f-MICL: Understanding and Generalizing InfoNCE-based Contrastive Learning

Anonymous authors
Paper under double-blind review

Abstract

In self-supervised contrastive learning, a widely-adopted objective function is InfoNCE, which uses the heuristic cosine similarity for the representation comparison, and is closely related to maximizing the Kullback–Leibler (KL)-based mutual information. In this paper, we aim at answering two intriguing questions: (1) Can we go beyond the KL-based objective? (2) Besides the popular cosine similarity, can we design a better similarity function? We provide answers to both questions by generalizing the KL-based mutual information to the f-**M**utual **I**nformation in **C**ontrastive **L**earning (f-MICL) using the f-divergences. To answer the first question, we provide a wide range of f-MICL objectives which share the nice properties of InfoNCE (e.g., alignment and uniformity), and meanwhile result in similar or even superior performance. For the second question, assuming that the joint feature distribution is proportional to the Gaussian kernel, we derive an **f-Gaussian similarity** with better interpretability and empirical performance. Finally, we identify close relationships between the f-MICL objective and several popular InfoNCE-based objectives. Using benchmark tasks from both vision and natural language, we empirically evaluate f-MICL with different f-divergences on various architectures (SimCLR, MoCo, and MoCo v3) and datasets. We observe that f-MICL generally outperforms the benchmarks and the best-performing f-divergence is task and dataset dependent.

1 Introduction

Recent advances in self-supervised learning aim at learning similar representations from different augmented views of the same data sample. However, naively implementing this idea would easily make representations converge to some trivial constant (i.e., feature collapse) in practice. To address this problem, researchers propose new algorithms either from the *model architecture* perspective or the *training objective* perspective. The former method (e.g., Grill et al. (2020); Chen & He (2021); Zhang et al. (2021)) applies techniques such as stop gradient or predictor module to create asymmetry in networks, while the latter method encourages the contrastiveness between similar (positive) and dissimilar (negative) sample pairs through the objective design.

In this paper, we intend to deepen our understanding of contrastive learning by generalizing the current objective design. To achieve self-supervised contrastive learning, existing objectives are proposed from different perspectives such as the mutual information (e.g., InfoNCE (Wu et al., 2018; van den Oord et al., 2018; Chen et al., 2020; Hénaff et al., 2020; He et al., 2020)), the information redundancy (e.g., Barlow Twins (Zbontar et al., 2021)), and the regularization (e.g., VICReg (Bardes et al., 2021)). In particular, the InfoNCE objective is widely used, which aims to maximize the probability of picking a similar sample pair among a batch of sample pairs, and can be interpreted as a lower bound of the mutual information (MI) between two views of samples (van den Oord et al., 2018; Bachman et al., 2019; Tian et al., 2020a; Tschannen et al., 2020). This is consistent with the well-known "InfoMax principle" (Linsker, 1988). To measure the similarity between sample pairs, cosine similarity is usually adopted.

To attain the aforementioned goals of better understanding and generalizing contrastive learning, we here focus on the widely-adopted InfoNCE objective, and aim at two questions regarding it:

(1) MI is essentially the Kullback–Leibler (KL) divergence between the joint distribution and the product of the marginal distributions. Is this KL-based objective optimal? If not, can we go beyond the KL-based objective?

(2) Besides the commonly used cosine similarity for measuring the distance between samples, can we provide a better similarity function with a theoretical basis?

To answer the above two questions, we generalize the KL-based mutual information to the broader f-divergence family (Ali & Silvey, 1966; Csiszár, 1967), and propose the benchmark of f-mutual information in contrastive learning (f-MICL). By searching through a wide range of f-divergences, we observe that the KL divergence is not always the best, and several other f-divergences in fact show similar or even superior performance in practice.

For the second question, while it is challenging to provide an answer based on the InfoNCE objective, it is possible to derive a proper similarity function under the f-MICL framework. By assuming that the joint feature distribution is proportional to the popularly-adopted Gaussian kernel, we propose a novel *f-Gaussian similarity* function that enjoys better empirical performance.

Finally, we show the generalization of f-MICL by drawing connections between the f-MICL objective and several popular InfoNCE-based objectives (e.g., SimCLR(Chen et al., 2020), MoCo(He et al., 2020), and Alignment and Uniformity (AU) (Wang & Isola, 2020)). We identify that those objectives are closely related to f-MICL: Alignment and Uniformity (AU) (Wang & Isola, 2020) can be treated as a special case, and SimCLR (Chen et al., 2020) and MoCo (He et al., 2020) are upper bounds for a transformed f-MICL. These results provide a different angle to better understand InfoNCE. Moreover, we show both theoretically and empirically that nice properties of InfoNCE (e.g., alignment and uniformity (Wang & Isola, 2020)) can be naturally extended to f-MICL.

We summarize our main contributions as follows:

- Motivated by InfoNCE, we propose a general framework for contrastive learning by extending the MI to the general f-MI, which provides a wide range of objective choices.
- Instead of using heuristic similarity functions, we provide a novel similarity function, called *f-Gaussian similarity*, based on the convex conjugate and an assumption on the joint feature distribution.
- We identify close relationships between our f-MICL objective and several InfoNCE-based contrastive learning objectives.
- Empirically, we show that f-MICL achieves notable improvement over benchmarks on various datasets, and the best-performing f-divergence depends on the specific task and dataset. In addition, our proposed f-Gaussian similarity consistently outperforms the cosine similarity.

2 f-Mutual Information

To provide answers to the above two questions regarding the InfoNCE objective, we first extend the KL-based mutual information to the more general f-mutual information. The definition of the f-mutual information (f-MI) is as follows:

Definition 1 (f-mutual information, Csiszár 1967). *Consider a pair of random variables (X, Y) with the density function $p(x, y)$. The f-mutual information I_f between X and Y is defined as*

$$I_f(X;Y) := \int f\left(\frac{p(x,y)}{p(x)p(y)}\right) p(x)p(y) \cdot \mathrm{d}\lambda(x,y), \tag{1}$$

where $f : \mathbb{R}_+ \to \mathbb{R}$ is (closed) convex with $f(1) = 0$.

Note that the f-MI is essentially the f-divergence between the joint distribution and the product of marginal distributions. It is well-known that the f-MI is non-negative and symmetric, and provided that f is strictly convex, $I_f(X;Y) = 0$ holds iff X and Y are independent (Ali & Silvey, 1966).

Table 1: Common choices of f-divergences. KL: Kullback–Leibler; JS: Jensen–Shannon; SH: Squared Hellinger; VLC: Vincze–Le Cam (Le Cam, 2012). For JS, we define $\varphi(u) = -(u+1)\log\frac{1+u}{2} + u\log u$. The Tsallis-$\alpha$ divergence is defined in Tsallis (1988). See Appendix A.1 for more details.

Divergence	$f(u)$	$f^*(t)$	$f'(u)$	$f^* \circ f'(u)$
KL	$u\log u$	$\exp(t-1)$	$\log u + 1$	u
JS	$\varphi(u)$	$-\log(2-e^t)$	$\log 2 + \log\frac{u}{1+u}$	$-\log\frac{2}{1+u}$
Pearson χ^2	$(u-1)^2$	$t^2/4+t$	$2(u-1)$	u^2-1
SH	$(\sqrt{u}-1)^2$	$\frac{t}{1-t}$	$1-u^{-1/2}$	$u^{1/2}-1$
Tsallis-α	$\frac{u^\alpha}{\alpha-1}$	$(\frac{\alpha-1}{\alpha}t)^{\frac{\alpha}{\alpha-1}}$	$\frac{\alpha}{\alpha-1}u^{\alpha-1}$	u^α
VLC	$\frac{(u-1)^2}{u+1}$	$4-t-4\sqrt{1-t}$	$1-\frac{4}{(u+1)^2}$	$3-\frac{4}{u+1}$

Since it is challenging to provide an accurate estimation of the f-divergences in high dimensions, Nguyen et al. (2010) used the convex conjugate as a lower bound for the f-divergences. With this result we can lower bound $I_f(X;Y)$ as follows:

$$I_f(X;Y) \geq \sup_{s\in\mathcal{F}} \Big(\mathop{\mathbb{E}}_{(x,y)\sim p(\cdot,\cdot)} s(x,y) - \mathop{\mathbb{E}}_{(x,y)\sim p(\cdot)p(\cdot)} f^* \circ s(x,y) \Big), \tag{2}$$

where $p(\cdot,\cdot)$ denotes the joint distribution, $p(\cdot)p(\cdot)$ stands for the product of marginals, and the symbol $\circ$ denotes the function composition. $f^*(t) := \sup_{x\in\mathbb{R}_+}(xt - f(x))$ refers to the convex conjugate of f and is *monotonically increasing*,[1] and $s(\cdot)$ belongs to $\mathcal{F}$, which is some function class. Using results in Nguyen et al. (2010), one can show that eq. (2) is equal to $I_f(X;Y)$ if there exists $s^* \in \mathcal{F}$ such that for any $(x,y) \in \mathrm{supp}(p(\cdot)) \times \mathrm{supp}(p(\cdot))$, where $\mathrm{supp}(\cdot)$ denotes the support of a distribution, we have:

$$s^*(x,y) = f'\left(\frac{p(x,y)}{p(x)p(y)}\right). \tag{3}$$

In other words, plugging the optimal $s^*(x,y)$ into eq. (2) we obtain equality. In Table 1 we list common choices of f-divergences, their conjugates, and the derivatives. We also include the composition $f^* \circ f'$ for later purposes (Theorem 5).

3 f-MICL

With the introduction of the more general f-MI we now proceed to the design of the objective and similarity function. Then we will analyze the property of the proposed framework and compare it with some existing InfoNCE-based benchmarks.

3.1 f-MICL objective

Contrastive learning is a popular *self-supervised* method for representation learning. In contrastive learning, we expect similar sample pairs to be close to each other in the embedding space, while dissimilar pairs to be far apart. Based on the f-MI introduced in § 2, we propose a general framework for contrastive learning, coined as f-MICL.

We denote $g : \mathcal{X} \to \mathbb{S}^{d-1}$ as the feature encoder (usually constructed by a neural network) from the input space $\mathcal{X}$ to the hypersphere, and we use the shorthands $x^g := g(x)$ and $y^g := g(y)$ to represent the feature embeddings. The notation p_d stands for the data distribution, and $p_\times := p_d \otimes p_d$ means its self product (product of marginals, e.g., different images). We denote p_+ as the distribution of *positive pairs*, *i.e.*, two samples with similar feature embeddings (joint distribution, e.g., the same image with different data augmentation). Using the lower bound of the f-MI in eq. (2), we have the general f-MICL objective as follows:

$$\sup_{s\in\mathcal{F}} \; \mathbb{E}_{(x,y)\sim p_+} s(x^g, y^g) - \mathbb{E}_{(x,y)\sim p_\times} f^* \circ s(x^g, y^g), \tag{4}$$

[1]Note that the domain of x is non-negative.

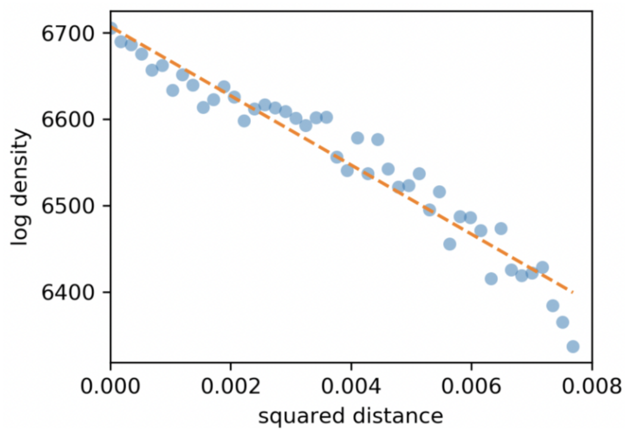

Figure 1: Experiment for verifying Assumption 3. Here we draw the relation between the squared distances $\|x^g - y^g\|^2$ and the averaged log likelihood $\log p_g$, with $\log p_g$ estimated by the flow model RealNVP (Dinh et al., 2017). The features are learned by SimCLR trained on CIFAR-10. See more details in Appendix D.3.

where $s(x^g, y^g)$ can be understood as the similarity measurement between two feature embeddings in the context of contrastive learning. Essentially, we are studying the variational lower bound eq. (2) in the *feature space*, with the feature embeddings learnable. We can treat the first term as the similarity score between *positive pairs* with similar feature embeddings, and the second term as the similarity score between two random samples, a.k.a. *negative pairs*. As f^* is an increasing function, maximizing the f-MI is equivalent to simultaneously maximizing the similarity between positive pairs and minimizing the similarity between negative pairs.

With eq. (4) we have answered the first question: there are a spectrum of f-MICL objectives that can be applied in contrastive learning by using different f functions. We will discuss how to choose the best f empirically in §4.

3.2 f-Gaussian similarity

Previous works in constrastive learning usually adopts some heuristic similarity function such as the cosine similarity function $s(x^g, y^g) = x^g \cdot y^g$. Although it shows promising performance in practice, our second question is that, can we provide a better similarity function than the popular cosine similarity?

Note that eq. (3) is a natural choice of $s(\cdot)$ from the perspective of deriving the f-MI. In the context of contrastive learning, by denoting the density functions of the marginal feature distributions as $p_g(x^g)$ and $p_g(y^g)$, and the density of the joint feature distribution as $p_g(x^g, y^g)$, from eq. (3) we have an optimal similarity function as follows:

Lemma 2 (*e.g.*, Nguyen et al. 2010, Lemma 1)**.** *Suppose f is differentiable, and the embedding function g is fixed. The similarity function s^* maximizes eq.* (4) *with:*

$$s^*(x^g, y^g) = f'\left(\frac{p_g(x^g, y^g)}{p_g(x^g)p_g(y^g)}\right). \tag{5}$$

Obviously, the optimal s^* in fact gives the f-MI on the feature space, which is a low bound of the original f-MI. Although eq. (5) provides an optimal similarity function, it nevertheless depends on the unknown density functions. How can we implement eq. (5) in practice? Among various known density functions, it is natural to choose a typical kernel function for structured data for validation (Balcan et al., 2008; Powell, 1987; Murphy, 2012), e.g., the Gaussian kernel.

Assumption 3 (**Gaussian kernel**)**.** *The joint feature distribution is proportional to a Gaussian kernel,*

$$p_g(x^g, y^g) \propto G_\sigma(\|x^g - y^g\|^2) = \mu \exp\left(-\tfrac{\|x^g - y^g\|^2}{2\sigma^2}\right),$$

where μ is a constant left to be determined.

Verifying Assumption 3: One may question that Assumption 3 can be too strong for practical usage. For example, replacing the Gaussian kernel G_σ with any other decreasing function would also provide a valid assumption. However, we found that among several popular choices only the Gaussian kernel works well in practice. Also, we can empirically verify that Assumption 3 approximately holds. To this end, it is sufficient to check whether the log density, *i.e.*, $\log p_g(x^g, y^g)$, is linear with the distance between each positive pair, *i.e.*, $\|x^g - y^g\|^2$. In Figure 1, we use the flow-based model RealNVP[2] (Dinh et al., 2017) to estimate the log density, and learn the feature encoder g from SimCLR (Chen et al., 2020). We observe that the linear relationship approximately holds.

Combining equation 5 and Assumption 3, we can write the similarity function with the Gaussian kernel as follows:

$$s_f(x^g, y^g) = f' \circ G_\sigma(\|x^g - y^g\|^2), \tag{6}$$

where we have used the following lemma:

Lemma 4 (uniform marginals). *Under Assumption 3, the marginal distributions $p_g(x^g)$ and $p_g(y^g)$ are uniform on the hypersphere $\mathbb{S}^{d-1}$, with d the feature dimension.*

The detailed proof can be seen in Appendix A.3, and we provide a sketch here. Note that we choose x^g and y^g to be both normalized such that when we marginalize either of them, the integral would be a constant due to rotational invariance. For example, there exists constant C such that:

$$p(x^g) = \int_{\mathbb{S}^{d-1}} C \exp\left(-\frac{\|x^g - y^g\|^2}{2\sigma^2}\right) dy^g = C \int_{\mathbb{S}^{d-1}} \exp\left(-\frac{1 - x^g \cdot y^g}{\sigma^2}\right) dy^g. \tag{7}$$

Whatever the value of x^g is, the marginal $p(x^g)$ remains the same, since any two x^g's on a unit sphere are equal. Thus, the product of marginals $p_g(x^g)p_g(y^g)$ is a constant and can be absorbed into a weighting parameter α that we will introduce in Section 3.3. We observe that $s_f(\cdot)$ depends on the choice of f as well, thus we call it ***f*-Gaussian similarity**. As a result, we have provided the answer to the second question, again from the f-MI perspective. We will empirically compare it with the cosine similarity in §4.

3.3 Implementation

With our designed f-Gaussian similarity s_f we now have an implementable f-MICL objective in eq. (4). Bringing the f-Gaussian Similarity s_f in eq. (6) into our objective eq. (4) we have a specific f-MICL objective:

$$\mathop{\mathbb{E}}_{(x,y)\sim p_+} s_f(x^g, y^g) - \mathop{\mathbb{E}}_{(x,y)\sim p_\times} f^* \circ s_f(x^g, y^g). \tag{8}$$

Given a batch of N samples, its empirical estimation is as follows:

$$\frac{1}{N}\sum_{i=1}^{N} s_f(x_i^g, y_i^g) - \frac{1}{N(N-1)}\sum_{i \neq j} f^* \circ s_f(x_i^g, x_j^g), \tag{9}$$

where x_i and y_i are two types of data augmentation of the i-th sample, and x_i and x_j are different samples of independently sampled data augmentations.

With the f-MICL objective in eq. (9) we propose our algorithm for contrastive learning in Algorithm 1. To balance the two terms in our objective, we additionally include a weighting parameter α in front of the second term. This change can still be incorporated within our f-MICL framework, as we show in Appendix A.2. Figure 2 gives a high-level summary of our f-MICL framework. Given a batch of samples (*e.g.*, images) we generate *positive pairs* via data augmentation and *negative pairs* using other augmented samples in the same batch. This sampling method follows SimCLR (Chen et al., 2020).

[2]RealNVP is a flow-based model, which applies real-valued non-volume preserving transformation for log-likelihood computation.

Algorithm 1: f-MICL

Input: batch size N, function f, weighting parameter α, constant μ (in G_σ), variance σ^2

for *each sampled mini-batch* $\{z_i\}_{i=1}^N$ **do**

  for k *in* $1, \ldots, N$ **do**

    randomly sample two augmentation functions t_1 and t_2

    $y_k \leftarrow t_1(z_k)$, $x_k \leftarrow t_2(z_k)$

  define $s_f(x^g, y^g) = f' \circ G_\sigma(\|x^g - y^g\|^2)$

  compute $-\mathcal{L}$ as

$$\frac{1}{N}\sum_{i=1}^{N} s_f(x_i^g, y_i^g) - \frac{\alpha}{N(N-1)}\sum_{i \neq j} f^* \circ s_f(x_i^g, x_j^g)$$

  update g by minimizing $\mathcal{L}$

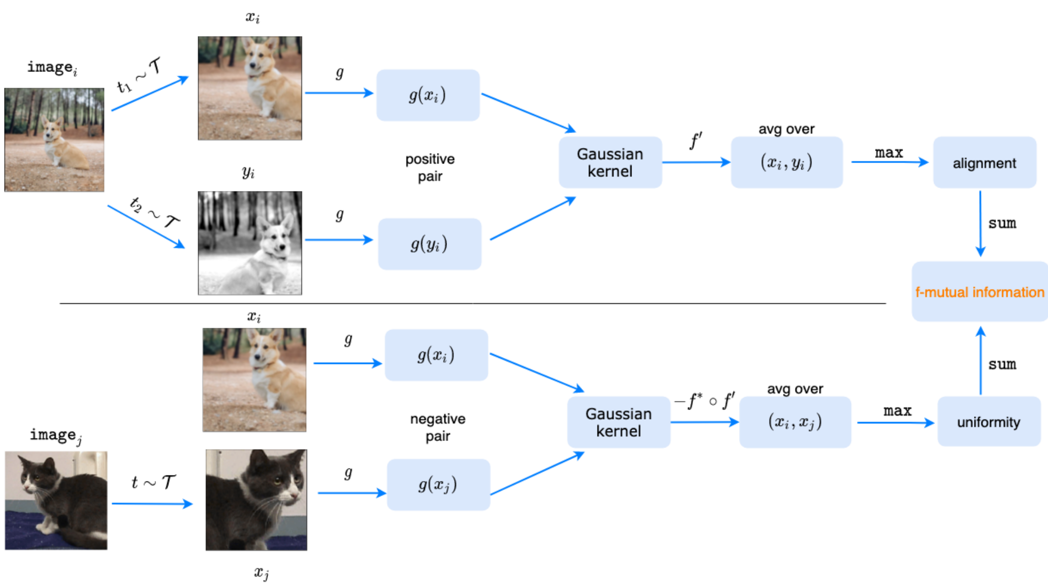

Figure 2: Network architecture of f-MICL. `image`$_i$: the i^{th} image in the current batch; f: the function used in the f-mutual information (§2); g: feature embedding; t, t_1, t_2: augmentation functions drawn from the same family $\mathcal{T}$ of augmentations; f': the derivative; f^*: the Fenchel conjugate. The symbol $\circ$ denotes the function composition. The sum of the two terms gives the variational lower bound of f-mutual information. x_i and y_i are two types of data augmentation of the i-th sample, and x_i and x_j are different samples of independently sampled data augmentations. `max` stands for maximization. See eq. (9) for more details.

3.4 f-MICL family

In this section, we will deepen the understanding of f-MICL by drawing connections with some popular constrastive learning methods.

Connection with InfoNCE: Firstly, we show that InfoNCE is an upper bound of f-MICL. Recall our f-MICL objective in eq. (4), and the popular InfoNCE objective $\mathcal{L}_{\text{InfoNCE}}$ as follows (here we take the maximization) (van den Oord et al., 2018):

$$\mathbb{E}_{(x,y)\sim p_+} s(x^g, y^g) - \mathbb{E}_{x\sim p_d} \log \mathbb{E}_{y\sim p_d} \exp(s(x^g, y^g)). \tag{10}$$

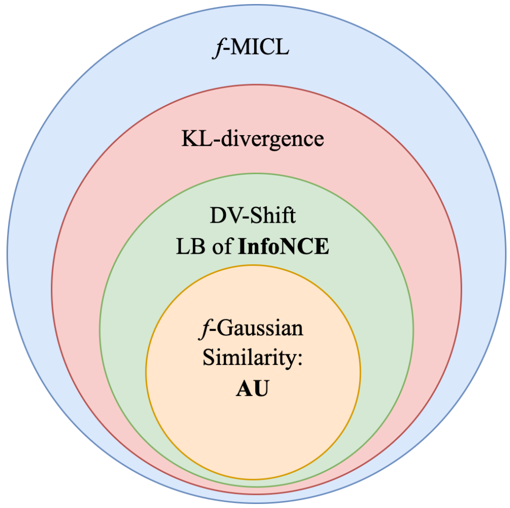

Figure 3: f-MICL generalizes InfoNCE-based objectives.

Consider that we perform a Donsker-Varadhan (DV) shift transformation v (Donsker & Varadhan, 1975; Tsai et al., 2021) from eq. (4) such that by taking the maximum over the transformation we have:

$$\sup_{v\in\mathbb{R}} \Big(\mathbb{E}_{(x,y)\sim p_+} s(x^g, y^g) - v - \mathbb{E}_{(x,y)\sim p_\times} f^* \circ (s(x^g, y^g) - v)\Big). \tag{11}$$

In practice, such a shift transformation can be approximated by a scaling factor (α in Algorithm 1) such that eq. (4) and eq. (11) are equivalent. Given that f is the KL divergence, thus $f(u) = u \log u$ and $f^*(t) = \exp(t-1)$ from Table 1, the maximizer of v in eq. (11) occurs at $v^* = \log(\mathbb{E}_{(x,y)\sim p_\times} s(x^g, y^g)) - 1)$. With v^* eq. (11) can be written as follows:

$$\mathbb{E}_{(x,y)\sim p_+} s(x^g, y^g) - \log \mathbb{E}_{(x,y)\sim p_\times} \exp(s(x^g, y^g)). \tag{12}$$

According to Jensen's inequality we have

$$\mathbb{E}_{x\sim p(\cdot)} \log \mathbb{E}_{y\sim p(\cdot)} \exp(s(x^g, y^g)) \le \log \mathbb{E}_{(x,y)\sim p_\times} \exp(s(x^g, y^g)). \tag{13}$$

Therefore,

$$\mathbb{E}_{(x,y)\sim p_+} s(x^g, y^g) - \log \mathbb{E}_{(x,y)\sim p_\times} \exp(s(x^g, y^g)) \le \mathcal{L}_{\text{InfoNCE}}. \tag{14}$$

The above transformation shows that the InfoNCE loss is an upper bound of the f-MICL objective. In other words, maximizing f-MICL can potentially increase the InfoNCE objective.

Connection with Alignment and Uniformity (AU) (Wang & Isola, 2020): We further show that the Alignment and Uniformity (AU) loss is a special cases of f-MICL. Wang & Isola (2020) shows that InfoNCE approximately aligns positive feature embeddings while encouraging uniformly distributed negative ones. Wang & Isola (2020) further proposes a new objective which quantifies such properties. Here we show that this new objective is essentially a subclass of the InfoNCE loss under the f-MICL framework. Concretely, applying the f-Gaussian similarity function for the KL divergence, we have $f'(u) = \log u + 1$ from Table 1 and thus $s_f(x^g, y^g) = -\|x^g - y^g\|^2$. Using $s_f(x^g, y^g)$ in eq. (12) we can recover the AU objective:

$$-\mathbb{E}_{(x,y)\sim p_+} \|x^g - y^g\|^2 - \log \mathbb{E}_{(x,y)\sim p_\times} \left[\exp(-\|x^g - y^g\|^2)\right]. \tag{15}$$

Note that for KL, this is equivalent to the cosine similarity with a scaling factor: $-\|x^g - y^g\|^2 = 2(x^g)^\top y^g - 2$.

Connection with the Spectral Contrastive Loss: Here we show that f-MICL objective is closely related to the Spectral Contrastive Loss (HaoChen et al., 2021). Recall our objective:

$$\mathbb{E}_{(x,y)\sim p_+} s(x^g, y^g) - \mathbb{E}_{(x,y)\sim p_\times} f^* \circ s(x^g, y^g), \tag{16}$$

where $s_f(x^g, y^g) = f' \circ G_\sigma(\|x^g - y^g\|^2)$. If we choose the Pearson χ^2 divergence, where $f(u) = (u-1)^2, f'(u) = 2(u-1), f^* \circ f'(u) = u^2 - 1$, we have our χ^2-MICL objective:

$$2\mathbb{E}_{(x,y)\sim p_+} G_\sigma(\|x^g - y^g\|^2) - \mathbb{E}_{(x,y)\sim p_\times} G_\sigma(\|x^g - y^g\|^2)^2 - 3. \tag{17}$$

This recovers the spectral contrastive loss exactly if we choose the proper hyperparameter and apply the cosine similarity instead. Thus we generalize the spectral contrastive loss as a special case of χ^2-MICL.

More on AU: Finally, based on our objective in eq. (8) we will show that the alignment and uniformity (AU) property of InfoNCE also extends to the general f-MICL family: (1) Alignment: In the ideal case, maximizing the first term of eq. (8) would yield $x^g = y^g$ for all $(x, y) \sim p_+$, *i.e.*, similar sample pairs should have aligned representations. Note that the derivative f' is increasing since f is convex. (2) Uniformity: We demonstrate the uniformity property by minimizing the second term of eq. (8), or more rigorously and realistically, its empirical version in eq. (9).

Theorem 5 (**Uniformity**). *Suppose that the batch size N satisfies $2 \leq N \leq d+1$, with d the dimension of the feature space. If the real function*

$$h(t) = f^* \circ f' \circ G_\sigma(t) \textit{ is strictly convex on } [0, 4], \tag{18}$$

then all minimizers of the second term of eq. (9), i.e., $\sum_{i \neq j} f^* \circ s_f(x_i^g, x_j^g)$, *satisfy the following condition: the feature representations of all samples are distributed uniformly on the unit hypersphere $\mathbb{S}^{d-1}$.*

In Theorem 5, the assumption $N \leq d+1$ is always satisfied in our experiments in §4. For instance, on CIFAR-10 we chose $N = d = 512$. Also, we claim that the samples are "distributed uniformly" if the feature vectors form a regular simplex, and thus the distances between all sample pairs are the same. Although minimizing the negative term gives uniformity, the positive term is also needed for aligning similar pairs, as we observe in §4. This implies the *tradeoff* between alignment and uniformity.

eq. (18) in fact provides us guidance to select proper f-divergences for uniformity. In Table 1, we list some common choices of f-divergences. By inspecting the last column and using the definition of G_σ, we can easily verify that they all satisfy eq. (18). However, this is not true for all f-divergences. In Appendix A.1 we also provide some counterexamples that violate eq. (18) and thus Theorem 5, such as the Reversed Kullback–Leibler (RKL) and the Neynman χ^2 divergences. Experimentally, we found that these divergences generally result in feature collapse (*i.e.*, all feature vectors are the same) and thus poor performance in downstream applications.

4 Experiments

In this section, we empirically evaluate the analysis on our provided answers: (1) Can we go beyond the KL-based objective (§4.2): we apply various f-MICL objectives to popular vision and language datasets. In particular, we show that under the same network architecture design, f-MICL can always provide a better choice of objective. We observe that the best-performing f-divergence is largely dataset dependent. (2) Can we design a better similarity function (§4.3): we show that the proposed f-Gaussian similarity is more powerful than the heuristic cosine similarity, regardless of the choice of f. Moreover, we confirm empirically that f-MICL extends the nice property of alignment and uniformity in §4.4.

4.1 Experimental settings

Our detailed settings can be found in Appendix D. In all our experiments, we change only the objective of different methods for fair comparison. We use the f-Gaussian similarity in f-MICL by default.

Vision task. Our vision datasets include CIFAR-10 (Krizhevsky et al., 2009), STL-10 (Coates et al., 2011), TinyImageNet (Chrabaszcz et al., 2017), and ImageNet (Deng et al., 2009) for image classification. After learning the feature embeddings, we evaluate the quality of representations using the test accuracy via a linear classifier. Note that we use $\alpha = 40$ across all vision experiments.
(1) Smaller datasets: For feature encoders, we use ResNet-18 (He et al., 2016) for CIFAR-10; ResNet-50 (He

Table 2: We compare the test accuracy (%) obtained with the linear evaluation on the vision datasets. On the Wikipedia dataset, we compare the semantic textual similarity (STS) via the Spearman's correlation. For each dataset and each method we take three different runs to get the mean and the standard derivation.

Dataset	Baselines			f-MICL			
	MoCo	SimCLR	AU	KL	JS	Pearson	VLC
CIFAR-10	$90.30_{\pm 0.19}$	$89.71_{\pm 0.37}$	$90.41_{\pm 0.26}$	$\mathbf{90.61}_{\pm \mathbf{0.47}}$	$89.66_{\pm 0.28}$	$89.35_{\pm 0.52}$	$89.13_{\pm 0.33}$
STL-10	$83.69_{\pm 0.22}$	$82.97_{\pm 0.32}$	$84.44_{\pm 0.19}$	$85.33_{\pm 0.39}$	$\mathbf{85.94}_{\pm \mathbf{0.17}}$	$82.64_{\pm 0.37}$	$\mathbf{85.94}_{\pm \mathbf{0.72}}$
TinyImageNet	$35.72_{\pm 0.17}$	$30.56_{\pm 0.28}$	$41.20_{\pm 0.19}$	$39.46_{\pm 0.20}$	$42.98_{\pm 0.18}$	$\mathbf{43.45}_{\pm \mathbf{0.54}}$	$38.65_{\pm 0.45}$
Wikipedia	$77.88_{\pm 0.15}$	$77.40_{\pm 0.12}$	$77.95_{\pm 0.08}$	$\mathbf{78.02}_{\pm \mathbf{0.13}}$	$76.76_{\pm 0.09}$	$77.59_{\pm 0.12}$	$55.07_{\pm 0.13}$

Table 3: We compare the test accuracy (%) with SOTA methods on ImageNet. We take three different runs to get the mean, where the standard derivations are less than 0.1% for f-MICL.

Dataset	Baselines						f-MICL		
	SwAV	BYOL	Barlow Twins	VICReg	RényiCL	MoCo v3	KL	JS	Pearson
ImageNet	75.3	74.3	73.2	73.2	**76.2**	73.2	73.9	**76.5**	74.6

et al., 2016) for the rest. Our implementation is based on SimCLR (Chen et al., 2020), where we used the same 3-layer projection head during training. All models are trained for 800 epochs.
(2) ImageNet: We choose Vision Transformer (ViT-S) (Dosovitskiy et al., 2020) as our feature encoder. We choose the smaller ViT-S model with 6 blocks because larger ViT models are extremely expensive to train on GPUs. Our implementations are based on MoCo V3 (Chen et al., 2021), where models are trained for 1000 epochs.

Language task. To show the wide applicability of our f-MICL framework, we also conduct experiments on a natural language dataset, English Wikipedia (Gao et al., 2021). We follow the experimental setting in (Gao et al., 2021), which applies BERT-based models to SimCLR (Devlin et al., 2019; Liu et al., 2019). Specifically, we choose the $\text{BERT}_{\texttt{base}}$ model due to limited computing resources. For f-MICL objectives, we choose $\alpha = 409600$. The application task is called semantic textual similarity (STS (Agirre et al., 2013)) and we report the averaged Spearman's correlation in Table 2 for comparison.

4.2 f-MICL objectives

Smaller datasets and language task. We first compare f-MICL with several InfoNCE-based contrastive learning algorithms (i.e., SimCLR (Chen et al., 2020), MoCo (He et al., 2020), and AU (Wang & Isola, 2020)) on smaller datasets and the language task in Table 2. Here we choose four f-divergences with the best overall performance. See Appendix D for results on other f-divergences.

From Table 2 we observe that: (1) As we have shown in §3.4 that f-MICL generalizes InfoNCE-based objectives, empirically KL-MICL achieves similar performance to the baselines. In practice, we can tune the hyperparameter α such that KL-MICL outperforms the InfoNCE-based objectives. (2) KL-MICL is not always the optimal choice. We can see that the best-performing f-MICL objectives refer to four different f-divergences on four datasets.

The above results indicate that f-MICL can provide a wide range of objective choices for the downstream tasks. Although how to derive an optimal f-divergence deserves more study in theory, in practice we can select the best f among several common f-divergences on a validation set.

Besides the f-divergences in Table 1, in Theorem 5 we have identified non-satisfying f-divergences. In our experiments, we found that applying these f-divergences such as the RKL and Neyman χ^2 divergences would result in feature collapse. For example, in Figure 4 we show that with RKL the features all collapse to a constant.

Table 4: Comparison between the cosine and f-Gaussian similarities on CIFAR-10 with the test accuracy (%). For the Tsallis-α divergence we take $\alpha = 3$.

Similarity	KL	JS	Pearson	SH	Tsallis-α	VLC
Cosine	89.95 $\pm$0.26	88.06 $\pm$0.33	87.79 $\pm$0.42	87.06 $\pm$0.55	88.55 $\pm$0.28	10.00 $\pm$0.00
Gaussian	**90.61** $\pm$0.47	**89.66** $\pm$0.28	**89.35** $\pm$0.52	**89.52** $\pm$0.25	**89.15** $\pm$0.42	**89.13** $\pm$0.33

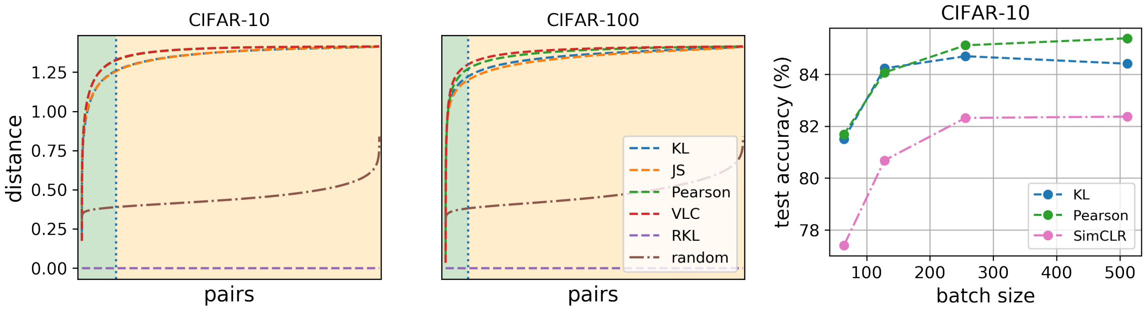

Figure 4: **(left and middle)** Distances between pairs of normalized features within a batch. **Green region:** similar pairs. **Orange region:** dissimilar pairs. f-MICL gives nearly uniform distances for dissimilar pairs for the f-divergences in Table 1. For non-satisfying f-divergences such as the RKL, the features collapse to a constant and thus the distances are zero. **(right)** The test accuracy v.s. the batch size after training 200 epochs for all algorithms.

ImageNet results: To further demonstrate the efficacy of f-MICL we then compare with several popular self-supervised learning methods, including both contrastive-based and non contrastive-based ones. These methods can be categorized into the ResNet-based (SwAV (Caron et al., 2020), BYOL (Huynh et al., 2020), Barlow Twins (Zbontar et al., 2021), VICReg (Bardes et al., 2021)) and RényiCL (Lee & Shin, 2022), and the ViT-based (Dosovitskiy et al., 2020). For the ResNet-based methods, we directly retrieve results from (Bardes et al., 2021), which are obtained by training ResNet-50 models for 1000 epochs. Chen et al. (Chen et al., 2021) show that these two types of methods are directly comparable in terms of the model size and supervised learning performance. For the ViT-based method and our f-MICL, we apply ViT-S for 1000 epochs. Specifically, our f-MICL follows MoCo v3 with only the objective changed for different choices of the f-divergence.

Results in Table 3 show that: (1) By only changing the objective function, our method improves MoCo v3 by 2.1% using JS-MICL. (2) f-MICL objectives are comparable with the ResNet-based methods (e.g., SwAV and RényiCL).

Overall, our experiments confirm that f-MICL can provide a better choice of objective than InfoNCE on a variety of datasets, tasks, and encoder architectures.

4.3 f-Gaussian similarity

Next, we want to examine the effect of our similarity function while fixing the f-divergences. In Table 4 we compare the cosine and f-Gaussian similarities for different f-divergences on CIFAR-10. It can be seen that under our f-MICL framework the f-Gaussian similarity consistently outperforms the cosine similarity for various f-divergences[3]. This agrees with our Theorem 2 and eq. (6), and also implies the validity of Assumption 3. In particular, we identify that without the theoretical guarantee, the heuristic cosine similarity would fail for certain f-MICL objectives (*e.g.*, VLC).

[3]As we have shown the equivalence between cosine and Gaussian similarity on KL, the difference results on KL just show the choice of different scaling factors.

4.4 Alignment and uniformity test

We empirically check the properties of alignment and uniformity for f-MICL by plotting the pairwise distance $\|x_i^g - x_j^g\|$ of the feature representations within the same batch on CIFAR-10. We compute the distances between the normalized features of every pair from a random batch, and then sort the pairs in increasing order. From Figure 4 we can see that f-MICL gives nearly uniform distances for dissimilar pairs (orange regions) on both datasets with various proper f-divergences. In contrast, a random initialized model gives a less uniform distribution for dissimilar pairs. Besides, we observe small pairwise distances for similar pairs (green regions).

4.5 Sensitivity to batch size

Finally, we study the sensitivity to the batch size of our f-MICL framework on CIFAR-10. On the right panel of Figure 4, we evaluate the classification accuracy by varying the batch size for different f-divergences and SimCLR. We can see that for all different batch sizes and with the proper choice of f-divergences, our performance is always better than SimCLR. In other words, we require fewer negative samples to achieve the same performance.

5 Related Work

Contrastive learning. Self-supervised contrastive learning learns representations by contrasting sample pairs. Recently it has been shown analytically that improving the contrastiveness can benefit the downstream applications (Saunshi et al., 2019; Tosh et al., 2021). For popular contrastive learning methods such as Contrastive Predictive Coding (CPC) (van den Oord et al., 2018), SimCLR (Chen et al., 2020), and MoCo (He et al., 2020), their loss functions can be interpreted as a lower bound of mutual information, which is essentially the KL divergence between the joint distribution and the product of margin distributions. Besides the KL divergence, other statistical divergences or distances have been individually studied under the context of contrastive learning, *e.g.*, the Wasserstein distance (Ozair et al., 2019), Pearson χ^2 divergence (Tsai et al., 2021), and Jensen–Shannon divergence (Hjelm et al., 2018).

f-divergences have been widely used in generative models (Nowozin et al., 2016) and domain adaptation (Acuna et al., 2021), for measuring the discrepancy of two distributions, where the variational lower bound is often employed for estimation. Compared to f-GAN (Nowozin et al., 2016) and f-DAL (Acuna et al., 2021) which minimize the f-divergence between two different distributions, our f-MICL objective is to maximize the f-divergence between the joint distribution and the product of marginal distributions. This agrees with our purpose of contrasting sample pairs. Moreover, we provide a theoretical criterion for choosing proper f-divergences.

Mutual Information also plays an important role in the context of deep representation learning (Tian et al., 2020a; Bachman et al., 2019; Hjelm et al., 2018; Tian et al., 2020b; Poole et al., 2019; Belghazi et al., 2018). *Loss function wise*, our losses partially cover the losses in the literature and generalizes them: e.g., Poole et al. (2019) considers several variational lower bounds of mutual information, where we generalize the DV objective; (b) *application-wise*, none of them considers contrastive learning: e.g., Poole et al. (2019) considers mutual information estimation, Belghazi et al. (2018) improves adversarial generative models, Hjelm et al. (2018) considers representation learning that maximizes local and global information.

Metric learning Our work is closely related to metric learning (Kaya & Bilge, 2019; Suárez-Díaz et al., 2018), which aims to learn a distance metric bringing similar objects closer and distancing dissimilar objects further. In contrastive learning, a pre-defined similarity metric, *e.g.*, the cosine similarity (Chen et al., 2020; He et al., 2020) or a bilinear function (van den Oord et al., 2018; Tian et al., 2020a; Hénaff et al., 2020) is commonly used to measure the sample similarity. These pre-designed metrics may not necessarily lead to satisfactory performance in practice. Comparably, the design of our similarity function is empirically tailored for contrastive learning.

6 Conclusion

We developed f-MICL for contrastive learning, which generalizes the KL-based mutual information to the f-mutual information. With f-MICL we provided a broad spectrum of objective choices with better downstream performance. We also proposed a novel f-Gaussian similarity function, which shows superior performance to the commonly used cosine similarity. In addition, we confirmed the generalization of f-MICL by comparing with popular InfoNCE-based objectives. Empirically, we exhibited the efficacy of f-MICL across a wide range of datasets from both vision and natural language.

Limitations and future work. While f-MICL provides a variety of objective functions, it is yet unclear how to choose an optimal f based on a task and a dataset in theory, such that we usually rely on a validation set in practice for selection. An interesting future work is to learn an optimal f-divergence using a parametrized neural network. Moreover, Lee & Shin (2022) applies Skew Rényi divergence for contrastive learning. However, we observe that applying Rényi-MICL naively leads to a large variance (similar to Section 4.1 in Lee & Shin (2022)), and we leave the discussion on skew divergences for future works.

References

David Acuna, Guojun Zhang, Marc T Law, and Sanja Fidler. f-domain-adversarial learning: Theory and algorithms. In *ICML*, 2021.

Eneko Agirre, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, and Weiwei Guo. *SEM 2013 shared task: Semantic textual similarity. In *Second joint conference on lexical and computational semantics (*SEM), volume 1: proceedings of the Main conference and the shared task: semantic textual similarity*, 2013.

Syed Mumtaz Ali and Samuel D Silvey. A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society: Series B (Methodological)*, 1966.

Philip Bachman, R Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. *NeurIPS*, 2019.

Maria-Florina Balcan, Avrim Blum, and Nathan Srebro. A theory of learning with similarity functions. *Machine Learning*, 72(1):89–112, 2008.

Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *arXiv preprint arXiv:2105.04906*, 2021.

Peter L Bartlett, Nick Harvey, Christopher Liaw, and Abbas Mehrabian. Nearly-tight VC-dimension and pseudodimension bounds for piecewise linear neural networks. *JMLR*, 2019.

Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeswar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and R Devon Hjelm. MINE: mutual information neural estimation. *arXiv preprint arXiv:1801.04062*, 2018.

Sergiy V Borodachov, Douglas P Hardin, and Edward B Saff. *Discrete energy on rectifiable sets*. Springer, 2019.

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

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020. URL `http://proceedings.mlr.press/v119/chen20j.html`.

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15750–15758, 2021.

Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9640–9649, 2021.

Patryk Chrabaszcz, Ilya Loshchilov, and Frank Hutter. A downsampled variant of ImageNet as an alternative to the CIFAR datasets. *arXiv preprint arXiv:1707.08819*, 2017.

Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 215–223. JMLR Workshop and Conference Proceedings, 2011.

Imre Csiszár. Information-type measures of difference of probability distributions and indirect observation. *studia scientiarum Mathematicarum Hungarica*, pp. 229–318, 1967.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*. IEEE, 2009.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019.

Laurent Dinh, David Krueger, and Yoshua Bengio. NICE: Non-linear independent components estimation. *arXiv preprint arXiv:1410.8516*, 2014.

Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real NVP. In *ICLR*, 2017.

Monroe D Donsker and SR Srinivasa Varadhan. Asymptotic evaluation of certain markov process expectations for large time, i. *Communications on Pure and Applied Mathematics*, 28(1):1–47, 1975.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. SimCSE: Simple contrastive learning of sentence embeddings. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2021.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in Neural Information Processing Systems*, 33:21271–21284, 2020.

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

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

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

Olivier J Hénaff, Aravind Srinivas, Jeffrey De Fauw, Ali Razavi, Carl Doersch, SM Eslami, and Aaron van den Oord. Data-efficient image recognition with contrastive predictive coding. In *ICML*. PMLR, 2020.

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

Tri Huynh, Simon Kornblith, Matthew R Walter, Michael Maire, and Maryam Khademi. Boosting contrastive self-supervised learning with false negative cancellation. *arXiv preprint arXiv:2011.11765*, 2020.

Mahmut Kaya and Hasan Şakir Bilge. Deep metric learning: A survey. *Symmetry*, 2019.

Diederik P Kingma and Prafulla Dhariwal. Glow: generative flow with invertible $1\times$ 1 convolutions. In *NeruIPS*, 2018.

Vladimir Koltchinskii. Rademacher penalties and structural risk minimization. *IEEE Transactions on Information Theory*, 47(5):1902–1914, 2001.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images, 2009. Technical report.

Lucien Le Cam. *Asymptotic methods in statistical decision theory*. Springer Science & Business Media, 2012.

Kyungmin Lee and Jinwoo Shin. Rényicl: Contrastive representation learning with skew rényi divergence. *Advances in Neural Information Processing Systems*, 35:6463–6477, 2022.

Ralph Linsker. Self-organization in a perceptual network. *Computer*, 1988.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. *International Conference on Learning Representations*, 2017.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2018.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2018.

Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

XuanLong Nguyen, Martin J Wainwright, and Michael I Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 2010.

Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. f-GAN: Training generative neural samplers using variational divergence minimization. In *NeurIPS*, 2016.

Sherjil Ozair, Corey Lynch, Yoshua Bengio, Aaron van den Oord, Sergey Levine, and Pierre Sermanet. Wasserstein dependency measure for representation learning. In *NeurIPS*, 2019.

Ben Poole, Sherjil Ozair, Aaron Van Den Oord, Alex Alemi, and George Tucker. On variational bounds of mutual information. In *International Conference on Machine Learning*, pp. 5171–5180. PMLR, 2019.

M. J. D. Powell. *Radial Basis Functions for Multivariable Interpolation: A Review*. Clarendon Press, USA, 1987. ISBN 0198536127.

Ralph Rockafellar. Characterization of the subdifferentials of convex functions. *Pacific Journal of Mathematics*, 1966.

Igal Sason and Sergio Verdú. f-divergence inequalities. *IEEE Transactions on Information Theory*, 2016.

Nikunj Saunshi, Orestis Plevrakis, Sanjeev Arora, Mikhail Khodak, and Hrishikesh Khandeparkar. A theoretical analysis of contrastive unsupervised representation learning. In *ICML*, 2019. URL `http://proceedings.mlr.press/v97/saunshi19a.html`.

Juan Luis Suárez-Díaz, Salvador García, and Francisco Herrera. A tutorial on distance metric learning: Mathematical foundations, algorithms, experimental analysis, prospects and challenges (with appendices on mathematical background and detailed algorithms explanation). *arXiv preprint arXiv:1812.05944*, 2018.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *ECCV*. Springer, 2020a.

Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning. In *NeurIPS*, 2020b. URL `https://proceedings.neurips.cc//paper_files/paper/2020/hash/4c2e5eaae9152079b9e95845750bb9ab-Abstract.html`.

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

Yao-Hung Hubert Tsai, Martin Q Ma, Muqiao Yang, Han Zhao, Louis-Philippe Morency, and Ruslan Salakhutdinov. Self-supervised representation learning with relative predictive coding. In *ICLR*, 2021.

Constantino Tsallis. Possible generalization of Boltzmann-Gibbs statistics. *Journal of statistical physics*, 1988.

Michael Tschannen, Josip Djolonga, Paul K. Rubenstein, Sylvain Gelly, and Mario Lucic. On mutual information maximization for representation learning. In *ICLR*, 2020.

Jean-Baptiste Hiriart Urruty and Claude Lemaréchal. *Convex analysis and minimization algorithms*. Springer-Verlag, 1993.

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

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

Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *CVPR*, 2018.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, pp. 12310–12320. PMLR, 2021.

Shaofeng Zhang, Feng Zhu, Junchi Yan, Rui Zhao, and Xiaokang Yang. Zero-cl: Instance and feature decorrelation for negative-free symmetric contrastive learning. In *International Conference on Learning Representations*, 2021.