Beyond I-Con: A Roadmap for Representation Learning Loss Discovery

Jasmine Shone* JASSHONE@MIT.EDU
Zhening Li* ZLI11010@CSAIL.MIT.EDU
Shaden Alshammari SHADEN@MIT.EDU
Mark Hamilton MARHAMIL@MIT.EDU
William Freeman BILLF@MIT.EDU
MIT CSAIL, Cambridge, MA

Editors: List of editors' names

Abstract

The Information Contrastive (I-Con) framework revealed that over 23 representation learning methods implicitly minimize KL divergence between data and learned distributions that encode similarities between data points. However, a KL-based loss may be misaligned with the true objective, and properties of KL divergence such as asymmetry and unboundedness may create optimization challenges. We present Beyond I-Con, a framework that enables systematic discovery of novel loss functions by exploring alternative statistical divergences. Key findings: (1) on unsupervised clustering of DINO-ViT embeddings, we achieve state-of-the-art results by modifying the PMI algorithm to use total variation (TV) distance; (2) supervised contrastive learning with Euclidean distance as the feature space metric is improved by replacing the standard loss function with Jenson-Shannon divergence (JSD); (3) on dimensionality reduction, we achieve superior qualitative results and better performance on downstream tasks than SNE by replacing KL with a bounded f-divergence. Our results highlight the importance of considering divergence choices in representation learning optimization.

Keywords: representation learning, contrastive learning, clustering, dimensionality reduction

1. Introduction

The choice of optimization objective fundamentally determines the success of representation learning methods, yet the field has largely focused on a single statistical divergence measure without systematic exploration of alternatives. The Information Contrastive (I-Con) framework recently revealed that over 23 diverse representation learning methods all implicitly minimize KL divergence between data and learned distributions that encode similarities between data points (Alshammari et al., 2025). The natural question emerges: if representation learning methods can be unified under minimizing the divergence between two distributions, what happens when we systematically explore alternative divergences?

Contributions. We present Beyond I-Con, making the following contributions: (1) We generalize I-Con by replacing KL divergence with alternative f-divergences, revealing that KL is not unique in enabling meaningful feature optimization; (2) We systematically explore f-divergences, uncovering novel loss functions with superior performance on unsupervised clustering, supervised contrastive learning, and dimensionality reduction.

* Equal contribution

2. Background: I-Con Overview

The I-Con framework unifies representation learning methods by framing them as "minimizing the average KL divergence between two conditional 'neighborhood distributions' that define transition probabilities between data points," which we index with $i = 1, \ldots, N$ (Alshammari et al., 2025). $p(j \mid i)$ is typically a fixed "supervisory" distribution, and $q_\phi(j \mid i)$ are learnable transition probabilities typically calculated using similarities between features — see Figure 2a of the original paper for an illustration. The core I-Con loss function (Equation 1 of the original paper) is

$$\mathcal{L}_{\text{I-Con}} = \mathbb{E}_{i \sim p(i)} \left[\mathcal{D}_{\text{KL}}\big(p(\cdot|i) \big\| q(\cdot|i)\big) \right]. \tag{1}$$

By varying how p is constructed from the dataset and how q is defined in terms of similarities between features, $\mathcal{L}_{\text{I-Con}}$ reproduces the loss functions of many existing representation learning methods.

3. Beyond I-Con Framework

We generalize the I-Con objective by replacing the KL divergence with any positive-definite divergence $\mathcal{D}$:

$$\mathcal{L}_{\text{Beyond I-Con}} = \mathbb{E}_{i \sim p(i)} \left[\mathcal{D}\big(p(\cdot|i) \big\| q(\cdot|i)\big) \right] \tag{2}$$

We focus on **f-divergences** including KL, Total Variation (TV), Jensen-Shannon (JSD), and Hellinger because they are most directly comparable to KL as a measure of distance between distributions — some losses such as JSD directly remedy weaknesses of KL such as asymmetry and unboundedness.

4. Experimental Results

4.1. Unsupervised Clustering

We modify the Pointwise Mutual Information (PMI) clustering algorithm (Adaloglou et al., 2023) by using different divergences than KL. We followed the same training setup as in the I-Con paper, which clusters DINO ViT embeddings (Caron et al., 2021) on ImageNet-1K (Deng et al., 2009) by training a linear classifier for 30 epochs with a batch size of 4096, a learning rate of 1×10^{-3}, and the Adam optimizer (Kingma, 2015). Data augmentation was applied to the training samples. Table 1 contains the results of this experiment. Clustering using PMI with TV outperforms the state-of-the-art on ViT-B/14 and ViT-L/14 embeddings.

4.2. Supervised Contrastive Learning

We trained ResNet-50 (He et al., 2016) models with supervised contrastive learning (Khosla et al., 2020) on CIFAR-10 (Krizhevsky, 2009). We used a Euclidean distance metric on features and trained for 150 epochs with a batch size of 2048 and learning rate of 1e-3. We systematically varied divergence measures. We used the learned features to perform classification by training a linear probe or applying k-nearest neighbors (k-NN). Results are shown in Table 2.

Method	DiNO ViT-S/14	DiNO ViT-B/14	DiNO ViT-L/14
k-Means	51.84	52.26	53.36
TEMI (Adaloglou et al., 2023)	56.84	58.62	—
Debiased InfoNCE Clustering	**57.8 ± 0.26**	64.75 ± 0.18	67.52 ± 0.28
JSD	53.50	63.80	66.60
TV	55.90	**65.13 ± 0.13**	**68.40 ± 0.29**
Hellinger	54.90	63.80	67.85

Table 1: Comparison of methods on ImageNet-1K clustering with respect to Hungarian Accuracy. TV outperforms the state-of-the-art for ViT-B and ViT-L.

Divergence	Linear probe test acc.	k-NN ($k = 7$) test acc.
KL	90.03 ± 0.14	89.61 ± 0.13
TV	83.23 ± 0.18	82.95 ± 0.16
Hellinger	90.47 ± 0.08	90.40 ± 0.09
JSD	**90.84 ± 0.11**	**90.62 ± 0.11**

Table 2: Downstream classification accuracy from SupCon-learned features on CIFAR-10. Errors are standard errors of the mean over 5 seeds.

Both Hellinger distance and Jenson-Shannon divergence (JSD) exhibit better performance than vanilla supervised contrastive learning (KL divergence). In particular, JSD achieves the best performance.

4.3. Dimensionality Reduction

We also ran SNE (Hinton and Roweis, 2002) with a CNN backbone on CIFAR-10 using different divergences to visualize qualitative differences between resulting image embeddings, as shown in Figure 1. Visually, while SNE with KL divergence creates highly overlapping clusters, the clusters resulting from SNE with the other divergences are more cleanly separated.

5. Analysis and Discussion

Across the tasks of unsupervised clustering, supervised contrastive learning, and dimensionality reduction, we observe that a non-KL loss outperforms KL. We hypothesize this is because optimizing for KL overly penalizes placing dissimilar points farther apart in the feature space (KL diverges to infinity as $q(j \mid i) \to 0$). This causes different clusters or classes to crowd together and start overlapping with each other.

For example, in dimensionality reduction, SNE (Hinton and Roweis, 2002) is known to have this crowding problem, which we also observe in Figure 1. Beyond I-Con provides a solution by replacing the KL divergence in the loss with other distances/divergences. TV,

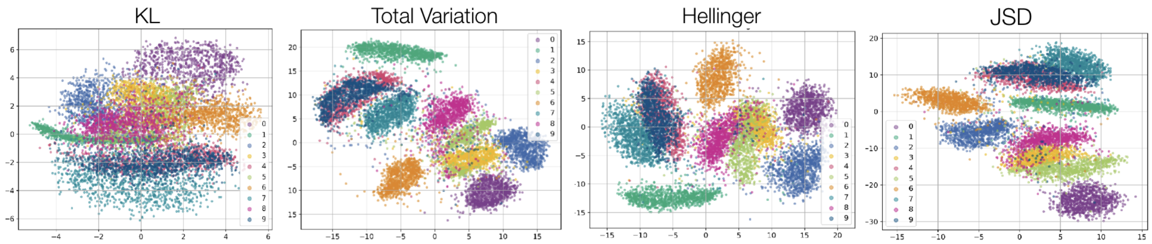

Figure 1: Results for running SNE on CIFAR-10 using different divergences, after 150 epochs with a CNN model architecture at learning rate 1e-3. Each color represents a class. KL divergence produces highly overlapping categories in the SNE visualization while other divergences achieve separation.

JSD, and Hellinger all penalize small values of $q(j \mid i)$ less than KL does, since they all remain bounded as $q(j \mid i) \to 0$.[1] Indeed, they all solve the crowding problem as seen in the resultant low-dimensional visualizations (Figure 1). Appendix B provides further analysis on the geometric arrangement of clusters in high- vs. low-dimensional spaces.

We also hypothesize that KL-based losses may be prone to unstable gradients during training, similar to previous findings (Lazić et al., 2021; Arjovsky et al., 2017). For example, in dimensionality reduction, our gradient norm plots during training (Figure 2) show large spikes in gradients near the beginning of training when the KL-based loss from vanilla SNE is used.

6. Conclusion

In this work, we extended the I-Con representation learning framework with a new dimension — the type of f-divergence in the loss function. By experimenting with alternative divergences, we achieve state-of-the-art ImageNet-1K clustering, surpass vanilla supervised contrastive learning on CIFAR-10, and outperform SNE on CIFAR-10. Beyond I-Con challenges the default reliance on KL divergence in representation learning by showing that alternative statistical divergences can yield superior performance and serve as a basis for novel loss discovery.

References

Nikolas Adaloglou, Felix Michels, Hamza Kalisch, and Markus Kollmann. Exploring the limits of deep image clustering using pretrained models. In *The British Machine Vision Conference*, pages 297–299, 2023. URL http://proceedings.bmvc2023.org/297/.

Shaden Naif Alshammari, Mark Hamilton, Axel Feldmann, John R. Hershey, and William T. Freeman. A unifying framework for representation learning. In *The Thirteenth Interna-*

1. t-SNE (Maaten and Hinton, 2008) provides an alternative solution to the crowding problem by replacing $q(\cdot \mid i)$ with one where $q(j \mid i)$ does not decrease as quickly as the distance between x_i and x_j increases.

tional Conference on Learning Representations, 2025. URL https://openreview.net/forum?id=WfaQrKCr4X.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 214–223, 2017. URL https://proceedings.mlr.press/v70/arjovsky17a.html.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9650–9660, 2021.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

Geoffrey E Hinton and Sam Roweis. Stochastic neighbor embedding. *Advances in Neural Information Processing Systems*, 15, 2002.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 18661–18673, 2020.

Diederik P Kingma. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 2015.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

Nevena Lazić, Botao Hao, Yasin Abbasi-Yadkori, Dale Schuurmans, and Csaba Szepesvári. Optimization issues in KL-constrained approximate policy iteration. *arXiv preprint arXiv:2102.06234*, 2021.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

C Ambrose Rogers. *Packing and Covering*, volume 54. University Press Cambridge, 1964.

Appendix A. Supplementary Figures and Tables

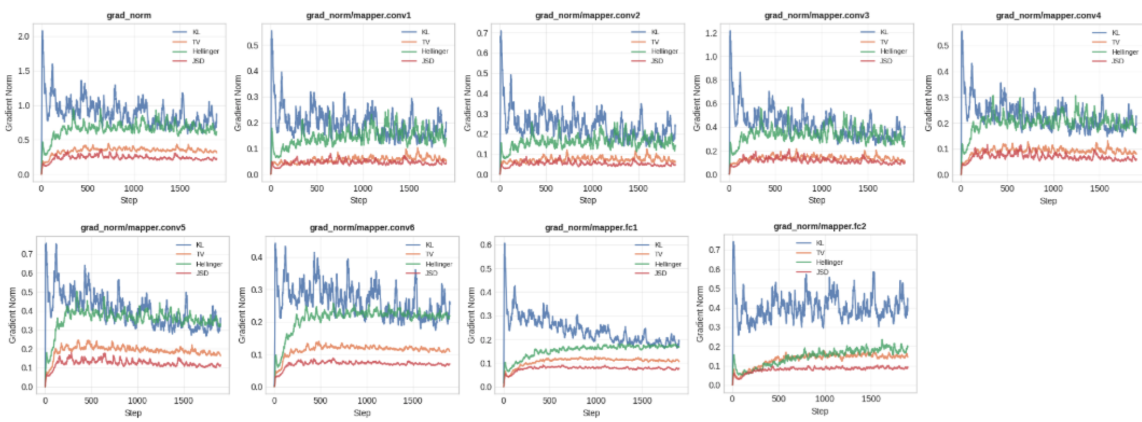

Figure 2: Gradient norms for each divergence from running SNE on CIFAR-10 images with a CNN backbone. KL's unbounded nature creates initialization instability that manifests consistently across all network layers, while bounded divergences (TV, Hellinger, JSD) provide more stable gradient behavior throughout training.

Appendix B. Geometric Analysis of the Crowding Problem

In this appendix, we illustrate with an example why the KL divergence loss results in overcrowding in SNE.

Let the original high-dimensional space have d dimensions. A d-dimensional space permits $d+1$ clusters/classes that are equidistant from each other — say they're distance 1 from each other. Now, mapping these clusters to a reduced d'-dimensional feature space ($d' \ll d$) while ensuring the clusters are still separated by distance at least 1 means that some clusters will be very far from each other ($\Omega\big(d^{1/d'}\big)$) (Rogers, 1964). This incurs a high KL penalty as $q(j \mid i) \ll p(j \mid i)$ for data points i, j from two clusters that are now much farther away each other. Thus, minimizing the KL loss inevitably leads to some clusters being brought too close together (crowding) in the low-dimensional feature space to prevent a high KL penalty from two clusters that are too far apart. Bounded divergences like TV, Hellinger, and JSD resolve this issue since $q(j \mid i) \ll p(j \mid i)$ incurs a lower penalty.