5th Symposium on Advances in Approximate Bayesian Inference, 2023 1–36

An Information-Theoretic Perspective on Variance-Invariance-Covariance Regularization

Ravid Shwartz-Ziv, *New York University* RAVID.SHWARTZ.ZIV@NYU.EDU
Randall Balestriero, *Meta AI, FAIR*
Kenji Kawaguchi, *National University of Singapore*
Tim G. J. Rudner, *New York University*
Yann LeCun, *New York University & Meta AI, FAIR*

Abstract

In this paper, we provide an information-theoretic perspective on Variance-Invariance-Covariance Regularization (VICReg) for self-supervised learning. To do so, we first demonstrate how information-theoretic quantities can be obtained for deterministic networks as an alternative to the commonly used unrealistic stochastic networks assumption. Next, we relate the VICReg objective to mutual information maximization and use it to highlight the underlying assumptions of the objective. Then, we derive a generalization bound for VICReg, providing generalization guarantees for downstream supervised learning tasks and presenting novel self-supervised learning methods derived from a mutual information maximization objective that outperform existing methods in terms of performance. This work provides a new information-theoretic perspective on self-supervised learning and Variance-Invariance-Covariance Regularization in particular and guides the way for improved transfer learning via information-theoretic self-supervised learning objectives.

1. Introduction

Information-theoretic methods have played a key role in several advances in deep learning—from practical applications in representation learning (Alemi et al., 2016) to theoretical investigations (Xu and Raginsky, 2017; Steinke and Zakynthinou, 2020; Shwartz-Ziv, 2022). Some works have attempted to use information theory for SSL, such as the InfoMax principle (Linsker, 1988) in SSL (Bachman et al., 2019). However, these works often present objective functions without rigorous justification, make implicit assumptions (Kahana and Hoshen, 2022; Wang et al., 2022; Lee et al., 2021), and explicitly assume that the deep neural network mappings are stochastic—which is rarely the case for modern neural networks. See Shwartz-Ziv and LeCun (2023) for a detailed review.

This paper presents an information-theoretic perspective on Variance-Invariance-Covariance Regularization (VICReg; Bardes et al. (2021)). We show that the VICReg objective is closely related to approximate mutual information maximization, derive a generalization bound for VICReg, and relate the generalization bound to information maximization. We show that under a series of assumptions about the data, which we validate empirically, our results apply to deterministic deep neural network training and do not require further stochasticity assumptions about the network. Our key contributions are as follows:

1. We relate the VICReg objective to information-theoretic quantities and use this relationship to highlight the underlying assumptions of the objective.
2. We study the relationship between the optimization of information-theoretic quantities and predictive performance in downstream tasks by introducing a generalization bound that connects VICReg, information theory, and downstream generalization.
3. We present new information-theoretic SSL methods and evaluate them empirically.

© R. Shwartz-Ziv, R. Balestriero, K. Kawaguchi, T.G.J. Rudner & Y. LeCun.

2. An Information-Theoretic Perspective on SSL in Deterministic DNNs

In order to better understand and develop SSL methods, we first present the general SSL goal from an information-theoretic perspective. This enables the analysis and comparison of SSL methods based on their ability to maximize the mutual information between representations, potentially leading to new SSL methods.

While information-theoretic methods have contributed to deep learning achievements (Alemi et al., 2016; Steinke and Zakynthinou, 2020; Shwartz-Ziv and Tishby, 2017b), a key problem is the source of randomness in deterministic deep neural networks. The mutual information between the input and representation is infinite, causing ill-posed optimization problems or piecewise constant (Amjad and Geiger, 2019; Goldfeld et al., 2018). To address this, researchers have proposed various solutions, including stochastic deep networks, additive noise injection, and considering data augmentation as the source of noise (Lee et al., 2021; Shwartz-Ziv and Alemi, 2020; Goldfeld et al., 2018; Dubois et al., 2021).

In this work, we assume that the stochasticity comes from the data itself, which is a less restrictive assumption and does not require changing current algorithms. We assume that any training sample $\boldsymbol{x}$ can be seen as coming from a single Gaussian distribution, $\boldsymbol{x} \sim \mathcal{N}(\mu_{\boldsymbol{x}}, \Sigma_{\boldsymbol{x}})$. From this, we show that the output of any DNN $f(\boldsymbol{x})$ corresponds to a mixture of truncated Gaussian distributions. This enables information measures to be applied to deterministic DNNs. See Appendices 2, 2.1 and 2.2 for more details and validation of our assumptions.

After presenting the framework for analyzing information in deterministic networks, we show how we can analyze current methods from an information-theoretic perspective. We start with the *MultiView InfoMax principle*, which aims to maximize the mutual information between the representations of two different views, X and X', and their corresponding representations, Z and Z'. To maximize their information, we maximize $I(Z;X')$ and $I(Z';X)$ using the lower bound:

$$I(Z,X') = H(Z) - H(Z|X') \geq H(Z) + \mathbb{E}_{x'}[\log q(z|x')] \tag{1}$$

where $H(Z)$ is the entropy of Z.

3. Information Optimization and Optimality

Next, we will show how SSL algorithms for deterministic networks can be derived from information-theoretic principles. According to Section 2, we want to maximize $I(Z;X')$ and $I(Z';X)$. Although this mutual information is intractable in general, we can obtain a tractable variational approximation. First, when the input noise is small, namely that the effective support of the Gaussian centered at x is contained within the region w of the DNN's input space partition, we can reduce the conditional output density to a single Gaussian: $(Z'|X'=x_n) \sim \mathcal{N}\left(\mu(x_n), \Sigma(x_n)\right)$, where $\mu(x_n) = \boldsymbol{A}_{\omega(\boldsymbol{x}_n)}\boldsymbol{x}_n + \boldsymbol{b}_{\omega(\boldsymbol{x}_n)}$ and $\Sigma(x_n) = \boldsymbol{A}^T_{\omega(\boldsymbol{x}_n)}\Sigma_{\boldsymbol{x}_n}\boldsymbol{A}_{\omega(\boldsymbol{x}_n)}$. Second, to compute the expected loss, we need to marginalize out the stochasticity in the output of the network. In general, training with squared loss is equivalent to maximum likelihood estimation in a Gaussian observation model, $p(z|z') \sim \mathcal{N}(z', \Sigma_r)$, where $\Sigma_r = I$. To compute the expected loss over samples of x', we need to marginalize out the stochasticity in Z': which means that the conditional decoder is a Gaussian: $(Z|X'=x_n) \sim \mathcal{N}(\mu(x_n), \Sigma_r + \Sigma(x_n))$. However, the expected log loss over samples of Z is hard to compute. We instead focus on a lower bound; the expected log loss over samples of Z'. For simplicity, let $\Sigma_r = I$. By Jensen's inequality, we then obtain the following lower bound on $\mathbb{E}_{x'}\left[\log q(z|x')\right]$:

$$\mathbb{E}_{x'}\left[\log q(z|x')\right] \geq \mathbb{E}_{z'|x'}\left[\log q(z|z')\right] = \frac{1}{2}(d\log 2\pi - \left(z-\mu(x')\right)^2 - \operatorname{Tr}\log\Sigma(x')). \tag{2}$$

Now, taking the expectation over Z, we obtain

$$\mathbb{E}_{z|x}\left[\mathbb{E}_{z'|x'}\left[\log q(z|z')\right]\right] = \frac{1}{2}(d\log 2\pi - \left(\mu(x)-\mu(x')\right)^2 - \log\left(|\Sigma(x)|\cdot|\Sigma(x')|\right)). \tag{3}$$

Full derivations of Equations (2) and (3) are given in Appendix 6. Combining all of the above then yields

$$\begin{aligned} I(Z;X') &\geq H(Z) + \mathbb{E}_{x,z|x,x',z'|x'}[\log q(z|z')] &(4)\\ &= H(Z) + \frac{d}{2}\log 2\pi - \frac{1}{2}\mathbb{E}_{x,x'}[(\mu(x)-\mu(x'))^2 + \log(|\Sigma(x)|\cdot|\Sigma(x')|)]. &(5)\end{aligned}$$

To optimize this objective in practice, we can approximate $p(x,x')$ using the empirical data distribution:

$$L \approx \frac{1}{N}\sum_{i=1}^{N} \underbrace{H(Z) - \log\left(|\Sigma(x_i)|\cdot|\Sigma(x_i')|\right)}_{\text{Regularizer}} - \underbrace{\frac{1}{2}\left(\mu(x_i)-\mu(x_i')\right)^2}_{\text{Invariance}}. \tag{6}$$

Next, we will discuss how estimating the intractable entropy $H(Z)$ changes the objective.

3.1. An Information-Theoretic Perspective on VICReg

In the previous section, we derived an objective function based on information-theoretical principles. The "invariance term" in Equation (6) is similar to the invariance loss of VICReg. However, computing the regularization term—and $H(Z)$ in particular—is challenging. Estimating the entropy of random variables is a classic problem in information theory, with the Gaussian mixture density being a popular representation. However, there is no closed-form solution to the differential entropy of Gaussian mixtures. Approximations, including loose upper and lower bounds (Huber et al., 2008) and Monte Carlo sampling, exist in the literature. Unfortunately, Monte Carlo sampling is computationally expensive and requires many samples in high dimensions (Brewer, 2017).

One of the simplest and straightforward approaches to approximating the entropy is to capture the first two moments of the distribution, which provides an upper bound on the entropy. However, minimizing an upper bound means that there is no guarantee that the original objective is being optimized. In practice, there have been cases where successful results have been achieved by minimizing an upper bound (Martinez et al., 2021; Nowozin et al., 2016). However, this may cause instability in the training process. For a detailed discussion and results on various entropy estimators, see Section 4. Letting Σ_Z be the covariance matrix of Z, we will use the first two moments to approximate the entropy we wish to maximize. Using this approach, we obtain the following approximation

$$L \approx \sum_{n=1}^{N} \log \frac{|\Sigma_Z|}{|\Sigma(x_i)| \cdot |\Sigma(x_i')|} - \frac{1}{2}(\mu(x) - \mu(x'))^2. \tag{7}$$

For a discussion of this approximation, see Appendix 3.

4. Self-Supervised Learning via Mutual Information Maximization

Implementing Equation (4) in practice requires various design choices. As shown in Section 4, VICReg approximates entropy based on certain assumptions. We now compare VICReg with other SSL methods, such as contrastive learning methods like SimCLR, and non-contrastive methods like BYOL and SimSiam, to examine their implementation of the information maximization objective.

By analyzing their assumptions and differing approaches, we propose new objective functions incorporating recent information and entropy estimators from the information theory literature. This helps improve SSL performance and enhances our understanding of the underlying learning mechanisms.

4.1. Alternative Entropy Estimators

The VICReg objective approximates the log determinant of the empirical covariance matrix using diagonal terms, but this can be problematic (Section 3.1). We instead employ alternative entropy estimators, such as the LogDet Entropy Estimator (Zhouyin and Liu, 2021), which provides a tighter upper bound. To address the limitations of the upper bound, we use a lower bound estimator based on pairwise distances of mixture components (Kolchinsky and Tracey, 2017). These estimators are computationally efficient, continuous, smooth, and converge to the exact solution for well-separated clusters. We compare these methods with VICReg, SimCLR (Chen et al., 2020), and Barlow Twin (Zbontar et al., 2021).

Setup. Our experiments are conducted on CIFAR-10 (Krizhevsky and Hinton, 2009), and ResNet-18 architecture (He et al., 2016) as the backbone. We use linear evaluation for the quality of the representation. For full details, see Appendix 13.

Results. It can be seen from Table 1 that the proposed estimators outperform both the original VICReg and SimCLR as well as Barlow Twin. By estimating the entropy with a more accurate estimator, we can improve the results of VICReg, and the pairwise distance estimator, which is a lower bound, achieves the best results. This aligns with the theory that we want to maximize a lower bound on true entropy. The results of our study suggest that a smart selection of entropy estimators, inspired by our framework, leads to better results.

5. Information Maximization for VICReg and Downstream Generalization

In the previous sections, we showed the connection between information-theoretic principles and the VICReg objective. Next, we will connect it to the downstream generalization of VICReg by deriving a generalization bound. These findings, along with the results from previous sections, connect generalization in VICReg to information maximization and implicit regularization.

Table 1: CIFAR-10 predictive accuracy on linear evaluation of SSL. The suggested Entropy estimators achieved better results on SSL than previous works.

Method	**Accuracy** (in %)
SimCLR	89.72 ± 0.05
Barlow Twins	88.81 ± 0.10
VICReg	89.32 ± 0.09
VICReg + Pairwise Distances Estimator (**ours**)	$\mathbf{90.09 \pm 0.09}$
VICReg + Log-Determinant Estimator (**ours**)	89.77 ± 0.08

Notation. Consider input points x, outputs $y \in \mathbb{R}^r$, labeled training data $S = ((x_i, y_i))_{i=1}^n$ of size n and unlabeled training data $\mathbf{S} = ((x_i^+, x_i^{++}))_{i=1}^m$ of size m, where x_i^+ and x_i^{++} share the same (unknown) label. With the unlabeled training data, we define the invariance loss

$$I_{\mathbf{S}}(f_\theta) = \frac{1}{m} \sum_{i=1}^m \|f_\theta(x_i^+) - f_\theta(x_i^{++})\| \tag{8}$$

where f_θ is the trained representation on the unlabeled data $\mathbf{S}$. We define the labeled loss $\ell_{x,y}(w) = \|W f_\theta(x) - y\|$ where $w = \text{vec}[W] \in \mathbb{R}^{dr}$ is the vectorization of the matrix $W \in \mathbb{R}^{r \times d}$. Let $w_S = \text{vec}[W_S]$ be the minimum norm solution as $W_S = \text{minimize}_{W'} \|W'\|_F$ such that

$$W' \in \operatorname*{arg\,min}_W \frac{1}{n} \sum_{i=1}^n \|W f_\theta(x_i) - y_i\|^2. \tag{9}$$

We also define the representation matrices

$$Z_S = [f(x_1), \ldots, f(x_n)] \in \mathbb{R}^{d \times n} \qquad \text{and} \qquad Z_{\mathbf{S}} = [f(x_1^+), \ldots, f(x_m^+)] \in \mathbb{R}^{d \times m},$$

and the projection matrices

$$\mathbf{P}_{Z_S} = I - {Z_S}^\intercal (Z_S {Z_S}^\intercal)^\dagger Z_S \qquad \text{and} \qquad \mathbf{P}_{Z_{\mathbf{S}}} = I - {Z_{\mathbf{S}}}^\intercal (Z_{\mathbf{S}} {Z_{\mathbf{S}}}^\intercal)^\dagger Z_{\mathbf{S}}.$$

We define the label matrix $Y_S = [y_1, \ldots, y_n]^\intercal \in \mathbb{R}^{n \times r}$ and the unknown label matrix $Y_{\mathbf{S}} = [y_1^+, \ldots, y_m^+]^\intercal \in \mathbb{R}^{m \times r}$, where y_i^+ is the unknown label of x_i^+. Let $\mathcal{F}$ be a hypothesis space of f_θ. For a given hypothesis space $\mathcal{F}$, we define the normalized Rademacher complexity

$$\tilde{\mathcal{R}}_m(\mathcal{F}) = \frac{1}{\sqrt{m}} \mathbb{E}_{\mathbf{S},\xi} \left[\sup_{f \in \mathcal{F}} \sum_{i=1}^m \xi_i \|f(x_i^+) - f(x_i^{++})\| \right],$$

where $\xi_1, \ldots, \xi_m$ are independent uniform random variables taking values in $\{-1, 1\}$. It is normalized such that $\tilde{\mathcal{R}}_m(\mathcal{F}) = O(1)$ as $m \to \infty$ for typical choices of hypothesis spaces $\mathcal{F}$, including deep neural networks (Bartlett et al., 2017; Kawaguchi et al., 2018).

5.1. A Generalization Bound for Variance-Invariance-Covariance Regularization

Theorem 1 shows that VICReg improves generalization on supervised downstream tasks. More specifically, minimizing the unlabeled invariance loss while controlling the covariance $Z_{\mathbf{S}} {Z_{\mathbf{S}}}^\intercal$ and the complexity of representations $\tilde{\mathcal{R}}_m(\mathcal{F})$ minimizes the expected *labeled loss*:

Theorem 1 *(Informal version). For any $\delta > 0$, with probability at least $1 - \delta$,*

$$\mathbb{E}_{x,y}[\ell_{x,y}(w_S)] \leq I_{\mathbf{S}}(f_\theta) + \frac{2}{\sqrt{m}}\|\mathbf{P}_{Z_{\mathbf{S}}}Y_{\mathbf{S}}\|_F + \frac{1}{\sqrt{n}}\|\mathbf{P}_{Z_S}Y_S\|_F + \frac{2\tilde{\mathcal{R}}_m(\mathcal{F})}{\sqrt{m}} + \mathcal{Q}_{m,n}, \tag{10}$$

where $\mathcal{Q}_{m,n} = O(G\sqrt{\ln(1/\delta)/m} + \sqrt{\ln(1/\delta)/n}) \to 0$ as $m, n \to \infty$. In $\mathcal{Q}_{m,n}$, the value of G for the term decaying at the rate $1/\sqrt{m}$ depends on the hypothesis space of f_θ and w whereas the term decaying at the rate $1/\sqrt{n}$ is independent of any hypothesis space.

Proof The complete version of Theorem 1 and its proof are presented in Appendix 14. ■

The term $\|\mathbf{P}_{Z_{\mathbf{S}}}Y_{\mathbf{S}}\|_F$ in Theorem 1 contains the unobservable label matrix $Y_{\mathbf{S}}$. However, we can minimize this term by using $\|\mathbf{P}_{Z_{\mathbf{S}}}Y_{\mathbf{S}}\|_F \leq \|\mathbf{P}_{Z_{\mathbf{S}}}\|_F\|Y_{\mathbf{S}}\|_F$ and by minimizing $\|\mathbf{P}_{Z_{\mathbf{S}}}\|_F$. The factor $\|\mathbf{P}_{Z_{\mathbf{S}}}\|_F$ is minimized when the rank of the covariance $Z_{\mathbf{S}}Z_{\mathbf{S}}{}^\intercal$ is maximized. Since a strictly diagonally dominant matrix is non-singular, this can be enforced by maximizing the diagonal entries while minimizing the off-diagonal entries, as is done in VICReg. For example, if $d \geq n$, then $\|\mathbf{P}_{Z_{\mathbf{S}}}\|_F = 0$ when the covariance $Z_{\mathbf{S}}Z_{\mathbf{S}}{}^\intercal$ is of full rank.

The term $\|\mathbf{P}_{Z_S}Y_S\|_F$ contains only observable variables, and we can directly measure the value of this term using training data. The term $\|\mathbf{P}_{Z_S}Y_S\|_F$ is also minimized when the rank of the covariance $Z_SZ_S{}^\intercal$ is maximized. Since the covariances $Z_SZ_S{}^\intercal$ and $Z_{\mathbf{S}}Z_{\mathbf{S}}{}^\intercal$ concentrate to each other via concentration inequalities with the error in the order of $O(\sqrt{(\ln(1/\delta))/n} + \tilde{\mathcal{R}}_m(\mathcal{F})\sqrt{(\ln(1/\delta))/m})$, we can also minimize the upper bound on $\|\mathbf{P}_{Z_S}Y_S\|_F$ by maximizing the diagonal entries of $Z_{\mathbf{S}}Z_{\mathbf{S}}{}^\intercal$ while minimizing its off-diagonal entries, as in VICReg.

Thus, VICReg can be understood as a method to minimize the generalization bound in Theorem 1 by minimizing the invariance loss while controlling the covariance $Z_{\mathbf{S}}Z_{\mathbf{S}}{}^\intercal$ to minimize the *label-agnostic* upper bounds on $\|\mathbf{P}_{Z_{\mathbf{S}}}Y_{\mathbf{S}}\|_F$ and $\|\mathbf{P}_{Z_S}Y_S\|_F$. If we know *partial* information about the label $Y_{\mathbf{S}}$ of the unlabeled data, we can use it to minimize $\|\mathbf{P}_{Z_{\mathbf{S}}}Y_{\mathbf{S}}\|_F$ and $\|\mathbf{P}_{Z_S}Y_S\|_F$ directly. This direction can be used to improve VICReg in future work for the partially observable setting. In Appendix 5, we compare this bound to other bounds and discuss how Theorem 1 can be understood via mutual information maximization.

6. Conclusions

In this study, we investigated the VICReg algorithm for SSL through an information-theoretic lens. By shifting the necessary stochasticity for information-theoretic analysis to the input distribution, we demonstrated how the VICReg objective can be derived from information-theoretic principles. This perspective allowed us to uncover implicit assumptions within its objective, derive a generalization bound for downstream tasks, and relate it to information maximization. Additionally, we leveraged the insights from our analysis to propose a novel VICReg-style SSL objective.

Our results indicate that VICReg's performance can be further enhanced in settings with partial label information by aligning the covariance matrix with the partially observable label matrix. This finding presents numerous opportunities for future research, such as developing improved estimators for information-theoretic quantities and exploring the appropriateness of various SSL methods based on specific data attributes.

References

Alessandro Achille and Stefano Soatto. Information dropout: Learning optimal representations through noisy computation. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2897–2905, 2018.

N.A. Ahmed and D.V. Gokhale. Entropy expressions and their estimators for multivariate distributions. *IEEE Transactions on Information Theory*, 35(3):688–692, 1989. doi: 10.1109/18.30996.

Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. Deep variational information bottleneck. *arXiv preprint arXiv:1612.00410*, 2016.

Rana Ali Amjad and Bernhard Claus Geiger. Learning representations for neural network-based classification using the information bottleneck principle. *IEEE transactions on pattern analysis and machine intelligence*, 2019.

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 Richard Baraniuk. A spline theory of deep networks. In *Proc. ICML*, volume 80, pages 374–383, Jul. 2018.

Randall Balestriero and Yann LeCun. Contrastive and non-contrastive self-supervised learning recover global and local spectral embedding methods, 2022. URL https://arxiv.org/abs/2205.11508.

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 and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

Peter L Bartlett, Dylan J Foster, and Matus J Telgarsky. Spectrally-normalized margin bounds for neural networks. *Advances in neural information processing systems*, 30, 2017.

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.

Itamar Ben-Ari and Ravid Shwartz-Ziv. Attentioned convolutional lstm inpaintingnetwork for anomaly detection in videos. *arXiv preprint arXiv:1811.10228*, 2018.

Brendon J Brewer. Computing entropies with nested sampling. *Entropy*, 19(8):422, 2017.

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 *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

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

Elliott Ward Cheney and William Allan Light. *A course in approximation theory*, volume 101. American Mathematical Soc., 2009.

Ralph B. D'Agostino. 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.

Zheng Dang, Kwang Moo Yi, Yinlin Hu, Fei Wang, Pascal Fua, and Mathieu Salzmann. Eigendecomposition-free training of deep networks with zero eigenvalue-based losses. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 768–783, 2018.

Yann Dubois, Benjamin Bloem-Reddy, Karen Ullrich, and Chris J Maddison. Lossy compression for lossless prediction. *Advances in Neural Information Processing Systems*, 34, 2021.

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

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

Marco Federici, Anjan Dutta, Patrick Forré, Nate Kushman, and Zeynep Akata. Learning robust representations via multi-view information bottleneck. In *International Conference on Learning Representations*, 2019.

Charles Fefferman, Sanjoy Mitter, and Hariharan Narayanan. Testing the manifold hypothesis. *Journal of the American Mathematical Society*, 29(4):983–1049, 2016.

Mike B. Giles. Collected matrix derivative results for forward and reverse mode algorithmic differentiation. In Christian H. Bischof, H. Martin Bücker, Paul Hovland, Uwe Naumann, and Jean Utke, editors, *Advances in Automatic Differentiation*, pages 35–44, Berlin, Heidelberg, 2008. Springer Berlin Heidelberg.

Z. Goldfeld, E. van den Berg, K. Greenewald, I. Melnyk, N. Nguyen, B. Kingsbury, and Y. Polyanskiy. Estimating Information Flow in Neural Networks. *ArXiv e-prints*, 2018.

Noah Golowich, Alexander Rakhlin, and Ohad Shamir. Size-independent sample complexity of neural networks. In *Conference On Learning Theory*, pages 297–299. PMLR, 2018.

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

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.

Suriya Gunasekar, Blake E Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nati Srebro. Implicit regularization in matrix factorization. In *Advances in Neural Information Processing Systems*, pages 6151–6159, 2017.

Suriya Gunasekar, Jason D Lee, Daniel Soudry, and Nati Srebro. Implicit bias of gradient descent on linear convolutional networks. In *Advances in Neural Information Processing Systems*, pages 9461–9471, 2018.

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.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 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. *arXiv preprint arXiv:1808.06670*, 2018.

Marco Huber, Tim Bailey, Hugh Durrant-Whyte, and Uwe Hanebeck. On entropy approximation for gaussian mixture random vectors. In *IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems*, pages 181 – 188, 09 2008. doi: 10.1109/MFI.2008.4648062.

Catalin Ionescu, Orestis Vantzos, and Cristian Sminchisescu. Matrix backpropagation for deep networks with structured layers. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2965–2973, 2015. doi: 10.1109/ICCV.2015.339.

Jonathan Kahana and Yedid Hoshen. A contrastive objective for learning disentangled representations. *arXiv preprint arXiv:2203.11284*, 2022.

Kenji Kawaguchi, Leslie Pack Kaelbling, and Yoshua Bengio. Generalization in deep learning. *MIT-CSAIL-TR-2018-014, Massachusetts Institute of Technology*, 2018.

Kenji Kawaguchi, Zhun Deng, Kyle Luh, and Jiaoyang Huang. Robustness Implies Generalization via Data-Dependent Generalization Bounds. In *International Conference on Machine Learning (ICML)*, 2022.

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

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

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical Report 0, University of Toronto, Toronto, Ontario, 2009.

Kuang-Huei Lee, Anurag Arnab, Sergio Guadarrama, John Canny, and Ian Fischer. Compressive visual representations. *Advances in Neural Information Processing Systems*, 34, 2021.

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

Julieta Martinez, Jashan Shewakramani, Ting Wei Liu, Ioan Andrei Bârsan, Wenyuan Zeng, and Raquel Urtasun. Permute, quantize, and fine-tune: Efficient compression of neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15699–15708, 2021.

Neeraj Misra, Harshinder Singh, and Eugene Demchuk. Estimation of the entropy of a multivariate normal distribution. *Journal of Multivariate Analysis*, 92(2):324–342, 2005. ISSN 0047-259X. doi: https://doi.org/10.1016/j.jmva.2003.10.003.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2012.

Guido F Montufar, Razvan Pascanu, Kyunghyun Cho, and Yoshua Bengio. On the number of linear regions of deep neural networks. In *Proc. NeurIPS*, pages 2924–2932, 2014.

Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. f-gan: Training generative neural samplers using variational divergence minimization. In *Advances in Neural Information Processing systems*, pages 271–279, 2016.

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

Zoe Piran, Ravid Shwartz-Ziv, and Naftali Tishby. The dual information bottleneck. *arXiv preprint arXiv:2006.04641*, 2020.

Nikunj Saunshi, Orestis Plevrakis, Sanjeev Arora, Mikhail Khodak, and Hrishikesh Khandeparkar. A theoretical analysis of contrastive unsupervised representation learning. In *International Conference on Machine Learning*, pages 5628–5637. PMLR, 2019.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Ravid Shwartz-Ziv. Information flow in deep neural networks. *arXiv preprint arXiv:2202.06749*, 2022.

Ravid Shwartz-Ziv and Alexander A Alemi. Information in infinite ensembles of infinitely-wide neural networks. In *Symposium on Advances in Approximate Bayesian Inference*, pages 1–17. PMLR, 2020.

Ravid Shwartz-Ziv and Yann LeCun. To compress or not to compress–self-supervised learning and information theory: A review. *arXiv preprint arXiv:2304.09355*, 2023.

Ravid Shwartz-Ziv and Naftali Tishby. Compression of deep neural networks via information, (2017). *arXiv preprint arXiv:1703.00810*, 2017a.

Ravid Shwartz-Ziv and Naftali Tishby. Opening the black box of deep neural networks via information. *arXiv preprint arXiv:1703.00810*, 2017b.

Ravid Shwartz-Ziv, Amichai Painsky, and Naftali Tishby. Representation compression and generalization in deep neural networks, 2018.

Ravid Shwartz-Ziv, Randall Balestriero, Kenji Kawaguchi, and Yann LeCun. What do we maximize in self-supervised learning and why does generalization emerge?, 2023. URL https://openreview.net/forum?id=tuE-MnjN7DV.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *The Journal of Machine Learning Research*, 19(1):2822–2878, 2018.

Thomas Steinke and Lydia Zakynthinou. Reasoning about generalization via conditional mutual information. In *Conference on Learning Theory*, pages 3437–3452. PMLR, 2020.

Haoqing Wang, Xun Guo, Zhi-Hong Deng, and Yan Lu. Rethinking minimal sufficient representation in contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16041–16050, 2022.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, pages 9929–9939. PMLR, 2020.

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

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*, pages 12310–12320. PMLR, 2021.

Zhanghao Zhouyin and Ding Liu. Understanding neural networks with logarithm determinant entropy estimator. *arXiv preprint arXiv:2105.03705*, 2021.

Roland S Zimmermann, Yash Sharma, Steffen Schneider, Matthias Bethge, and Wieland Brendel. Contrastive learning inverts the data generating process. In *International Conference on Machine Learning*, pages 12979–12990. PMLR, 2021.

Supplementary Material

1. Background & Preliminaries

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 $0,\dots,k-1$. As we will focus on affine splines ($k=1$), we define this case only for concreteness. A K-dimensional affine spline f produces its output via

$$f(\boldsymbol{z}) = \sum\nolimits_{\omega \in \Omega} (\boldsymbol{A}_\omega \boldsymbol{z} + \boldsymbol{b}_\omega) \mathbb{1}_{\{\boldsymbol{z} \in \omega\}}, \tag{1.1}$$

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 widely used in function approximation theory (Cheney and Light, 2009), optimal control (Egerstedt and Martin, 2009), statistics (Fantuzzi et al., 2002), and related fields.

Deep Neural Networks as CPA Mappings. A deep neural network (DNN) is a (non-linear) operator f_Θ with parameters Θ that map a *input* $\boldsymbol{x} \in \mathbb{R}^D$ to a *prediction* $\boldsymbol{y} \in \mathbb{R}^K$. The precise definitions of DNN operators can be found in Goodfellow et al. (2016). To avoid cluttering notation, we will omit Θ unless needed for clarity. The only assumption we require for our analysis is that the non-linearities present in the DNN are CPA mappings—as is the case with (leaky-) ReLU, absolute value, and max-pooling operators. The entire input–output mapping then becomes a CPA spline with an implicit partition Ω, the function of the weights and architecture of the network (Montufar et al., 2014; Balestriero and Baraniuk, 2018). For smooth nonlinearities, our results hold by employing a first-order Taylor approximation argument.

Self-Supervised Learning. Joint embedding methods learn DNN parameters Θ without the need for supervision and input reconstruction. The difficulty of self-supervised learning (SSL) is generating a good representation for downstream tasks whose labels are unavailable during self-supervised training while avoiding trivial solutions where the model maps all inputs to a constant output. Many methods have been proposed to solve this problem (see Balestriero and LeCun (2022) for a summary and connections between methods). *Contrastive methods*, such as SimCLR (Chen et al., 2020) and its InfoNCE criterion (Oord et al., 2018), learn representations by contrasting positive and negative examples. In contrast, *non-contrastive methods* employ different regularization methods to prevent collapsing of the representation and do not explicitly rely on negative samples. Some methods use stop-gradients and extra predictors to avoid collapse (Chen and He, 2021; Grill et al., 2020) while Caron et al. (2020) use an additional clustering step. Of particular interest to us is the *Variance-Invariance-Covariance Regularization* method(VICReg; 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 $[\boldsymbol{Z}, \boldsymbol{Z}']$, the VICReg triplet loss is given by

$$\mathcal{L} = \frac{1}{K}\sum_{k=1}^{K}(\alpha \mathrm{Var}(Z_k) + \beta \mathrm{Cov}(Z_k, Z_{k'})) + \gamma \mathrm{Inv}(Z_k, Z_{k'}), \tag{1.2}$$

where

$$\mathrm{Var}(Z_k) = \max(0, \gamma - \sqrt{\boldsymbol{C}_{k,k} + \epsilon}) \tag{1.3}$$

$$\mathrm{Cov}(Z_k, Z_{k'}) = \sum\nolimits_{k' \neq k} \left(\boldsymbol{C}_{k,k'}\right)^2 \tag{1.4}$$

$$\mathrm{Inv}(Z_k, Z_{k'}) = \|\boldsymbol{Z}_k - \boldsymbol{Z}_{k'}\|_F^2 / N. \tag{1.5}$$

Deep Networks and Information-Theory. Recently, information-theoretic methods have played an essential role in advancing deep learning (Alemi et al., 2016; Xu and Raginsky, 2017; Steinke and Zakynthinou, 2020; Shwartz-Ziv and Tishby, 2017b) by developing and applying information-theoretic estimators and learning principles to DNN training (Hjelm et al., 2018; Belghazi et al., 2018; Piran et al., 2020; Shwartz-Ziv et al., 2018). However, information-theoretic objectives for deterministic DNNs often exhibit a common pitfall: They assume that DNN mappings are stochastic- an assumption that is usually violated. As a result, the mutual information between the input and the DNN representation in such objectives would be infinite, resulting in ill-posed optimization problems. To avoid this problem, stochastic DNNs with variational bounds could be used, where the output of the deterministic network is used as the parameters of the conditional distribution (Lee et al., 2021; Shwartz-Ziv and Alemi, 2020). Dubois et al. (2021) assumed that the randomness of data augmentation among the two views is the source of stochasticity in the network. Other work assumed a random input, but without making any assumptions about the properties of the distribution of the network's output, to analyze the objective and relied on general lower bounds (Wang and Isola, 2020; Zimmermann et al., 2021). For supervised learning, Goldfeld et al. (2018) introduced an auxiliary (noisy) DNN by injecting additive noise into the model and demonstrated that the resulting model is a good proxy for the original (deterministic) DNN in terms of both performance and representation. Finally, Achille and Soatto (2018) found that minimizing a stochastic network with a regularizer is equivalent to minimizing the cross-entropy over deterministic DNNs with multiplicative noise. All of these methods assume that the source of randomness comes from the DNN, contradicting common practice.

2. Assumptions

2.1. Data Distribution Hypothesis

First, we examine the way the output random variables of the network are represented and assume a distribution over the data. Under the 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 (Fefferman et al., 2016). Therefore, throughout this study, we will consider 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 as 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

$$X \sim \sum_{n=1}^{N} \mathcal{N}(\boldsymbol{x}_n^*, \Sigma_{\boldsymbol{x}_n^*})^{1_{\{T=n\}}}, T \sim \text{Cat}(N), \tag{2.6}$$

with T the uniform Categorical random variable. For simplicity, 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, where the effective support is defined as $\{x \in \mathbb{R}^D : p(x) > \epsilon\}$. Therefore, we have that.

$$p(\boldsymbol{x}) \approx \mathcal{N}\left(\boldsymbol{x}; \boldsymbol{x}_{n(\boldsymbol{x})}^*, \Sigma_{\boldsymbol{x}_{n(\boldsymbol{x})}^*}\right)/N, \tag{2.7}$$

where $\mathcal{N}(\boldsymbol{x}; ., .)$ is the Gaussian density 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 and could be extended to the general case if needed.

2.2. Data Distribution Under the Deep Neural Network Transformation

Consider an affine spline operator f (Equation (1.1)) that goes from a space of dimension D to a space of dimension K with $K \geq D$. The span, which we denote as an image, of this mapping is given by

$$\text{Im}(f) \triangleq \{f(\boldsymbol{x}) : \boldsymbol{x} \in \mathbb{R}^D\} = \bigcup\nolimits_{\omega \in \Omega} \text{Aff}(\omega; \boldsymbol{A}_\omega, \boldsymbol{b}_\omega) \tag{2.8}$$

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. The practical computation of the per-region affine mapping can be obtained by setting $\boldsymbol{A}_\omega$ to the Jacobian matrix of the network at the corresponding input x, and b to be defined as $f(x) - \boldsymbol{A}_\omega x$. Therefore, the DNN 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, the density of $f(X)$ is intractable. However, given the disjoint support assumption from Appendix 2.1, we can arbitrarily increase the representation power of the density by increasing the number of prototypes N. By doing so, the support of each Gaussian is included within the region ω in which its means lie, leading to this result:

Theorem 2 *Given the setting of Equation (2.7), the unconditional DNN output density, Z, is approximately a mixture of the affinely transformed distributions $\boldsymbol{x}|\boldsymbol{x}_{n(\boldsymbol{x})}^*$:*

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

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.

Proof See Appendix 7. ■

2.3. Validation of Assumptions

Based on the theory outlined in Appendix 2.2, the conditional output density $p_{\boldsymbol{z}|x=i}$ can be reduced to a single Gaussian with decreasing input noise. To validate this, we used a ResNet-18 model trained with either SimCLR or VICReg objectives on the CIFAR-10 and CIFAR-100 datasets (Krizhevsky, 2009). We sampled 512 Gaussian samples for each image from the test dataset, and analyzed whether each sample remained Gaussian in the penultimate layer of the DNN. We then used the D'Agostino and Pearson's test (D'Agostino, 1971) to determine the validity of this assumption. Figure 1 (left) shows the p-value as a function of the normalized standard deviation. For small noise, we can reject the hypothesis that the conditional output density of the network is not Gaussian with a probability of 85% for VICReg. However, as the input noise increases, the network's output becomes less Gaussian and even for the small noise regime, there is a 15% chance of a Type I error.

Next, to confirm our assumption that the model of the data distribution has non-overlapping effective support, we calculated the distribution of pairwise l_2 distances between images for seven datasets: MNIST, CIFAR10, CIFAR100, Flowers102, Food101, FGVAircaft. Figure 1 (right) shows that even for raw pixels, the pairwise distances are far from zero, which means that we can use a small Gaussian around each point without overlapping. Since the effective support of these datasets is non-overlapping, our assumption is realistic.

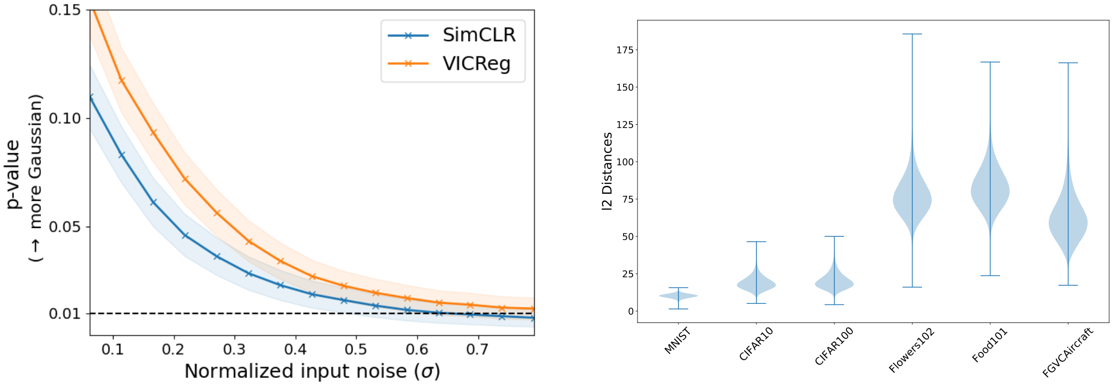

Figure 1: **Left: The network output SSL training is more Gaussian for small input noise**. The p-value of the normality test for different SSL models trained on CIFAR-10 for different input noise levels. The dashed line represents the point at which the null hypothesis (Gaussian distribution) can be rejected with 99% confidence. **Right: The Gaussians around each point are not overlapping** The plots show the $l2$ distances between raw images for different datasets. As can be seen, the distances are largest for more complex real-world datasets.

3. Discussion of Approximation

A standard fact in linear algebra is that the determinant of a matrix is the product of its eigenvalues. Therefore, maximizing the sum of the log eigenvalues implies maximizing the log determinant of Z. Many works have considered this problem (Giles, 2008; Ionescu et al., 2015; Dang et al., 2018). One approach is to find the solutions using the eigendecomposition, which leads to numerical instability (Dang et al., 2018). An alternative approach is to diagonalize the covariance matrix and increase its diagonal elements. Because the eigenvalues of a diagonal matrix are the diagonal entries, increasing the sum of the log-diagonal terms is equivalent to increasing the sum of the log eigenvalues. One way to do this is to push the off-diagonal terms of Σ_Z to be zero and maximize the sum of its log diagonal. This can be done using the covariance term of VICReg. Even though this approach is simple and efficient, the values on the diagonal may become close to zero, which may cause instability when we calculate the logarithm. Therefore, we use an upper bound and calculate the sum of the diagonal elements directly, which is the variance term of VICReg. In conclusion, we see the connection between the information-theoretic objective and the three terms of CIVReg. An exciting research direction is to maximize the eigenvalues of Z using more sophisticated methods, such as using a differential expression for eigendecomposition.

4. VICReg vs. SimCLR

Contrastive Learning with SimCLR. Lee et al. (2021) connect the SimCLR objective (Chen et al., 2020) to the variational bound on the information between representations by using the von Mises-Fisher distribution as the conditional variational family. By applying our analysis for information in deterministic networks with their work, we can compare the differences between SimCLR and VICReg, and identify two main differences: (i) **Conditional distribution:** SimCLR assumes a von Mises-Fisher distribution for the encoder, while VICReg assumes a Gaussian distribution. (ii) **Entropy estimation:** The entropy term in SimCLR is approximate and based on the finite sum of the input samples. In contrast, VICReg estimates the entropy of Z solely based on the second moment. Creating self-supervised methods that combine these two differences would be an interesting future research direction.

Empirical comparison. As we saw in previous sections, the different methods use different objective functions to optimize the entropy of their representation. Next, we compare the SSL methods and check directly their entropy. To do so, we trained ResNet-18 architecture (He et al., 2016) on CIFAR-10 (Krizhevsky and Hinton, 2009) for VICReg, SimCLR and BYOL. we used the *pairwise distances* entropy estimator based on the distances of the individual mixture component (Kolchinsky and Tracey, 2017). Even though this quantity is just an estimator for the entropy, it is shown as a tight estimator and is directly optimized by neither one of the methods. Therefore, we can treat it as a outsource validation of the entropy for the different methods. For more details on this and other entropy estimators see Section 4.1. In Figure 2, we see that, as expected from our analysis before, all the entropy decreased during the training for all the methods. Additionally, we see that SimCLR has the lowest entropy during the training, while VICReg has the highest one.

5. Discussion of Generalization Bound

5.1. Comparison of Generalization Bounds

The SimCLR generalization bound (Saunshi et al., 2019) requires the number of label classes to go infinity to close the generalization gap, whereas the VICReg bound in Theorem 1 does *not* require the number of label classes to approach infinity for the generalization gap to go to zero. This reflects the fact that, unlike SimCLR, VICReg does not use negative pairs and thus does not use a loss function that is based on the implicit expectation that the labels of a negative pair (y^+, y^-) are different. Another difference is that our VICReg bound improves as n increases, while the previous bound of SimCLR (Saunshi et al., 2019) does not depend on n. This is because Saunshi et al. (2019) assume partial access to the true distribution $p(x \mid y)$ per class for setting W, which removes the importance of labeled data size n and is not assumed in our study.

Consequently, the generalization bound in Theorem 1 provides a new insight for VICReg regarding the ratio of the effects of m v.s. n through $G\sqrt{\ln(1/\delta)/m} + \sqrt{\ln(1/\delta)/n}$. Finally, Theorem 1 also illuminates the advantages of VICReg over standard supervised training. That is, with standard training, the generalization bound via the Rademacher complexity requires the complexities of hypothesis spaces, $\tilde{\mathcal{R}}_n(\mathcal{W})/\sqrt{n}$ and $\tilde{\mathcal{R}}_n(\mathcal{F})/\sqrt{n}$, with respect to the size of labeled data n, instead of the size of unlabeled data m.

Thus, Theorem 1 shows that using self-supervised learning, we can replace all the complexities of hypothesis spaces in terms of n with those in terms of m. Since the number of unlabeled data points is typically much larger than the number of labeled data points, this illuminates the benefit of self-supervised learning.

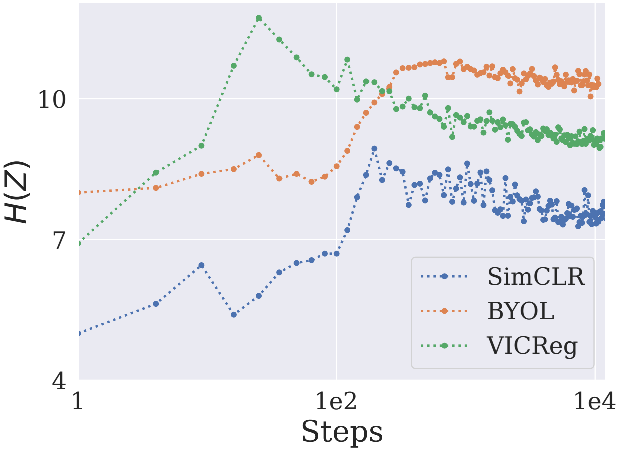

Figure 2: **The entropy for the SSL models VICReg decreased during the training.** The entropy (measured by the LogDet Entropy estimator) as a function of the number of steps during training for VICReg and SimCLR and BYOL. Additionally, SimCLR entropy estimation is tighter compared to the others.

5.2. Understanding Theorem 2 via Mutual Information Maximization

Theorem 1 together with the result of the previous section shows that, for generalization in the downstream task, it is helpful to maximize the mutual information $I(Z;X')$ in SSL via minimizing the invariance loss $I_{\mathbf{S}}(f_\theta)$ while controlling the covariance $Z_{\mathbf{S}} Z_{\mathbf{S}}{}^\intercal$. The term $2\tilde{\mathcal{R}}_m(\mathcal{F})/\sqrt{m}$ captures the importance of controlling the complexity of the representations f_θ. To understand this term further in terms of mutual information, let us consider a discretization of the parameter space of $\mathcal{F}$ to have finite $|\mathcal{F}| < \infty$ (indeed, a computer always implements some discretization of continuous variables). Then, by Massart's Finite Class Lemma, we have that $\tilde{\mathcal{R}}_m(\mathcal{F}) \leq C\sqrt{\ln|\mathcal{F}|}$ for some constant $C > 0$. Moreover, Shwartz-Ziv (2022) shows that we can approximate $\ln|\mathcal{F}|$ by $2^{I(Z;X)}$. Thus, in Theorem 1, the term $I_{\mathbf{S}}(f_\theta) + \frac{2}{\sqrt{m}}\|\mathbf{P}_{Z_{\mathbf{S}}} Y_{\mathbf{S}}\|_F + \frac{1}{\sqrt{n}}\|\mathbf{P}_{Z_S} Y_S\|_F$ corresponds to $I(Z;X')$ while the term of $2\tilde{\mathcal{R}}_m(\mathcal{F})/\sqrt{m}$ corresponds to $I(Z;X)$. Recall that the information can be decomposed as

$$I(Z;X) = I(Z;X') + I(Z;X|X'). \tag{5.9}$$

where we want to maximize the predictive information $I(Z;X')$, while minimizing $I(Z;X)$ (Federici et al., 2019; Shwartz-Ziv and Tishby, 2017a). Thus, to improve generalization, we also need to control $2\tilde{\mathcal{R}}_m(\mathcal{F})/\sqrt{m}$ to restrict the superfluous information $I(Z;X|X')$, in addition to minimizing $I_{\mathbf{S}}(f_\theta) + \frac{2}{\sqrt{m}}\|\mathbf{P}_{Z_{\mathbf{S}}} Y_{\mathbf{S}}\|_F + \frac{1}{\sqrt{n}}\|\mathbf{P}_{Z_S} Y_S\|_F$ that corresponded to maximize the predictive information $I(Z;X')$. Although we can explicitly add regularization on $I(Z;X|X')$ to control $2\tilde{\mathcal{R}}_m(\mathcal{F})/\sqrt{m}$, it is possible that $I(Z;X|X')$ and $2\tilde{\mathcal{R}}_m(\mathcal{F})/\sqrt{m}$ are implicitly regularized via implicit bias through e design choises (Gunasekar et al., 2017; Soudry et al., 2018; Gunasekar et al., 2018). Thus, Theorem 1 connects the information-theoretic understanding of VICReg with the probabilistic guarantee on downstream generalization.

6. Lower bounds on $\mathbb{E}_{x'}\left[\log q(z|x')\right]$

In this section of the supplementary material, we present the full derivation of the lower bound on $\mathbb{E}_{x'}\left[\log q(z|x')\right]$. Because $Z'|X'$ is a Gaussian, we can write it as $Z' = \mu(x') + L(x')\epsilon$ where $\epsilon \sim \mathcal{N}(0,1)$ and $L(x')^T L(x') = \Sigma(x')$. Now, setting $\Sigma_r = I$, will give us:

$$\begin{aligned}&\mathbb{E}_{x'}\left[\log q(z|x')\right]\\ \geq&\mathbb{E}_{z'|x'}\left[\log q(z|z')\right]\end{aligned} \tag{6.10}$$

$$=\mathbb{E}_{z'|x'}\left[\frac{d}{2}\log 2\pi - \frac{1}{2}\left(z-z'\right)^T (I))^{-1}\left(z-z'\right)\right] \tag{6.11}$$

$$=\frac{d}{2}\log 2\pi - \frac{1}{2}\mathbb{E}_{z'|x',}\left[\left(z-z'\right)^2\right] \tag{6.12}$$

$$=\frac{d}{2}\log 2\pi - \frac{1}{2}\mathbb{E}_{\epsilon}\left[\left(z-\mu(x')-L(x')\epsilon\right)^2\right] \tag{6.13}$$

$$=\frac{d}{2}\log 2\pi - \frac{1}{2}\mathbb{E}_{\epsilon}\left[\left(z-\mu(x')\right)^2 - 2\left(z-\mu(x') * L(x')\epsilon\right) + \left(\left(L(x')\epsilon\right)^T\left(L(x')\epsilon\right)\right)\right] \tag{6.14}$$

$$=\frac{d}{2}\log 2\pi - \frac{1}{2}\mathbb{E}_{\epsilon}\left[\left(z-\mu(x')\right)^2\right] + \left(z-\mu(x')L(x')\right)\mathbb{E}_{\epsilon}\left[\epsilon\right] - \frac{1}{2}\mathbb{E}_{\epsilon}\left[\epsilon^T L(x')^T L(x')\epsilon\right] \tag{6.15}$$

$$=\frac{d}{2}\log 2\pi - \frac{1}{2}\left(z-\mu(x')\right)^2 - \frac{1}{2}Tr\log\Sigma(x') \tag{6.16}$$

where $\mathbb{E}_{x'}\left[\log q(z|x')\right] = \mathbb{E}_{x'}\left[\log \mathbb{E}_{z'|x'}\left[q(z|z')\right]\right] \geq \mathbb{E}_{z'}\left[\log q(z|z')\right]$ by Jensen's inequality, $\mathbb{E}_{\epsilon}[\epsilon] = 0$ and $\mathbb{E}_{\epsilon}\left[\epsilon\left(L(x')^T L(x')\,\epsilon\right] = Tr\log\Sigma(x')$ by the Hutchinson's estimator.

$$\mathbb{E}_{z|x}\left[\mathbb{E}_{z'|x'}\left[\log q(z|z')\right]\right] =\mathbb{E}_{z|x}\left[\frac{d}{2}\log 2\pi - \frac{1}{2}\left(z-\mu(x')\right)^2 - \frac{1}{2}Tr\log\Sigma(x')\right] \tag{6.17}$$

$$=\frac{d}{2}\log 2\pi - \frac{1}{2}\mathbb{E}_{z|x}\left[\left(z-\mu(x')\right)^2\right] - \frac{1}{2}Tr\log\Sigma(x') \tag{6.18}$$

$$=\frac{d}{2}\log 2\pi - \frac{1}{2}\mathbb{E}_{\epsilon}\left[\left(\mu(x)+L(x)\epsilon-\mu(x')\right)^2\right] - \frac{1}{2}Tr\log\Sigma(x') \tag{6.19}$$

$$\begin{aligned}=&\frac{d}{2}\log 2\pi - \frac{1}{2}\mathbb{E}_{\epsilon}\left[\left(\mu(x)-\mu(x')\right)^2\right] + \mathbb{E}_{\epsilon}\left[\left(\mu(x)-\mu(x')\right)L(x)\epsilon\right]\\ &- \frac{1}{2}\mathbb{E}_{\epsilon}\left[\epsilon^T L(x)^T L(x)\epsilon\right] - \frac{1}{2}Tr\log\Sigma(x')\end{aligned} \tag{6.20}$$

$$=\frac{d}{2}\log 2\pi - \frac{1}{2}\left(\mu(x)-\mu(x')\right)^2 - \frac{1}{2}Tr\log\Sigma(x) - \frac{1}{2}Tr\log\Sigma(x') \tag{6.21}$$

$$=\frac{d}{2}\log 2\pi - \frac{1}{2}\left(\mu(x)-\mu(x')\right)^2 - \frac{1}{2}\log\left(|\Sigma(x)|\cdot|\Sigma(x')|\right) \tag{6.22}$$

7. Data Distribution after Deep Network Transformation

Theorem 3 *Given the setting of eq. (2.7) the unconditional DNN output density denoted as Z approximates (given the truncation of the Gaussian on its effective support that is* *included within a single region ω of the DN's input space partition) a mixture of the affinely transformed distributions $\boldsymbol{x}|\boldsymbol{x}^*_{n(\boldsymbol{x})}$ e.g. for the Gaussian case*

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

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.

Proof We know 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. Therefore, any sample from p will almost surely lie within a single region $\omega \in \Omega$ and therefore the entire mapping can be considered linear with respect to p. Thus, the output distribution is a linear transformation of the input distribution based on the per-region affine mapping. ■

8. Generalization Bound

The following theorem is the complete version of Theorem 1:

Theorem 4 *For any $\delta > 0$, with probability at least $1 - \delta$, the following holds:*

$$\mathbb{E}_{x,y}[\ell_{x,y}(w_S)] \leq cI_{\mathbf{S}}(f_\theta) + \frac{2}{\sqrt{m}}\|\mathbf{P}_{Z_{\mathbf{S}}}Y_{\mathbf{S}}\|_F + \frac{1}{\sqrt{n}}\|\mathbf{P}_{Z_S}Y_S\|_F + Q_{m,n}, \tag{8.23}$$

where

$$\begin{aligned} Q_{m,n} = c&\left(\frac{2\tilde{\mathcal{R}}_m(\mathcal{F})}{\sqrt{m}} + \tau\sqrt{\frac{\ln(3/\delta)}{2m}} + \tau_{\mathbf{S}}\sqrt{\frac{\ln(3/\delta)}{2n}}\right) \\ &+ \kappa_S\sqrt{\frac{2\ln(6|\mathcal{Y}|/\delta)}{2n}}\sum_{y\in\mathcal{Y}}\left(\sqrt{\hat{p}(y)} + \sqrt{p(y)}\right) \\ &+ \frac{4\mathcal{R}_m(\mathcal{W}\circ\mathcal{F})}{\sqrt{m}} + 2\kappa\sqrt{\frac{\ln(4/\delta)}{2m}} + 2\kappa_{\mathbf{S}}\sqrt{\frac{\ln(4/\delta)}{2n}}. \end{aligned}$$

Proof The complete proof is presented in Appendix 8.1. ■

The bound in the complete version of Theorem 4 is better than the one in the informal version of Theorem 1, because of the factor c. The factor c measures the difference between the minimum norm solution W_S of the labeled training data and the minimum norm solution $W_{\mathbf{S}}$ of the unlabeled training data. Thus, the factor c also decreases towards zero as n and m increase. Moreover, if the labeled and unlabeled training data are similar, the value of c is small, decreasing the generalization bound further, which makes sense. Thus, we can view the factor c as a measure on the distance between the labeled training data and the unlabeled training data.

We obtain the informal version from the complete version of Theorem 1 by the following reasoning to simplify the notation in the main text. We have that $cI_{\mathbf{S}}(f_\theta) + c\frac{2\tilde{\mathcal{R}}_m(\mathcal{F})}{\sqrt{m}} =$

$I_{\mathbf{S}}(f_\theta) + \frac{2\tilde{\mathcal{R}}_m(\mathcal{F})}{\sqrt{m}} + Q$, where $Q = (c-1)(I_{\mathbf{S}}(f_\theta) + \frac{2\tilde{\mathcal{R}}_m(\mathcal{F})}{\sqrt{m}}) \leq \varsigma \to 0$ as as $m, n \to \infty$, since $c \to 0$ as $m, n \to \infty$. However, this reasoning is used only to simplify the notation in the main text. The bound in the complete version of Theorem 1 is more accurate and indeed tighter than the one in the informal version.

In Theorem 1, $Q_{m,n} \to 0$ as $m, n \to \infty$ if $\frac{\tilde{\mathcal{R}}_m(\mathcal{F})}{\sqrt{m}} \to 0$ as $m \to \infty$. Indeed, this typically holds because $\tilde{\mathcal{R}}_m(\mathcal{F}) = O(1)$ as $m \to \infty$ for typical choices of $\mathcal{F}$, including deep neural networks (Bartlett et al., 2017; Kawaguchi et al., 2018; Golowich et al., 2018) as well as other common machine learning models (Bartlett and Mendelson, 2002; Mohri et al., 2012; Shalev-Shwartz and Ben-David, 2014; Shwartz-Ziv et al., 2023).

8.1. Proof of Theorem 1

Proof [Proof of Theorem 1] Let $W = W_S$ where W_S is the the minimum norm solution as $W_S = \text{minimize}_{W'} \|W'\|_F$ s.t. $W' \in \arg\min_W \frac{1}{n}\sum_{i=1}^n \|Wf_\theta(x_i) - y_i\|^2$. Let $W^* = W_{\mathbf{S}}$ where $W_{\mathbf{S}}$ is the minimum norm solution as $W^* = W_{\mathbf{S}} = \text{minimize}_{W'} \|W'\|_F$ s.t. $W' \in \arg\min_W \frac{1}{m}\sum_{i=1}^m \|Wf_\theta(x_i^+) - g^*(x_i^+)\|^2$. Since $y = g^*(x)$,

$$y = g^*(x) \pm W^*f_\theta(x) = W^*f_\theta(x) + (g^*(x) - W^*f_\theta(x)) = W^*f_\theta(x) + \varphi(x)$$

where $\varphi(x) = g^*(x) - W^*f_\theta(x)$. Define $L_S(w) = \frac{1}{n}\sum_{i=1}^n \|Wf_\theta(x_i) - y_i\|$. Using these,

$$\begin{aligned}
L_S(w) &= \frac{1}{n}\sum_{i=1}^n \|Wf_\theta(x_i) - y_i\| \\
&= \frac{1}{n}\sum_{i=1}^n \|Wf_\theta(x_i) - W^*f_\theta(x_i) - \varphi(x_i)\| \\
&\geq \frac{1}{n}\sum_{i=1}^n \|Wf_\theta(x_i) - W^*f_\theta(x_i)\| - \frac{1}{n}\sum_{i=1}^n \|\varphi(x_i)\| \\
&= \frac{1}{n}\sum_{i=1}^n \|\tilde{W}f_\theta(x_i)\| - \frac{1}{n}\sum_{i=1}^n \|\varphi(x_i)\|
\end{aligned}$$

where $\tilde{W} = W - W^*$. We now consider new fresh samples $\bar{x}_i \sim \mathcal{D}_{y_i}$ for $i = 1, \ldots, n$ to rewrite the above further as:

$$\begin{aligned}
L_S(w) &\geq \frac{1}{n}\sum_{i=1}^n \|\tilde{W}f_\theta(x_i) \pm \tilde{W}f_\theta(\bar{x}_i)\| - \frac{1}{n}\sum_{i=1}^n \|\varphi(x_i)\| \\
&= \frac{1}{n}\sum_{i=1}^n \|\tilde{W}f_\theta(\bar{x}_i) - (\tilde{W}f_\theta(\bar{x}_i) - \tilde{W}f_\theta(x_i))\| - \frac{1}{n}\sum_{i=1}^n \|\varphi(x_i)\| \\
&\geq \frac{1}{n}\sum_{i=1}^n \|\tilde{W}f_\theta(\bar{x}_i)\| - \frac{1}{n}\sum_{i=1}^n \|\tilde{W}f_\theta(\bar{x}_i) - \tilde{W}f_\theta(x_i)\| - \frac{1}{n}\sum_{i=1}^n \|\varphi(x_i)\| \\
&= \frac{1}{n}\sum_{i=1}^n \|\tilde{W}f_\theta(\bar{x}_i)\| - \frac{1}{n}\sum_{i=1}^n \|\tilde{W}(f_\theta(\bar{x}_i) - f_\theta(x_i))\| - \frac{1}{n}\sum_{i=1}^n \|\varphi(x_i)\|
\end{aligned}$$

This implies that

$$\frac{1}{n}\sum_{i=1}^{n}\|\tilde{W}f_\theta(\bar{x}_i)\| \leq L_S(w) + \frac{1}{n}\sum_{i=1}^{n}\|\tilde{W}(f_\theta(\bar{x}_i) - f_\theta(x_i))\| + \frac{1}{n}\sum_{i=1}^{n}\|\varphi(x_i)\|.$$

Furthermore, since $y = W^* f_\theta(x) + \varphi(x)$, by writing $\bar{y}_i = W^* f_\theta(\bar{x}_i) + \varphi(\bar{x}_i)$ (where $\bar{y}_i = y_i$ since $\bar{x}_i \sim \mathcal{D}_{y_i}$ for $i = 1, \dots, n$),

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\|\tilde{W}f_\theta(\bar{x}_i)\| &= \frac{1}{n}\sum_{i=1}^{n}\|Wf_\theta(\bar{x}_i) - W^* f_\theta(\bar{x}_i)\| \\ &= \frac{1}{n}\sum_{i=1}^{n}\|Wf_\theta(\bar{x}_i) - \bar{y}_i + \varphi(\bar{x}_i)\| \\ &\geq \frac{1}{n}\sum_{i=1}^{n}\|Wf_\theta(\bar{x}_i) - \bar{y}_i\| - \frac{1}{n}\sum_{i=1}^{n}\|\varphi(\bar{x}_i)\|\end{aligned}$$

Combining these, we have that

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\|Wf_\theta(\bar{x}_i) - \bar{y}_i\| \leq L_S(w) &+ \frac{1}{n}\sum_{i=1}^{n}\|\tilde{W}(f_\theta(\bar{x}_i) - f_\theta(x_i))\| \\ &+ \frac{1}{n}\sum_{i=1}^{n}\|\varphi(x_i)\| + \frac{1}{n}\sum_{i=1}^{n}\|\varphi(\bar{x}_i)\|.\end{aligned} \tag{8.24}$$

To bound the left-hand side of (8.24), we now analyze the following random variable:

$$\mathbb{E}_{X,Y}[\|W_S f_\theta(X) - Y\|] - \frac{1}{n}\sum_{i=1}^{n}\|W_S f_\theta(\bar{x}_i) - \bar{y}_i\|, \tag{8.25}$$

where $\bar{y}_i = y_i$ since $\bar{x}_i \sim \mathcal{D}_{y_i}$ for $i = 1, \dots, n$. Importantly, this means that as W_S depends on y_i, W_S depends on $\bar{y}_i$. Thus, the collection of random variables $\|W_S f_\theta(\bar{x}_1) - \bar{y}_1\|, \dots, \|W_S f_\theta(n_n) - \bar{y}_n\|$ is *not* independent. Accordingly, we cannot apply standard concentration inequality to bound (8.25). A standard approach in learning theory is to first bound (8.25) by $\mathbb{E}_{x,y}\|W_S f_\theta(x) - y\| - \frac{1}{n}\sum_{i=1}^{n}\|W_S f_\theta(\bar{x}_i) - \bar{y}_i\| \leq \sup_{W\in\mathcal{W}}\mathbb{E}_{x,y}\|Wf_\theta(x) - y\| - \frac{1}{n}\sum_{i=1}^{n}\|Wf_\theta(\bar{x}_i) - \bar{y}_i\|$ for some hypothesis space $\mathcal{W}$ (that is independent of S) and realize that the right-hand side now contains the collection of independent random variables $\|Wf_\theta(\bar{x}_1) - \bar{y}_1\|, \dots, \|Wf_\theta(n_n) - \bar{y}_n\|$, for which we can utilize standard concentration inequalities. This reasoning leads to the Rademacher complexity of the hypothesis space $\mathcal{W}$. However, the complexity of the hypothesis space $\mathcal{W}$ can be very large, resulting into a loose bound. In this proof, we show that we can avoid the dependency on hypothesis space $\mathcal{W}$ by using a very different approach with conditional expectations to take care the dependent random variables $\|W_S f_\theta(\bar{x}_1) - \bar{y}_1\|, \dots, \|W_S f_\theta(n_n) - \bar{y}_n\|$. Intuitively, we utilize the fact that for these dependent random variables, there are a structure of conditional independence, conditioned on each $y \in \mathcal{Y}$.

We first write the expected loss as the sum of the conditional expected loss:

$$\begin{aligned}\mathbb{E}_{X,Y}[\|W_S f_\theta(X) - Y\|] &= \sum_{y\in\mathcal{Y}} \mathbb{E}_{X,Y}[\|W_S f_\theta(X) - Y\| \mid Y = y]\mathbb{P}(Y = y) \\ &= \sum_{y\in\mathcal{Y}} \mathbb{E}_{X_y}[\|W_S f_\theta(X_y) - y\|]\mathbb{P}(Y = y),\end{aligned}$$

where X_y is the random variable for the conditional with $Y = y$. Using this, we decompose (8.25) into two terms:

$$\begin{aligned}&\mathbb{E}_{X,Y}[\|W_S f_\theta(X) - Y\|] - \frac{1}{n}\sum_{i=1}^{n} \|W_S f_\theta(\bar{x}_i) - \bar{y}_i\| \qquad (8.26)\\ &= \left(\sum_{y\in\mathcal{Y}} \mathbb{E}_{X_y}[\|W_S f_\theta(X_y) - y\|]\frac{|\mathcal{I}_y|}{n} - \frac{1}{n}\sum_{i=1}^{n} \|W_S f_\theta(\bar{x}_i) - \bar{y}_i\|\right) \\ &\quad + \sum_{y\in\mathcal{Y}} \mathbb{E}_{X_y}[\|W_S f_\theta(X_y) - y\|]\left(\mathbb{P}(Y = y) - \frac{|\mathcal{I}_y|}{n}\right),\end{aligned}$$

where

$$\mathcal{I}_y = \{i \in [n] : y_i = y\}.$$

The first term in the right-hand side of (8.26) is further simplified by using

$$\frac{1}{n}\sum_{i=1}^{n} \|W_S f_\theta(\bar{x}_i) - \bar{y}_i\| = \frac{1}{n}\sum_{y\in\mathcal{Y}}\sum_{i\in\mathcal{I}_y} \|W_S f_\theta(\bar{x}_i) - y\|,$$

as

$$\begin{aligned}&\sum_{y\in\mathcal{Y}} \mathbb{E}_{X_y}[\|W_S f_\theta(X_y) - y\|]\frac{|\mathcal{I}_y|}{n} - \frac{1}{n}\sum_{i=1}^{n} \|W_S f_\theta(\bar{x}_i) - \bar{y}_i\| \\ &= \frac{1}{n}\sum_{y\in\tilde{\mathcal{Y}}} |\mathcal{I}_y| \left(\mathbb{E}_{X_y}[\|W_S f_\theta(X_y) - y\|] - \frac{1}{|\mathcal{I}_y|}\sum_{i\in\mathcal{I}_y} \|W_S f_\theta(\bar{x}_i) - y\|\right),\end{aligned}$$

where $\tilde{\mathcal{Y}} = \{y \in \mathcal{Y} : |\mathcal{I}_y| \neq 0\}$. Substituting these into equation (8.26) yields

$$\begin{aligned}&\mathbb{E}_{X,Y}[\|W_S f_\theta(X) - Y\|] - \frac{1}{n}\sum_{i=1}^{n} \|W_S f_\theta(\bar{x}_i) - \bar{y}_i\| \qquad (8.27)\\ &= \frac{1}{n}\sum_{y\in\tilde{\mathcal{Y}}} |\mathcal{I}_y| \left(\mathbb{E}_{X_y}[\|W_S f_\theta(X_y) - y\|] - \frac{1}{|\mathcal{I}_y|}\sum_{i\in\mathcal{I}_y} \|W_S f_\theta(\bar{x}_i) - y\|\right) \\ &\quad + \sum_{y\in\mathcal{Y}} \mathbb{E}_{X_y}[\|W_S f_\theta(X_y) - y\|]\left(\mathbb{P}(Y = y) - \frac{|\mathcal{I}_y|}{n}\right)\end{aligned}$$

Importantly, while $\|W_S f_\theta(\bar{x}_1)-\bar{y}_1\|, \dots, \|W_S f_\theta(\bar{x}_n)-\bar{y}_n\|$ on the right-hand side of (8.27) are dependent random variables, $\|W_S f_\theta(\bar{x}_1) - y\|, \dots, \|W_S f_\theta(\bar{x}_n) - y\|$ are independent random variables since W_S and $\bar{x}_i$ are independent and y is fixed here. Thus, by using Hoeffding's inequality (Lemma 1), and taking union bounds over $y \in \tilde{\mathcal{Y}}$, we have that with probability at least $1 - \delta$, the following holds for all $y \in \tilde{\mathcal{Y}}$:

$$\mathbb{E}_{X_y}[\|W_S f_\theta(X_y) - y\|] - \frac{1}{|\mathcal{I}_y|} \sum_{i \in \mathcal{I}_y} \|W_S f_\theta(\bar{x}_i) - y\| \le \kappa_S \sqrt{\frac{\ln(|\tilde{\mathcal{Y}}|/\delta)}{2|\mathcal{I}_y|}}.$$

This implies that with probability at least $1 - \delta$,

$$\begin{aligned}
&\frac{1}{n} \sum_{y \in \tilde{\mathcal{Y}}} |\mathcal{I}_y| \left(\mathbb{E}_{X_y}[\|W_S f_\theta(X_y) - y\|] - \frac{1}{|\mathcal{I}_y|} \sum_{i \in \mathcal{I}_y} \|W_S f_\theta(\bar{x}_i) - y\| \right) \\
&\le \frac{\kappa_S}{n} \sum_{y \in \tilde{\mathcal{Y}}} |\mathcal{I}_y| \sqrt{\frac{\ln(|\tilde{\mathcal{Y}}|/\delta)}{2|\mathcal{I}_y|}} \\
&= \kappa_S \left(\sum_{y \in \tilde{\mathcal{Y}}} \sqrt{\frac{|\mathcal{I}_y|}{n}} \right) \sqrt{\frac{\ln(|\tilde{\mathcal{Y}}|/\delta)}{2n}}.
\end{aligned}$$

Substituting this bound into (8.27), we have that with probability at least $1 - \delta$,

$$\begin{aligned}
&\mathbb{E}_{X,Y}[\|W_S f_\theta(X) - Y\|] - \frac{1}{n} \sum_{i=1}^{n} \|W_S f_\theta(\bar{x}_i) - \bar{y}_i\| \qquad (8.28) \\
&\le \kappa_S \left(\sum_{y \in \tilde{\mathcal{Y}}} \sqrt{\hat{p}(y)} \right) \sqrt{\frac{\ln(|\tilde{\mathcal{Y}}|/\delta)}{2n}} + \sum_{y \in \mathcal{Y}} \mathbb{E}_{X_y}[\|W_S f_\theta(X_y) - y\|] \left(\mathbb{P}(Y = y) - \frac{|\mathcal{I}_y|}{n} \right)
\end{aligned}$$

where

$$\hat{p}(y) = \frac{|\mathcal{I}_y|}{n}.$$

Moreover, for the second term on the right-hand side of (8.28), by using Lemma 1 of (Kawaguchi et al., 2022), we have that with probability at least $1 - \delta$,

$$\begin{aligned}
&\sum_{y \in \mathcal{Y}} \mathbb{E}_{X_y}[\|W_S f_\theta(X_y) - y\|] \left(\mathbb{P}(Y = y) - \frac{|\mathcal{I}_y|}{n} \right) \\
&\le \left(\sum_{y \in \mathcal{Y}} \sqrt{p(y)} \mathbb{E}_{X_y}[\|W_S f_\theta(X_y) - y\| \right) \sqrt{\frac{2 \ln(|\mathcal{Y}|/\delta)}{2n}} \\
&\le \kappa_S \left(\sum_{y \in \mathcal{Y}} \sqrt{p(y)} \right) \sqrt{\frac{2 \ln(|\mathcal{Y}|/\delta)}{2n}}
\end{aligned}$$

where $p(y) = \mathbb{P}(Y = y)$. Substituting this bound into (8.28) with the union bound, we have that with probability at least $1 - \delta$,

$$\begin{aligned}
&\mathbb{E}_{X,Y}[\|W_S f_\theta(X) - Y\|] - \frac{1}{n}\sum_{i=1}^{n} \|W_S f_\theta(\bar{x}_i) - \bar{y}_i\| \qquad (8.29)\\
&\le \kappa_S \left(\sum_{y \in \tilde{\mathcal{Y}}} \sqrt{\hat{p}(y)}\right) \sqrt{\frac{\ln(2|\tilde{\mathcal{Y}}|/\delta)}{2n}} + \kappa_S \left(\sum_{y \in \mathcal{Y}} \sqrt{p(y)}\right) \sqrt{\frac{2\ln(2|\mathcal{Y}|/\delta)}{2n}}\\
&\le \left(\sum_{y \in \mathcal{Y}} \sqrt{\hat{p}(y)}\right) \kappa_S \sqrt{\frac{2\ln(2|\mathcal{Y}|/\delta)}{2n}} + \left(\sum_{y \in \mathcal{Y}} \sqrt{p(y)}\right) \kappa_S \sqrt{\frac{2\ln(2|\mathcal{Y}|/\delta)}{2n}}\\
&\le \kappa_S \sqrt{\frac{2\ln(2|\mathcal{Y}|/\delta)}{2n}} \sum_{y \in \mathcal{Y}} \left(\sqrt{\hat{p}(y)} + \sqrt{p(y)}\right)
\end{aligned}$$

Combining (8.24) and (8.29) implies that with probability at least $1 - \delta$,

$$\begin{aligned}
&\mathbb{E}_{X,Y}[\|W_S f_\theta(X) - Y\|] \qquad (8.30)\\
&\le \frac{1}{n}\sum_{i=1}^{n} \|W_S f_\theta(\bar{x}_i) - \bar{y}_i\| + \kappa_S \sqrt{\frac{2\ln(2|\mathcal{Y}|/\delta)}{2n}} \sum_{y \in \mathcal{Y}} \left(\sqrt{\hat{p}(y)} + \sqrt{p(y)}\right)\\
&\le L_S(w_S) + \frac{1}{n}\sum_{i=1}^{n} \|\tilde{W}(f_\theta(\bar{x}_i) - f_\theta(x_i))\|\\
&\quad + \frac{1}{n}\sum_{i=1}^{n} \|\varphi(x_i)\| + \frac{1}{n}\sum_{i=1}^{n} \|\varphi(\bar{x}_i)\| + \kappa_S \sqrt{\frac{2\ln(2|\mathcal{Y}|/\delta)}{2n}} \sum_{y \in \mathcal{Y}} \left(\sqrt{\hat{p}(y)} + \sqrt{p(y)}\right).
\end{aligned}$$

We will now analyze the term $\frac{1}{n}\sum_{i=1}^{n} \|\varphi(x_i)\| + \frac{1}{n}\sum_{i=1}^{n} \|\varphi(\bar{x}_i)\|$ on the right-hand side of (8.30). Since $W^* = W_{\mathbf{S}}$,

$$\frac{1}{n}\sum_{i=1}^{n} \|\varphi(x_i)\| = \frac{1}{n}\sum_{i=1}^{n} \|g^*(x_i) - W_{\mathbf{S}} f_\theta(x_i)\|.$$

By using Hoeffding's inequality (Lemma 1), we have that for any $\delta > 0$, with probability at least $1 - \delta$,

$$\frac{1}{n}\sum_{i=1}^{n} \|\varphi(x_i)\| \le \frac{1}{n}\sum_{i=1}^{n} \|g^*(x_i) - W_{\mathbf{S}} f_\theta(x_i)\| \le \mathbb{E}_{x^+}[\|g^*(x^+) - W_{\mathbf{S}} f_\theta(x^+)\|] + \kappa_{\mathbf{S}} \sqrt{\frac{\ln(1/\delta)}{2n}}.$$

Moreover, by using (Mohri et al., 2012, Theorem 3.1) with the loss function $x^+ \mapsto \|g^*(x^+) - Wf(x^+)\|$ (i.e., Lemma 2), we have that for any $\delta > 0$, with probability at least $1 - \delta$,

$$\mathbb{E}_{x^+}[\|g^*(x^+) - W_{\mathbf{S}} f_\theta(x^+)\|] \le \frac{1}{m}\sum_{i=1}^{m} \|g^*(x_i^+) - W_{\mathbf{S}} f_\theta(x_i^+)\| + \frac{2\tilde{\mathcal{R}}_m(\mathcal{W} \circ \mathcal{F})}{\sqrt{m}} + \kappa \sqrt{\frac{\ln(1/\delta)}{2m}} \qquad (8.31)$$

where $\tilde{\mathcal{R}}_m(\mathcal{W} \circ \mathcal{F}) = \frac{1}{\sqrt{m}} \mathbb{E}_{\mathbf{S},\xi}[\sup_{W \in \mathcal{W}, f \in \mathcal{F}} \sum_{i=1}^m \xi_i \|g^*(x_i^+) - Wf(x_i^+)\|]$ is the normalized Rademacher complexity of the set $\{x^+ \mapsto \|g^*(x^+) - Wf(x^+)\| : W \in \mathcal{W}, f \in \mathcal{F}\}$ (it is normalized such that $\tilde{\mathcal{R}}_m(\mathcal{F}) = O(1)$ as $m \to \infty$ for typical choices of $\mathcal{F}$), and $\xi_1, \dots, \xi_m$ are independent uniform random variables taking values in $\{-1, 1\}$. Takinng union bounds, we have that for any $\delta > 0$, with probability at least $1 - \delta$,

$$\frac{1}{n} \sum_{i=1}^n \|\varphi(x_i)\| \leq \frac{1}{m} \sum_{i=1}^m \|g^*(x_i^+) - W_{\mathbf{S}} f_\theta(x_i^+)\| + \frac{2\tilde{\mathcal{R}}_m(\mathcal{W} \circ \mathcal{F})}{\sqrt{m}} + \kappa \sqrt{\frac{\ln(2/\delta)}{2m}} + \kappa_{\mathbf{S}} \sqrt{\frac{\ln(2/\delta)}{2n}}$$

Similarly, for any $\delta > 0$, with probability at least $1 - \delta$,

$$\frac{1}{n} \sum_{i=1}^n \|\varphi(\bar{x}_i)\| \leq \frac{1}{m} \sum_{i=1}^m \|g^*(x_i^+) - W_{\mathbf{S}} f_\theta(x_i^+)\| + \frac{2\tilde{\mathcal{R}}_m(\mathcal{W} \circ \mathcal{F})}{\sqrt{m}} + \kappa \sqrt{\frac{\ln(2/\delta)}{2m}} + \kappa_{\mathbf{S}} \sqrt{\frac{\ln(2/\delta)}{2n}}.$$

Thus, by taking union bounds, we have that for any $\delta > 0$, with probability at least $1 - \delta$,

$$\begin{aligned} &\frac{1}{n} \sum_{i=1}^n \|\varphi(x_i)\| + \frac{1}{n} \sum_{i=1}^n \|\varphi(\bar{x}_i)\| \\ &\leq \frac{2}{m} \sum_{i=1}^m \|g^*(x_i^+) - W_{\mathbf{S}} f_\theta(x_i^+)\| + \frac{4\mathcal{R}_m(\mathcal{W} \circ \mathcal{F})}{\sqrt{m}} + 2\kappa \sqrt{\frac{\ln(4/\delta)}{2m}} + 2\kappa_{\mathbf{S}} \sqrt{\frac{\ln(4/\delta)}{2n}} \end{aligned} \tag{8.32}$$

To analyze the first term on the right-hand side of (8.32), recall that

$$W_{\mathbf{S}} = \underset{W'}{\text{minimize}} \|W'\|_F \text{ s.t. } W' \in \arg\min_W \frac{1}{m} \sum_{i=1}^m \|W f_\theta(x_i^+) - g^*(x_i^+)\|^2. \tag{8.33}$$

Here, since $W f_\theta(x_i^+) \in \mathbb{R}^r$, we have that

$$W f_\theta(x_i^+) = \text{vec}[W f_\theta(x_i^+)] = [f_\theta(x_i^+)^\intercal \otimes I_r] \text{vec}[W] \in \mathbb{R}^r,$$

where $I_r \in \mathbb{R}^{r \times r}$ is the identity matrix, and $[f_\theta(x_i^+)^\intercal \otimes I_r] \in \mathbb{R}^{r \times dr}$ is the Kronecker product of the two matrices, and $\text{vec}[W] \in \mathbb{R}^{dr}$ is the vectorization of the matrix $W \in \mathbb{R}^{r \times d}$. Thus, by defining $A_i = [f_\theta(x_i^+)^\intercal \otimes I_r] \in \mathbb{R}^{r \times dr}$ and using the notation of $w = \text{vec}[W]$ and its inverse $W = \text{vec}^{-1}[w]$ (i.e., the inverse of the vectorization from $\mathbb{R}^{r \times d}$ to $\mathbb{R}^{dr}$ with a fixed ordering), we can rewrite (8.33) by

$$W_{\mathbf{S}} = \text{vec}^{-1}[w_{\mathbf{S}}] \quad \text{where} \quad w_{\mathbf{S}} = \underset{w'}{\text{minimize}} \|w'\|_F \text{ s.t. } w' \in \arg\min_w \sum_{i=1}^m \|g_i - A_i w\|^2,$$

with $g_i = g^*(x_i^+) \in \mathbb{R}^r$. Since the function $w \mapsto \sum_{i=1}^m \|g_i - A_i w\|^2$ is convex, a necessary and sufficient condition of the minimizer of this function is obtained by

$$0 = \nabla_w \sum_{i=1}^m \|g_i - A_i w\|^2 = 2 \sum_{i=1}^m A_i{}^\intercal (g_i - A_i w) \in \mathbb{R}^{dr}$$

This implies that

$$\sum_{i=1}^{m} {A_i}^\intercal A_i w = \sum_{i=1}^{m} {A_i}^\intercal g_i.$$

In other words,

$$A^\intercal A w = A^\intercal g \qquad \text{where } A = \begin{bmatrix} A_1 \\ A_2 \\ \vdots \\ A_m \end{bmatrix} \in \mathbb{R}^{mr \times dr} \text{ and } g = \begin{bmatrix} g_1 \\ g_2 \\ \vdots \\ g_m \end{bmatrix} \in \mathbb{R}^{mr}$$

Thus,

$$w' \in \arg\min_{w} \sum_{i=1}^{m} \|g_i - A_i w\|^2 = \{(A^\intercal A)^\dagger A^\intercal g + v : v \in \mathrm{Null}(A)\}$$

where $(A^\intercal A)^\dagger$ is the Moore–Penrose inverse of the matrix $A^\intercal A$ and $\mathrm{Null}(A)$ is the null space of the matrix A. Thus, the minimum norm solution is obtained by

$$\mathrm{vec}[W_{\mathbf{S}}] = w_{\mathbf{S}} = (A^\intercal A)^\dagger A^\intercal g.$$

Thus, by using this $W_{\mathbf{S}}$, we have that

$$\begin{aligned} \frac{1}{m}\sum_{i=1}^{m} \|g^*(x_i^+) - W_{\mathbf{S}} f_\theta(x_i^+)\| &= \frac{1}{m}\sum_{i=1}^{m} \sqrt{\sum_{k=1}^{r}((g_i - A_i w_{\mathbf{S}})_k)^2} \\ &\le \sqrt{\frac{1}{m}\sum_{i=1}^{m}\sum_{k=1}^{r}((g_i - A_i w_{\mathbf{S}})_k)^2} \\ &= \frac{1}{\sqrt{m}}\sqrt{\sum_{i=1}^{m}\sum_{k=1}^{r}((g_i - A_i w_{\mathbf{S}})_k)^2} \\ &= \frac{1}{\sqrt{m}}\|g - A w_{\mathbf{S}}\|_2 \\ &= \frac{1}{\sqrt{m}}\|g - A(A^\intercal A)^\dagger A^\intercal g\|_2 = \frac{1}{\sqrt{m}}\|(I - A(A^\intercal A)^\dagger A^\intercal)g\|_2 \end{aligned}$$

where the inequality follows from the Jensen's inequality and the concavity of the square root function. Thus, we have that

$$\begin{aligned} &\frac{1}{n}\sum_{i=1}^{n}\|\varphi(x_i)\| + \frac{1}{n}\sum_{i=1}^{n}\|\varphi(\bar{x}_i)\| \qquad (8.34) \\ &\le \frac{2}{\sqrt{m}}\|(I - A(A^\intercal A)^\dagger A^\intercal)g\|_2 + \frac{4\mathcal{R}_m(\mathcal{W}\circ\mathcal{F})}{\sqrt{m}} + 2\kappa\sqrt{\frac{\ln(4/\delta)}{2m}} + 2\kappa_{\mathbf{S}}\sqrt{\frac{\ln(4/\delta)}{2n}} \end{aligned}$$

By combining (8.30) and (8.34) with union bound, we have that

$$\begin{aligned}
&\mathbb{E}_{X,Y}[\|W_S f_\theta(X) - Y\|] \qquad (8.35)\\
&\le L_S(w_S) + \frac{1}{n}\sum_{i=1}^n \|\tilde{W}(f_\theta(\bar{x}_i) - f_\theta(x_i))\| + \frac{2}{\sqrt{m}}\|\mathbf{P}_A g\|_2 \\
&\quad + \frac{4\mathcal{R}_m(\mathcal{W}\circ\mathcal{F})}{\sqrt{m}} + 2\kappa\sqrt{\frac{\ln(8/\delta)}{2m}} + 2\kappa_{\mathbf{S}}\sqrt{\frac{\ln(8/\delta)}{2n}} \\
&\quad + \kappa_S\sqrt{\frac{2\ln(4|\mathcal{Y}|/\delta)}{2n}}\sum_{y\in\mathcal{Y}}\left(\sqrt{\hat{p}(y)} + \sqrt{p(y)}\right).
\end{aligned}$$

where $\tilde{W} = W_S - W^*$ and $\mathbf{P}_A = I - A(A^\intercal A)^\dagger A^\intercal$.

We will now analyze the second term on the right-hand side of (8.35):

$$\frac{1}{n}\sum_{i=1}^n \|\tilde{W}(f_\theta(\bar{x}_i) - f_\theta(x_i))\| \le \|\tilde{W}\|_2 \left(\frac{1}{n}\sum_{i=1}^n \|f_\theta(\bar{x}_i) - f_\theta(x_i)\|\right), \qquad (8.36)$$

where $\|\tilde{W}\|_2$ is the spectral norm of $\tilde{W}$. Since $\bar{x}_i$ shares the same label with x_i as $\bar{x}_i \sim \mathcal{D}_{y_i}$ (and $x_i \sim \mathcal{D}_{y_i}$), and because f_θ is trained with the unlabeled data $\mathbf{S}$, using Hoeffding's inequality (Lemma 1) implies that with probability at least $1-\delta$,

$$\frac{1}{n}\sum_{i=1}^n \|f_\theta(\bar{x}_i) - f_\theta(x_i)\| \le \mathbb{E}_{y\sim\rho}\mathbb{E}_{\bar{x},x\sim\mathcal{D}_y^2}[\|f_\theta(\bar{x}) - f_\theta(x)\|] + \tau_{\mathbf{S}}\sqrt{\frac{\ln(1/\delta)}{2n}}. \qquad (8.37)$$

Moreover, by using (Mohri et al., 2012, Theorem 3.1) with the loss function $(x,\bar{x}) \mapsto \|f_\theta(\bar{x}) - f_\theta(x)\|$ (i.e., Lemma 2), we have that with probability at least $1-\delta$,

$$\mathbb{E}_{y\sim\rho}\mathbb{E}_{\bar{x},x\sim\mathcal{D}_y^2}[\|f_\theta(\bar{x}) - f_\theta(x)\|] \le \frac{1}{m}\sum_{i=1}^m \|f_\theta(x_i^+) - f_\theta(x_i^{++})\| + \frac{2\tilde{\mathcal{R}}_m(\mathcal{F})}{\sqrt{m}} + \tau\sqrt{\frac{\ln(1/\delta)}{2m}} \qquad (8.38)$$

where $\tilde{\mathcal{R}}_m(\mathcal{F}) = \frac{1}{\sqrt{m}}\mathbb{E}_{\mathbf{S},\xi}[\sup_{f\in\mathcal{F}}\sum_{i=1}^m \xi_i\|f(x_i^+) - f(x_i^{++})\|]$ is the normalized Rademacher complexity of the set $\{(x^+, x^{++}) \mapsto \|f(x^+) - f(x^{++})\| : f \in \mathcal{F}\}$ (it is normalized such that $\tilde{\mathcal{R}}_m(\mathcal{F}) = O(1)$ as $m \to \infty$ for typical choices of $\mathcal{F}$), and $\xi_1, \ldots, \xi_m$ are independent uniform random variables taking values in $\{-1, 1\}$. Thus, taking union bound, we have that for any $\delta > 0$, with probability at least $1-\delta$,

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n \|\tilde{W}(f_\theta(\bar{x}_i) - f_\theta(x_i))\| \qquad (8.39)\\
&\le \|\tilde{W}\|_2\left(\frac{1}{m}\sum_{i=1}^m \|f_\theta(x_i^+) - f_\theta(x_i^{++})\| + \frac{2\tilde{\mathcal{R}}_m(\mathcal{F})}{\sqrt{m}} + \tau\sqrt{\frac{\ln(2/\delta)}{2m}} + +\tau_{\mathbf{S}}\sqrt{\frac{\ln(2/\delta)}{2n}}\right).
\end{aligned}$$

By combining (8.35) and (8.39) using the union bound, we have that with probability at least $1-\delta$,

$$\begin{aligned}
&\mathbb{E}_{X,Y}[\|W_S f_\theta(X) - Y\|] \qquad (8.40)\\
&\le L_S(w_S) + \|\tilde{W}\|_2 \left(\frac{1}{m}\sum_{i=1}^{m} \|f_\theta(x_i^+) - f_\theta(x_i^{++})\| + \frac{2\tilde{\mathcal{R}}_m(\mathcal{F})}{\sqrt{m}} + \tau\sqrt{\frac{\ln(4/\delta)}{2m}} + \tau_{\mathbf{S}}\sqrt{\frac{\ln(4/\delta)}{2n}} \right)\\
&\quad + \frac{2}{\sqrt{m}}\|\mathbf{P}_A g\|_2 + \frac{4\mathcal{R}_m(\mathcal{W}\circ\mathcal{F})}{\sqrt{m}} + 2\kappa\sqrt{\frac{\ln(16/\delta)}{2m}} + 2\kappa_{\mathbf{S}}\sqrt{\frac{\ln(16/\delta)}{2n}}\\
&\quad + \kappa_S\sqrt{\frac{2\ln(8|\mathcal{Y}|/\delta)}{2n}} \sum_{y\in\mathcal{Y}} \left(\sqrt{\hat{p}(y)} + \sqrt{p(y)}\right)\\
&= L_S(w_S) + \|\tilde{W}\|_2 \left(\frac{1}{m}\sum_{i=1}^{m} \|f_\theta(x_i^+) - f_\theta(x_i^{++})\| \right) + \frac{2}{\sqrt{m}}\|\mathbf{P}_A g\|_2 + Q_{m,n}
\end{aligned}$$

where

$$\begin{aligned}
Q_{m,n} = \|\tilde{W}\|_2 &\left(\frac{2\tilde{\mathcal{R}}_m(\mathcal{F})}{\sqrt{m}} + \tau\sqrt{\frac{\ln(3/\delta)}{2m}} + \tau_{\mathbf{S}}\sqrt{\frac{\ln(3/\delta)}{2n}} \right)\\
&+ \kappa_S\sqrt{\frac{2\ln(6|\mathcal{Y}|/\delta)}{2n}} \sum_{y\in\mathcal{Y}} \left(\sqrt{\hat{p}(y)} + \sqrt{p(y)}\right)\\
&+ \frac{4\mathcal{R}_m(\mathcal{W}\circ\mathcal{F})}{\sqrt{m}} + 2\kappa\sqrt{\frac{\ln(4/\delta)}{2m}} + 2\kappa_{\mathbf{S}}\sqrt{\frac{\ln(4/\delta)}{2n}}.
\end{aligned}$$

Define $Z_{\mathbf{S}} = [f(x_1^+), \dots, f(x_m^+)] \in \mathbb{R}^{d\times m}$. Then, we have $A = [Z_{\mathbf{S}}{}^\top \otimes I_r]$. Thus,

$$\mathbf{P}_A = I - [Z_{\mathbf{S}}{}^\top \otimes I_r][Z_{\mathbf{S}} Z_{\mathbf{S}}{}^\top \otimes I_r]^\dagger [Z_{\mathbf{S}} \otimes I_r] = I - [Z_{\mathbf{S}}{}^\top (Z_{\mathbf{S}} Z_{\mathbf{S}}{}^\top)^\dagger Z_{\mathbf{S}} \otimes I_r] = [\mathbf{P}_{Z_{\mathbf{S}}} \otimes I_r]$$

where $\mathbf{P}_{Z_{\mathbf{S}}} = I_m - Z_{\mathbf{S}}{}^\top (Z_{\mathbf{S}} Z_{\mathbf{S}}{}^\top)^\dagger Z_{\mathbf{S}} \in \mathbb{R}^{m\times m}$. By defining $Y_{\mathbf{S}} = [g^*(x_1^+), \dots, g^*(x_m^+)]^\top \in \mathbb{R}^{m\times r}$, since $g = \operatorname{vec}[Y_{\mathbf{S}}{}^\top]$,

$$\|\mathbf{P}_A g\|_2 = \|[\mathbf{P}_{Z_{\mathbf{S}}} \otimes I_r]\operatorname{vec}[Y_{\mathbf{S}}{}^\top]\|_2 = \|\operatorname{vec}[Y_{\mathbf{S}}{}^\top \mathbf{P}_{Z_{\mathbf{S}}}]\|_2 = \|\mathbf{P}_{Z_{\mathbf{S}}} Y_{\mathbf{S}}\|_F \qquad (8.41)$$

On the other hand, recall that W_S is the minimum norm solution as

$$W_S = \underset{W'}{\text{minimize}}\, \|W'\|_F \text{ s.t. } W' \in \arg\min_W \frac{1}{n}\sum_{i=1}^{n} \|W f_\theta(x_i) - y_i\|^2.$$

By solving this, we have

$$W_S = Y^\top Z_S{}^\top (Z_S Z_S{}^\top)^\dagger,$$

where $Z_S = [f(x_1), \dots, f(x_n)] \in \mathbb{R}^{d \times n}$ and $Y_S = [y_1, \dots, y_n]^\intercal \in \mathbb{R}^{n \times r}$. Then,

$$\begin{aligned} L_S(w_S) = \frac{1}{n}\sum_{i=1}^{n} \|W_S f_\theta(x_i) - y_i\| &= \frac{1}{n}\sum_{i=1}^{n} \sqrt{\sum_{k=1}^{r} ((W_S f_\theta(x_i) - y_i)_k)^2} \\ &\leq \sqrt{\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{r} ((W_S f_\theta(x_i) - y_i)_k)^2} \\ &= \frac{1}{\sqrt{n}} \|W_S Z_S - Y^\intercal\|_F \\ &= \frac{1}{\sqrt{n}} \|Y^\intercal (Z_S{}^\intercal (Z_S Z_S{}^\intercal)^\dagger Z_S - I)\|_F \\ &= \frac{1}{\sqrt{n}} \|(I - Z_S{}^\intercal (Z_S Z_S{}^\intercal)^\dagger Z_S) Y\|_F \end{aligned}$$

Thus,

$$L_S(w_S) = \frac{1}{\sqrt{n}} \|\mathbf{P}_{Z_S} Y\|_F \tag{8.42}$$

where $\mathbf{P}_{Z_S} = I - Z_S{}^\intercal (Z_S Z_S{}^\intercal)^\dagger Z_S$.

By combining (8.40)–(8.42) and using $1 \leq \sqrt{2}$, we have that with probability at least $1 - \delta$,

$$\mathbb{E}_{X,Y}[\|W_S f_\theta(X) - Y\|] \leq c I_{\mathbf{S}}(f_\theta) + \frac{2}{\sqrt{m}} \|\mathbf{P}_{Z_{\mathbf{S}}} Y_{\mathbf{S}}\|_F + \frac{1}{\sqrt{n}} \|\mathbf{P}_{Z_S} Y_S\|_F + Q_{m,n}, \tag{8.43}$$

where

$$\begin{aligned} Q_{m,n} = c &\left(\frac{2\tilde{\mathcal{R}}_m(\mathcal{F})}{\sqrt{m}} + \tau \sqrt{\frac{\ln(3/\delta)}{2m}} + \tau_{\mathbf{S}} \sqrt{\frac{\ln(3/\delta)}{2n}} \right) \\ &+ \kappa_S \sqrt{\frac{2\ln(6|\mathcal{Y}|/\delta)}{2n}} \sum_{y \in \mathcal{Y}} \left(\sqrt{\hat{p}(y)} + \sqrt{p(y)} \right) \\ &+ \frac{4\mathcal{R}_m(\mathcal{W} \circ \mathcal{F})}{\sqrt{m}} + 2\kappa \sqrt{\frac{\ln(4/\delta)}{2m}} + 2\kappa_{\mathbf{S}} \sqrt{\frac{\ln(4/\delta)}{2n}}. \end{aligned}$$

■

9. Known Lemmas

We use the following well-known theorems as lemmas in our proof. We put these below for completeness. These are classical results and *not* our results.

Lemma 1 (Hoeffding's inequality) *Let $X_1, ..., X_n$ be independent random variables such that $a \leq X_i \leq b$ almost surely. Consider the average of these random variables, $S_n = \frac{1}{n}(X_1 + \cdots + X_n)$. Then, for all $t > 0$,*

$$\mathbb{P}_S \left(\mathrm{E}\left[S_n\right] - S_n \geq (b-a)\sqrt{\frac{\ln(1/\delta)}{2n}} \right) \leq \delta,$$

and

$$\mathbb{P}_S \left(S_n - \mathrm{E}\left[S_n\right] \geq (b-a)\sqrt{\frac{\ln(1/\delta)}{2n}} \right) \leq \delta.$$

Proof By using Hoeffding's inequality, we have that for all $t > 0$,

$$\mathbb{P}_S \left(\mathrm{E}\left[S_n\right] - S_n \geq t \right) \leq \exp\left(-\frac{2nt^2}{(b-a)^2} \right),$$

and

$$\mathbb{P}_S \left(S_n - \mathrm{E}\left[S_n\right] \geq t \right) \leq \exp\left(-\frac{2nt^2}{(b-a)^2} \right),$$

Setting $\delta = \exp\left(-\frac{2nt^2}{(b-a)^2}\right)$ and solving for $t > 0$,

$$\begin{aligned} 1/\delta &= \exp\left(\frac{2nt^2}{(b-a)^2}\right) \\ \implies \ln(1/\delta) &= \frac{2nt^2}{(b-a)^2} \\ \implies \frac{(b-a)^2 \ln(1/\delta)}{2n} &= t^2 \\ \implies t &= (b-a)\sqrt{\frac{\ln(1/\delta)}{2n}} \end{aligned}$$

■

It has been shown that generalization bounds can be obtained via Rademacher complexity (Bartlett and Mendelson, 2002; Mohri et al., 2012; Shalev-Shwartz and Ben-David, 2014). The following is a trivial modification of (Mohri et al., 2012, Theorem 3.1) for a one-sided bound on the nonnegative general loss functions:

Lemma 2 *Let $\mathcal{G}$ be a set of functions with the codomain $[0, M]$. Then, for any $\delta > 0$, with probability at least $1 - \delta$ over an i.i.d. draw of m samples $S = (q_i)_{i=1}^m$, the following holds for all $\psi \in \mathcal{G}$:*

$$\mathbb{E}_q[\psi(q)] \leq \frac{1}{m}\sum_{i=1}^{m} \psi(q_i) + 2\mathcal{R}_m(\mathcal{G}) + M\sqrt{\frac{\ln(1/\delta)}{2m}}, \tag{9.44}$$

where $\mathcal{R}_m(\mathcal{G}) := \mathbb{E}_{S,\xi}[\sup_{\psi\in\mathcal{G}} \frac{1}{m}\sum_{i=1}^m \xi_i\psi(q_i)]$ *and* $\xi_1, \ldots, \xi_m$ *are independent uniform random variables taking values in* $\{-1, 1\}$.

Proof Let $S = (q_i)_{i=1}^m$ and $S' = (q_i')_{i=1}^m$. Define

$$\varphi(S) = \sup_{\psi\in\mathcal{G}} \mathbb{E}_{x,y}[\psi(q)] - \frac{1}{m}\sum_{i=1}^m \psi(q_i). \tag{9.45}$$

To apply McDiarmid's inequality to $\varphi(S)$, we compute an upper bound on $|\varphi(S) - \varphi(S')|$ where S and S' be two test datasets differing by exactly one point of an arbitrary index i_0; i.e., $S_i = S_i'$ for all $i \neq i_0$ and $S_{i_0} \neq S_{i_0}'$. Then,

$$\varphi(S') - \varphi(S) \leq \sup_{\psi\in\mathcal{G}} \frac{\psi(q_{i_0}) - \psi(q_{i_0}')}{m} \leq \frac{M}{m}. \tag{9.46}$$

Similarly, $\varphi(S) - \varphi(S') \leq \frac{M}{m}$. Thus, by McDiarmid's inequality, for any $\delta > 0$, with probability at least $1 - \delta$,

$$\varphi(S) \leq \mathbb{E}_S[\varphi(S)] + M\sqrt{\frac{\ln(1/\delta)}{2m}}. \tag{9.47}$$

Moreover,

$$\mathbb{E}_S[\varphi(S)] \tag{9.48}$$

$$= \mathbb{E}_S\left[\sup_{\psi\in\mathcal{G}} \mathbb{E}_{S'}\left[\frac{1}{m}\sum_{i=1}^m \psi(q_i')\right] - \frac{1}{m}\sum_{i=1}^m \psi(q_i)\right] \tag{9.49}$$

$$\leq \mathbb{E}_{S,S'}\left[\sup_{\psi\in\mathcal{G}} \frac{1}{m}\sum_{i=1}^m (\psi(q_i') - \psi(q_i))\right] \tag{9.50}$$

$$\leq \mathbb{E}_{\xi,S,S'}\left[\sup_{\psi\in\mathcal{G}} \frac{1}{m}\sum_{i=1}^m \xi_i(\psi(q_i') - \psi(q_i))\right] \tag{9.51}$$

$$\leq 2\mathbb{E}_{\xi,S}\left[\sup_{\psi\in\mathcal{G}} \frac{1}{m}\sum_{i=1}^m \xi_i\psi(q_i)\right] = 2\mathcal{R}_m(\mathcal{G}) \tag{9.52}$$

where the first line follows the definitions of each term, the second line uses Jensen's inequality and the convexity of the supremum, and the third line follows that for each $\xi_i \in \{-1, +1\}$, the distribution of each term $\xi_i(\ell(f(x_i'), y_i') - \ell(f(x_i), y_i))$ is the distribution of $(\ell(f(x_i'), y_i') - \ell(f(x_i), y_i))$ since S and S' are drawn iid with the same distribution. The fourth line uses the subadditivity of supremum. ■

10. SimCLR

In contrastive learning, different augmented views of the same image are attracted (positive pairs), while different augmented views are repelled (negative pairs). MoCo (He et al., 2020) and SimCLR (Chen et al., 2020) are recent examples of self-supervised visual representation learning that reduce the gap between self-supervised and fully-supervised learning. SimCLR applies randomized augmentations to an image to create two different views, x and y, and encodes both of them with a shared encoder, producing representations r_x and r_y. Both r_x and r_y are $l2$-normalized. The SimCLR version of the InfoNCE objective is:

$$\mathbb{E}_{x,y}\left[-\log\left(\frac{e^{\frac{1}{\eta}r_y^T r_x}}{\sum_{k=1}^{K} e^{\frac{1}{\eta}r_{y_k}^T r_x}}\right)\right],$$

where η is a temperature term and K is the number of views in a minibatch.

11. Entropy Estimators

Entropy estimation is one of the classical problems in information theory, where Gaussian mixture density is one of the most popular representations. With a sufficient number of components, they can approximate any smooth function with arbitrary accuracy. For Gaussian mixtures, there is, however, no closed-form solution to differential entropy. There exist several approximations in the literature, including loose upper and lower bounds (Huber et al., 2008). Monte Carlo (MC) sampling is one way to approximate Gaussian mixture entropy. With sufficient MC samples, an unbiased estimate of entropy with an arbitrarily accurate can be obtained. Unfortunately, MC sampling is a very computationally expensive and typically requires a large number of samples, especially in high dimensions (Brewer, 2017). Using the first two moments of the empirical distribution, VIGCreg used one of the most straightforward approaches for approximating the entropy. Despite this, previous studies have found that this method is a poor approximation of the entropy in many cases Huber et al. (2008). Another options is to use the LogDet function. Several estimators have been proposed to implement it, including uniformly minimum variance unbiased (UMVU) (Ahmed and Gokhale, 1989), and bayesian methods Misra et al. (2005). These methods, however, often require complex optimizations. The LogDet estimator presented in Zhouyin and Liu (2021) used the differential entropy α order entropy using scaled noise. They demonstrated that it can be applied to high-dimensional features and is robust to random noise. Based on Taylor-series expansions, Huber et al. (2008) presented a lower bound for the entropy of Gaussian mixture random vectors. They use Taylor-series expansions of the logarithm of each Gaussian mixture component to get an analytical evaluation of the entropy measure. In addition, they present a technique for splitting Gaussian densities to avoid components with high variance, which would require computationally expensive calculations. Kolchinsky and Tracey (2017) introduce a novel family of estimators for the mixture entropy. For this family, a pairwise-distance function between component densities defined for each member. These estimators are computationally efficient, as long as the pairwise-distance function and the entropy of each component distribution are easy to compute. Moreover, the estimator is continuous and smooth and is therefore useful for optimization problems. In addition, they presented both lower bound (using Chernoff distance) and an upper bound (using the

KL divergence) on the entropy, which are are exact when the component distributions are grouped into well-separated clusters,

12. EM, Information and collapsing

Let us examine a toy dataset on the pattern of two intertwining moons to illustrate the collapse phenomenon under GMM (Figure 1 - right). We begin by training a classical GMM with maximum likelihood, where the means are initialized based on random samples, and the covariance is used as the identity matrix. A red dot represents the Gaussian's mean after training, while a blue dot represents the data points. In the presence of fixed input samples, we observe that there is no collapsing and that the entropy of the centers is high (Figure 4 - left, in the Appendix). However, when we make the input samples trainable and optimize their location, all the points collapse into a single point, resulting in a sharp decrease in entropy (Figure 4 - right, in the Appendix).

To prevent collapse, we follow the K-means algorithm in enforcing sparse posteriors, i.e. using small initial standard deviations and learning only the mean. This forces a one-to-one mapping which leads all points to be closest to the mean without collapsing, resulting in high entropy (Figure 4 - middle, in the Appendix). Another option to prevent collapse is to use different learning rates for input and parameters. Using this setting, the collapsing of the parameters does not maximize the likelihood. Figure 1 (right) shows the results of GMM with different learning rates for learned inputs and parameters. When the parameter learning rate is sufficiently high in comparison to the input learning rate, the entropy decreases much more slowly and no collapse occurs.

13. Experimental Verification of Information-Based Bound Optimization

Setup Our experiments are conducted on CIFAR-10 Krizhevsky and Hinton (2009). We use ResNet-18 (He et al., 2016) as our backbone. Each model is trained with 512 batch size for 800 epochs. We use linear evaluation to assess the quality of the representation. Once the model has been pre-trained, we follow the same fine-tuning procedures as for the baseline methods (Caron et al., 2020).

14. On Benefits of Information Maximization for Generalization

In this Appendix, we present the complete version of Theorem 1 along with its proof and additional discussions.

14.1. Additional Notation and details

We start to introduce additional notation and details. We use the notation of $x \in \mathcal{X}$ for an input and $y \in \mathcal{Y} \subseteq \mathbb{R}^r$ for an output. Define $p(y) = \mathbb{P}(Y = y)$ to be the probability of getting label y and $\hat{p}(y) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{y_i = y\}$ to be the empirical estimate of $p(y)$. Let ζ be an upper bound on the norm of the label as $\|y\|_2 \leq \zeta$ for all $y \in \mathcal{Y}$. Define the minimum norm solution $W_{\mathbf{S}}$ of the unlabeled data as $W_{\mathbf{S}} = \text{minimize}_{W'} \|W'\|_F$ s.t. $W' \in \arg\min_W \frac{1}{m}\sum_{i=1}^{m} \|Wf_\theta(x_i^+) - g^*(x_i^+)\|^2$. Let κ_S be a data-dependent upper bound on the per-sample Euclidian norm loss with the trained model as $\|W_S f_\theta(x) - y\| \leq \kappa_S$ for

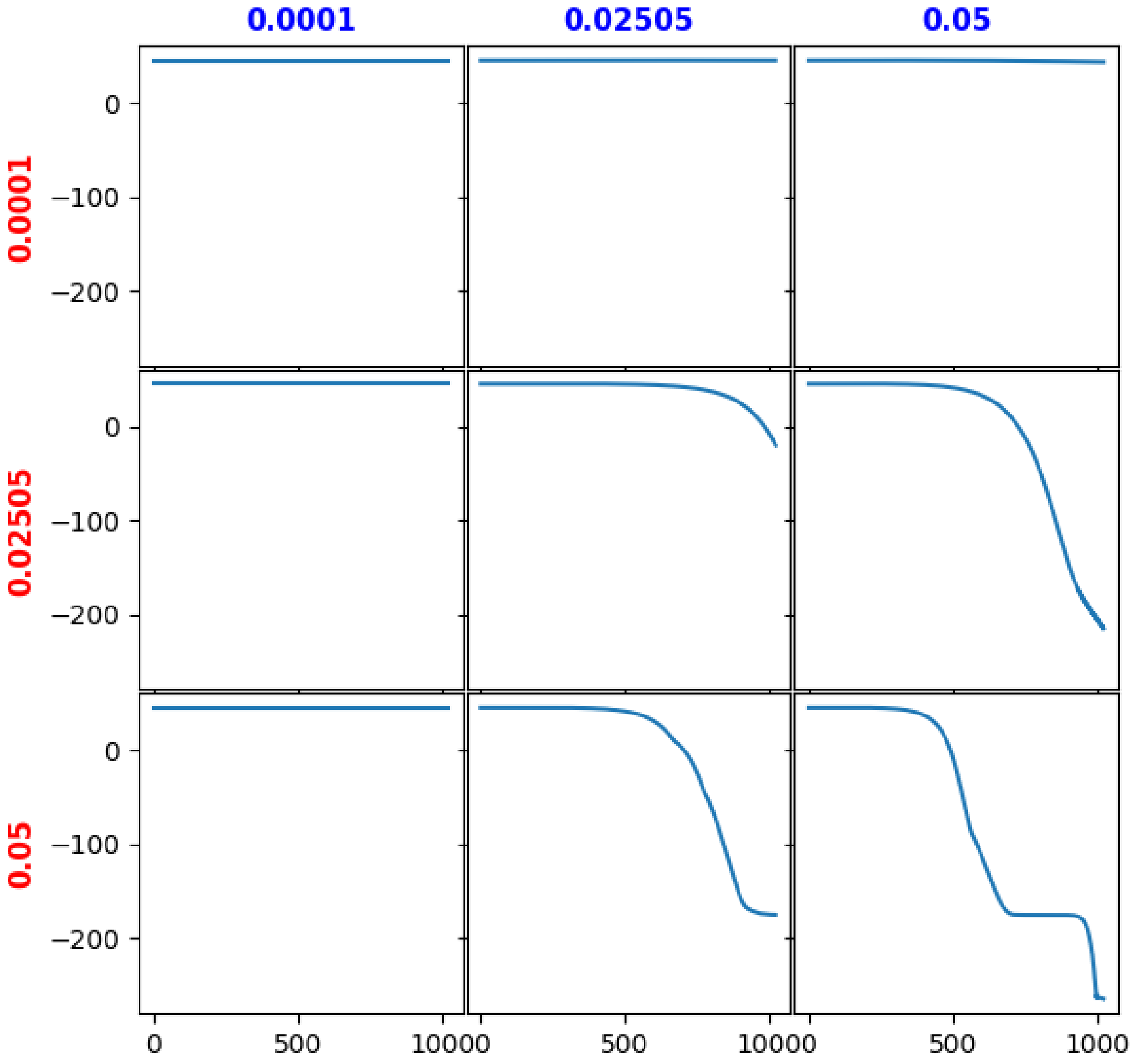

Figure 3: Evolution of the entropy for each of the learning rate configurations showing that the impact of picking the incorrect learning rate for the data and/or centroids lead to a collapse of the samples.

all $(x, y) \in \mathcal{X} \times \mathcal{Y}$. Similarly, let $\kappa_{\mathbf{S}}$ be a data-dependent upper bound on the per-sample Euclidian norm loss as $\|W_{\mathbf{S}} f_\theta(x) - y\| \leq \kappa_{\mathbf{S}}$ for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$. Define the difference between W_S and $W_{\mathbf{S}}$ by $c = \|W_S - W_{\mathbf{S}}\|_2$. Let $\mathcal{W}$ be a hypothesis space of W such that $W_{\mathbf{S}} \in \mathcal{W}$. We denote by $\tilde{\mathcal{R}}_m(\mathcal{W} \circ \mathcal{F}) = \frac{1}{\sqrt{m}} \mathbb{E}_{\mathbf{S},\xi}[\sup_{W \in \mathcal{W}, f \in \mathcal{F}} \sum_{i=1}^m \xi_i \|g^*(x_i^+) - W f(x_i^+)\|]$ the normalized Rademacher complexity of the set $\{x^+ \mapsto \|g^*(x^+) - W f(x^+)\| : W \in \mathcal{W}, f \in \mathcal{F}\}$.

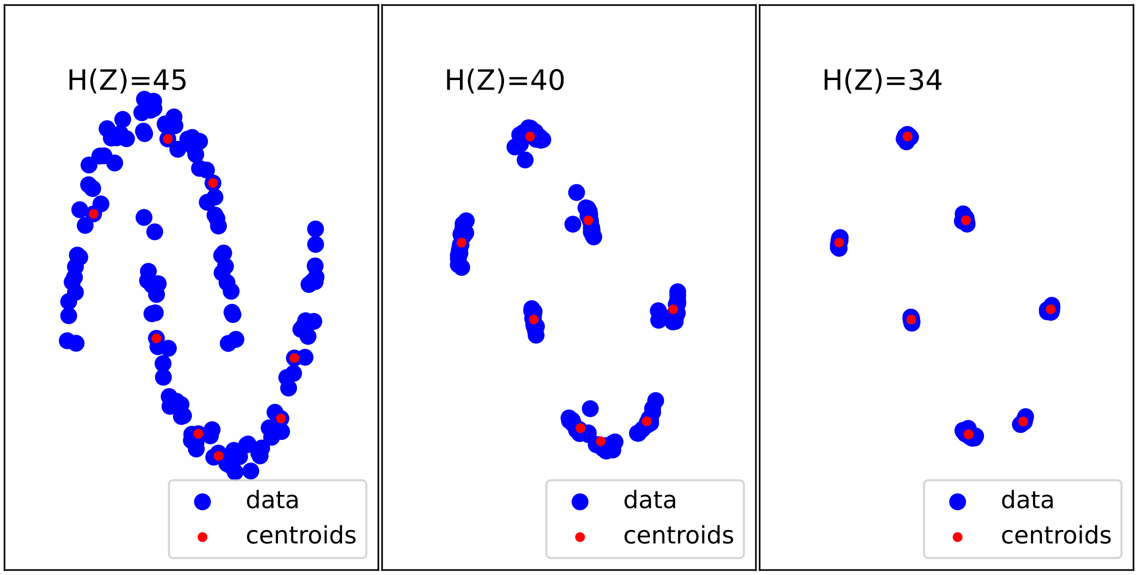

Figure 4: **Evolution of GMM training when enforcing a one-to-one mapping between the data and centroids akin to K-means i.e. using a small and fixed covariance matrix. We see that collapse does not occur.** Left - In the presence of fixed input samples, we observe that there is no collapsing and that the entropy of the centers is high. Right - when we make the input samples trainable and optimize their location, all the points collapse into a single point, resulting in a sharp decrease in entropy.

we denote by κ a upper bound on the per-sample Euclidian norm loss as $\|Wf(x) - y\| \leq \kappa$ for all $(x, y, W, f) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{W} \times \mathcal{F}$.

We adopt the following data-generating process model that is used in the previous paper on analyzing contrastive learning (Saunshi et al., 2019; Ben-Ari and Shwartz-Ziv, 2018). For the labeled data, first, y is drawn from the distritbuion ρ on $\mathcal{Y}$, and then x is drawn from the conditional distribution $\mathcal{D}_y$ conditioned on the label y. That is, we have the join distribution $\mathcal{D}(x, y) = \mathcal{D}_y(x)\rho(y)$ with $((x_i, y_i))_{i=1}^n \sim \mathcal{D}^n$. For the unlabeled data, first, each of the *unknown* labels y^+ and y^- is drawn from the distritbuion ρ, and then each of the positive examples x^+ and x^{++} is drawn from the conditional distribution $\mathcal{D}_{y^+}$ while the negative example x^- is drawn from the $\mathcal{D}_{y^-}$. Unlike the analysis of contrastive learning, we do not require the negative samples. Let $\tau_{\mathbf{S}}$ be a data-dependent upper bound on the invariance loss with the trained representation as $\|f_\theta(\bar{x}) - f_\theta(x)\| \leq \tau_{\mathbf{S}}$ for all $(\bar{x}, x) \sim \mathcal{D}_y^2$ and $y \in \mathcal{Y}$. Let τ be a data-independent upper bound on the invariance loss with the trained representation as$\|f(\bar{x}) - f(x)\| \leq \tau$ for all $(\bar{x}, x) \sim \mathcal{D}_y^2$, $y \in \mathcal{Y}$, and $f \in \mathcal{F}$. For the simplicity, we assume that there exists a function g^* such that $y = g^*(x) \in \mathbb{R}^r$ for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$. Discarding this assumption adds the average of label noises to the final result, which goes to zero as the sample sizes n and m increase, assuming that the mean of the label noise is zero.