WHAT DO WE MAXIMIZE IN SELF-SUPERVISED LEARNING AND WHY DOES GENERALIZATION EMERGE?

Anonymous authors
Paper under double-blind review

ABSTRACT

In this paper, we provide an information-theoretic (IT) understanding of self-supervised learning methods, their construction, and optimality. As a first step, we demonstrate how IT quantities can be obtained for deterministic networks, as an alternative to the commonly used unrealistic stochastic networks assumption. Secondly, we demonstrate how different SSL models can be (re)discovered based on first principles and highlight what the underlying assumptions of different SSL variants are. Third, we derive a novel generalization bound based on our IT understanding of SSL methods, providing generalization guarantees for the downstream supervised learning task. As a result of this bound, along with our unified view of SSL, we can compare the different approaches and provide general guidelines to practitioners. Consequently, our derivation and insights can contribute to a better understanding of SSL and transfer learning from a theoretical and practical perspective.

1 INTRODUCTION

Self-Supervised Learning methods (SSL) learn representations using a surrogate objective (i.e. a self-supervised objective) between inputs and self-defined signals. In SimCLR (Chen et al., 2020), for example, a contrastive loss is defined that makes representations for different versions of the same image similar, while making the representations for different images different. After optimizing the surrogate objective, the pre-trained model is used as a feature extractor. The features are used as inputs to a downstream supervised task, such as image classification, object detection, instance segmentation, and pose estimation (Caron et al., 2021; Chen et al., 2020; Misra & Maaten, 2020). However, despite success in practice, only a few small number of authors (Arora et al., 2019; Lee et al., 2021a) have sought to provide theoretical insights about the effectiveness of SSL.

In recent years, information theory methods have played a key role in several deep learning achievements, from practical applications in representation learning as the variational information bottleneck (Alemi et al., 2016), to theoretical investigations as generalization bounds (Xu & Raginsky, 2017; Steinke & Zakynthinou, 2020; Shwartz-Ziv, 2022). Moreover, different deep learning problems have been successfully approached by developing and applying novel estimators and learning principles derived from information-theoretic quantities, such as mutual information estimation. Specifically, many works have attempted to analyze SSL from an information theory perspective. An example is the use of the renowned information maximization (InfoMax) principle (Linsker, 1988) in self-supervised representation learning (Bachman et al., 2019). However, looking at these works may be confusing. Numerous objective functions are presented without a rigorous justification, some contradicting each other, as well as many implicit assumptions. Moreover, these works rely on a crucial assumption: a stochastic (often Gaussian) DN mapping, which is rarely the case nowadays.

This paper presents a unified framework for SSL methods from an information theory perspective which can be applied to deterministic DN training. We summarize our contributions into three points: (i) First, in order to study deterministic DNs from an information theory perspective, we shift stochasticity to the DN input, which is a much more faithful assumption for current training techniques. (ii) Second, based on this formulation, we analyze how current SSL methods that use deterministic networks optimize information-theoretic quantities. (iii) Third, we study how the

optimization of information-theoretic quantities is related to the final performance in the downstream task using a new generalization bound.

2 BACKGROUND

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

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

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

Deep Networks. A deep network (DN) is a (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 DNs operators can be found in Goodfellow et al. (2016). We will omit the Θ notation for clarity unless needed. The only assumption we require for our study is that the non-linearities present in the DN are CPA, as is the case with (leaky-) ReLU, absolute value, and max-pooling. In that case, the entire input-output mapping becomes a CPA spline with an implicit partition Ω, the function of the weights and architecture of the network (Montufar et al., 2014; Balestriero & Baraniuk, 2018). For smooth nonlinearities, our results hold from a first-order Taylor approximation argument.

Self-Supervised Learning. Joint embedding methods learn the DN parameters Θ without supervision and input reconstruction. The difficulty of SSL is to produce a good representation for downstream tasks whose labels are not available during training —while avoiding a trivially simple solution where the model maps all inputs to constant output. Many methods have been proposed to solve this problem, see Balestriero & LeCun (2022) for a summary and connections between methods. *Contrastive methods* learn representations by contrasting positive and negative examples, e.g. SimCLR (Chen et al., 2020) and its InfoNCE criterion (Oord et al., 2018). Other recent work introduced *non-contrastive methods* that employ different regularization methods to prevent collapsing of the representation. Several papers used stop-gradients and extra predictors to avoid collapse (Chen & He, 2021; Grill et al., 2020) while Caron et al. (2020) uses an additional clustering step. As opposed to contrastive methods, noncontrastive methods do not explicitly rely on negative samples. Of particular interest to us is the *VICReg* method (Bardes et al., 2021) that considers two embedding batches $\boldsymbol{Z} = [f(\boldsymbol{x}_1), \ldots, f(\boldsymbol{x}_N)]$ and $\boldsymbol{Z}' = [f(\boldsymbol{x}'_1), \ldots, f(\boldsymbol{x}'_N)]$ each of size $(N \times K)$. Denoting by $\boldsymbol{C}$ the $(K \times K)$ covariance matrix obtained from $[\boldsymbol{Z}, \boldsymbol{Z}']$ we obtain the VICReg triplet loss

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

Deep Networks and Information-Theory. Recently, information-theoretic methods have played a key role in several remarkable deep learning achievements (Alemi et al., 2016; Xu & Raginsky, 2017; Steinke & Zakynthinou, 2020; Shwartz-Ziv & Tishby, 2017). Moreover, different deep learning problems have been successfully approached by developing and applying information-theoretic estimators and learning principles (Hjelm et al., 2018; Belghazi et al., 2018; Piran et al., 2020; Shwartz-Ziv et al., 2018). There is, however, a major problem when it comes to analyzing information-theoretic objectives in deterministic deep neural networks: the source of randomness. The mutual information between the input and the representation in such networks is infinite, resulting in ill-posed optimization problems or piecewise constant, making gradient-based optimization methods ineffective (Amjad & Geiger, 2019). To solve these problems, researchers have proposed several solutions. For SSL, stochastic deep networks with variational bounds could be used, where the output of the deterministic network is used as parameters of the conditional distribution (Lee et al., 2021b; Shwartz-Ziv & Alemi, 2020). Dubois et al. (2021) suggested another option, which assumed

that the randomness of data augmentation among the two views is the source of stochasticity in the network. For supervised learning, Goldfeld et al. (2018) introduced an auxiliary (noisy) DN framework by injecting additive noise into the model and demonstrated that it is a good proxy for the original (deterministic) DN in terms of both performance and representation. Finally, Achille and Soatto (Achille & Soatto, 2018) found that minimizing a stochastic network with a regularizer is equivalent to minimizing cross-entropy over deterministic DNs with multiplicative noise. However, all of these methods assume that the noise comes from the model itself, which contradicts current training methods. In this work, we explicitly assume that the stochasticity comes from the data, which is a less restrictive assumption and does not require changing current algorithms.

3 Information Theory for Deterministic Deep Networks

This section first sets up notation and assumption on the information-theoretic challenges in SSL (section 3.1) and on our assumptions regarding the data distribution (section 3.2) so that any training sample $\boldsymbol{x}$ can be seen as coming from a single Gaussian distribution as in $\boldsymbol{x} \sim \mathcal{N}(\mu_{\boldsymbol{x}}, \Sigma_{\boldsymbol{x}})$. From this we obtain that the output of any deep network $f(\boldsymbol{x})$ corresponds to a mixture of truncated Gaussian (section 3.3). In particular, it can fall back to a single Gaussian under small noise ($\det(\Sigma) \to \epsilon$) assumptions. These results will enable information measures to be applied to deterministic DNs. We then recover known SSL methods (Bardes et al., 2021; Chen et al., 2020) by making different assumptions about the data distribution and estimating their information.

3.1 SSL as an Information-Theoretic Problem

To better grasp the difference between key SSL methods, we first formulate the general SSL goal from an information-theoretical perspective. We start with the *MultiView InfoMax* principle, i.e., maximizing the mutual information between the representations of the two views. To do so, as shown in Federici et al. (2020), we maximize $I(Z; X')$ and $I(Z'; X)$ using the lower bound

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

where $H(Z)$ is the entropy of Z. In supervised learning, where we need to maximize $I(Z; Y)$, the labels (Y) are fixed, the entropy term $H(Y)$ is constant, and you only need to optimize the log-loss $E[\log q(z|x)]$ (cross-entropy or square loss). However, it is well known that for Siamese networks there exists a degenerate solution, in which all outputs "collapse" into an undesired value (Chen et al., 2020). By looking on eq. (2) we can see that the entropies $H(Z)$ and $H(Z')$ are not constant and can be optimized throughout the learning process. Therefore, only maximizing $E[\log q(z|x')]$ will cause it to collapse to the trivial solution of making the representations constant (where the entropy goes to zero). To regularize these entropies, i.e., prevent collapse, different methods utilize different approaches to implicit regularizing information. To recover them in section 4, we must first introduce the notation and results around the data distribution (section 3.2) and how a DN transforms that distribution (section 3.3).

3.2 Data Distribution Hypothesis

Our first step is to assess how the output random variables of the network are represented, assuming a distribution on the data itself. Under 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. Therefore, we will consider throughout this study the conditioning of a latent representation with respect to the mean of the observation, i.e., $X|\boldsymbol{x}^* \sim \mathcal{N}(\boldsymbol{x}^*, \Sigma_{\boldsymbol{x}^*})$ where the eigenvectors of $\Sigma_{\boldsymbol{x}^*}$ are in the same linear subspace than the tangent space of the data manifold at $\boldsymbol{x}^*$, which varies with the position of $\boldsymbol{x}^*$ in space.

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

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

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. This keeps things general, as it is enough to cover

the domain of the data manifold overall, without overlap between different Gaussians. Hence, we have that.

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

where $\mathcal{N}(\boldsymbol{x}; ., .)$ is the 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. Note that this is not restrictive since, given a sufficiently large N, the above can represent any manifold with an arbitrarily good approximation.

3.3 Data Distribution After Deep Network Transformation

Consider an affine spline operator f (Eq. 1) that goes from a space of dimension D to a space of dimension K with $K \geq D$. The span, that we denote as image, of this mapping is given by

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

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

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

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

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

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. The proof of of Theorem 1 is presented in Appendix A. □

4 Information Optimization and Optimality

Next, we will show how SSL algorithms for deterministic networks can be derived. According to Section 3.1, we want to maximize $I(Z; X')$ and $I(Z'; X)$. When our input noise is small, we can reduce the conditional output density $p_{\boldsymbol{z}|\boldsymbol{x}*}$ to a single Gaussian.

$$(Z'|X' = x_n) \sim \mathcal{N}(\mu_n, \Sigma_n),$$

where we abbreviated the parameters.Combine with the result from section 3.1, we see that one should optimize both $H(Z|X')$ and $H(Z)$. As in a standard regression task, we assume a Gaussian observation model, which means $p(z|z') \sim \mathcal{N}(z', \Sigma_r)$. Using the mean square error as a loss function in regression tasks is a particular application of this assumption, where $\Sigma_r = I$. To compute the expected loss, we need to marginalize out the stochasticity in Z', which means that the conditional decoder is a Gaussian:

$$q(z|X' = x^*_n) = \int q(z|z') p(z'|x^*_n) dz', \tag{6}$$

which gives the distribution $\mathcal{N}(\mu_n, \Sigma_r + \Sigma_n)$ meaning that we can lower bound the mutual information with

$$I(Z; X') \geq H(Z) - \frac{d}{2} \log 2\pi\Sigma_r - \sum_{n=1}^{N} \frac{1}{2} (z_n - z'_n)^T \Sigma_r^{-1} (z_n - z'_n) - \log |\Sigma_n|. \tag{7}$$

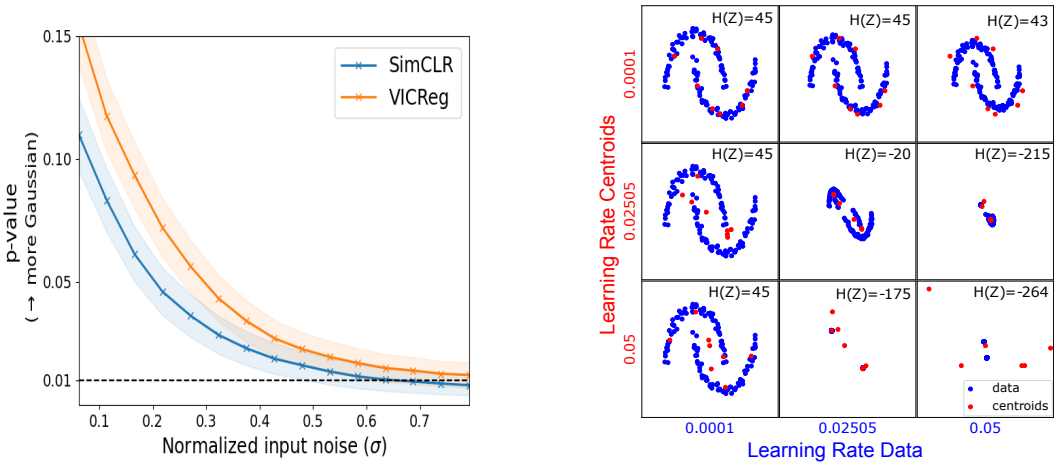

Figure 1: **Left: The network output SSL training is more Gaussian for small input noise**. The P-value of the normality test for different SSL models trained on CIFAR-10 for different input noise levels. The dashed line represents the point at which the null hypothesis (Gaussian distribution) can be rejected with 99% confidence. **Right: Smaller learning rates prevent collapsing.** GMM points where in black is the entropy estimate based on a Kernel Density Estimator, in blue and red are the data points and GMM centroids respectively, with the corresponding learning rate for each used to maximize the model's likelihood.

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

4.1 Deriving VICReg From First Principles

As a result of the above information-theoretic principles, we can now reconstruct VICReg from first principles. We are seeking to estimate the entropy $H(Z)$ in eq. (7), where Z is a Gaussian mixture. Due to the logarithm of a sum of exponential functions, this quantity cannot be determined for a mixture of Gaussian densities, except in the case of a single Gaussian density. Nevertheless, there are several approximations in the literature that provide upper and lower bounds. Among these methods, some employ the logarithmic sum of the probability (Kolchinsky & Tracey, 2017), while others employ entropy-adjusted logarithmic probabilities (Huber et al., 2008).

A simpler solution is to approximate the entire mixture as a single Gaussian by only capturing the first two moments. This method provides an upper bound approximation of our entropy of interest $H(Z)$ since the Gaussian distribution maximizes entropy for a given covariance matrix. Using Σ_Z as the covariance matrix of Z, we find that we should maximize:

$$\max_{Z,Z'} \sum_{n=1}^{N} \log \frac{|\Sigma_Z|}{|\Sigma_n|} - \frac{1}{2}(z_n - z'_n)^T \Sigma_r^{-1} (z_n - z'_n) - \frac{d}{2} \log 2\pi\Sigma_r$$

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

the maximum eigenvalue is greater than or equal to each diagonal element. Therefore, under the constraint of the eigenvalues of the matrix, the most efficient way is to decrease the off-diagonal terms and increase the diagonal terms. By setting $\Sigma_r = I$, *we therefore fully recover the VICReg objective*.

4.2 SimCLR vs VICReg

Lee et al. (2021b) connected 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. Based on our analysis in Section 4.1, we can identify two main differences between SimCLR and VICReg: (i) **The conditional distribution $\boldsymbol{p(z|x')}$**; 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 based on the finite sum of the input samples. In contrast, VICReg uses a different approach and estimates the entropy of Z solely from the first second moment. Creating self-supervised methods that combine these two differences would be an interesting future research direction.

4.3 Empirical Evaluation

Next, we would like to verify the validity of our assumptions. Based on the theory presented in Section 3.3, the conditional output density $p_{\boldsymbol{z}|x=i}$ reduces to a single Gaussian with decreasing input noise. We validated it using a ResNet-18 model trained with SimCLR or VICReg on the CIFAR-10 dataset (Krizhevsky, 2009). From the test dataset, we sample 512 Gaussian samples for each image and analyzed whether these samples remain Gaussian (for each image) in the penultimate layer of the DN, independently for each output dimension. Then, we employ the D'Agostino and Pearson's test (D'Agostino, 1971) and compute the p-value under the null hypothesis that the sample represents a normal distribution. In this test, the deviation from normality is measured, and the test aims to determine whether the sample represents a normally distributed population. A kurtosis and skewness transformation is used to perform the test. The process is repeated for different noise standard deviations. Figure 1 (left) shows the p-value as a function of the normalized standard deviation. We can observe that the network's output is indeed Gaussian with a high probability for small input noise. As we increase the input noise, the network's output becomes less Gaussian until the noise distribution can be rejected with 99% confidence. Moreover, we can see that VICReg is, interestingly, more "Gaussian" than SimCLR, which may have to do with the fact that it optimizes only the second moments of the density distribution to regularize $H(Z)$.

5 Self Supervised Learning, EM and Information

Several SSL methods have been developed recently that employ the stop gradient operator and only train with positive pairs of data (Grill et al., 2020; Chen & He, 2021). According to (Chen & He, 2021), presetting the stop gradient operation implicitly involves presenting two sets of variables where the algorithm alternates between optimizing each set. Next, we formalize these SSL methods as generalized EM optimization problems, link them to information theory, and analyze how specific design choices affect their collapse.

5.1 The EM algorithm and self supervised learning

The classical approach to learning with hidden variables is based on the Expectation Maximization (EM) algorithm (Dempster et al., 1977). This algorithm performs a greedy search of the likelihood surface and is proven to converge to a local stationary point. Neal & Hinton (1998) showed that we can view it as a dual optimization where both steps are seen as maximizing the same function, $F(\tilde{P}, \theta)$. Using the same notation, we can write the functional as $F(\tilde{P}, \theta) = \mathbb{E}_{\tilde{P}}\left[P(Z, Z'|\theta)\right] + H(\tilde{P})$ where $H(\tilde{P}) = -\mathbb{E}_{\tilde{P}}\left[\log \tilde{P}(z')\right]$ is the entropy of the empirical distribution $\tilde{P}$ and $\mathbb{E}_{\tilde{P}}\left[P(Z, Z'|\theta)\right]$ is the regular likelihood. Using this formulation, Neal & Hinton (1998) showed that the (G)EM algorithm maximizes a variational lower bound on the log likelihood. However, as discussed in section 3.1, optimize the likelihood can be problematic when both variables are changing. Unlike the classic

EM algorithm, for SSL, our input variable Z changes in each iteration, and the optimization is with respect to both Z and Z'.

5.2 Preventing Point Collapse under the EM algorithm

For Gaussian mixture models (GMMs), clustering consists of estimating the parameters that maximize its likelihood function, followed by assigning to each data point the cluster corresponding to its most likely multivariate Gaussian distribution. Chen & He (2021) suggested that the SimSiam method can be viewed as the K-means algorithm, which can be derived by reducing the GMMs by: (a) fixing all covariances a priori to be the identity matrix, and (b) assigning each data point to a single cluster.

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 3 - 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 3 - right, in the Appendix).

What can we do to prevent this collapse? We follow the K-means algorithm in enforcing sparse posteriors, i.e. using very small initial standard deviations for the Gaussian distribution and learning only the mean. This forces a one-to-one mapping between the data points and the centers. Forcing the one-to-one mapping leads all points to be closest to the mean without collapsing, resulting in high entropy (Figure 3 - 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.

6 Benefits of Information Maximization for Generalization

Section 4.1 shows that VICReg maximizes the mutual information $I(Z;X') = H(Z) - H(Z|X')$ for the corresponding data assumption and entropy estimation. Maximizing the negative conditional entropy $-H(Z|X')$ corresponds to minimizing the invariance loss of VICReg, while maximizing the entropy $H(Z)$ corresponds to controlling the covariance matrix. The purpose of this section is to further connect the invariance loss, the covariance matrix, and the information with the input $I(Z;X)$ to the generalization ability of the model by deriving a novel generalization bound. Together with the results from the previous sections, this provides a mathematical understanding of the benefits of SSL through maximization of mutual information with the implicit information-regularization.

6.1 Notation

We use the notation of x for an input and $y \in \mathbb{R}^r$ for an output. We are given a labeled training data $S = ((x_i, y_i))_{i=1}^n$ of size n and an unlabeled training data $\bar{S} = ((x_i^+, x_i^{++}))_{i=1}^m$ of size m, where x_i^+ and x_i^{++} share the same (unknown) label. With the unlabeled training data, we define the invariance loss $I_{\bar{S}}(f_\theta) = \frac{1}{m}\sum_{i=1}^m \|f_\theta(x_i^+) - f_\theta(x_i^{++})\|$ where f_θ is the representation that has been trained with the unlabeled training data $\bar{S}$. We define a labeled loss $\ell_{x,y}(w) = \|Wf_\theta(x) - y\|$ 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 $W_S = \text{minimize}_{W'} \|W'\|_F$ s.t. $W' \in \arg\min_W \frac{1}{n}\sum_{i=1}^n \|Wf_\theta(x_i) - y_i\|^2$. We also define the representation matrices $Z_S = [f(x_1), \ldots, f(x_n)] \in \mathbb{R}^{d\times n}$ and $Z_{\bar{S}} = [f(x_1^+), \ldots, f(x_m^+)] \in \mathbb{R}^{d\times m}$, and the projection matrices $\mathbf{P}_{Z_S} = I - Z_S\mathcal{T}(Z_S Z_S \mathcal{T})^\dagger Z_S$ and $\mathbf{P}_{Z_{\bar{S}}} = I - Z_{\bar{S}}\mathcal{T}(Z_{\bar{S}} Z_{\bar{S}} \mathcal{T})^\dagger Z_{\bar{S}}$. Similarly, we define the label matrix $Y_S = [y_1, \ldots, y_n]\mathcal{T} \in \mathbb{R}^{n\times r}$ and the unknown label matrix $Y_{\bar{S}} = [y_1^+, \ldots, y_m^+]\mathcal{T} \in \mathbb{R}^{m\times r}$, 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 $\tilde{\mathcal{R}}_m(\mathcal{F}) = \frac{1}{\sqrt{m}}\mathbb{E}_{\bar{S},\xi}[\sup_{f\in\mathcal{F}} \sum_{i=1}^m \xi_i \|f(x_i^+) - f(x_i^{++})\|]$, 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 DNs (Bartlett et al., 2017; Kawaguchi et al., 2018; Golowich et al., 2018) and other common machine learning models (Bartlett & Mendelson, 2002; Mohri et al., 2012; Shalev-Shwartz & Ben-David, 2014).

6.2 Generalization Bound for VICReg

We now show that SSL via VICReg can be mathematically understood to improve the generalization ability for the supervised objective in the downstream task. Namely, Theorem 2 shows that the expected labeled loss $\mathbb{E}_{x,y}[\ell_{x,y}(w_S)]$ is minimized when we minimize the unlabeled invariance loss $I_{\bar{S}}(f_\theta)$ while controlling the covariance $Z_{\bar{S}}Z_{\bar{S}}\mathcal{T}$ (to minimize $\|\mathbf{P}_{Z_{\bar{S}}}Y_{\bar{S}}\|_F$ and $\|\mathbf{P}_{Z_S}Y_S\|_F$) and the complexity of representations $\tilde{\mathcal{R}}_m(\mathcal{F})$:

Theorem 2. *(Informal version). For any $\delta > 0$, with probability at least $1-\delta$, the following holds:*

$$\mathbb{E}_{x,y}[\ell_{x,y}(w_S)] \leq I_{\bar{S}}(f_\theta) + \frac{2}{\sqrt{m}}\|\mathbf{P}_{Z_{\bar{S}}}Y_{\bar{S}}\|_F + \frac{1}{\sqrt{n}}\|\mathbf{P}_{Z_S}Y_S\|_F + \frac{2\tilde{\mathcal{R}}_m(\mathcal{F})}{\sqrt{m}} + \mathcal{Q}_{m,n}, \quad (8)$$

where $\mathcal{Q}_{m,n} = O(G\sqrt{\frac{\ln(1/\delta)}{m}} + \sqrt{\frac{\ln(1/\delta)}{n}}) \to 0$ 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 2 and its proof are presented in Appendix C. □

The term $\|\mathbf{P}_{Z_{\bar{S}}}Y_{\bar{S}}\|_F$ in Theorem 2 contains the unobservable label matrix $Y_{\bar{S}}$. However, we can minimize this term by using $\|\mathbf{P}_{Z_{\bar{S}}}Y_{\bar{S}}\|_F \leq \|\mathbf{P}_{Z_{\bar{S}}}\|_F\|Y_{\bar{S}}\|_F$ and by minimizing $\|\mathbf{P}_{Z_{\bar{S}}}\|_F$. The factor $\|\mathbf{P}_{Z_{\bar{S}}}\|_F$ is minimized when the rank of the covariance $Z_{\bar{S}}Z_{\bar{S}}\mathcal{T}$ is maximized, because $\mathbf{P}_{Z_{\bar{S}}}$ is the projection matrix onto the null space of $Z_{\bar{S}}$. 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 $\|\mathbf{P}_{Z_{\bar{S}}}\|_F = 0$ when the covariance $Z_{\bar{S}}Z_{\bar{S}}\mathcal{T}$ is of full rank. On the other hand, 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. In addition, the term $\|\mathbf{P}_{Z_S}Y_S\|_F$ is also minimized when the rank of the covariance $Z_SZ_S\mathcal{T}$ is maximized. Since the covariances $Z_SZ_S\mathcal{T}$ and $Z_{\bar{S}}Z_{\bar{S}}\mathcal{T}$ 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 the term $\|\mathbf{P}_{Z_S}Y_S\|_F$ by maximizing the diagonal entries of $Z_{\bar{S}}Z_{\bar{S}}\mathcal{T}$ while minimizing its off-diagonal entries, as is done in VICReg.

Thus, VICReg can be understood as a method to minimize the generalization bound in Theorem 2 by minimizing the invariance loss while controlling the covariance $Z_{\bar{S}}Z_{\bar{S}}\mathcal{T}$ to minimize the *label-agnostic* upper bounds on $\|\mathbf{P}_{Z_{\bar{S}}}Y_{\bar{S}}\|_F$ and $\|\mathbf{P}_{Z_S}Y_S\|_F$. In this view, if we know some *partial* information about the label $Y_{\bar{S}}$ of the unlabeled training data, we can use this partial information to minimize $\|\mathbf{P}_{Z_{\bar{S}}}Y_{\bar{S}}\|_F$ and $\|\mathbf{P}_{Z_S}Y_S\|_F$ directly. This direction can be used to improve VICReg in future work for the partial observable setting.

6.3 Understanding via Mutual Information

Theorem 2 together with the result of the previous section shows that, for generalization in the downstream supervised learning task, it is helpful to maximize the mutual information $I(Z; X')$ in SSL via minimizing the invariance loss $I_{\bar{S}}(f_\theta)$ while controlling the covariance matrix $Z_{\bar{S}}Z_{\bar{S}}\mathcal{T}$. In addition, the term $\frac{2\tilde{\mathcal{R}}_m(\mathcal{F})}{\sqrt{m}}$ 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, the proof of Theorem 2 in Shwartz-Ziv (2022) shows that we can approximate $\ln|\mathcal{F}|$ by $2^{I(Z;X)}$ where $I(Z; X)$ is what we want to minimize in the information bottleneck principle (Tishby et al., 2000).

Thus, in Theorem 2, the term of $I_{\bar{S}}(f_\theta) + \frac{2}{\sqrt{m}}\|\mathbf{P}_{Z_{\bar{S}}}Y_{\bar{S}}\|_F + \frac{1}{\sqrt{n}}\|\mathbf{P}_{Z_S}Y_S\|_F$ corresponds to $I(Z;X')$ while the term of $\frac{2\tilde{\mathcal{R}}_m(\mathcal{F})}{\sqrt{m}}$ corresponds to $I(Z;X)$. To understand why it makes sense to have these two terms $I(Z;X')$ and $I(Z;X)$, recall that the mutual information $I(Z;X)$ can be decomposed as

$$I(Z;X) = I(Z;X') + I(Z;X|X'). \tag{9}$$

Here, from the multi-view information bottleneck principle (Federici et al., 2019), we want to maximize the predictive information $I(Z;X')$, while minimizing the information bottleneck $I(Z;X)$. Thus, in order to improve generalization, we also need to control $\frac{2\tilde{\mathcal{R}}_m(\mathcal{F})}{\sqrt{m}}$ to restrict the superfluous information $I(Z;X|X')$, in addition to minimize $I_{\bar{S}}(f_\theta) + \frac{2}{\sqrt{m}}\|\mathbf{P}_{Z_{\bar{S}}}Y_{\bar{S}}\|_F + \frac{1}{\sqrt{n}}\|\mathbf{P}_{Z_S}Y_S\|_F$ that corresponded to maximize the predictive information $I(Z;X')$. Although we can explicitly add some regularization on $I(Z;X|X')$ to control $\frac{2\tilde{\mathcal{R}}_m(\mathcal{F})}{\sqrt{m}}$, it is possible that $I(Z;X|X')$ and $\frac{2\tilde{\mathcal{R}}_m(\mathcal{F})}{\sqrt{m}}$ are implicitly regularized via implicit bias through the architecture design and a choice of a learning algorithm (Gunasekar et al., 2017; Soudry et al., 2018; Gunasekar et al., 2018; Woodworth et al., 2020; Moroshko et al., 2020; Kawaguchi, 2021). Thus, Theorem 2 connects the information-theoretic understanding of SSL with the probabilistic guarantee on the generalization ability for the downstream task.

6.4 Comparing Generalization Bounds

The generalization bound of SimCLR (Saunshi et al., 2019) requires the number of label classes to approach infinity to make the generalization gap decrease towards zero. This is because the bound includes the additive term of $\mathbb{P}(y^+ = y^-)/(1 - \mathbb{P}(y^+ = y^-))$ where $\mathbb{P}(y^+ = y^-)$ is the probability that the labels of a negative pair (y^+, y^-) are the same. This term goes to zero as the number of label classes to approach infinity since $\mathbb{P}(y^+ = y^-) \to 0$ as the number approach infinity. In contrast, the bound on VICReg in Theorem 2 does *not* require the number of label classes to approach infinity to let the generalization gap 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 bound of VICReg improves as n increases, which is consistent with practice, while the previous bound of SimCLR (Saunshi et al., 2019) does not depend on n. This is because the previous work assumes partial access to the true distribution of x given y per class for setting W, which removes the importance of labeled data size n and is not assumed in our study. Consequently, our bound 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 2 also illuminates the advantages of VICReg over the 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. Here, $\tilde{\mathcal{R}}_n(\mathcal{W})$ is the the normalized Rademacher complexity for the hypothesis space of w. Thus, Theorem 2 shows that using SSL, we can replace all the complexities of the hypothesis spaces in terms of n with those in terms of m. Since m is typically much larger than n, this illuminates the benefit of SSL.

7 Conclusions

In this study, we examine SSL's objective function from an information-theoretic perspective. Based on transfering of the required stochasticity to the input distribution, we show how SSL objectives can be derived. Thus, even when using deterministic DNs, it is possible to perform an information-theoretic analysis. The second part of the paper rediscovered SSL loss functions from first principles and demonstrated their implicit assumptions. We empirically validated our analysis and confirmed the validity of our novel understanding. Furthermore, we derived a generalization bound on the downstream task, tight it to known information objeective terms and demonstrate that VICReg minimizes it. In addition, our work opens many new avenues for future research, including a better estimation of information-theoretic quantities that are consistent with our assumptions and identifying which SSL method is the most appropriate according to data characteristics. In addition, our probabilistic guarantee suggests that VICReg can be further improved for the setting of partial label information by aligning the covariance matrix with the partially observable label matrix.

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.

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.

Sanjeev Arora, Hrishikesh Khandeparkar, Mikhail Khodak, Orestis Plevrakis, and Nikunj Saunshi. A theoretical analysis of contrastive unsupervised representation learning. *arXiv preprint arXiv:1902.09229*, 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, pp. 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.

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.

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

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

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

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.

Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1): 1–22, 1977.

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.

Marco Federici, Anjan Dutta, Patrick Forré, Nate Kushman, and Zeynep Akata. Learning robust representations via multi-view information bottleneck. *arXiv preprint arXiv:2002.07017*, 2020.

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*, pp. 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*, pp. 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*, pp. 9461–9471, 2018.

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*, pp. 181 – 188, 09 2008. doi: 10.1109/MFI.2008.4648062.

Kenji Kawaguchi. On the theory of implicit deep learning: Global convergence with implicit layers. In *International Conference on Learning Representations (ICLR)*, 2021.

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.

Jason D Lee, Qi Lei, Nikunj Saunshi, and Jiacheng Zhuo. Predicting what you already know helps: Provable self-supervised learning. *Advances in Neural Information Processing Systems*, 34, 2021a.

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

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

Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6707–6717, 2020.

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*, pp. 2924–2932, 2014.

Edward Moroshko, Suriya Gunasekar, Blake Woodworth, Jason D Lee, Nathan Srebro, and Daniel Soudry. Implicit bias in deep linear classification: Initialization scale vs training accuracy. *arXiv preprint arXiv:2007.06738*, 2020.

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

Radford M Neal and Geoffrey E Hinton. A view of the em algorithm that justifies incremental, sparse, and other variants. In *Learning in graphical models*, pp. 355–368. Springer, 1998.

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*, pp. 5628–5637. PMLR, 2019.

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

Tao Shi, Mikhail Belkin, and Bin Yu. Data spectroscopy: Eigenspaces of convolution operators and clustering. *The Annals of Statistics*, pp. 3960–3984, 2009.

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*, pp. 1–17. PMLR, 2020.

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

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

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*, pp. 3437–3452. PMLR, 2020.

Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

Blake Woodworth, Suriya Gunasekar, Jason D Lee, Edward Moroshko, Pedro Savarese, Itay Golan, Daniel Soudry, and Nathan Srebro. Kernel and rich regimes in overparametrized models. *arXiv preprint arXiv:2002.09277*, 2020.

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

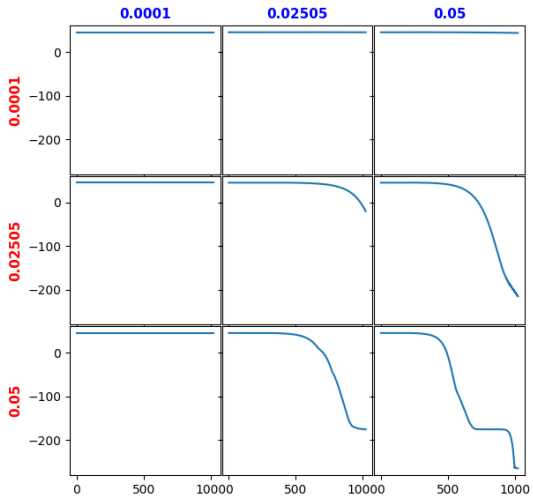

Figure 2: Evolution of the entropy for each of the learning rate configurations showing that the impact of picking the incorrect learning rate for the data and/or centroids lead to a collapse of the samples.

A DATA DISTRIBUTION AFTER DEEP NETWORK TRANSFORMATION

Theorem 3. *Given the setting of eq.* (4) *the unconditional DN output density denoted as Z is a mixture of the affinely transformed distributions $\boldsymbol{x}|\boldsymbol{x}^*_{n(\boldsymbol{x})}$ e.g. for the Gaussian case*

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

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. □

B EXPECTATION MAXIMIZATION AND COLLAPSING

C ON BENEFITS OF INFORMATION MAXIMIZATION FOR GENERALIZATION

In this Appendix, we present the complete version of Theorem 2 along with its proof and additional discussions.

C.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 $W_{\bar{S}}$ of the unlabeled data as $W_{\bar{S}} = \text{minimize}_{W'} \|W'\|_F$ 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 all $(x, y) \in \mathcal{X} \times \mathcal{Y}$. Similarly, let $\kappa_{\bar{S}}$ be a data-dependent upper bound on the per-sample Euclidian norm loss as $\|W_{\bar{S}} f_\theta(x) - y\| \leq \kappa_{\bar{S}}$ for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$. Define the

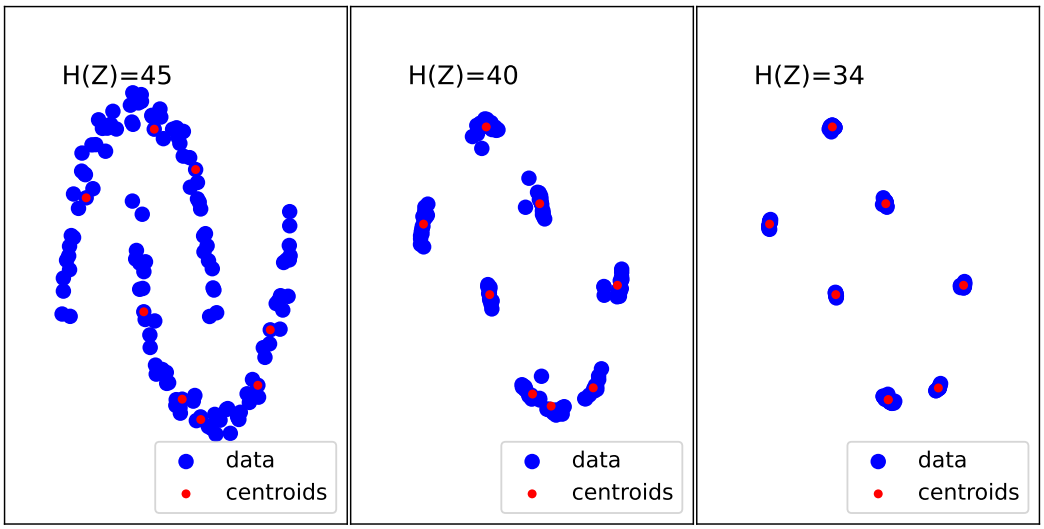

Figure 3: **Evolution of GMM training when enforcing a one-to-one mapping between the data and centroids akin to K-means i.e. using a small and fixed covariance matrix. We see that collapse does not occur.** Left - In the presence of fixed input samples, we observe that there is no collapsing and that 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.

difference between W_S and $W_{\bar{S}}$ by $c = \|W_S - W_{\bar{S}}\|_2$. Let $\mathcal{W}$ be a hypothesis space of W such that $W_{\bar{S}} \in \mathcal{W}$. We denote by $\tilde{\mathcal{R}}_m(\mathcal{W} \circ \mathcal{F}) = \frac{1}{\sqrt{m}}\mathbb{E}_{\bar{S},\xi}[\sup_{W\in\mathcal{W},f\in\mathcal{F}} \sum_{i=1}^{m} \xi_i \|g^*(x_i^+) - Wf(x_i^+)\|]$ the normalized Rademacher complexity of the set $\{x^+ \mapsto \|g^*(x^+) - Wf(x^+)\| : W \in \mathcal{W}, f \in \mathcal{F}\}$. 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). For the labeled data, first, y is drawn from the distribuion 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 distribuion 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 $\tau_{\bar{S}}$ be a data-dependent upper bound on the invariance loss with the trained representation as $\|f_\theta(\bar{x}) - f_\theta(x)\| \leq \tau_{\bar{S}}$ 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.

C.2 Result

The following theorem is the complete version of Theorem 2:

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

$$\mathbb{E}_{x,y}[\ell_{x,y}(w_S)] \leq cI_{\bar{S}}(f_\theta) + \frac{2}{\sqrt{m}}\|\mathbf{P}_{Z_{\bar{S}}} Y_{\bar{S}}\|_F + \frac{1}{\sqrt{n}}\|\mathbf{P}_{Z_S} Y_S\|_F + Q_{m,n}, \tag{10}$$

where

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

Proof. The complete proof is presented in Appendix C.3. □

The bound in the complete version of Theorem 1 is better than the one in the informal version of Theorem 2, 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 $W_{\bar{S}}$ 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 2 by the following reasoning to simplify the notation in the main text. We have that $cI_{\bar{S}}(f_\theta) + c\frac{2\tilde{\mathcal{R}}_m(\mathcal{F})}{\sqrt{m}} = I_{\bar{S}}(f_\theta) + \frac{2\tilde{\mathcal{R}}_m(\mathcal{F})}{\sqrt{m}} + Q$, where $Q = (c-1)(I_{\bar{S}}(f_\theta) + \frac{2\tilde{\mathcal{R}}_m(\mathcal{F})}{\sqrt{m}}) \leq \varsigma \to 0$ 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 2 is more accurate and indeed tighter than the one in the informal version.

In Theorem 2, $Q_{m,n} \to 0$ 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 & Mendelson, 2002; Mohri et al., 2012; Shalev-Shwartz & Ben-David, 2014).

C.3 Proof of Theorem 2

Proof of Theorem 2. Let $W = W_S$ where W_S is the the minimum norm solution as $W_S = \text{minimize}_{W'} \|W'\|_F$ s.t. $W' \in \arg\min_W \frac{1}{n} \sum_{i=1}^n \|W f_\theta(x_i) - y_i\|^2$. Let $W^* = W_{\bar{S}}$ where $W_{\bar{S}}$ is the minimum norm solution as $W^* = W_{\bar{S}} = \text{minimize}_{W'} \|W'\|_F$ s.t. $W' \in \arg\min_W \frac{1}{m} \sum_{i=1}^m \|W f_\theta(x_i^+) - g^*(x_i^+)\|^2$. 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 $L_S(w) = \frac{1}{n} \sum_{i=1}^n \|W f_\theta(x_i) - y_i\|$. Using these,

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

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 $i = 1, \ldots, n$),

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

Combining these, we have that

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} \|W f_\theta(\bar{x}_i) - \bar{y}_i\| \leq L_S(w) &+ \frac{1}{n}\sum_{i=1}^{n} \|\tilde{W}(f_\theta(\bar{x}_i) - f_\theta(x_i))\| \\
&+ \frac{1}{n}\sum_{i=1}^{n} \|\varphi(x_i)\| + \frac{1}{n}\sum_{i=1}^{n} \|\varphi(\bar{x}_i)\|.
\end{aligned} \tag{11}$$

To bound the left-hand side of equation 11, we now analyze the following random variable:

$$\mathbb{E}_{X,Y}[\|W_S f_\theta(X) - Y\|] - \frac{1}{n}\sum_{i=1}^{n} \|W_S f_\theta(\bar{x}_i) - \bar{y}_i\|, \tag{12}$$

where $\bar{y}_i = y_i$ since $\bar{x}_i \sim \mathcal{D}_{y_i}$ for $i = 1, \ldots, n$. Importantly, this means that as W_S depends on y_i, W_S depends on $\bar{y}_i$. Thus, the collection of random variables $\|W_S f_\theta(\bar{x}_1) - \bar{y}_1\|, \ldots, \|W_S f_\theta(n_n) - \bar{y}_n\|$ is *not* independent. Accordingly, we cannot apply standard concentration inequality to bound equation 12. A standard approach in learning theory is to first bound equation 12 by $\mathbb{E}_{x,y}\|W_S f_\theta(x) - y\| - \frac{1}{n}\sum_{i=1}^{n} \|W_S f_\theta(\bar{x}_i) - \bar{y}_i\| \leq \sup_{W \in \mathcal{W}} \mathbb{E}_{x,y}\|W f_\theta(x) - y\| - \frac{1}{n}\sum_{i=1}^{n} \|W f_\theta(\bar{x}_i) - \bar{y}_i\|$ 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 $\|W f_\theta(\bar{x}_1) - \bar{y}_1\|, \ldots, \|W f_\theta(n_n) - \bar{y}_n\|$, 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 $\|W_S f_\theta(\bar{x}_1) - \bar{y}_1\|, \ldots, \|W_S f_\theta(n_n) - \bar{y}_n\|$. 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 equation 12 into two terms:

$$\begin{aligned}&\mathbb{E}_{X,Y}[\|W_S f_\theta(X) - Y\|] - \frac{1}{n}\sum_{i=1}^{n} \|W_S f_\theta(\bar{x}_i) - \bar{y}_i\| \\ &= \left(\sum_{y\in\mathcal{Y}} \mathbb{E}_{X_y}[\|W_S f_\theta(X_y) - y\|]\frac{|\mathcal{I}_y|}{n} - \frac{1}{n}\sum_{i=1}^{n} \|W_S f_\theta(\bar{x}_i) - \bar{y}_i\|\right) \\ &\quad + \sum_{y\in\mathcal{Y}} \mathbb{E}_{X_y}[\|W_S f_\theta(X_y) - y\|]\left(\mathbb{P}(Y = y) - \frac{|\mathcal{I}_y|}{n}\right),\end{aligned} \tag{13}$$

where

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

The first term in the right-hand side of equation 13 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 equation 13 yields

$$\begin{aligned}&\mathbb{E}_{X,Y}[\|W_S f_\theta(X) - Y\|] - \frac{1}{n}\sum_{i=1}^{n} \|W_S f_\theta(\bar{x}_i) - \bar{y}_i\| \\ &= \frac{1}{n}\sum_{y\in\tilde{\mathcal{Y}}} |\mathcal{I}_y| \left(\mathbb{E}_{X_y}[\|W_S f_\theta(X_y) - y\|] - \frac{1}{|\mathcal{I}_y|}\sum_{i\in\mathcal{I}_y} \|W_S f_\theta(\bar{x}_i) - y\|\right) \\ &\quad + \sum_{y\in\mathcal{Y}} \mathbb{E}_{X_y}[\|W_S f_\theta(X_y) - y\|]\left(\mathbb{P}(Y = y) - \frac{|\mathcal{I}_y|}{n}\right)\end{aligned} \tag{14}$$

Importantly, while $\|W_S f_\theta(\bar{x}_1) - \bar{y}_1\|, \ldots, \|W_S f_\theta(\bar{x}_n) - \bar{y}_n\|$ on the right-hand side of equation 14 are dependent random variables, $\|W_S f_\theta(\bar{x}_1) - y\|, \ldots, \|W_S f_\theta(\bar{x}_n) - y\|$ 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 $y \in \tilde{\mathcal{Y}}$:

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

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

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

Substituting this bound into equation 14, we have that with probability at least $1-\delta$,

$$\begin{aligned}
&\mathbb{E}_{X,Y}[\|W_S f_\theta(X)-Y\|]-\frac{1}{n}\sum_{i=1}^{n}\|W_S f_\theta(\bar{x}_i)-\bar{y}_i\| \qquad (15)\\
&\leq \kappa_S\left(\sum_{y\in\tilde{\mathcal{Y}}}\sqrt{\hat{p}(y)}\right)\sqrt{\frac{\ln(|\tilde{\mathcal{Y}}|/\delta)}{2n}}+\sum_{y\in\mathcal{Y}}\mathbb{E}_{X_y}[\|W_S f_\theta(X_y)-y\|]\left(\mathbb{P}(Y=y)-\frac{|\mathcal{I}_y|}{n}\right)
\end{aligned}$$

where

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

Moreover, for the second term on the right-hand side of equation 15, by using Lemma 1 of (Kawaguchi et al., 2022), we have that with probability at least $1-\delta$,

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

where $p(y)=\mathbb{P}(Y=y)$. Substituting this bound into equation 15 with the union bound, we have that with probability at least $1-\delta$,

$$\begin{aligned}
&\mathbb{E}_{X,Y}[\|W_S f_\theta(X)-Y\|]-\frac{1}{n}\sum_{i=1}^{n}\|W_S f_\theta(\bar{x}_i)-\bar{y}_i\| \qquad (16)\\
&\leq\kappa_S\left(\sum_{y\in\tilde{\mathcal{Y}}}\sqrt{\hat{p}(y)}\right)\sqrt{\frac{\ln(2|\tilde{\mathcal{Y}}|/\delta)}{2n}}+\kappa_S\left(\sum_{y\in\mathcal{Y}}\sqrt{p(y)}\right)\sqrt{\frac{2\ln(2|\mathcal{Y}|/\delta)}{2n}}\\
&\leq\left(\sum_{y\in\mathcal{Y}}\sqrt{\hat{p}(y)}\right)\kappa_S\sqrt{\frac{2\ln(2|\mathcal{Y}|/\delta)}{2n}}+\left(\sum_{y\in\mathcal{Y}}\sqrt{p(y)}\right)\kappa_S\sqrt{\frac{2\ln(2|\mathcal{Y}|/\delta)}{2n}}\\
&\leq\kappa_S\sqrt{\frac{2\ln(2|\mathcal{Y}|/\delta)}{2n}}\sum_{y\in\mathcal{Y}}\left(\sqrt{\hat{p}(y)}+\sqrt{p(y)}\right)
\end{aligned}$$

Combining equation 11 and equation 16 implies that with probability at least $1-\delta$,

$$\begin{aligned}
&\mathbb{E}_{X,Y}[\|W_S f_\theta(X) - Y\|] \\
&\leq \frac{1}{n}\sum_{i=1}^{n}\|W_S f_\theta(\bar{x}_i) - \bar{y}_i\| + \kappa_S\sqrt{\frac{2\ln(2|\mathcal{Y}|/\delta)}{2n}}\sum_{y\in\mathcal{Y}}\left(\sqrt{\hat{p}(y)} + \sqrt{p(y)}\right) \\
&\leq L_S(w_S) + \frac{1}{n}\sum_{i=1}^{n}\|\tilde{W}(f_\theta(\bar{x}_i) - f_\theta(x_i))\| \\
&\quad + \frac{1}{n}\sum_{i=1}^{n}\|\varphi(x_i)\| + \frac{1}{n}\sum_{i=1}^{n}\|\varphi(\bar{x}_i)\| + \kappa_S\sqrt{\frac{2\ln(2|\mathcal{Y}|/\delta)}{2n}}\sum_{y\in\mathcal{Y}}\left(\sqrt{\hat{p}(y)} + \sqrt{p(y)}\right).
\end{aligned} \tag{17}$$

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 equation 17. Since $W^* = W_{\bar{S}}$,

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

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

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

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 $1-\delta$,

$$\mathbb{E}_{x^+}[\|g^*(x^+) - W_{\bar{S}} f_\theta(x^+)\|] \leq \frac{1}{m}\sum_{i=1}^{m}\|g^*(x_i^+) - W_{\bar{S}} f_\theta(x_i^+)\| + \frac{2\tilde{\mathcal{R}}_m(\mathcal{W}\circ\mathcal{F})}{\sqrt{m}} + \kappa\sqrt{\frac{\ln(1/\delta)}{2m}} \tag{18}$$

where $\tilde{\mathcal{R}}_m(\mathcal{W}\circ\mathcal{F}) = \frac{1}{\sqrt{m}}\mathbb{E}_{\bar{S},\xi}[\sup_{W\in\mathcal{W}, f\in\mathcal{F}}\sum_{i=1}^{m}\xi_i\|g^*(x_i^+) - Wf(x_i^+)\|]$ 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 $\xi_1, \ldots, \xi_m$ are independent uniform random variables taking values in $\{-1, 1\}$. Takinng union bounds, we have that for any $\delta > 0$, with probability at least $1-\delta$,

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

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

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

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

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\|\varphi(x_i)\| + \frac{1}{n}\sum_{i=1}^{n}\|\varphi(\bar{x}_i)\| \\
&\leq \frac{2}{m}\sum_{i=1}^{m}\|g^*(x_i^+) - W_{\bar{S}} f_\theta(x_i^+)\| + \frac{4\mathcal{R}_m(\mathcal{W}\circ\mathcal{F})}{\sqrt{m}} + 2\kappa\sqrt{\frac{\ln(4/\delta)}{2m}} + 2\kappa_{\bar{S}}\sqrt{\frac{\ln(4/\delta)}{2n}}
\end{aligned} \tag{19}$$

To analyze the first term on the right-hand side of equation 19, recall that

$$W_{\bar{S}} = \underset{W'}{\text{minimize}}\,\|W'\|_F \text{ s.t. } W' \in \arg\min_{W}\frac{1}{m}\sum_{i=1}^{m}\|Wf_\theta(x_i^+) - g^*(x_i^+)\|^2. \tag{20}$$

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

$$Wf_\theta(x_i^+) = \mathrm{vec}[Wf_\theta(x_i^+)] = [f_\theta(x_i^+)\mathcal{T} \otimes I_r]\,\mathrm{vec}[W] \in \mathbb{R}^r,$$

where $I_r \in \mathbb{R}^{r\times r}$ is the identity matrix, and $[f_\theta(x_i^+)\mathcal{T} \otimes I_r] \in \mathbb{R}^{r\times dr}$ is the Kronecker product of the two matrices, and $\mathrm{vec}[W] \in \mathbb{R}^{dr}$ is the vectorization of the matrix $W \in \mathbb{R}^{r\times d}$. Thus, by defining $A_i = [f_\theta(x_i^+)\mathcal{T} \otimes I_r] \in \mathbb{R}^{r\times dr}$ and using the notation of $w = \mathrm{vec}[W]$ and its inverse $W = \mathrm{vec}^{-1}[w]$ (i.e., the inverse of the vectorization from $\mathbb{R}^{r\times d}$ to $\mathbb{R}^{dr}$ with a fixed ordering), we can rewrite equation 20 by

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

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

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

This implies that

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

In other words,

$$A\mathcal{T}Aw = A\mathcal{T}g \quad \text{where } A = \begin{bmatrix} A_1 \\ A_2 \\ \vdots \\ A_m \end{bmatrix} \in \mathbb{R}^{mr\times dr} \text{ and } g = \begin{bmatrix} g_1 \\ g_2 \\ \vdots \\ g_m \end{bmatrix} \in \mathbb{R}^{mr}$$

Thus,

$$w' \in \arg\min_w \sum_{i=1}^m \|g_i - A_i w\|^2 = \{(A\mathcal{T}A)^\dagger A\mathcal{T}g + v : v \in \mathrm{Null}(A)\}$$

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

$$\mathrm{vec}[W_{\bar{S}}] = w_{\bar{S}} = (A\mathcal{T}A)^\dagger A\mathcal{T}g.$$

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

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

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

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n \|\varphi(x_i)\| + \frac{1}{n}\sum_{i=1}^n \|\varphi(\bar{x}_i)\| \qquad (21) \\
&\le \frac{2}{\sqrt{m}}\|(I - A(A\mathcal{T}A)^\dagger A\mathcal{T})g\|_2 + \frac{4\mathcal{R}_m(\mathcal{W}\circ\mathcal{F})}{\sqrt{m}} + 2\kappa\sqrt{\frac{\ln(4/\delta)}{2m}} + 2\kappa_{\bar{S}}\sqrt{\frac{\ln(4/\delta)}{2n}}
\end{aligned}$$

By combining equation 17 and equation 21 with union bound, we have that

$$\begin{aligned}
&\mathbb{E}_{X,Y}[\|W_S f_\theta(X) - Y\|] \qquad (22)\\
&\leq L_S(w_S) + \frac{1}{n}\sum_{i=1}^{n}\|\tilde{W}(f_\theta(\bar{x}_i) - f_\theta(x_i))\| + \frac{2}{\sqrt{m}}\|\mathbf{P}_A g\|_2\\
&\quad + \frac{4\mathcal{R}_m(\mathcal{W}\circ\mathcal{F})}{\sqrt{m}} + 2\kappa\sqrt{\frac{\ln(8/\delta)}{2m}} + 2\kappa_{\bar{S}}\sqrt{\frac{\ln(8/\delta)}{2n}}\\
&\quad + \kappa_S\sqrt{\frac{2\ln(4|\mathcal{Y}|/\delta)}{2n}}\sum_{y\in\mathcal{Y}}\left(\sqrt{\hat{p}(y)} + \sqrt{p(y)}\right).
\end{aligned}$$

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

We will now analyze the second term on the right-hand side of equation 22:

$$\frac{1}{n}\sum_{i=1}^{n}\|\tilde{W}(f_\theta(\bar{x}_i) - f_\theta(x_i))\| \leq \|\tilde{W}\|_2\left(\frac{1}{n}\sum_{i=1}^{n}\|f_\theta(\bar{x}_i) - f_\theta(x_i)\|\right), \qquad (23)$$

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 $\bar{S}$, using Hoeffding's inequality (Lemma 1) implies that with probability at least $1-\delta$,

$$\frac{1}{n}\sum_{i=1}^{n}\|f_\theta(\bar{x}_i) - f_\theta(x_i)\| \leq \mathbb{E}_{y\sim\rho}\mathbb{E}_{\bar{x},x\sim\mathcal{D}_y^2}[\|f_\theta(\bar{x}) - f_\theta(x)\|] + \tau_{\bar{S}}\sqrt{\frac{\ln(1/\delta)}{2n}}. \qquad (24)$$

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 $1-\delta$,

$$\mathbb{E}_{y\sim\rho}\mathbb{E}_{\bar{x},x\sim\mathcal{D}_y^2}[\|f_\theta(\bar{x}) - f_\theta(x)\|] \leq \frac{1}{m}\sum_{i=1}^{m}\|f_\theta(x_i^+) - f_\theta(x_i^{++})\| + \frac{2\tilde{\mathcal{R}}_m(\mathcal{F})}{\sqrt{m}} + \tau\sqrt{\frac{\ln(1/\delta)}{2m}} \qquad (25)$$

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

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\|\tilde{W}(f_\theta(\bar{x}_i) - f_\theta(x_i))\| \qquad (26)\\
&\leq \|\tilde{W}\|_2\left(\frac{1}{m}\sum_{i=1}^{m}\|f_\theta(x_i^+) - f_\theta(x_i^{++})\| + \frac{2\tilde{\mathcal{R}}_m(\mathcal{F})}{\sqrt{m}} + \tau\sqrt{\frac{\ln(2/\delta)}{2m}} + +\tau_{\bar{S}}\sqrt{\frac{\ln(2/\delta)}{2n}}\right).
\end{aligned}$$

By combining equation 22 and equation 26 using the union bound, we have that with probability at least $1-\delta$,

$$\begin{aligned}
&\mathbb{E}_{X,Y}[\|W_S f_\theta(X) - Y\|] \qquad (27)\\
&\leq L_S(w_S) + \|\tilde{W}\|_2\left(\frac{1}{m}\sum_{i=1}^{m}\|f_\theta(x_i^+) - f_\theta(x_i^{++})\| + \frac{2\tilde{\mathcal{R}}_m(\mathcal{F})}{\sqrt{m}} + \tau\sqrt{\frac{\ln(4/\delta)}{2m}} + \tau_{\bar{S}}\sqrt{\frac{\ln(4/\delta)}{2n}}\right)\\
&\quad + \frac{2}{\sqrt{m}}\|\mathbf{P}_A g\|_2 + \frac{4\mathcal{R}_m(\mathcal{W}\circ\mathcal{F})}{\sqrt{m}} + 2\kappa\sqrt{\frac{\ln(16/\delta)}{2m}} + 2\kappa_{\bar{S}}\sqrt{\frac{\ln(16/\delta)}{2n}}\\
&\quad + \kappa_S\sqrt{\frac{2\ln(8|\mathcal{Y}|/\delta)}{2n}}\sum_{y\in\mathcal{Y}}\left(\sqrt{\hat{p}(y)} + \sqrt{p(y)}\right)\\
&= L_S(w_S) + \|\tilde{W}\|_2\left(\frac{1}{m}\sum_{i=1}^{m}\|f_\theta(x_i^+) - f_\theta(x_i^{++})\|\right) + \frac{2}{\sqrt{m}}\|\mathbf{P}_A g\|_2 + Q_{m,n}
\end{aligned}$$

where

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

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

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

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

$$\|\mathbf{P}_A g\|_2 = \|[\mathbf{P}_{Z_{\bar{S}}} \otimes I_r]\,\mathrm{vec}[Y_{\bar{S}}\mathcal{T}]\|_2 = \|\,\mathrm{vec}[Y_{\bar{S}}\mathcal{T}\mathbf{P}_{Z_{\bar{S}}}]\|_2 = \|\mathbf{P}_{Z_{\bar{S}}}Y_{\bar{S}}\|_F \tag{28}$$

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

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

By solving this, we have

$$W_S = Y\mathcal{T}Z_S\mathcal{T}(Z_SZ_S\mathcal{T})^\dagger,$$

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

$$\begin{aligned} L_S(w_S) = \frac{1}{n}\sum_{i=1}^n \|W_Sf_\theta(x_i) - y_i\| &= \frac{1}{n}\sum_{i=1}^n \sqrt{\sum_{k=1}^r ((W_Sf_\theta(x_i) - y_i)_k)^2} \\ &\le \sqrt{\frac{1}{n}\sum_{i=1}^n\sum_{k=1}^r ((W_Sf_\theta(x_i) - y_i)_k)^2} \\ &= \frac{1}{\sqrt{n}}\|W_SZ_S - Y\mathcal{T}\|_F \\ &= \frac{1}{\sqrt{n}}\|Y\mathcal{T}(Z_S\mathcal{T}(Z_SZ_S\mathcal{T})^\dagger Z_S - I)\|_F \\ &= \frac{1}{\sqrt{n}}\|(I - Z_S\mathcal{T}(Z_SZ_S\mathcal{T})^\dagger Z_S)Y\|_F \end{aligned}$$

Thus,

$$L_S(w_S) = \frac{1}{\sqrt{n}}\|\mathbf{P}_{Z_S}Y\|_F \tag{29}$$

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

By combining equation 27–equation 29 and using $1 \le \sqrt{2}$, we have that with probability at least $1-\delta$,

$$\mathbb{E}_{X,Y}[\|W_Sf_\theta(X) - Y\|] \le cI_{\bar{S}}(f_\theta) + \frac{2}{\sqrt{m}}\|\mathbf{P}_{Z_{\bar{S}}}Y_{\bar{S}}\|_F + \frac{1}{\sqrt{n}}\|\mathbf{P}_{Z_S}Y_S\|_F + Q_{m,n}, \tag{30}$$

where

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

□

D KNOWN LEMMAS

We use the following well-known theorems as lemmas in our proof. We put these below for the 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 & Mendelson, 2002; Mohri et al., 2012; Shalev-Shwartz & 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 $\psi \in \mathcal{G}$:*

$$\mathbb{E}_q[\psi(q)] \leq \frac{1}{m}\sum_{i=1}^m \psi(q_i) + 2\mathcal{R}_m(\mathcal{G}) + M\sqrt{\frac{\ln(1/\delta)}{2m}}, \tag{31}$$

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

$$\varphi(S) = \sup_{\psi\in\mathcal{G}} \mathbb{E}_{x,y}[\psi(q)] - \frac{1}{m}\sum_{i=1}^m \psi(q_i). \tag{32}$$

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,

$$\varphi(S') - \varphi(S) \leq \sup_{\psi \in \mathcal{G}} \frac{\psi(q_{i_0}) - \psi(q'_{i_0})}{m} \leq \frac{M}{m}. \tag{33}$$

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

$$\varphi(S) \leq \mathbb{E}_S[\varphi(S)] + M\sqrt{\frac{\ln(1/\delta)}{2m}}. \tag{34}$$

Moreover,

$$\begin{aligned}
&\mathbb{E}_S[\varphi(S)] && (35)\\
&= \mathbb{E}_S\left[\sup_{\psi \in \mathcal{G}} \mathbb{E}_{S'}\left[\frac{1}{m}\sum_{i=1}^{m}\psi(q'_i)\right] - \frac{1}{m}\sum_{i=1}^{m}\psi(q_i)\right] && (36)\\
&\leq \mathbb{E}_{S,S'}\left[\sup_{\psi \in \mathcal{G}} \frac{1}{m}\sum_{i=1}^{m}(\psi(q'_i) - \psi(q_i))\right] && (37)\\
&\leq \mathbb{E}_{\xi,S,S'}\left[\sup_{\psi \in \mathcal{G}} \frac{1}{m}\sum_{i=1}^{m}\xi_i(\psi(q'_i) - \psi(q_i))\right] && (38)\\
&\leq 2\mathbb{E}_{\xi,S}\left[\sup_{\psi \in \mathcal{G}} \frac{1}{m}\sum_{i=1}^{m}\xi_i\psi(q_i)\right] = 2\mathcal{R}_m(\mathcal{G}) && (39)
\end{aligned}$$

where the fist line follows the definitions of each term, the second line uses the 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 forth line uses the subadditivity of supremum. □