I-CON: A UNIFYING FRAMEWORK FOR REPRESENTATION LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

As the field of unsupervised learning grows, there has been a proliferation of different loss functions to solve different classes of problems. We find that a large collection of modern loss functions can be generalized by a single equation rooted in information theory. In particular, we introduce I-Con, a framework that shows that several broad classes of machine learning methods are precisely minimizing an integrated KL divergence between two conditional distributions: the supervisory and learned representations. This viewpoint exposes a hidden information geometry underlying clustering, spectral methods, dimensionality reduction, contrastive learning, and supervised learning. I-Con enables the development of new loss functions by combining successful techniques from across the literature. We not only present a wide array of proofs, connecting over 11 different approaches, but we also leverage these theoretical results to create state of the art unsupervised image classifiers that achieve a +8% improvement over the prior state-of-the-art on unsupervised classification on ImageNet-1K.

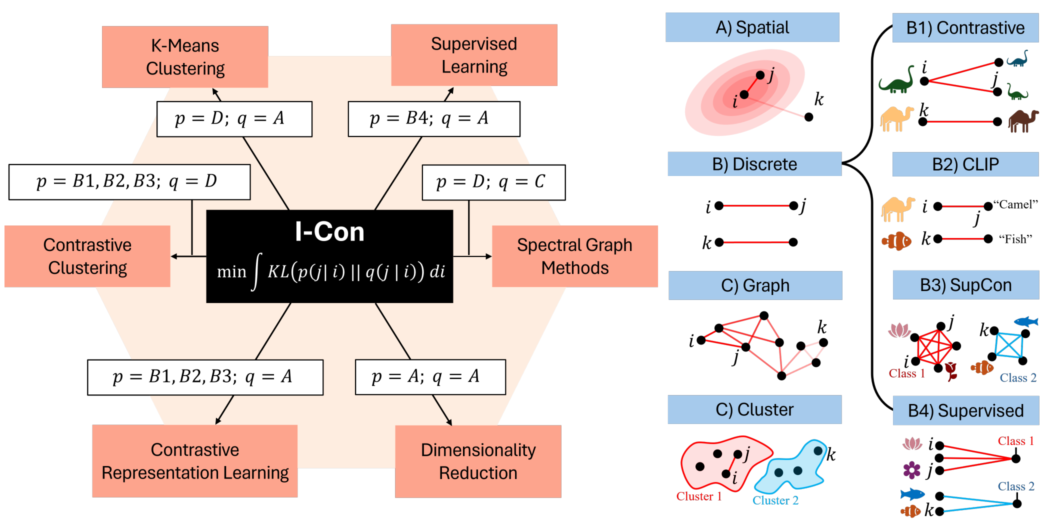

Figure 1: **I-Con unifies representation learning methods.** By choosing different types of conditional probability distributions over neighbors, I-Con generalizes over 11 commonly used representation learning methods.

1 INTRODUCTION

In the past 10 years the field of representation learning has flourished, with new techniques, architectures, and loss functions emerging daily. These advances have powered humanities' most intelligent models and have enabled machines to rely less and less on human supervision. However, as the field of unsupervised learning grows, the number of distinct and specific loss functions grows in turn making it difficult to understand which particular loss function to choose for a given problem or domain. In this work we propose a novel mathematical framework that unifies several broad classes of supervised and unsupervised representation learning systems with a single information-theoretic equation. Our framework, which we call Information Contrastive learning (I-Con) demonstrates that many methods in clustering, spectral graph theory, contrastive learning, dimensionality reduction, and supervised learning are all specific instances of the same underlying loss function. Though some specific connections implied by I-Con have been documented or approximated in the literature Balestriero & LeCun (2022); Yang et al. (2022); Böhm et al. (2022); Hu et al. (2022), to the best of the authors knowledge this is the first time the theory has been described in general. I-Con not only unifies a broad swath of literature but provides a framework to build and discover new loss functions and learning paradigms. In particular, the framework allows us to move techniques and results from any given method, to improve every other method in the broader class. We use this technique to derive new loss functions for unsupervised image classification that significantly outperform the prior art on several standard datasets. We summarize our contributions:

- We present I-Con, a single equation that unifies several broad classes of methods in representation learning
- We prove 9 theorems which connect a variety of methods to the I-Con framework
- We use I-Con to derive new improvements for unsupervised image classification and achieve a +8% increase in unsupervised ImageNet-1K accuracy over the prior state-of-the-art
- We carefully ablate our discovered improvements, demonstrating their efficacy.

2 Related Work

Representation learning is a vast field with thousands of methods, we overview some of the key methods that I-Con leverages and generalizes. We refer the interested reader to Le-Khac et al. (2020); Bengio et al. (2013); Weng (2021) for more complete reviews of the representation and contrastive learning literature.

Feature Learning aims to learn informative low-dimensional continuous vectors from high dimensional data. Feature learners come in a variety of flavors, using supervisory signals like distance in a high dimensional space, nearest neighbors, known positive and negative pairs, auxiliary supervised losses, and reconstruction loss. The most common methods learn directly from distances between points in high dimensional space such as PCA Pearson (1901) which optimizes for reconstruction error, MDS Kruskal (1964) which preserves distances between points. Other approaches try to match pairwise high-dimensional distances, neighborhoods, or topological structure with low dimensional vectors. Techniques include UMAP McInnes et al. (2018) which preserves a soft topology of the points, and SNE/t-SNE Hinton & Roweis (2002); Van der Maaten & Hinton (2008) that use a KL divergence to align joint distributions across low and high dimensional views of the data. SNE and t-SNE were some of the first works to explicitly phrase their optimization in terms of KL minimization between two joint distributions, which is the central idea of I-Con. Methods like SimCLR Chen et al. (2020a), CMC Tian et al. (2020), CLIP Radford et al. (2021), MoCo v3 Chen* et al. (2021), and others use positive and negative pairs of data, often formed through augmentations or aligned corpora to drive feature learning. I-Con generalizes all of these frameworks and through our analysis the subtle differences in how they implicitly construct their losses becomes apparent. Finally, one of the most famous approaches for representation learning in fields like computer vision, uses the penultimate activations of a supervised classifier as informative features Krizhevsky et al. (2017). Interestingly, I-Con generalizes this case as well if we consider discrete class labels as "points" in a contrastive learning setup. This provides an intuitive justification for why penultimate activations of supervised activations make high quality representations.

Clustering aims to learn a discrete representation for data, again using distances in ambient space, nearest neighbors, or contrastive supervision. Classic methods like k-Means Macqueen (1967) and EM Dempster et al. (1977) implicitly fit cluster distributions to data points to maximize data likelihood. Spectral Clustering Shi & Malik (2000) uses the spectra of the graph Laplacian to cut a graph into two strongly connected components. Methods like IIC Ji et al. (2019) and Contrastive Clustering Li et al. (2021) use augmentation invariance to drive learning. SCAN Gansbeke et al. (2020) realized that including nearest neighbors as contrastive positive pairs could improve clustering and

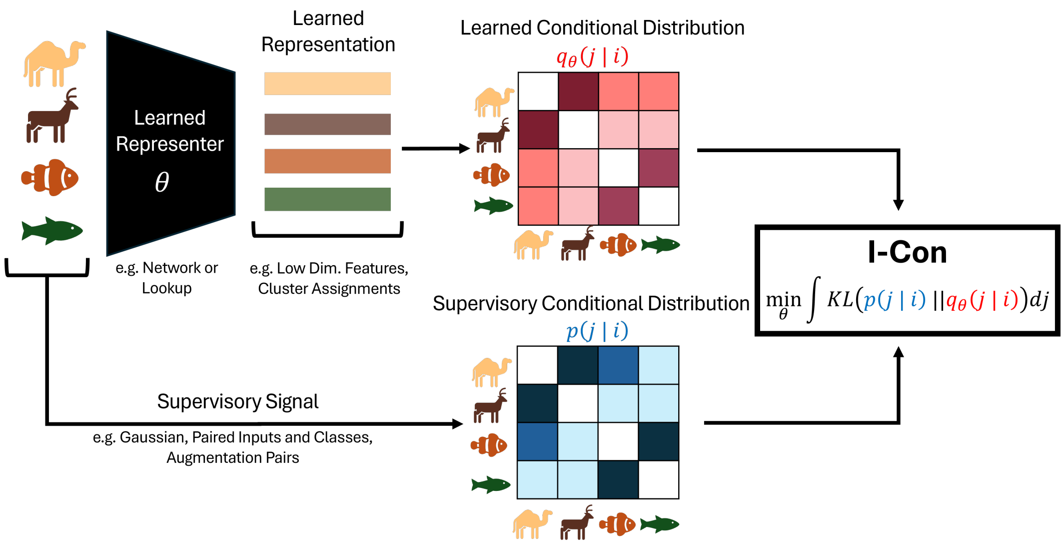

Figure 2: Architecture diagram of an I-Con model. I-Con aligns a parameterized neighborhood distribution computed from a learned representation, with a supervisory neighborhood distribution chosen by the method designer.

most recently TEMI Adaloglou et al. (2023) shows that student-teacher style EMA architectures Chen et al. (2020b) can also improve outcomes. I-Con generalizes many of these methods, by aligning a cluster-induced joint distribution with a supervisory joint distribution derived from either distances, the graph Laplacian, or contrastive pairs. Improvements like EMA-style architectures can be included naturally in I-Con as different parameterizations of the clusters that are optimized with the central I-Con loss.

Unifying unsupervised learning methods has been a goal of several existing and seminal works in the literature. Hu et al. (2022) discovered that contrastive learning and t-SNE could be seen as two different aspects of the same loss. Yang et al. (2022) found that cross-entropy and contrastive learning could be unified by considering different kinds of neighbor relations between points. Balestriero & LeCun (2022) found approximate connections between spectral methods and contrastive learning. Grosse et al. (2012) showed how many common classical unsupervised learners can be derived with Bayesian tensor factorization grammars. These prior works are elegant and impactful, however to the best of our knowledge we are the first to describe the unification of supervised, contrastive, dimensionality reduction, spectral graph, and clustering methods using a single KL divergence loss.

3 Methods

The I-Con framework unifies several representation learning methods using a single loss function: minimizing the KL divergence between a pair of conditional "neighborhood distributions" which measure the probability of transition from a data point i to a point j. I-Con's single information-theoretic objective generalizes methods from the fields of clustering, contrastive learning, dimensionality reduction, spectral graph theory, and supervised learning. By choosing how we construct the supervisory neighborhood distribution, and parameterize the neighborhood distribution of the learned representation, we can create a broad class of existing and novel methods using I-Con. We first introduce I-Con, then use the framework to aggregate important techniques from across the broader literature to make a novel state of the art unsupervised image classification method.

3.1 Information Contrastive Learning

We begin by defining the mathematical objects of interest. Let $i, j \in \mathcal{X}$ represent two abstract points of a broader set $\mathcal{X}$. We can then form a probabilistic "neighborhood" around a point j using a function $p(j|i)$. Intuitively, this function measures the probability to "transition" from a point i to another point $j \in \mathcal{X}$. To ensure this function is a proper probability density over $\mathcal{X}$ it should be non-negative: $p(j|i) \geq 0$, and sum to unity: $\int_{j \in \mathcal{X}} p(j|i) = 1$. Here we use the measure-theoretic integral, which includes both the continuous integral and discrete summation depending on the choice of the space $\mathcal{X}$. Next, we parameterize this neighborhood distribution by abstract parameters, $\theta \in \Theta$. Note that $p_\theta(j|i)$ should be a distribution for all $\theta \in \Theta$. This parameterization transforms p into a *learnable* distribution that can adapt the neighborhoods around each point. Next, let $q_\phi(j|i)$ be a similarly defined family of distributions parameterized by an abstract parameter space $\phi \in \Phi$. With these two families of neighborhood distributions defined we can write the main loss function of I-Con:

$$\mathcal{L}(\theta, \phi) = \int_{i \in \mathcal{X}} D_{\mathrm{KL}}\left(p_\theta(\cdot|i) || q_\phi(\cdot|i)\right) = \int_{i \in \mathcal{X}} \int_{j \in \mathcal{X}} p_\theta(j|i) \log \frac{p_\theta(j|i)}{q_\phi(j|i)} \tag{3.1.1}$$

Where for clarity, D_{KL}, represents the Kullback-Leibler divergence Kullback & Leibler (1951). Intuitively this loss function measures the average similarity between the two parameterized neighborhood distributions and is minimized when $p_\theta(j|i) = q_\phi(j|i)$. In practice, one of the distributions usually p, is set to a fixed "supervisory" distribution with no optimizable parameters θ. We will sometimes omit the parameterization in this case refer to it as $p(j|i)$. In these scenarios the remaining distribution, q_ϕ, is parameterized by a comparison of deep network representations or a comparison of prototypes, clusters, or per-point representations. We illustrate this architecture in Figure 2. Minimizing equation 3.1.1 aligns this "learned" distribution q_θ to the "supervisory" distribution p by minimizing the average KL divergence between p and q. In the next section, we will show that by selecting different types of parameterized neighborhood distributions p and q, several common methods in the literature emerge as special cases. Interestingly, we note that it is possible to optimize both p_θ and q_ϕ even though no existing methods use this generality. This is an interesting avenue for future study, though we caution that if a uniform distribution is possible in both families of distributions, the optimization will find a trivial solution. Nevertheless, it could be possible to choose the families of distributions p_θ and q_ϕ carefully so that useful behavior emerges.

3.2 Unifying Representation Learning Algorithms with I-Con

Despite the incredible simplicity of Equation 3.1.1, this equation is rich enough to generalize several existing methods in the literature simply through the choice of parameterized neighborhood distributions p_θ and q_ϕ. Table 1 summarizes some key choices which recreate popular methods from contrastive learning (SimCLI, MOCOv3, SupCon, CMC, CLIP), dimensionality reduction (SNE, t-SNE), clustering (K-Means, Spectral, TEMI), and supervised learning (Cross-Entropy and Mean Squared Error). Due to limited space, we defer proofs of each of these theorems to the supplemental material. We also note that Table 1 is most certainly not exhaustive, and encourage the community to explore whether other unsupervised learning frameworks implicitly minimize Equation 3.1.1 for some choice of p and q.

Though there are too many methods unified by I-Con to explain each in detail here, we give an intuitive explanation of how the various choices of p and q generalize SNE and InfoNCE to help the reader gain intuition. The simplest and most direct method to generalize with the I-Con loss is SNE, which was originally phrased as a KL divergence minimization problem. Given a n-dimensional dataset of d vectors, $x \in \mathbb{R}^{d \times n}$, SNE aims to learn a m-dimensional vector representation, $\phi \in \mathbb{R}^{d \times m}$, such that local relationships between high dimensional datapoints are approximately preserved in the low-dimensional representation. The challenge is that the representation dimension, m, is usually much smaller than the data dimensionality n, so the learned representation is significantly constrained. More formally, SNE constructs a probabilistic neighborhood function, $p(j|i)$, around a point x_i, by placing a symmetric Gaussian at x_i and evaluating this distribution at candidate neighbors x_j. It does the same in the low dimensional space to create $q_\phi(j|i)$ by placing a Gaussian at the learned representation vector ϕ_i and comparing to ϕ_j. Finally, SNE learns the representation parameters ϕ to minimize the average KL Divergence, which is exactly the I-Con loss function.

We only need to slightly modify SNE to derive the InfoNCE loss used in contrastive learning approaches like SimCLR and MocoV3. The first difference is that contrastive learners don't typically

Method	Choice of $p_\theta(j\|i)$	Choice of $q_\phi(j\|i)$
SNE Hinton & Roweis (2002) Theorem 1	Gaussians on Datapoints, x_i $\frac{\exp\left(-\\|x_i - x_j\\|^2/2\sigma_i^2\right)}{\sum_{k\neq i}\exp\left(-\\|x_i - x_k\\|^2/2\sigma_i^2\right)}$	Gaussians on Learned Vectors, ϕ_i $\frac{\exp\left(-\\|\phi_i - \phi_j\\|^2\right)}{\sum_{k\neq i}\exp\left(-\\|\phi_i - \phi_k\\|^2\right)}$
tSNE Van der Maaten & Hinton (2008) Theorem 2	Perplexity-sized Gaussians on Datapoints, x_i $\frac{\exp\left(-\\|x_i - x_j\\|^2/2\sigma_i^2\right)}{\sum_{k\neq i}\exp\left(-\\|x_i - x_k\\|^2/2\sigma_i^2\right)}$	Cauchy Distributions on Learned Vectors, ϕ_i $\frac{\left(1+\\|\phi_i - \phi_j\\|^2\right)^{-1}}{\sum_{k\neq i}\left(1+\\|\phi_i - \phi_k\\|^2\right)^{-1}}$
InfoNCE Bachman et al. (2019) Theorem 3	Uniform over positive pairs $\frac{1}{Z}\mathbb{1}[i \text{ and } j \text{ are positive pairs }]$	Gaussian based on deep features, $f_\phi(x_i)$ $\frac{\exp\left(-\\|f_\phi(x_i) - f_\phi(x_j)\\|^2\right)}{\sum_{k\neq i}\exp\left(-\\|f_\phi(x_i) - f_\phi(x_k)\\|^2\right)}$
SupCon Khosla et al. (2020) Theorem 3	Uniform over classes $\frac{1}{Z}\mathbb{1}[i \text{ and } j \text{ have the same class }]$	Gaussian based on deep features, $f_\phi(x_i)$ $\frac{\exp\left(-\\|f_\phi(x_i) - f_\phi(x_j)\\|^2\right)}{\sum_{k\neq i}\exp\left(-\\|f_\phi(x_i) - f_\phi(x_k)\\|^2\right)}$
InfoNCE Clustering (New in this work)	Uniform over positive pairs $\frac{1}{Z}\mathbb{1}[i \text{ and } j \text{ are positive pairs }]$	Shared cluster likelihood by point $\frac{f_\phi(x_i)\cdot f_\phi(x_j)}{\mathbb{E}[\text{size of } x_i\text{'s cluster w.r.t } f_\phi]}$
Probabilistic k-Means Macqueen (1967) Theorem 6	Shared cluster likelihood by cluster $\sum_{c=1}^{m}\frac{p(f_\theta(x_i) \text{ and } f_\theta(x_j) \text{ are in cluster } c)}{\mathbb{E}[\text{size of cluster } c]}$	Gaussians on Datapoints, x_i $\frac{\exp\left(-\\|x_i - x_j\\|^2/2\sigma_i^2\right)}{\sum_{k\neq i}\exp\left(-\\|x_i - x_k\\|^2/2\sigma_i^2\right)}$
Normalized Cuts Shi & Malik (2000) Theorem 8	Degree-weighted shared cluster likelihood by cluster $\sum_{c=1}^{m}\frac{p(f_\theta(x_i) \text{ and } f_\theta(x_j) \text{ are in cluster } c)\cdot d_j}{\mathbb{E}[\text{degree of members of cluster } c]}$	Gaussians on edge weights $\frac{\exp(w_{ij}/d_j)}{\sum_k \exp(w_{ik}/d_k)}$
Mutual Information Clustering Adaloglou et al. (2023) Theorem 7	Uniform over nearest neighbors $\frac{1}{k}\mathbb{1}[j \text{ is a k-nearest neighbor of } i]$	Shared cluster likelihood by cluster $\sum_{c=1}^{m}\frac{p(f_\phi(x_i) \text{ and } f_\phi(x_j) \text{ are in cluster } c)}{\mathbb{E}[\text{size of cluster } c]}$
CMC & CLIP Tian et al. (2020) Theorem 4	Uniform over positive pairs from different modalities V $\frac{1}{Z}\mathbb{1}[i \text{ and } j \text{ are positive pairs and } V_i \neq V_j]$	Gaussian based on deep features, $f_\phi(x_i)$ $\frac{\exp\left(-\\|f_\phi(x_i) - f_\phi(x_j)\\|^2\right)}{\sum_{k\in V_j}\exp\left(-\\|f_\phi(x_i) - f_\phi(x_k)\\|^2\right)}$
Cross Entropy Good (1963) Corollary 1	Indicator function over Labels $\frac{1}{\|C\|}\mathbb{1}[i \in D \text{ is a data point in a class } j \in C]$	Gaussian based on deep features $\frac{\exp\left(f_\phi(x_i)\cdot\phi_j\right)}{\sum_{k\in C}\exp\left(f_\phi(x_i)\cdot\phi_k\right)}$

Table 1: **I-Con unifies representation learners** under different choices of $p_\theta(j|i)$ and $q_\phi(j|i)$. Proofs of the propositions in this table can be found in the supplement.

learn a separate representation ϕ_i for every datapoint x_i, but instead learn a parameterized representation function $f_\phi(x_i)$ to create representations for data. Secondly, these methods don't rely on Gaussian neighborhoods in the original data space, instead they use a discrete neighborhood of known positive pairs for each point x_i. In practice these positive pairs are usually formed by augmenting or transforming data, such as horizontally flipping or blurring images. When the KL divergence is taken between this discrete neighborhood and the Gaussian neighborhoods in deep feature space, we precisely re-derive the InfoNCE loss function. To create MocoV3, we use a student model $f_\phi(x_i)$ to featurize a point and an exponential moving averaged teacher model $g(x_j)$ to represent the neighboring point.

3.3 CREATING NEW REPRESENTATION LEARNERS WITH I-CON

I-Con not only allows one to generalize many methods with a single equation but allows one to transfer insights across different domains of representation learning. This allows techniques from one area, like contrastive learning, to improve methods in another area like clustering. In this work we show that by surveying modern dimensionality and representation learners we can create new clustering and unsupervised classification methods that perform much better than the prior art. In particular, we transfer intuitions from spectral clustering, t-SNE, debiased contrastive learning Chuang et al. (2020) and SCAN to create a state-of-the-art unsupervised image classification system.

Adaptive neighborhoods like SNE and t-SNE take a great advantage from having adaptive neighborhood specified by, "perplexity", which is an effective measure of local neighbors instead of using fixed-variance Gaussians which might affect the representation of points in a high density neighborhoods, and other approximations of t-SNE approximate the Gaussian distribution directly with a uniform distribution over KNN neighbors to handle large datasets. Similarly, Table 1 shows that by swapping k-Means' Gaussian neighborhoods to (degree-weighted) KNN neighborhoods we rederive Spectral clustering, which is well known for its flexibility and quality. We leverage this insight, and Tables 3 and 4 shows that training a contrastive learner with KNN-based neighborhood distributions yields a significant improvement on unsupervised image classification.

Debiased Contrastive Learning aims to correct for the fact that contrastive learning typically uses random points as negative samples. If a dataset has a small number of underlying classes, this approach significantly overestimates the negative terms in contrastive learning. Chuang et al. (2020) show that by solving this problem one can improve contrastive representation learning across backbones and datasets. We leverage this technique in I-Con by adding a 'debiasing' neighborhood to the original contrastive training neighborhood $p(j|i)$:

$$\tilde{p}(j|i) = (1-\alpha)p(j|i) + \frac{\alpha}{N} \tag{3.3.1}$$

Where α controls the amount of debiasing and N is the number of points in the neighborhood of point i. Here the factor of N ensures the neighborhood probability distribution stays normalized. Intuitively this modification adds an $\frac{\alpha}{N}$ amount of probability to each negative pair, counteracting the aforementioned bias effects. Unlike the original formulation in Chuang et al. (2020), this technique can now apply to any other class of methods I-Con generalizes including clusters and dimensionality reducers. In Tables 3 and Figures 4 and 3 we show that this has a net positive effect across all experiments and batch sizes tested. It also has the effect of relaxing the stiff clustering optimization, similar to how label smoothing Szegedy et al. (2016) can improve model distillation and generalization. We also explore debiasing the learned distribution as well as the supervisory distribution, which also yields a performance improvement. This is in direct analogy to t-SNE's long tailed Cauchy distributions in the learned neighborhoods. Like in t-SNE this addition helps the optimization find good local minima and avoid saturated solutions with vanishing gradients. This can be seen both quantitatively and qualitatively in Figure 3

Neighbor Propagation as Kernel Smoothing in I-Con Another widespread technique in dimensionality reduction, clustering, and contrastive learning is the use of nearest neighbors in deep feature space to form positive pairs. Within the I-Con framework, this can be seen as an additional form of kernel smoothing. For example, in contrastive learning, the target probabilities can be smoothed by not only considering augmentations but also nearest neighbors in deep feature space. We can also form walks on the nearest neighbor graph similiar to the successful Word-Graph2Vec algorithm Li et al. (2023). We refer to this as *neighbor propagation*, and note that it significantly improves performance.

The conditional distribution matrix P, which defines the probability of selecting x_j as a neighbor of x_i (i.e., $P_{ij} = p(x_j|x_i)$), can be interpreted as an adjacency matrix for the training data. A neighbor propagation smoothing of this target distribution is established by considering the number of walks of length at most k between points x_i and x_j:

$$\tilde{P} \propto P + P^2 + \cdots + P^k$$

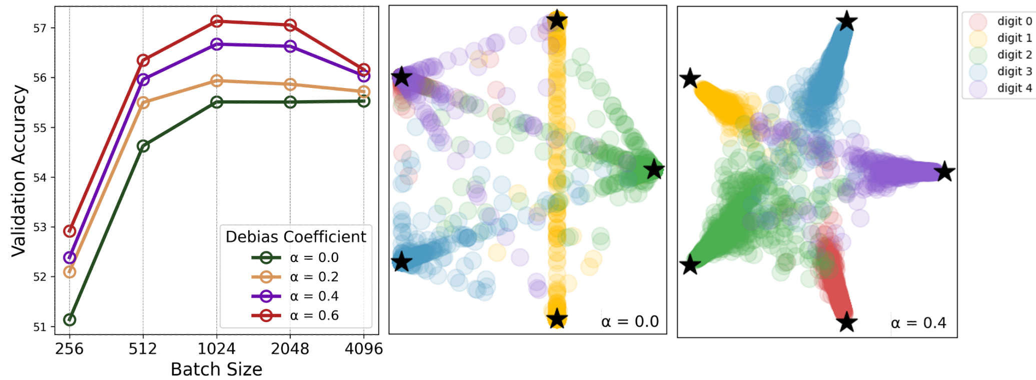

Figure 3: Left: Debiasing cluster learning improves performance across on ImageNet-1K batch sizes. Center and Right: Debiased training reduces optimization stiffness and yields solutions that are less likely to encounter saturated logits and vanishing gradients. These visualizations use a subset of MNIST Deng (2012).

Further smoothing can be applied by transforming the probabilities into a uniform distribution over neighbors reachable within k steps, leading to the following transformation:

$$\tilde{P_U} \propto I[P + P^2 + \cdots + P^k > 0]$$

This type of smoothing, based on propagation through nearest neighbors and walk-based approaches, effectively broadens the neighborhood structure considered during learning, allowing the model to capture richer relationships within the data.

4 Experiments

The primary objective of this work is to demonstrate that the I-Con framework offers testable hypotheses and practical insights into self-supervised and unsupervised learning. Rather than aiming only for state-of-the-art performance, our goal is to show how I-Con can enhance existing unsupervised learning methods by leveraging a unified information-theoretic approach. Through this framework, we also highlight the potential for cross-pollination between techniques in varied machine learning domains, such as clustering, contrastive learning, and dimensionality reduction. This transfer of techniques, enabled by I-Con, can significantly improve existing methodologies and open new avenues for exploration.

We focus our experiments on clustering because it is relatively understudied compared to contrastive learning and there are a variety of techniques that can now be adapted to this task. By connecting established methods such as k-Means, SimCLR, and t-SNE within the I-Con framework, we uncover a wide range of possibilities for improving clustering methods. We validate these theoretical insights experimentally, demonstrating the practical impact of I-Con.

We evaluate the I-Con framework using the ImageNet-1K dataset Deng et al. (2009), which consists of 1,000 classes and over one million high-resolution images. This dataset is considered one of the most challenging benchmarks for unsupervised image classification due to its scale and complexity. To ensure fair comparison with prior work, we strictly adhere to the experimental protocol introduced by Adaloglou et al. (2023). The primary metric for evaluating clustering performance is Hungarian Accuracy, which measures the quality of cluster assignments by finding the optimal alignment between predicted clusters and ground truth labels via the Hungarian algorithm Ji et al. (2019). This approach provides a robust measure of clustering performance in an unsupervised context, where direct label supervision is absent during training.

For feature extraction, we utilize the DiNO pre-trained Vision Transformer (ViT) models in three variants: ViT-S/14, ViT-B/14, and ViT-L/14 Caron et al. (2021). These models are chosen to ensure comparability with previous work and to explore how the I-Con framework performs across varying

Method	DiNO ViT-S/14	DiNO ViT-B/14	DiNO ViT-L/14
k-Means	51.84	52.26	53.36
Contrastive Clustering	47.35	55.64	59.84
SCAN	49.20	55.60	60.15
TEMI	56.84	58.62	–
I-Con (Ours)	**58.52**	**65.03**	**68.01**

Table 2: Comparison of methods on ImageNet-1K clustering with respect to Hungarian Accuracy. I-Con significantly outperforms the prior state-of-the-art TEMI. Note that TEMI does not report results for ViT-L.

model capacities. The experimental setup, including training protocols, optimization strategies, and data augmentation, mirrors those used in TEMI to ensure consistency in methodology.

The training process involved optimizing a linear classifier on top of the features extracted by the DiNO models. Each model was trained for 30 epochs, using ADAM Kingma & Ba (2017) with batch size of 4096 and an initial learning rate of 1e-3. The learning rate was decayed by a factor of 0.5 every 10 epochs to allow for stable convergence. Notably, no additional normalization was applied to the feature vectors. During training, we applied a variety of data augmentation techniques, including random re-scaling, cropping, color jittering, and Gaussian blurring, to create robust feature representations. Furthermore, to enhance the clustering performance, we pre-computed global nearest neighbors for each image in the dataset using cosine similarity. This allowed us to sample two augmentations and two nearest neighbors for each image in every training batch, thus incorporating both local and global information into the learned representations. We refer to our approach we derived in Table 2 as I-Con. In particular we use a supervisory neighborhood comprised of augmentations, KNNs ($k = 3$), and KNN walks of length 1. We use the 11shared cluster likelihood by cluster" neighbourhood from k-Means (See table 1 for Equation) as our learned neighborhood function to drive cluster learning.

4.1 Baselines

We compare our method against several state-of-the-art clustering methods, including TEMI, SCAN, IIC, and Contrastive Clustering. These methods rely on augmentations and learned representations but often require additional regularization terms or loss adjustments, such as controlling cluster size or reducing the weight of affinity losses. In contrast, our I-Con-based loss function is self-balancing and does not require such manual tuning, making it a cleaner, more theoretically grounded approach. This allows us to achieve higher accuracy and more stable convergence across three different sized backbones.

4.2 Results

Table 2 shows the Hungarian accuracy of I-Con across different DiNO variants (ViT-S/14, ViT-B/14, ViT-L/14) and compares it with several state-of-the-art clustering methods. The I-Con framework consistently outperforms the other state-of-the-art methods across all model sizes. Specifically, for the DiNO ViT-B/14 and ViT-L/14 models, I-Con achieves significant performance gains of +4.5% and +7.8% in Hungarian accuracy compared to TEMI, the prior state-of-the-art ImageNet clusterer. The improvements in performance can be attributed to two main factors:

Self-Balancing Loss: Unlike TEMI or SCAN, which require hand-tuned regularizations (e.g., balancing cluster sizes or managing the weight of affinity losses), I-Con's loss function automatically balances these factors without additional hyper-parameter tuning as we are using the exact same clustering kernel used by k-Means. This theoretical underpinning leads to more robust and accurate clusters.

Cross-Domain Insights: I-Con leverages insights from contrastive learning to refine clustering by looking at pairs of images based on their embeddings, treating augmentations and neighbors similarly. This approach, originally successful in contrastive learning, translates well into clustering and leads to improved performance in high-dimensional, noisy image data.

Method	DiNO ViT-S/14	DiNO ViT-B/14	DiNO ViT-L/14
Baseline	55.51	63.03	65.70
+ Debiasing	57.05	63.77	66.69
+ KNN Propagation	**58.52**	64.87	67.35
+ EMA	57.62	**65.03**	**68.01**

Table 3: Ablation study of new techniques discovered through the I-Con framework. We compare ImageNet-1K clustering accuracy across different sized backbones.

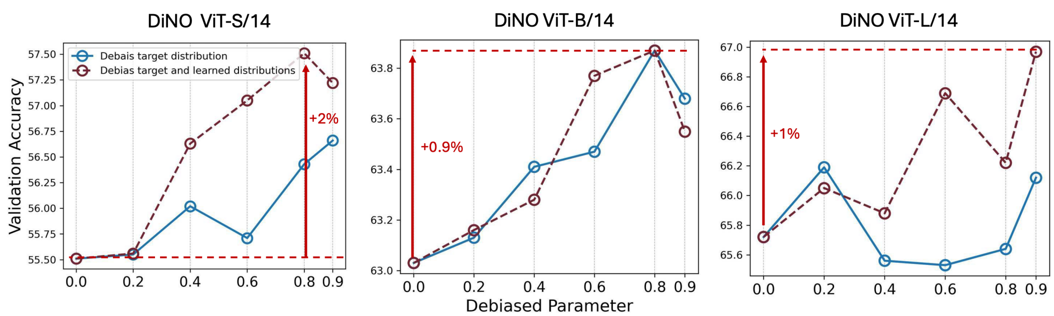

Figure 4: Effects of increasing the debias weight α on the supervisory neighborhood (blue line) and both the learned and supervisory neighborhood (red line). Adding some amount of debiasing helps in all cases, with a double debiasing yielding the largest improvements.

Method	DiNO ViT-S/14	DiNO ViT-B/14	DiNO ViT-L/14
Baseline	55.51	63.03	65.72
+ KNNs	56.43	64.26	65.70
+ 1-walks on KNN	**58.09**	**64.29**	65.97
+ 2-walks on KNN	57.84	64.27	**67.26**
+ 3-walks on KNN	57.82	64.15	67.02

Table 4: Ablation Study on Neighbor Propagation. Adding both KNNs and walks of length 1 or 2 on the KNN graph achieves the best performance.

4.3 Ablations

We conduct several ablation studies to experimentally justify the architectural improvements that emerged from analyzing contrastive clustering through the I-Con framework. These ablations focus on two key areas: the effect of incorporating debiasing into the target and embedding spaces and the impact of neighbor propagation strategies which are both kernel smoothing methods.

We perform experiments with different levels of debiasing in the target distribution, denoted by the parameter α, and test configurations where debiasing is applied on either the target side, both sides (target and learned representations), or none. As seen in Figure 4, adding debiasing improves performance, with the optimal value typically around $\alpha = 0.6$ to $\alpha = 0.8$, particularly when applied to both sides of the learning process. This method is similar to how debiasing work in contrastive learning by assuming that each negative sample has a non-zero probability (α/N) of being incorrect. Figure 3 shows how changing the value of α improves performance across different batch sizes.

In a second set of experiments, shown in Table 4, we examine the impact of neighbor propagation strategies. We evaluate clustering performance when local and global neighbors are included in the contrastive loss computation. Neighbor propagation, especially at small scales ($s = 1$ and $s = 2$), significantly boosts performance across all model sizes, showing the importance of capturing local structure in the embedding space. Larger neighbor propagation values (e.g., $s = 3$) offer diminishing returns, suggesting that over-propagating neighbors may dilute the information from the nearest, most relevant points. Note that only DiNO-L/14 showed preference for large step size, and this is likely due to its higher k-nearest neighbor ability, so the augmented links are correct.

Our ablation studies highlight that small adjustments in the debiasing parameter and neighbor propagation can lead to notable improvements that achieve a state-of-the-art result with a simple loss function. Additionally, sensitivity to α and propagation size varies across models, with larger models generally benefiting more from increased propagation but requiring fine-tuning of α for optimal performance. We recommend using $\alpha \approx 0.6$ to $\alpha \approx 0.8$ and limiting neighbor propagation to small values for a balance between performance and computational efficiency.

5 Conclusion

In summary, we have developed I-Con: a single information theoretic equation that unifies a broad class of machine learning methods. We provided over 9 theorems that prove this assertion for many of the most popular loss functions used in clustering, spectral graph theory, supervised and unsupervised contrastive learning, dimensionality reduction, and supervised classification and regression. We not only theoretically unify these algorithms but show that our connections can help us discover new state-of-the-art methods, and apply improvements discovered for a particular method to any other method in the class. We illustrate this by creating a new method for unsupervised image classification that achieves a +8% improvement over the prior art. We believe that the results presented in this work represent just a fraction of the methods that are potentially unify-able with I-Con, and we hope the community can use this viewpoint to improve collaboration and analysis across algorithms and machine learning disciplines.

References

Nikolas Adaloglou, Felix Michels, Hamza Kalisch, and Markus Kollmann. Exploring the limits of deep image clustering using pretrained models. *arXiv preprint arXiv:2303.17896*, 2023.

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. *Advances in Neural Information Processing Systems*, 35:26671–26685, 2022.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

Jan Niklas Böhm, Philipp Berens, and Dmitry Kobak. Unsupervised visualization of image datasets using contrastive learning. *arXiv preprint arXiv:2210.09879*, 2022.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers, 2021. URL `https://arxiv.org/abs/2104.14294`.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020a.

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

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

Ching-Yao Chuang, Joshua Robinson, Lin Yen-Chen, Antonio Torralba, and Stefanie Jegelka. Debiased contrastive learning, 2020. URL `https://arxiv.org/abs/2007.00224`.

Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society: series B (methodological)*, 39(1): 1–22, 1977.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Li Deng. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, Marc Proesmans, and Luc Van Gool. Scan: Learning to classify images without labels, 2020. URL `https://arxiv.org/abs/2005.12320`.

Irving J Good. Maximum entropy for hypothesis formulation, especially for multidimensional contingency tables. *The Annals of Mathematical Statistics*, pp. 911–934, 1963.

Roger Grosse, Ruslan R Salakhutdinov, William T Freeman, and Joshua B Tenenbaum. Exploiting compositionality to explore a large space of model structures. *arXiv preprint arXiv:1210.4856*, 2012.

Geoffrey E Hinton and Sam Roweis. Stochastic neighbor embedding. *Advances in neural information processing systems*, 15, 2002.

Tianyang Hu, Zhili Liu, Fengwei Zhou, Wenjia Wang, and Weiran Huang. Your contrastive learning is secretly doing stochastic neighbor embedding. *arXiv preprint arXiv:2205.14814*, 2022.

Xu Ji, Joao F Henriques, and Andrea Vedaldi. Invariant information clustering for unsupervised image classification and segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9865–9874, 2019.

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.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017. URL `https://arxiv.org/abs/1412.6980`.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

Joseph B Kruskal. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29(1):1–27, 1964.

Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

Phuc H Le-Khac, Graham Healy, and Alan F Smeaton. Contrastive representation learning: A framework and review. *Ieee Access*, 8:193907–193934, 2020.

Wenting Li, Jiahong Xue, Xi Zhang, Huacan Chen, Zeyu Chen, Feijuan Huang, and Yuanzhe Cai. Word-graph2vec: An efficient word embedding approach on word co-occurrence graph using random walk technique, 2023. URL `https://arxiv.org/abs/2301.04312`.

Yunfan Li, Peng Hu, Zitao Liu, Dezhong Peng, Joey Tianyi Zhou, and Xi Peng. Contrastive clustering. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 8547–8555, 2021.

J Macqueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability/University of California Press*, 1967.

Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

Karl Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, 2(11):559–572, 1901.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. URL `https://arxiv.org/abs/2103.00020`.

Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Transactions on pattern analysis and machine intelligence*, 22(8):888–905, 2000.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pp. 776–794. Springer, 2020.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

Lilian Weng. Contrastive representation learning. *lilianweng.github.io*, May 2021. URL `https://lilianweng.github.io/posts/2021-05-31-contrastive/`.

Jianwei Yang, Chunyuan Li, Pengchuan Zhang, Bin Xiao, Ce Liu, Lu Yuan, and Jianfeng Gao. Unified contrastive learning in image-text-label space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19163–19173, 2022.

A Dimensionality Reduction Methods

We begin by defining the setup for dimensionality reduction methods in the context of I-Con. Let $x_i \in \mathbb{R}^d$ represent high-dimensional data points, and $\phi_i \in \mathbb{R}^m$ represent their corresponding low-dimensional embeddings, where $m \ll d$. The goal of dimensionality reduction methods, such as Stochastic Neighbor Embedding (SNE) and t-Distributed Stochastic Neighbor Embedding (t-SNE), is to learn these embeddings such that neighborhood structures in the high-dimensional space are preserved in the low-dimensional space. In this context, the low-dimensional embeddings ϕ_i can be interpreted as the outputs of a mapping function $f_\theta(x_i)$, where f_θ is essentially an embedding matrix or look-up table. The I-Con framework is well-suited to express this relationship through a KL divergence loss between two neighborhood distributions: one in the high-dimensional space and one in the low-dimensional space.

Theorem 1. *Stochastic Neighbor Embedding (SNE) Hinton & Roweis (2002) is an instance of the I-Con framework.*

Proof. This is one of the most straightforward proofs in this paper, essentially based on the definition of SNE. The target distribution (supervised part), described by the neighborhood distribution in the high-dimensional space, is given by:

$$p_\theta(j|i) = \frac{\exp\left(-\|x_i - x_j\|^2 / 2\sigma_i^2\right)}{\sum_{k\neq i} \exp\left(-\|x_i - x_k\|^2 / 2\sigma_i^2\right)},$$

while the learned low-dimensional neighborhood distribution is:

$$q_\phi(j|i) = \frac{\exp\left(-\|\phi_i - \phi_j\|^2\right)}{\sum_{k\neq i} \exp\left(-\|\phi_i - \phi_k\|^2\right)}.$$

The objective is to minimize the KL divergence between these distributions:

$$\mathcal{L} = \sum_i D_{\mathrm{KL}}(p_\theta(\cdot|i) \| q_\phi(\cdot|i)) = \sum_i \sum_j p_\theta(j|i) \log \frac{p_\theta(j|i)}{q_\phi(j|i)}.$$

The embeddings θ_i are learned implicitly by minimizing $\mathcal{L}$. The mapper is an embedding matrix, as SNE is a non-parametric optimization. Therefore, SNE is a special case of the I-Con framework, where $p_\theta(j|i)$ and $q_\phi(j|i)$ represent the neighborhood probabilities in the high- and low-dimensional spaces, respectively. □

Theorem 2 (t-SNE Van der Maaten & Hinton (2008))**.** *t-SNE is an instance of the I-Con framework.*

Proof. The proof is similar to the one for SNE. While the high-dimensional target distribution $p_\theta(j|i)$ remains unchanged, t-SNE modifies the low-dimensional distribution to a Student's t-distribution with one degree of freedom (Cauchy distribution):

$$q_\phi(j|i) = \frac{(1 + \|\phi_i - \phi_j\|^2)^{-1}}{\sum_{k\neq i}(1 + \|\phi_i - \phi_k\|^2)^{-1}}.$$

The objective remains to minimize the KL divergence. Therefore, t-SNE is an instance of the I-Con framework. □

B Feature Learning Methods

We now extend the I-Con framework to feature learning methods commonly used in contrastive learning. Let $x_i \in \mathbb{R}^d$ be the input data points, and $f_\phi(x_i) \in \mathbb{R}^m$ be their learned feature embedding. In contrastive learning, the goal is to learn these embeddings such that similar data points (positive pairs) are close in the embedding space, while dissimilar points (negative pairs) are far apart. This setup can be expressed using a neighborhood distribution in the original space, where ”neighbors” are defined not by proximity in Euclidean space, but by predefined relationships such as data augmentations or class membership. The learned embeddings $f_\phi(x_i)$ define a new distribution over neighbors, typically using a Gaussian kernel in the learned feature space. We show that InfoNCE is a natural instance of the I-Con framework, and many other methods, such as SupCon, CMC, and Cross Entropy, follow from this.

Theorem 3 (InfoNCE Bachman et al. (2019))**.** *InfoNCE is an instance of the I-Con framework.*

Proof. InfoNCE aims to maximize the similarity between positive pairs while minimizing it for negative pairs in the learned feature space. In the I-Con framework, this can be interpreted as minimizing the divergence between two distributions: the neighborhood distribution in the original space and the learned distribution in the embedding space.

The neighborhood distribution $p_\theta(j|i)$ is uniform over the positive pairs, defined as:

$$p_\theta(j|i) = \begin{cases} \frac{1}{k} & \text{if } x_j \text{ is among the } k \text{ positive views of } x_i, \\ 0 & \text{otherwise.} \end{cases}$$

where k is the number of positive pairs for x_i.

The learned distribution $q_\phi(j|i)$ is based on the similarities between the embeddings $f_\phi(x_i)$ and $f_\phi(x_j)$, constrained to unit norm ($\|f_\phi(x_i)\| = 1$). Using a temperature-scaled Gaussian kernel, this distribution is given by:

$$q_\phi(j|i) = \frac{\exp\left(f_\phi(x_i) \cdot f_\phi(x_j)/\tau\right)}{\sum_{k \neq i} \exp\left(f_\phi(x_i) \cdot f_\phi(x_k)/\tau\right)},$$

where τ is the temperature parameter controlling the sharpness of the distribution. Since $\|f_\phi(x_i)\| = 1$, the Euclidean distance between $f_\phi(x_i)$ and $f_\phi(x_j)$ is $2 - 2(f_\phi(x_i) \cdot f_\phi(x_j))$.

The InfoNCE loss can be written in its standard form:

$$\mathcal{L}_{\text{InfoNCE}} = -\sum_i \log \frac{\exp\left(f_\phi(x_i) \cdot f_\phi(x_i^+)/\tau\right)}{\sum_k \exp\left(f_\phi(x_i) \cdot f_\phi(x_k)/\tau\right)},$$

where j^+ is the index of a positive pair for i. Alternatively, in terms of cross-entropy, the loss becomes:

$$\mathcal{L}_{\text{InfoNCE}} \propto \sum_i \sum_j p_\theta(j|i) \log q_\phi(j|i) = H(p_\theta, q_\phi),$$

where $H(p_\theta, q_\phi)$ denotes the cross-entropy between the two distributions. Since $p_\theta(j|i)$ is fixed, minimizing the cross-entropy $H(p_\theta, q_\phi)$ is equivalent to minimizing the KL divergence $D_{\text{KL}}(p_\theta \| q_\phi)$. By aligning the learned distribution $q_\phi(j|i)$ with the target distribution $p_\theta(j|i)$, InfoNCE operates within the I-Con framework, where the neighborhood structure in the original space is preserved in the embedding space. Thus, InfoNCE is a direct instance of I-Con, optimizing the same divergence-based objective. $\square$

Theorem 4. *Contrastive Multiview Coding (CMC) and CLIP are instances of the I-Con framework.*

Proof. Since we have already established that InfoNCE is an instance of the I-Con framework, this corollary follows naturally. The key difference in Contrastive Multiview Coding (CMC) and CLIP is that they optimize alignment across different modalities. The target probability distribution $p_\theta(j|i)$ can be expressed as:

$$p_\theta(j|i) = \frac{1}{Z} \mathbb{1}[i \text{ and } j \text{ are positive pairs and } V_i \neq V_j],$$

where V_i and V_j represent the modality sets of x_i and x_j, respectively. Here, $p_\theta(j|i)$ assigns uniform probability over positive pairs drawn from different modalities.

The learned distribution $q_\phi(j|i)$, in this case, is based on a Gaussian similarity between deep features, but conditioned on points from the opposite modality set. Thus, the learned distribution is defined as:

$$q_\phi(j|i) = \frac{\exp\left(-\|f_\phi(x_i) - f_\phi(x_j)\|^2\right)}{\sum_{k \in V_j} \exp\left(-\|f_\phi(x_i) - f_\phi(x_k)\|^2\right)}.$$

This formulation shows that CMC and CLIP follow the same principles as InfoNCE but apply them in a multiview setting, fitting seamlessly within the I-Con framework by minimizing the divergence between the target and learned distributions across different modalities. $\square$

Corollary 1. *Cross-Entropy classification is an instance of the I-Con framework.*

Proof. Cross-Entropy can be viewed as a special case of the CMC loss, where one "view" corresponds to the data point features and the other to the class logits. The affinity between a data point and a class is based on whether the point belongs to that class. This interpretation has been explored in prior work, where Cross-Entropy was shown to be related to the CLIP loss Yang et al. (2022). ☐

C Clustering Methods

The connections between clustering and the I-Con framework are more intricate compared to the dimensionality reduction methods discussed earlier. To establish these links, we first introduce a probabilistic formulation of K-means and demonstrate its equivalence to the classical K-means algorithm, showing that it is a zero-gap relaxation. Building upon this, we reveal how probabilistic K-means can be viewed as an instance of I-Con, leading to a novel clustering kernel. Finally, we show that several clustering methods implicitly approximate and optimize for this kernel.

Definition 1 (Classical K-means)**.** *Let $x_1, x_2, \ldots, x_N \in \mathbb{R}^n$ denote the data points, and $\mu_1, \mu_2, \ldots, \mu_m \in \mathbb{R}^n$ be the cluster centers.*

The objective of classical K-means is to minimize the following loss function:

$$\mathcal{L}_{\textit{k-Means}} = \sum_{i=1}^{N} \sum_{c=1}^{m} \mathbb{1}(c^{(i)} = c) \|x_i - \mu_c\|^2,$$

where $c^{(i)}$ represents the cluster assignment for data point x_i, and is defined as:

$$c^{(i)} = \arg\min_c \|x_i - \mu_c\|^2.$$

C.1 Probabilistic K-means Relaxation

In probabilistic K-means, the cluster assignments are relaxed by assuming that each data point x_i belongs to a cluster c with probability ϕ_{ic}. In other words, ϕ_i represents the cluster assignments vector for x_i

Proposition 1. *The relaxed loss function for probabilistic K-means is given by:*

$$\mathcal{L}_{\textit{Prob-k-Means}} = \sum_{i=1}^{N} \sum_{c=1}^{m} \phi_{ic} \|x_i - \mu_c\|^2,$$

and is equivalent to the original K-means objective $\mathcal{L}_{\textit{k-Means}}$. The optimal assignment probabilities ϕ_{ic} are deterministic, assigning probability 1 to the closest cluster and 0 to others.

Proof. For each data point x_i, the term $\sum_{c=1}^{m} \phi_{ic} \|x_i - \mu_c\|^2$ is minimized when the assignment probabilities ϕ_{ic} are deterministic, i.e.,

$$\phi_{ic} = \begin{cases} 1 & \text{if } c = \arg\min_j \|x_i - \mu_j\|^2, \\ 0 & \text{otherwise.} \end{cases}$$

With these deterministic probabilities, $\mathcal{L}_{\text{Prob-k-Means}}$ simplifies to the classical K-means objective, confirming that the relaxation introduces no gap. ☐

C.1.1 Contrastive Formulation of Probabilistic K-means

Definition 2. *Let $\{x_i\}_{i=1}^N$ be a set of data points. Define the conditional probablity $q_\phi(j|i)$ as:*

$$q_\phi(j|i) = \sum_{c=1}^{m} \frac{\phi_{ic}\phi_{jc}}{\sum_{k=1}^{N} \phi_{kc}},$$

where ϕ_i is the soft-cluster assignments for x_i.

Given $q_\phi(j|i)$, we can reformulate probabilistic K-means as a contrastive loss:

Theorem 5. *Let $\{x_i\}_{i=1}^N \in \mathbb{R}^n$ and $\{\phi_{ic}\}_{i=1}^N$ be the corresponding assignment probabilities. Define the objective function $\mathcal{L}$ as:*

$$\mathcal{L} = -\sum_{i,j} (x_i \cdot x_j)\, q_\phi(j|i).$$

Minimizing $\mathcal{L}$ with respect to the assignment probabilities $\{\phi_{ic}\}$ yields optimal cluster assignments equivalent to those obtained by K-means.

Proof. The relaxed probabilistic K-means objective $\mathcal{L}_{\text{Prob-k-Means}}$ is:

$$\mathcal{L}_{\text{Prob-k-Means}} = \sum_{i=1}^N \sum_{c=1}^m \phi_{ic} \|x_i - \mu_c\|^2.$$

Expanding this, we obtain:

$$\mathcal{L}_{\text{Prob-k-Means}} = \sum_{c=1}^m \left(\sum_{i=1}^N \phi_{ic} \right) \|\mu_c\|^2 - 2 \sum_{c=1}^m \left(\sum_{i=1}^N \phi_{ic} x_i \right) \cdot \mu_c + \sum_{i=1}^N \|x_i\|^2.$$

The cluster centers μ_c that minimize this loss are given by:

$$\mu_c = \frac{\sum_{i=1}^N \phi_{ic} x_i}{\sum_{i=1}^N \phi_{ic}}.$$

Substituting μ_c back into the loss function, we get:

$$\mathcal{L} = -\sum_{i,j} (x_i \cdot x_j)\, q_\phi(j|i),$$

$$\mathcal{L} = -\sum_{i,j} A_{ij} q_\phi(j|i)$$

$$W = XX^\top$$

$$W = V\Lambda V^\top$$

$$\mathcal{L} = -\sum_{i,j} (v_i \cdot v_j)\, q_\phi(j|i),$$

$$\mathcal{L} = -\sum_{i,j} \tilde{W}_{ij} q_\phi(j|i)$$

$$\tilde{W} = V_k V_k^\top$$

which proves that minimizing this contrastive formulation leads to the same clustering assignments as classical K-means. □

Corollary 2. *The alternative loss function:*

$$\mathcal{L} = -\sum_{i,j} \|x_i - x_j\|^2\, q_\phi(j|i),$$

yields the same optimal clustering assignments when minimized with respect to $\{\phi_{ic}\}$.

Proof. Expanding the squared norm in the loss function gives:

$$\mathcal{L} = -\sum_{i,j} \left(\|x_i\|^2 - 2x_i \cdot x_j + \|x_j\|^2 \right) q_\phi(j|i).$$

The terms involving $\|x_i\|^2$ and $\|x_j\|^2$ simplify since $\sum_j q_\phi(j|i) = 1$, reducing the loss to:

$$\mathcal{L} = 2 \left(-\sum_{i,j} x_i \cdot x_j q_\phi(j|i) \right),$$

which is equivalent to the objective in the previous theorem. □

C.2 Probabilistic K-means as an I-Con Method

In the I-Con framework, the target and learned distributions represent affinities between data points based on specific measures. For instance, in SNE, these measures are Euclidean distances in high- and low-dimensional spaces, while in SupCon, the distances reflect whether data points belong to the same class. Similarly, we can define a measure of neighborhood probabilities in the context of clustering, where two points are considered neighbors if they belong to the same cluster. The probability of selecting x_j as x_i's neighbor is the probability that a point, chosen uniformly at random from x_i's cluster, is x_j. More explicitly, let $q_\phi(j|i)$ represent the probability that x_j is selected uniformly at random from x_i's cluster:

$$q_\phi(j|i) = \sum_{c=1}^{m} \frac{\phi_{ic}\phi_{jc}}{\sum_{k=1}^{N} \phi_{kc}}.$$

Theorem 6 (K-means as an instance of I-Con). *Given data points $\{x_i\}_{i=1}^N$, define the neighborhood probabilities $p_\theta(j|i)$ and $q_\phi(j|i)$ as:*

$$p_\theta(j|i) = \frac{\exp\left(-\|x_i - x_j\|^2/2\sigma^2\right)}{\sum_k \exp\left(-\|x_i - x_k\|^2/2\sigma^2\right)}, \quad q_\phi(j|i) = \sum_{c=1}^{m} \frac{\phi_{ic}\phi_{jc}}{\sum_{k=1}^{N} \phi_{kc}}.$$

Let the loss function $\mathcal{L}_{\text{c-SNE}}$ be the sum of KL divergences between the distributions $q_\phi(j|i)$ and $p_\theta(j|i)$:

$$\mathcal{L}_{\text{c-SNE}} = \sum_i D_{KL}(q_\phi(\cdot|i)\|p_\theta(\cdot|i)).$$

Then,

$$\mathcal{L}_{\text{c-SNE}} = \frac{1}{2\sigma^2}\mathcal{L}_{\text{Prob-k-Means}} - \sum_i H(q_\phi(\cdot|i)),$$

where $H(q_\phi(\cdot|i))$ is the entropy of $q_\phi(\cdot|i)$.

Proof. For simplicity, assume that $2\sigma^2 = 1$. Denote $\sum_k \exp\left(-\|x_i - x_k\|^2\right)$ by Z_i. Then we have:

$$\log p_\theta(j|i) = -\|x_i - x_j\|^2 - \log Z_i.$$

Let $\mathcal{L}_i$ be defined as $-\sum_j \|x_i - x_j\|^2 q_\phi(j|i)$. Using the equation above, $\mathcal{L}_i$ can be rewritten as:

$$\begin{aligned}
\mathcal{L}_i &= -\sum_j \|x_i - x_j\|^2 q_\phi(j|i) && \text{(C.2.1)}\\
&= \sum_j (\log(p_\theta(j|i)) + \log(Z_i)) q_\phi(j|i) && \text{(C.2.2)}\\
&= \sum_j q_\phi(j|i)\log(p_\theta(j|i)) + \sum_j q_\phi(j|i)\log(Z_i) && \text{(C.2.3)}\\
&= \sum_j q_\phi(j|i)\log(p_\theta(j|i)) + \log(Z_i) && \text{(C.2.4)}\\
&= H(q_\phi(\cdot|i), p_\theta(\cdot|i)) + \log(Z_i) && \text{(C.2.5)}\\
&= D_{\text{KL}}(q_\phi(\cdot|i)\|p_\theta(\cdot|i)) + H(q_\phi(\cdot|i)) + \log(Z_i). && \text{(C.2.6)}
\end{aligned}$$

Therefore, $\mathcal{L}_{\text{Prob-KMeans}}$, as defined in Corollary 2, can be rewritten as:

$$\begin{aligned}
\mathcal{L}_{\text{Prob-KMeans}} &= -\sum_i \sum_j \|x_i - x_j\|^2 q_\phi(j|i) = \sum_i \mathcal{L}_i && \text{(C.2.7)}\\
&= \sum_i D_{\text{KL}}(q_\phi(\cdot|i)\|p_\theta(\cdot|i)) + H(q_\phi(\cdot|i)) + \log(Z_i) && \text{(C.2.8)}\\
&= \mathcal{L}_{\text{c-SNE}} + \sum_i H(q_\phi(\cdot|i)) + \text{constant}. && \text{(C.2.9)}
\end{aligned}$$

Therefore,

$$\mathcal{L}_{\text{c-SNE}} = \mathcal{L}_{\text{Prob-KMeans}} - \sum_i H(q_\phi(\cdot|i)).$$

If we allow σ to take any value, the entropy penalty will be weighted accordingly:

$$\mathcal{L}_{\text{c-SNE}} = \frac{1}{2\sigma^2}\mathcal{L}_{\text{Prob-KMeans}} - \sum_i H(q_\phi(\cdot|i)).$$

Note that the relation above is up to an additive constant. This implies that minimizing the contrastive probabilistic K-means loss with entropy regularization minimizes the sum of KL divergences between $q_\phi(\cdot|i)$ and $p_\theta(\cdot|i)$. □

Theorem 7. *Mutual Information Clustering is an instance of I-Con.*

Proof. Given the connection established between SimCLR, K-Means, and the I-Con framework, this result follows naturally. Specifically, the target distribution $p_\theta(j|i)$ (the supervised part) is a uniform distribution over observed positive pairs:

$$p_\theta(j|i) = \begin{cases} \frac{1}{k} & \text{if } x_j \text{ is among the } k \text{ positive views of } x_i, \\ 0 & \text{otherwise.} \end{cases}$$

On the other hand, the learned embeddings ϕ_i represent the probabilistic assignments of x_i into clusters. Therefore, similar to the analysis of the K-Means connection, the learned distribution is modeled as:

$$q_\phi(j|i) = \sum_{c=1}^{m} \frac{\phi_{ic}\phi_{jc}}{\sum_{k=1}^{N} \phi_{kc}}.$$

This shows that Mutual Information Clustering can be viewed as a method within the I-Con framework, where the learned distribution $q_\phi(j|i)$ aligns with the target distribution $p_\theta(j|i)$, completing the proof. □

Theorem 8. *Normalized Cuts Shi & Malik (2000) is an instance of I-Con.*

Proof. The proof for this follows naturally from our work on K-Means analysis. The loss function for normalized cuts is defined as:

$$\mathcal{L}_{\text{NormCuts}} = \sum_{c=1}^{m} \frac{\text{cut}(A_c, \overline{A}_c)}{\text{vol}(A_c)},$$

where A_c is a subset of the data corresponding to cluster c, $\overline{A}_c$ is its complement, and $\text{cut}(A_c, \overline{A}_c)$ represents the sum of edge weights between A_c and $\overline{A}_c$, while $\text{vol}(A_c)$ is the total volume of cluster A_c, i.e., the sum of edge weights within A_c.

Similar to K-Means, by reformulating this in a contrastive style with soft-assignments, the learned distribution can be expressed using the probabilistic cluster assignments $\phi_{ic} = p(c|x_i)$ as:

$$q_\phi(j|i) = \sum_{c=1}^{m} \frac{\phi_{ic}\phi_{jc}d_j}{\sum_{k=1}^{N} \phi_{kc}d_k},$$

where d_j is the degree of node x_j, and the volume and cut terms can be viewed as weighted sums over the soft-assignments of data points to clusters.

This reformulation shows that normalized cuts can be written in a manner consistent with the I-Con framework, where the target distribution $p_\theta(j|i)$ and the learned distribution $q_\phi(j|i)$ represent affinity relationships based on graph structure and cluster assignments.

Thus, normalized cuts is an instance of I-Con, where the loss function optimizes the neighborhood structure based on the cut and volume of clusters in a manner similar to K-Means and its probabilistic relaxations. □