{x}}). \tag{13}$$

Afterwards, we will push the newest feature $z = f_\theta(x)$ to the memory bank, and drop the oldest one if exceeded. In this way, we could align the new feature with multiple old features that are less collapsed, which, in turn, also prevents the collapse of the updated features.

Result Analysis. We empirically evaluate this method on CIFAR-10 with ResNet-18 backbone (details in Appendix B.1). From Figure 2a, we can see that as we adopt a larger memory bank with a large aggregation step s, the linear probing accuracy continues to grow larger and larger. Although the overall accuracy is still relatively low, we can observe a clear benefit of multi-stage aggregation for preventing feature collapse, which previously has only been possible via asymmetric architectural designs. The success of applying multi-stage aggregation to contrastive learning not only suggests a new approach for negative-free contrastive learning, but also helps verify our established connection between contrastive learning and MP-GNNs. In Section D, we also show that multi-stage aggeregation can also bring clear benefits when combined with existing methods.

3.3.2 New Connection II: Graph Attention

A noticeable defect of existing contrastive loss is that they assign an equal weight for different positive pairs. As these positive samples are generated by input-agnostic random data augmentations, there could be a large difference in semantics between different positive samples. As marked in red (Figure 2b), some augmented views may even lose the center objects, and aligning with these bad cases will bridge samples from different classes together and distort the decision boundary. Therefore, in the alignment update, it is not proper to assign equal weights as in vanilla alignment loss (Eq. 5). Instead, we should adaptively assign different weights to different pairs based on *semantic similarities*.

Revisiting MP-GNNs, we find that there are well-established techniques to address this issue. A well-known extension of graph convolution is *graph attention* proposed in GAT (Veličković et al., 2018), and Transformers can be seen as a special case of GAT applied to a fully connected graph. GAT extends GCN by taking the semantic similarities between neighborhood features (corresponding to positive pairs in the augmentation graph) into consideration. Specifically, for a node x with neighbors

Table 1: The comparison of linear accuracy between the original InfoNCE loss and our proposed reweighting loss on CIFAR-10, CIFAR-100 and ImageNet-100 with different backbones including ResNet-18 and ResNet-50. Our promoted loss has a significant improvement in most situations.

Backbone	Method	CIFAR-10	CIFAR-100	ImageNet-100
ResNet-18	SimCLR	84.5	56.1	62.3
	SimCLR-Attn	**85.4**	**56.9**	**63.1**
ResNet-50	SimCLR	88.2	59.8	66.0
	SimCLR-Attn	**89.4**	**60.7**	**66.7**

$\mathcal{N}_x$, GAT aggregates neighbor features with the attention scores as

$$H_x^{(t+1)} = \sum_{x^+ \in \mathcal{N}_x} \alpha(x, x^+) H_{x^+}^{(t)}, \text{ where } \alpha(x, x^+) = \frac{\exp(e(x, x^+))}{\sum_{x' \in \mathcal{N}_i} \exp(e(x, x'))}. \tag{14}$$

In GAT, they adopt additive attention to compute the so-called attention coefficients $e(x, x')$. Alternatively, we can adopt the dot-product attention as in Transformers, *e.g.,* $e(x, x') = H_x^{(t)\top} H_{x'}^{(t)}$.

Motivated by the design of graph attention, we revise the vanilla alignment loss by reweighting positive pairs with their estimated semantic similarities. Following Transformer, we adopt the dot-product form to estimate the attention coefficient as $e(x, x^+) = f_\theta(x)^\top f_\theta(x^+)$. Indeed, Figure 2c shows that the contrastive encoder can provide decent estimation of semantic similarities between positive samples. Therefore, we adopt this attention coefficient to adaptively reweight the positive pairs, and obtain the following *attention alignment* loss (with the uniformity loss unchanged),

$$\mathcal{L}_{\mathrm{attn-align}}(\theta) = \frac{1}{2} \mathbb{E}_{x, x^+} \alpha(x, x^+) \| f_\theta(x) - f_\theta(x^+) \|^2. \tag{15}$$

Specifically, we assign the attention score $\alpha(x, x^+)$ as $\alpha(x, x^+) = \frac{\exp(\beta \cdot f_\theta(x)^\top f_\theta(x^+))}{\sum_{x'} \exp(\beta \cdot f_\theta(x)^\top f_\theta(x'))}$, where we simply adopt M negative samples in the mini-batch for computing the normalization, and β is a hyperparameter to control the degree of reweighting. We also detach the gradients from the weights as they mainly serve as loss coefficients. The corresponding *attention alignment update* is

$$\text{(local update with attention) } F_x \leftarrow (1 - 2\alpha) F_x + 2\alpha \sum_{x^+}^{N} \bar{A}_{xx^+} \alpha(x, x^+) F_{x^+}. \tag{16}$$

Compared to the vanilla alignment update (Eq. 5), the attention alignment update adaptively aligns different positive samples according to the *feature-level semantic similarity* $\alpha(x, x^+)$. Thus, our proposed attention alignment can help correct the potential sampling bias of existing data augmentation strategies, and learn more semantic-consistent features.

Result Analysis. We follow the basic setup of SimCLR (Chen et al., 2020b) and consider two popular backbones: ResNet-18 and ResNet-50. We consider image classification tasks on CIFAR-10, CIFAR-100 and ImageNet-100. We train 100 epochs on ImageNet-100 and 200 epochs on CIFAR-10 and CIFAR-100. More details can be found in Appendix B.2. As shown in Table 1, the proposed attention alignment leads to a consistent 1% increase in linear probing accuracy among all settings, which helps verify that our attentional reweighting of the positive samples indeed lead to consistently better performance compared to the vanilla equally averaged alignment loss.

4 Conclusion

In this paper, we proposed a new message passing perspective for understanding the learning dynamics of contrastive learning. We showed that the alignment update corresponds to a message passing scheme on the augmentation graph, while the uniformity update is a reversed message passing on the affinity graph, and the two message passing rules strike a balance at the equilibrium where the two graphs agree. Based on this perspective as a bridge, we further developed a deep connection between contrastive learning and MP-GNNs. On the one hand, we established formal equivalences between their message passing rules, and pointed out the inherent connections between some existing techniques developed in each domain. On the other hand, borrowing techniques from the MP-GNN literature, we developed two new methods for contrastive learning, and verified their effectiveness on real-world data. Overall, we believe that the discovered connection between contrastive learning and message passing could inspire more studies from both sides to benefit from each other, both theoretically and empirically.

REFERENCES

Rie Ando and Tong Zhang. Learning on graph with laplacian regularization. In *NeurIPS*, 2006.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021.

Ming Chen, Zhewei Wei, Zengfeng Huang, Bolin Ding, and Yaliang Li. Simple and deep graph convolutional networks. *ICML*, 2020a.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020b.

Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. Big self-supervised models are strong semi-supervised learners. In *NeurIPS*, 2020c.

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *CVPR*, 2021.

Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020d.

Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. *arXiv preprint arXiv:2104.02057*, 2021.

Fan RK Chung. *Spectral graph theory*. American Mathematical Society, 1997.

Yihe Dong, Jean-Baptiste Cordonnier, and Andreas Loukas. Attention is not all you need: Pure attention loses rank doubly exponentially with depth. In *ICML*, 2021.

Jean-Bastien Grill, Florian Strub, Florent Altché, C. Tallec, Pierre H. Richemond, Elena Buchatskaya, C. Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, B. Piot, K. Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: a new approach to self-supervised learning. In *NeurIPS*, 2020.

Jeff Z HaoChen, Colin Wei, Adrien Gaidon, and Tengyu Ma. Provable guarantees for self-supervised deep learning with spectral contrastive loss. In *NeurIPS*, 2021.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020.

Geoffrey E Hinton, Peter Dayan, Brendan J Frey, and Radford M Neal. The" wake-sleep" algorithm for unsupervised neural networks. *Science*, 268(5214):1158–1161, 1995.

Tianyu Hua, Wenxiao Wang, Zihui Xue, Sucheng Ren, Yue Wang, and Hang Zhao. On feature decorrelation in self-supervised learning. In *ICCV*, 2021.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, 2015.

Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *ICLR*, 2017.

Johannes Klicpera, Aleksandar Bojchevski, and Stephan Günnemann. Predict then propagate: Graph neural networks meet personalized pagerank. In *ICLR*, 2019.

Qimai Li, Zhichao Han, and Xiao-Ming Wu. Deeper insights into graph convolutional networks for semi-supervised learning. *AAAI*, 2018.

Tianyang Lin, Yuxin Wang, Xiangyang Liu, and Xipeng Qiu. A survey of transformers. *arXiv preprint arXiv:2106.04554*, 2021.

Dustin G Mixon, Hans Parshall, and Jianzong Pi. Neural collapse with unconstrained features. *Sampling Theory, Signal Processing, and Data Analysis*, 20(2):1–13, 2022.

Andrew Ng, Michael Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In *NeurIPS*, 2001.

Kenta Oono and Taiji Suzuki. Graph neural networks exponentially lose expressive power for node classification. In *ICLR*, 2020.

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

Emanuele Rossi, Fabrizio Frasca, Ben Chamberlain, Davide Eynard, Michael Bronstein, and Federico Monti. SIGN: Scalable inception graph neural networks. *arXiv preprint arXiv:2004.11198*, 2020.

Nikunj Saunshi, Orestis Plevrakis, Sanjeev Arora, Mikhail Khodak, and Hrishikesh Khandeparkar. A theoretical analysis of contrastive unsupervised representation learning. In *ICML*, 2019.

Han Shi, Jiahui Gao, Hang Xu, Xiaodan Liang, Zhenguo Li, Lingpeng Kong, Stephen Lee, and James T Kwok. Revisiting over-smoothing in bert from the perspective of graph. 2022.

Yuandong Tian, Xinlei Chen, and Surya Ganguli. Understanding self-supervised learning dynamics without contrastive pairs. In *ICML*, 2021.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. In *ICLR*, 2018.

Feng Wang and Huaping Liu. Understanding the behaviour of contrastive loss. In *CVPR*, 2021.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere, 2020.

Yifei Wang, Yisen Wang, Jiansheng Yang, and Zhouchen Lin. Dissecting the diffusion process in linear graph convolutional networks. In *NeurIPS*, 2021.

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. In *ICLR*, 2022.

Zixin Wen and Yuanzhi Li. Toward understanding the feature learning process of self-supervised contrastive learning. In *ICML*, 2021.

Jim Winkens, Rudy Bunel, Abhijit Guha Roy, Robert Stanforth, Vivek Natarajan, Joseph R Ledsam, Patricia MacWilliams, Pushmeet Kohli, Alan Karthikesalingam, Simon Kohl, et al. Contrastive training for improved out-of-distribution detection. *arXiv preprint arXiv:2007.05566*, 2020.

Keyulu Xu, Chengtao Li, Yonglong Tian, Tomohiro Sonobe, Ken-ichi Kawarabayashi, and Stefanie Jegelka. Representation learning on graphs with jumping knowledge networks. In *ICML*, 2018.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In *ICLR*, 2019.

Lingxiao Zhao and Leman Akoglu. PairNorm: Tackling oversmoothing in gnns. *ICLR*, 2020.

Kuangqi Zhou, Yanfei Dong, Kaixin Wang, Wee Sun Lee, Bryan Hooi, Huan Xu, and Jiashi Feng. Understanding and resolving performance degradation in graph convolutional networks. In *CIKM*, 2021.

Zhihui Zhu, Tianyu Ding, Jinxin Zhou, Xiao Li, Chong You, Jeremias Sulam, and Qing Qu. A geometric analysis of neural collapse with unconstrained features. In *NeurIPS*, 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*, pp. 12979–12990. PMLR, 2021.

A PROOF OF MAIN THEORETICAL RESULTS

Proposition 1. *When the x-th row of F is defined as $F_x = \sqrt{w_x} f_\theta(x)$, we have:*

$$\mathcal{L}_{\text{nce}}(\theta) = \mathcal{L}_{\text{align}}(F_\theta) + \mathcal{L}_{\text{unif}}(F_\theta). \tag{17}$$

Proof. We first prove the two equations Eq.(2a) and Eq.(2b), which states the relations between F and $f(x)$.

$$\mathcal{L}_{\text{align}}(F) = \text{Tr}(F^\top L F) = \frac{1}{2}\mathbb{E}_{x,x^+}\|f(x) - f(x^+)\|^2, \tag{2a}$$

$$\mathcal{L}_{\text{unif}}(F) = \text{LSE}(FF^\top) - \|F\|^2 = \mathbb{E}_x \log \mathbb{E}_{x'} \exp(f(x)^\top f(x')) - \mathbb{E}_x\|f(x)\|^2, \tag{2b}$$

For the alignment loss, we have

$$\begin{aligned}
\text{Tr}(F^\top L F) &= \text{Tr}(F^\top (I - \bar{A}) F) \\
&= -\sum_{x,x'} \sqrt{w_x}\sqrt{w_{x'}} \bar{A}_{xx'} f(x)^\top f(x') + \sum_x w_x f(x)^\top f(x) \\
&= -\sum_x \sum_{x'} A_{xx'} f(x)^\top f(x') + \sum_x \sum_{x'} A_{xx'} f(x)^\top f(x) \\
&= -\sum_x \sum_{x'} A_{xx'} f(x)^\top f(x') + \frac{1}{2}\sum_x \sum_{x'} (A_{xx'} f(x)^\top f(x) + A_{x'x} f(x')^\top f(x')) \\
&= -\sum_x \sum_{x'} A_{xx'} f(x)^\top f(x') + \frac{1}{2}\sum_x \sum_{x'} (A_{xx'} f(x)^\top f(x) + A_{xx'} f(x')^\top f(x')) \\
&= \frac{1}{2}\sum_x \sum_{x'} A_{xx'} \|f(x) - f(x')\|^2 \\
&= \frac{1}{2}\mathbb{E}_{x,x^+}\|f(x) - f(x^+)\|^2
\end{aligned} \tag{19}$$

For the uniformity loss, we have

$$\begin{aligned}
\text{LSE}(FF^\top) - \|F\|^2 &= \text{Tr}(D \log(D \deg(\exp(D^{-1/2} F F^\top D^{-1/2})))) - \|F\|^2 \\
&= \sum_x w_x \log \sum_{x'} w_{x'} \exp(f(x)^\top f(x')) - \sum_x w_x \|f(x)\|^2 \\
&= \mathbb{E}_x \log \mathbb{E}_{x'} \exp(f(x)^\top f(x')) - \mathbb{E}_x \|f(x)\|^2.
\end{aligned} \tag{20}$$

Therefore, we have

$$\begin{aligned}
\mathcal{L}_{\text{nce}}(\theta) &= -\mathbb{E}_{x,x^+}[f(x)^\top f(x^+)] + \mathbb{E}_x \log \mathbb{E}_{x'}[\exp(f(x)^\top f(x'))] \\
&= -\mathbb{E}_{x,x^+}[f(x)^\top f(x^+)] + \mathbb{E}_x(f(x)^\top f(x)) + \mathbb{E}_x \log \mathbb{E}_{x'}[\exp(f(x)^\top f(x'))] - \mathbb{E}_x\|f(x)\|^2 \\
&= \frac{1}{2}\mathbb{E}_{x,x^+}\|f(x) - f(x^+)\|^2 + \mathbb{E}_x \log \mathbb{E}_{x'}[\exp(f(x)^\top f(x'))] - \mathbb{E}_x\|f(x)\|^2 \\
&= \mathcal{L}_{\text{align}}(F_\theta) + \mathcal{L}_{\text{unif}}(F_\theta).
\end{aligned} \tag{21}$$

□

Proposition 2. *Denote the distance between a feature matrix X and a subspace Ω by $d_\Omega(X) := \inf\{\|X - Y\|_F \mid Y \in \Omega\}$. Assume that data augmentations are label-preserving. Then, for samples in each class k, denoting their adjacency matrix as $\bar{A}_k$ and their features at t-th step as $F_k^{(t)}$, the message passing in Eq. 4 leads to a smaller or equal distance to an one-dimensional subspace $\mathcal{M}$*

$$d_{\mathcal{M}}(F_k^{(t+1)}) \leq |1 - 2\alpha\lambda_{\mathcal{G}_k}| \cdot d_{\mathcal{M}}(F_k^{(t)}), \tag{22}$$

where $\lambda_{\mathcal{G}_k} \in [0, 2]$ denotes the intra-class algebraic connectivity of $\mathcal{G}$ (Chung, 1997), and $\mathcal{M} = \{e_1 \otimes y | y \in \mathbb{R}^c\}$, and e_1 is the eigenvector corresponding to the largest eigenvalue of $\bar{A}_k$.

Proof. Under the condition that the data augmentations are label-preserving, there are no edges between inter-class nodes. Thus, the propagation of the alignment update only happens within each class. We denote the propagation matrix as $Q_k = (1-2\alpha)I + 2\alpha\bar{A}_k$. Then we have $F_k^{(t+1)} = Q_k F_k^{(t)}$. Since $\bar{A}_k$ is a symmetrically normalized adjacent matrix, we can denote the spectrum of $\bar{A}_k$ as $1 = \lambda_1 \geq \cdots \geq \lambda_n \geq -1$. Then the spectrum of Q_k is $q_i = 1 - 2\alpha + 2\alpha\lambda_i, i = 1, 2, \ldots, N$, and specifically, $q_1 = 1$. We denote the corresponding orthonormal basis of Q_k as $[e_1, \ldots, e_n]$ since Q_k is a symmetric matrix.

Accordingly, we could decompose $F_k^{(t)}$ according to the orthonormal basis as

$$F_k^{(t)} = \sum_{i=1}^{n} e_i \otimes h_i, h_i \in \mathbb{R}^m. \tag{23}$$

Therefore, the projection of $F_k^{(t)}$ to the space $\mathcal{M} = \{e_1 \otimes y | y \in \mathbb{R}^c\}$ gives the first component of the decomposition (Eq.23), *i.e.*, $e_1 \otimes h_1$, and the residual is the remaining components orthogonal to e_1, *i.e.*, $\sum_{i=2}^{n} e_i \otimes h_i$, with the following distance,

$$d^2_{\mathcal{M}}(F_k^{(t)}) = \left\| \sum_{i=2}^{n} e_i \otimes h_i, h_i \in \mathbb{R}^m \right\|_F^2 = \sum_{i=2}^{n} \|h_i\|_2^2. \tag{24}$$

Similarly, we can also deduce the distance of the updated features. Since $Qe_i = q_i e_i$, we have

$$QF_k^{(t)} = e_1 \otimes h_1 + \sum_{i=2}^{n} e_i \otimes (q_i h_i). \tag{25}$$

Thus, its projection to $\mathcal{M}$ is also $e_1 \otimes h_1$, and we have

$$d^2_{\mathcal{M}}(QF_k^{(t)}) = \left\| \sum_{i=2}^{n} e_i \otimes (q_i h_i) \right\|_F^2 = \sum_{i=2}^{n} \|q_i h_i\|_2^2. \tag{26}$$

Combining Eqs.24,26, we arrive at

$$d^2_{\mathcal{M}}(F_k^{(t+1)}) = d^2_{\mathcal{M}}(QF_k^{(t)}) = \sum_{i=2}^{n} \|q_i h_i\|_2^2 \leq \sum_{i=2}^{n} q_2^2 \|h_i\|_2^2 = q_2^2 d^2_{\mathcal{M}}(F_k^{(t)}). \tag{27}$$

Note that $1 - \lambda_2$ is just the algebraic connectivity of $\mathcal{G}$. Therefore, with $q_2 = 1 - 2\alpha + 2\lambda_2$, we have

$$d^2_{\mathcal{M}}(F_k^{(t+1)}) \leq (1 - 2\alpha\lambda^{in}_{\mathcal{G}})^2 d^2_{\mathcal{M}}(F_k^{(t)}) \leq d^2_{\mathcal{M}}(F_k^{(t)}), \tag{28}$$

By taking square root on this inequality, we arrive at the final result

$$d_{\mathcal{M}}(F_k^{(t+1)}) \leq |(1 - 2\alpha\lambda^{in}_{\mathcal{G}})| \cdot d_{\mathcal{M}}(F_k^{(t)}) \leq d_{\mathcal{M}}(F_k^{(t)}), \tag{29}$$

which completes the proof. □

Proposition 3. *The contrastive learning dynamics with the InfoNCE loss saturates when the real and the fake adjacency matrices agree, i.e.,* $\bar{A} = \bar{A}'_{\text{nce}}$*, which means that element-wisely, we have:*

$$\forall\, x, x^+ \in \mathcal{X},\ \ \mathcal{P}_d(x^+|x) = \mathcal{P}_\theta(x^+|x) \triangleq \frac{\exp(f_\theta(x)^\top f_\theta(x^+))}{\sum_{x'} \exp(f_\theta(x)^\top f_\theta(x'))}. \tag{30}$$

That is, the ground-truth data conditional distribution $\mathcal{P}_d(x^+|x)$ *equals to the Boltzmann distribution* $\mathcal{P}_\theta(x^+|x)$ *estimated from the encoder features.*

Proof. Notice that the contrastive update takes the form of

$$F^{(t+1)} = F^{(t)} - \alpha \nabla_{F^{(t)}} \mathcal{L}_{\text{nce}}(F^{(t)}) = F^{(t)} + 2\alpha(\bar{A} - \bar{A}'_{\text{nce}})F^{(t)}. \tag{31}$$

Therefore, when $\bar{A} = \bar{A}'_{\text{nce}}$, we have $F^{(t+1)} = F^{(t)}$, which means that the contrastive learning dynamics with the InfoNCE loss saturates. At this time, we have

$$\bar{A}(i,j) = \bar{A}_{\text{nce}}(i,j) = \frac{\exp(f(x)^\top f(x'))}{\sum_k \exp(f(x)^\top f(x_k))}, \tag{32}$$

which implies the result in the proposition. □

B EXPERIMENTAL DETAILS

B.1 DETAILS OF MULTI-STAGE GRAPH AGGREGATION

We follow the default settings of SimSiam (Chen & He, 2021), except the predictor. We use ResNet-50 as the backbone and have a two-layer projector following that. We pretrain the encoder for 100 epochs and then evaluate linear accuracy on the pretrained encoder. During the training process, we use memory banks to store the feature of latest k epochs. When aggregating the stored features, we set the weight of each feature as $\frac{1}{r}$.

B.2 DETAILS OF GRAPH ATTENTION EXPERIMENTS

We use SimCLR (Chen et al., 2020b) as our baseline and ResNet-18,ResNet-50 as the backbones. We follow the backbone with a projection MLP to map the features to the 128-dimensional projection space. We consider image classification tasks on CIFAR-10, CIFAR-100 and ImageNet-100. There are two stages in contrastive learning, *i.e.,* self-supervised pretraining and linear evaluation. We pretrain the encoder for 100 epochs on ImageNet-100 and for 200 epochs on CIFAR-10 and CIFAR-100. We use the LARS optimizer with cosine annealed learning rate schedule and 512 batch size. After the pretraining process, we train a linear classifier following the frozen backbones using the SGD optimizer. To obtain a more accurate estimation of semantic similarity, we use the features before the projection layer and warmup the model for 30 epochs.

C DISCUSSIONS ON RELATED WORK

Our work is closely to Wang & Isola (2020) that propose the alignment-and-uniformity (A+U) perspective on contrastive learning. Nevertheless, they only show that if an encoder with perfect A+U *exists*, it would be the minimizer of the InfoNCE loss (see Theorem 1 (Wang & Isola, 2020)). But what we are actually interested is the **opposite**: will the minimization of the InfoNCE loss guide us to the desired A+U properties? This is exactly the focus of our work by studying the learning dynamics of contrastive learning. Besides, we also extend the analysis A+U on the pretraining task (Wang & Isola, 2020) to its theoretical effect on downstream generalization as discussed in Propositions 2. Comparing to Wang et al. (2022), they only *assume* the existence of perfect alignment, while our analysis shows that the contrastive learning process helps improve feature clustering.

Another closely related work is HaoChen et al. (2021) that develops the augmentation graph framework that we build upon. Different from ours, they only characterize the optimal solution of the spectral loss $\mathcal{L}_{\mathrm{sp}}$ directly through eigen-decomposition (ED), which leads to following drawbacks. First, because N can be exponentially large, ED ($\mathcal{O}(N^3)$ complexity) is not directly applicable in practice. Instead, our analysis focuses on the scalable first-order optimization methods that are adopted in practice with good performance in limited training steps. Second, their theory only applies to the spectral loss and fails to characterize the widely adopted InfoNCE loss, which does not admit a closed-form solution. In comparison, our message passing analysis naturally incorporates the InfoNCE loss, as shown in Section 2.2.2.

Besides, Wen & Li (2021) study the contrastive learning process, but only with single-layer networks under toy problems, which is hardly practical. Tian et al. (2021) study the dynamics of non-contrastive learning, but only focus on the predictor parameters. Different from theirs, we adopt a feature space view and analyze the overall evolution of sample features for contrastive learning in practice. Many works before like Wang & Liu (2021) also intuitively discuss the behaviors of contrastive learning gradients. Instead, we provide a formal and systematic study of the learning dynamics of the samples features by characterizing its effect on the downstream performance. In particular, we establish an intriguing connection between contrastive learning and message passing, which is the first time that the two distinctive techniques from different areas are formally bridged together.

At last, similar to our Proposition 3, Zimmermann et al. (2021) also show that contrastive learning can recover the data distribution from a nonlinear ICA perspective. The main difference is that they derive this result based on the learning objective, while ours focuses on how the learning dynamics attains this equilibrium. Besides, to obtain this distribution recovery property, Zimmermann et al. (2021) rely heavily on the data generation process assumption, while ours does not.

D INCORPORATING MULTI-STAGE AGGREGATION TO EXISTING METHODS

In Section 3.3.1, we have shown that multi-stage aggeragtion, when used alone, can alleviate feature collapse without asymmetric modules. Nevertheless, we find that this technique alone is insufficient to attain state-of-the-art results. In view of this property, in this section, we further explore combining multi-stage aggregation with existing techniques to alleviate feature collapse, such as the predictor module in SimSiam (Chen & He, 2021). Hopefully, combining the two effective techniques can better avoid feature collapse and learn diverse features.

Setup. We use ResNet-18 as the backbone and train the model for 200 epochs. We use the SGD optimizer with 512 batch size and 0.0005 weight decay. For the multi-stage variant, after warmup for 50 epochs, we replace the canonical positive feature in SimSiam with a multi-stage aggeragated feature (Eq. 13). Specifically, we average positive features from the current epoch and the last epoch.

Table 2: Linear probing accuracy with 200-epoch training on CIFAR-10 using ResNet-18 backbone.

Method	SimSiam	SimSiam (multi-stage)
Linear Accuracy (%)	83.82	**84.75 (+0.93)**

Results. As shown in Table 2, we find that the linear accuracy increases from 83.82% to 84.75%, which clearly demonstrates the benefits of multi-stage aggregation on improving existing non-contrastive methods by further alleviating their feature collapse issue.