What Do We Maximize in Self-Supervised Learning?

Anonymous Authors[1]

Abstract

This paper analyses self-supervised learning (SSL) methods, VICReg in particular, to provide for the first time an information-theoretical understanding of its construction. As a first step, we demonstrate how information-theoretic quantities can be obtained for a deterministic network, offering a possible alternative to prior works that rely on stochastic models. This enables us to demonstrate how VICReg can be (re-)discovered from first principles and its assumptions about data distribution. Furthermore, we demonstrated the validity of our assumptions empirically, confirming our novel understanding of VICReg. Finally, we believe that the derivation and insights we obtain can be generalized to many other SSL methods, opening new avenues for theoretical and practical understanding of SSL and transfer learning.

1. Introduction

Self-Supervised Learning (SSL) algorithms (Bromley et al., 1993) learn representations using a proxy objective (i.e., SSL objective) between inputs and self-defined signals. The results indicate that the learned representations can generalize well to a wide range of downstream tasks (Chen et al., 2020; Misra & Maaten, 2020), even when the SSL objective does not use downstream supervision during training. In SimCLR (Chen et al., 2020), for instance, a contrastive loss is defined between images with different augmentations (i.e., one as input and the other as a self-supervised signal). Then, we take our pre-learned model as a features extractor and adopt the features to various applications, spanning image classification, object detection, instance segmentation, and pose estimation (Caron et al., 2021). However, despite success in practice, only a few works (Arora et al., 2019; Lee et al., 2021) provide theoretical insights into the learning efficacy of SSL.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the First Workshop of Pre-training: Perspectives, Pitfalls, and Paths Forward at ICML 2022. Do not distribute.

In the last few years, information-theoretic methods have played a key role in several notable deep learning achievements, from practical applications in representation learning as the variational information bottleneck (Alemi et al., 2016)) to theoretical investigations (e.g., the generalization bound induced by mutual information (Xu & Raginsky, 2017; Steinke & Zakynthinou, 2020). Moreover, different deep learning problems have been successfully approached by developing and applying novel estimators and learning principles derived from information-theoretic quantities, such as mutual information estimation. Many works have attempted to analyze SSL from an information theory perspective. One example of this is using mutual information neural estimator (MINE) (Belghazi et al., 2018) in representation learning (Hjelm et al., 2018) in conjunction with the renowned information maximization (InfoMax) principle (Linsker, 1988a). But looking at these works may be confusing. Numerous objective functions, some contradicting each other, are presented, as well as many implicit assumptions. Moreover, these works rely on a crucial assumption: a stochastic (often Gaussian) DNN mapping, which is rarely the case in most cases these days.

This paper presents a unified framework for SSL methods from an information theory perspective which can be applied to deterministic DNN training. We summarize our contributions into two points: (i) First, we develop a method to study deterministic DNNs from an information theory perspective by shifting the stochasticity to the DNN input, which is a much more faithful assumption currently valid for most methods. (ii) Second, based on this formulation, we analyze how current SSL methods that use deterministic networks optimize information-theoretic quantities.

2. Background

Continuous Piecewise Affine (CPA) Mappings. A rich class of functions emerges from piecewise polynomials: spline operators. In short, given a partition, Ω of a domain $\mathbb{R}^D$, a spline of order k is a mapping defined by a polynomial of order k on each region $\omega \in \Omega$ with continuity constraints on the entire domain for the derivatives of order $0,\ldots,k-1$. As we will focus on affine splines ($k = 1$), we only define this case for concreteness. An K-dimensional affine spline

f produces its output via

$$f(\boldsymbol{z}) = \sum_{\omega \in \Omega} (\boldsymbol{A}_\omega \boldsymbol{z} + \boldsymbol{b}_\omega) \mathbb{1}_{\{\boldsymbol{z} \in \omega\}}, \tag{1}$$

with input $\boldsymbol{z} \in \mathbb{R}^D$ and $\boldsymbol{A}_\omega \in \mathbb{R}^{K \times D}, \boldsymbol{b}_\omega \in \mathbb{R}^K, \forall \omega \in \Omega$ the per-region *slope* and *offset* parameters respectively, with the key constraint that the entire mapping is continuous over the domain $f \in \mathcal{C}^0(\mathbb{R}^D)$. Spline operators and especially affine spline operators have been extensively used in function approximation theory (Cheney & Light, 2009), optimal control (Egerstedt & Martin, 2009), statistics (Fantuzzi et al., 2002), and related fields.

Deep Networks. A deep network (DN) is a (nonlinear) operator f_Θ with parameters Θ mapping a *input* $\boldsymbol{x} \in \mathbb{R}^D$ to a *prediction* $\boldsymbol{y} \in \mathbb{R}^K$. Precise definitions of DNs operators can be found in (Goodfellow et al., 2016). We will omit the Θ notation for clarity unless needed. The only assumption we require for our study is that the nonlinearities present in the DN are CPA, as is the case with (leaky-)ReLU, absolute value, and max-pooling. In that case, the entire input-output mapping becomes a CPA spline with an implicit partition Ω, the function of the network weights and architecture (Montufar et al., 2014; Balestriero & Baraniuk, 2018). For smooth nonlinearities, our results hold from a first-order Taylor approximation argument.

Self-Supervised Learning (joint-embedding) methods learn the DN parameters Θ without supervision and input reconstruction. Due to this formulation, the difficulty of SSL is to produce a good representation for downstream tasks whose labels are not available during training —while avoiding a trivially simple solution where the model maps all inputs to constant output. Many methods have been proposed to solve this problem. *Contrastive methods* learn representations by contrasting positive and negative examples, e.g., SimCLR (Chen et al., 2020) and its the InfoNCE criterion (Oord et al., 2018). Other recent works introduced *non-contrastive methods* that employ different regularization methods to prevent collapsing of the representation. Several papers used stop-gradients and extra predictors to avoid collapse (Chen & He, 2021; Grill et al., 2020) while (Caron et al., 2020) uses an additional clustering step. As opposed to contrastive methods, non-contrastive methods do not rely explicitly on negative samples. Of particular interest for us is the *VICReg* method (Bardes et al., 2021) that considers two embedding batches $\boldsymbol{Z} = [f(\boldsymbol{x}_1), \dots, f(\boldsymbol{x}_N)]$ and $\boldsymbol{Z}' = [f(\boldsymbol{x}'_1), \dots, f(\boldsymbol{x}'_N)]$ each of size $(N \times K)$. Denoting by $\boldsymbol{C}$ the $(K \times K)$ covariance matrix obtained from $[\boldsymbol{Z}, \boldsymbol{Z}']$ we obtain the VICReg triplet loss

$$\mathcal{L} = \frac{1}{K} \sum_{k=1}^{K} \left(\alpha \max\left(0, \gamma - \sqrt{\boldsymbol{C}_{k,k} + \epsilon}\right) + \beta \sum_{k' \neq k} (\boldsymbol{C}_{k,k'})^2 \right) + \gamma \|\boldsymbol{Z} - \boldsymbol{Z}'\|_F^2 / N.$$

Our goal will now be to formulate SSL as an information-theoretic problem from which we will be able to precisely relate VICReg to known methods even with a deterministic network.

3. Information Maximization of Deep Networks Outputs

This section first sets up notation and assumption on the information-theoretic challenges in self-supervised learning (Section 3.1) and on our assumptions regarding the data distribution (Section 3.2) so that any training sample $\boldsymbol{x}$ can be seen as coming from a single Gaussian distribution as in $\boldsymbol{x} \sim \mathcal{N}(\mu_{\boldsymbol{x}}, \Sigma_{\boldsymbol{x}})$. From this, we obtain that the output of any deep network $f(\boldsymbol{x})$ corresponds to a mixture of truncated Gaussian (Section 3.3). In particular, it can fall back to a single Gaussian under some small noise ($\det(\Sigma) \to \epsilon$) assumptions. These results will enable information measures to be applied to deterministic DNs. We then recover known SSL methods (Bardes et al., 2021) by making different assumptions on the data distribution and estimating their information.

3.1. SSL as an Information-Theoretic Problem

To better grasp the difference between key SSL methods, we first formulate the general SSL goal from an information-theoretical perspective.

Let's start with the InfoMax principle, (Linsker, 1988b) i.e., maximizing the mutual information between the representations of random variables Z and Z' of the two views. We can do so from a lower bound using

$$I(Z, Z') = H(Z) - H(Z|Z') \geq H(Z) + E[\log q(z|z')] \tag{2}$$

where $H(Z)$ is the entropy of Z. As opposed to supervised learning that employs fixed labels, in SSL, the entropy $H(Z)$ and $H(Z')$ are not constant and can be optimized throughout the learning process. Hence, only maximizing the above, we collapse to the trivial solution of making the representations constant. To regularize these entropies, i.e., prevent collapse, different methods utilize different approaches to implicit regularizing information. In order to recover them in Section 4, we must first introduce some notations and results around the data distribution (Section 3.2) and how a DN transforms that distribution (Section 3.3).

3.2. Data Distribution Hypothesis

Our first step is to assess how the output random variables of the network are represented, assuming a distribution on the data itself. Under a manifold hypothesis, any point can be seen as a Gaussian random variable with a low-rank covariance matrix in the direction of the manifold tangent space of the data. Hence, we will consider throughout this study the conditioning of a latent representation with respect to the mean of the observation, i.e., $X|\boldsymbol{x}^* \sim \mathcal{N}(\boldsymbol{x}^*, \Sigma_{\boldsymbol{x}^*})$ where the eigenvectors of $\Sigma_{\boldsymbol{x}^*}$ are in the same linear subspace than

the tangent space of the data manifold at $\boldsymbol{x}^*$, which varies with the position of $\boldsymbol{x}^*$ in space.

Hence a dataset is considered to be a collection of $\{\boldsymbol{x}_n^*, n = 1, \ldots, N\}$ and the full data distribution to be a sum of low-rank covariance Gaussian densities as in

$$X \sim \sum_{n=1}^{N} \mathcal{N}(\boldsymbol{x}_n^*, \Sigma_{\boldsymbol{x}_n^*})^{T=n}, T \sim \text{Cat}(N), \tag{3}$$

with T the uniform Categorical random variable. To keep things simple, and without loss of generality, we consider that the effective support of $\mathcal{N}(\boldsymbol{x}_i^*, \Sigma_{\boldsymbol{x}_i^*})$ and $\mathcal{N}(\boldsymbol{x}_j^*, \Sigma_{\boldsymbol{x}_j^*})$ do not overlap. This keeps things general as it is enough to cover the domain of the data manifold overall, without overlap between different Gaussians. Hence in general, we have that.

$$p(\boldsymbol{x}) \approx \mathcal{N}\left(\boldsymbol{x}; \boldsymbol{x}_{n(\boldsymbol{x})}^*, \Sigma_{\boldsymbol{x}_{n(\boldsymbol{x})}^*}\right)/N, \tag{4}$$

where $\mathcal{N}(\boldsymbol{x}; ., .)$ is the density of the Gaussian at $\boldsymbol{x}$, and with $n(\boldsymbol{x}) = \arg\min_n (\boldsymbol{x} - \boldsymbol{x}_n^*)^T \Sigma_{\boldsymbol{x}_n^*} (\boldsymbol{x} - \boldsymbol{x}_n^*)$. This assumption that a dataset is a mixture of Gaussians with non-overlapping support will simplify our derivations below, which could be extended to the general case if needed. Note that this is not restrictive since, given a large enough N, the above can represent any manifold with arbitrarily good approximation.

3.3. Data Distribution After Deep Network Transformation

Consider an affine spline operator f (Eq. 1) going from a space of dimension D to a space of dimension K with $K \geq D$ (I think that we can reduce this assumption as long as the covariance around the sample is low-rank (k), in that case. I think we would only need $K \geq k$).

The span —that we denote as image— of this mapping is given by

$$Im(f) \triangleq \{f(\boldsymbol{x}) : \boldsymbol{x} \in \mathbb{R}^D\} = \bigcup_{\omega \in \Omega} \text{Aff}(\omega; \boldsymbol{A}_\omega, \boldsymbol{b}_\omega) \tag{5}$$

with $\text{Aff}(\omega; \boldsymbol{A}_\omega, \boldsymbol{b}_\omega) = \{\boldsymbol{A}_\omega \boldsymbol{x} + \boldsymbol{b}_\omega : \boldsymbol{x} \in \omega\}$ the affine transformation of region ω by the per-region parameters $\boldsymbol{A}_\omega, \boldsymbol{b}_\omega$, and with Ω the partition of the input space in which $\boldsymbol{x}$ lives in. We also provide that analytical form of the per-region affine mappings in Section 2. Hence, the DN mapping consists of affine transformations on each input space partition region $\omega \in \Omega$ based on the coordinate change induced by $\boldsymbol{A}_\omega$ and the shift induced by $\boldsymbol{b}_\omega$.

When the input space is equipped with a density distribution, this density is transformed by the mapping f. In general, finding the density of $f(X)$ is an intractable task. However, given our disjoint support assumption provided in Section 3.2, we can arbitrarily increase the representation power of the density by increasing the number of prototypes N. In doing so, the support of each Gaussian is included with the region ω in which its means lie in, leading to the following result.

Theorem 1. *Given the setting of Equation (4) the unconditional DN output density denoted as Z is a mixture of the affinely transformed distributions $\boldsymbol{x}|\boldsymbol{x}_{n(\boldsymbol{x})}^*$ e.g. for the Gaussian case*

$$Z \sim \sum_{n=1}^{N} \mathcal{N}\Big(\boldsymbol{A}_{\omega(\boldsymbol{x}_n^*)} \boldsymbol{x}_n^* + \boldsymbol{b}_{\omega(\boldsymbol{x}_n^*)}, \boldsymbol{A}_{\omega(\boldsymbol{x}_n^*)}^T \Sigma_{\boldsymbol{x}_n^*} \boldsymbol{A}_{\omega(\boldsymbol{x}_n^*)}\Big)^{T=n},$$

where $\omega(\boldsymbol{x}_n^*) = \omega \in \Omega \iff \boldsymbol{x}_n^* \in \omega$ is the partition region in which the prototype $\boldsymbol{x}_n^*$ lives in.

The proof of the above involves that fact that if $\int_\omega p(\boldsymbol{x}|\boldsymbol{x}_{n(\boldsymbol{x})}^*) d\boldsymbol{x} \approx 1$ then f is linear within the effective support of p. Thus the output distribution is a linear transformation of the input distribution based on the per-region affine mapping.

4. Information Optimization and Optimality

Based on our analysis, we then show how specific algorithms can be derived for SSL. As we saw, when the noise in our input is small, the conditional output density $p_{\boldsymbol{z}|\boldsymbol{x}^*}$ reduces to a single Gaussian:

$$(Z|T = n) \sim \mathcal{N}(\mu_n, \Sigma_n),$$

where we abbreviated the parameters. Using that, and the result from Section 3.1, we see that one should optimize both $H(Z|X')$ and $H(Z)$. As in standard regression task, let's assume a Gaussian observation model, which mean $p(z|z') \sim \mathcal{N}(0, 1)$. we need to marginalize the stochasticity in Z'. To compute our expected loss, we can marginalize out Z' to obtain that the conditional decoding map is a Gaussian.

$$q(z|T' = n') = \int q(z|z', T' = n') p(z'|T' = n') dz', \tag{6}$$

which gives the distribution $\mathcal{N}(\mu_{n'}, I + \Sigma_{n'}))$ meaning we can lower bound the mutual information with

$$I(Z; X') \geq H(Z) + E[\log(q(z|x')] = H(Z) + \frac{d}{2} \log 2\pi - \sum_{i=1}^{N} \frac{1}{2} (z_i - z_i')^T \Sigma_r^{-1} (z_i - z_i') + \log |\Sigma_i|. \tag{7}$$

What happens if you attempt to optimize this objective? The only intractable component is the entropy of Z. We will begin by examining Z itself. It is natural to ask why the entropy of Z will not increase to infinity. Intuitively, the answer is that $H(Z)$ and $H(Z|X')$ are tied together, and one can not increase without the other. Now, recalling that under our distribution assumption, Z is a mixture of Gaussian (recall Thm. 1), we can see how existing upper and lower bounds could be used for this case; for example, the ones in (Moshksar & Khandani, 2016).

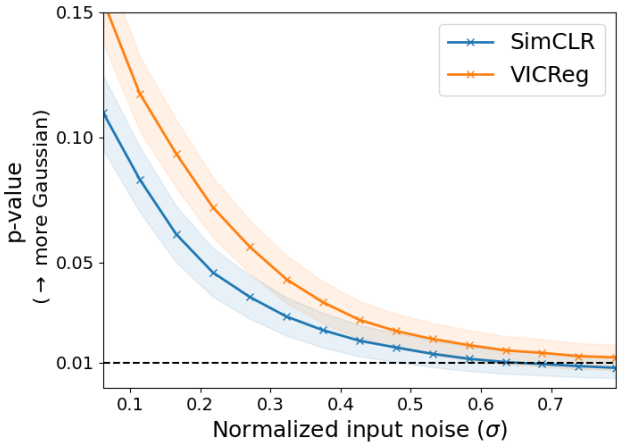

Figure 1. **The network output with VICReg training is more gaussian for small input noise**. The P-value of the normality test for different SSL models trained on CIFAR-10 for different input noise levels. The x-axis is the coefficient that multiplies the data distribution standard deviation to obtain the Gaussian standard deviation that samples around each image. The dashed line represents the point at which the null hypothesis (Gaussian distribution of the network output) can be rejected with 99% confidence.

4.1. Deriving VICReg From First Principles

We now propose to recover VICReg from the first principles per the above information-theoretic principle.

Recall that our goal is to estimate the entropy $H(Z)$ in 7, where Z is a Gaussian mixture. This quantity is not known for a mixture of Gaussians due to the logarithm of a sum of exponential functions, except for the special case of a single Gaussian density. There are, however, several approximations in the literature, including both upper and lower bounds. Among the methods, some use the log sum of the probabilities (Kolchinsky & Tracey, 2017), and some use an entropy-adjusted of the log probabilities (Huber et al., 2008).

An even simpler solution is to approximate the entire mixture as a single Gaussian by capturing only the first two moments of the mixture distribution. Since Gaussian distribution maximizes the entropy for a given covariance matrix, this method provides an upper bound approximation to our entropy of interest $H(Z)$. In this case, denoting by Σ_Z is the covariance matrix of Z, we get

$$I(Z;X') \geq \sum_{i=1}^{N} \log \frac{|\Sigma_Z|}{|\Sigma_i| + |\Sigma_r|} - \frac{1}{2}(z_i - z_i')^T \Sigma_r^{-1} (z_i - z_i')$$

where $|\Sigma_r|$ is constant with respect to our optimization process, and the second term is the prediction performance of one representation from the other one. A key result from Shi et al. (2009) connects the eigenvectors and eigenvalues of Σ_Z and the ones of each component $\Sigma_i, \forall i$, and showed that under the assumption that the separation $(\mu_i - \mu_j)\Sigma_i^{-1}(\mu_i - \mu_j)^T$ between the different components is large enough —which holds true in our case as per our data distribution model— the eigenfunctions of $\Sigma_i, \forall i$ are approximately the eigenfunctions of Σ_Z. Hence in our case, this means that since all those eigenvalues are tied, we only need to find the most efficient way to maximize $|\Sigma_Z|$.

We know that the determinant of the matrix is the product of its eigenvalues. For every positive matrix, the maximum eigenvalue is bigger or equal to each diagonal element. Therefore, under the constraint of the eigenvalues of the matrix, the most efficient way is to decrease the off-diagonal terms and increase the diagonal terms. By setting $\Sigma_r = I$, *we therefore fully recover the VICReg objective*.

4.2. Empirical Evaluation

The next step is to verify the validity of our assumptions. Based on the theory we presented3.3, the conditional output density $p_{\boldsymbol{z}|T=i}$ reduces to a single Gaussian with decreasing input noise. We validate it using a ResNet-18 model trained using SimCLR or VICReg on the CIFAR-10 dataset (Krizhevsky, 2009). From the test dataset, we sample 512 Gaussian samples for each image and analyze whether these samples remain Gaussian (for each image) at the penultimate layer of the DN, i.e., before the linear classification layer, independently for each output dimension. Then, we employ D'Agostino, and Pearson's test (D'Agostino, 1971) in order to compute the p-value of the normality test under the null hypothesis that the sample represents a normal distribution. The process is repeated for different noise standard deviations. Fig. 1 shows the p-value as a function of the normalized standard deviation. We can observe that the network's output is indeed Gaussian with a high probability for small input noise. VICReg is, interestingly, more "gaussian" than SimCLR, which may have to do with the fact that it optimizes only the second moments of the density distribution in order to regularize $H(Z)$.

5. Conclusions

In this study, we examine SSL's objective function from an information-theoretic perspective. Our analysis, based on transferring the required stochasticity to the input distribution, shows how to derive SSL objectives. Therefore, it is possible to obtain an information-theoretic analysis even when using deterministic DNs. In the second part, we rediscovered VICReg's loss function from first principles and showed its implicit assumptions. In short, VICReg performs a crude lower bound estimate of the output density entropy by means of approximating this distribution with a Gaussian matching the first two moments. Finally, we empirically validated that our assumptions hold in practice, thus confirming the validity of our novel understanding of VICReg. Our work opens many new paths for future research; A better estimation of information-theoretic quantities fits our assumptions. Another exciting research direction is identifying which SSL method is preferred based on data properties.

References

Alemi, A. A., Fischer, I., Dillon, J. V., and Murphy, K. Deep variational information bottleneck. *arXiv preprint arXiv:1612.00410*, 2016.

Arora, S., Khandeparkar, H., Khodak, M., Plevrakis, O., and Saunshi, N. A theoretical analysis of contrastive unsupervised representation learning. *arXiv preprint arXiv:1902.09229*, 2019.

Balestriero, R. and Baraniuk, R. A spline theory of deep networks. In *Proc. ICML*, volume 80, pp. 374–383, Jul. 2018.

Bardes, A., Ponce, J., and LeCun, Y. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *arXiv preprint arXiv:2105.04906*, 2021.

Belghazi, M. I., Baratin, A., Rajeswar, S., Ozair, S., Bengio, Y., Courville, A., and Hjelm, R. D. Mine: mutual information neural estimation. *arXiv preprint arXiv:1801.04062*, 2018.

Bromley, J., Guyon, I., LeCun, Y., Säckinger, E., and Shah, R. Signature verification using a" siamese" time delay neural network. *Advances in neural information processing systems*, 6, 1993.

Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., and Joulin, A. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in Neural Information Processing Systems*, 33:9912–9924, 2020.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., and Joulin, A. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9650–9660, 2021.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020.

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

Cheney, E. W. and Light, W. A. *A course in approximation theory*, volume 101. American Mathematical Soc., 2009.

D'Agostino, R. B. An omnibus test of normality for moderate and large size samples. *Biometrika*, 58(2):341–348, 1971. ISSN 00063444. URL http://www.jstor.org/stable/2334522.

Egerstedt, M. and Martin, C. *Control theoretic splines: optimal control, statistics, and path planning*. Princeton University Press, 2009.

Fantuzzi, C., Simani, S., Beghelli, S., and Rovatti, R. Identification of piecewise affine models in noisy environment. *International Journal of Control*, 75(18):1472–1485, 2002.

Goodfellow, I., Bengio, Y., and Courville, A. *Deep Learning*, volume 1. MIT Press, 2016.

Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in Neural Information Processing Systems*, 33:21271–21284, 2020.

Hjelm, R. D., Fedorov, A., Lavoie-Marchildon, S., Grewal, K., Bachman, P., Trischler, A., and Bengio, Y. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*, 2018.

Huber, M., Bailey, T., Durrant-Whyte, H., and Hanebeck, U. On entropy approximation for gaussian mixture random vectors. pp. 181 – 188, 09 2008. doi: 10.1109/MFI.2008.4648062.

Kolchinsky, A. and Tracey, B. D. Estimating mixture entropy with pairwise distances. *Entropy*, 19(7):361, 2017.

Krizhevsky, A. Learning multiple layers of features from tiny images. Technical report, 2009.

Lee, J. D., Lei, Q., Saunshi, N., and Zhuo, J. Predicting what you already know helps: Provable self-supervised learning. *Advances in Neural Information Processing Systems*, 34, 2021.

Linsker, R. Self-organization in a perceptual network. *Computer*, 21(3):105–117, 1988a.

Linsker, R. Self-organization in a perceptual network. *Computer*, 21(3):105–117, 1988b. doi: 10.1109/2.36.

Misra, I. and Maaten, L. v. d. Self-supervised learning of pretext-invariant representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6707–6717, 2020.

Montufar, G. F., Pascanu, R., Cho, K., and Bengio, Y. On the number of linear regions of deep neural networks. In *Proc. NeurIPS*, pp. 2924–2932, 2014.

Moshksar, K. and Khandani, A. K. Arbitrarily tight bounds on differential entropy of gaussian mixtures. *IEEE Transactions on Information Theory*, 62(6):3340–3354, 2016. doi: 10.1109/TIT.2016.2553147.

Oord, A. v. d., Li, Y., and Vinyals, O. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Shi, T., Belkin, M., and Yu, B. Data spectroscopy: Eigenspaces of convolution operators and clustering. *The Annals of Statistics*, pp. 3960–3984, 2009.

Steinke, T. and Zakynthinou, L. Reasoning about generalization via conditional mutual information. In *Conference on Learning Theory*, pp. 3437–3452. PMLR, 2020.

Xu, A. and Raginsky, M. Information-theoretic analysis of generalization capability of learning algorithms. *Advances in Neural Information Processing Systems*, 30, 2017.