Functional Regularization for Representation Learning: A Unified Theoretical Perspective

Siddhant Garg*
Amazon Alexa AI Search
Manhattan Beach, CA, USA
sidgarg@amazon.com

Yingyu Liang
Department of Computer Sciences
University of Wisconsin-Madison
yliang@cs.wisc.edu

Abstract

Unsupervised and self-supervised learning approaches have become a crucial tool to learn representations for downstream prediction tasks. We present a unifying theoretical perspective where several such approaches can be viewed as imposing a regularization on the representation via a *learnable* function using unlabeled data. We propose a discriminative framework, inspired from [3], for analyzing the sample complexity of these approaches. Our sample complexity bounds show that, with carefully chosen hypothesis classes to exploit the structure in the data, such functional regularization can prune the hypothesis space and help reduce the labeled data needed. We provide two concrete examples of functional regularization, using auto-encoders and masked self-supervision, and apply the framework to quantify the reduction in the sample complexity bound. We also provide complementary empirical results to support our analysis on synthetic and real data.

1 Introduction

Advancements in machine learning have resulted in large prediction models, which need large amounts of labeled data for effective learning. Expensive label annotation costs have increased the popularity of self-supervised representation learning techniques using additional unlabeled data. These techniques learn a representation function on the input, and a prediction function over the representation for the target prediction task. Unlabeled data is utilised by posing an auxiliary learning task on the representation, e.g., using the representation to reconstruct the input. Empirical results in several recent works show that learning representations using unlabeled data can drastically reduce the size of labeled data needed for the prediction task. In contrast to the popularity and impressive practical gains of these representation learning approaches, there have been far fewer theoretical studies towards understanding them, most of which have been specific to individual approaches. Theoretically, there is still ambiguity over questions like *"When can the auxiliary task over the unlabeled data help? How much can it reduce the sample size of the labeled data by?"*

In this work, we take a step to improve the theoretical understanding of the benefits of learning representations for the target prediction task via an auxiliary task. We focus on analyzing the sample complexity of labeled and unlabeled data for this representation learning paradigm. Our contribution is to propose a unified perspective where several representation learning approaches can be viewed as if they impose a regularization on the representation via a learnable regularization function. Under this paradigm, representations are learned jointly on unlabeled and labeled data. The former is used in the auxiliary task to jointly learn the representation and the regularization function. The latter is used in the target prediction task to learn the representation and the prediction function. Henceforth, we refer to this paradigm as *representation learning via functional regularization*.

In particular, we present a PAC-style discriminative framework [53] to bound the sample complexities of labeled and unlabeled data under different assumptions on the models and data distributions. This

* Work completed at the University of Wisconsin-Madison

NeurIPS 2020 Workshop: Self-Supervised Learning - Theory and Practice, Vancouver, Canada.

is inspired by [3] which analyzes semi-supervised learning and shows that unlabeled data can act as regularization. Our generalized framework allows *learnable* regularization functions and thus unifies multiple representation learning approaches for prediction. It shows that functional regularization using unlabeled data can prune the model hypothesis class for learning representations, leading to a reduction of the labeled data required for the prediction task.

To demonstrate the application of our framework, we construct two concrete examples of functional regularization, one using auto-encoder and the other using masked self-supervision. These specific functional regularization settings allow us to quantify the reduction in the sample bounds of labeled data more explicitly. We also provide complementary empirical support to our theoretical results through experiments on synthetic and real data in Appendix E-F.

2 Problem Formulation

Consider labeled data $S = \{(x_i, y_i)\}_{i=1}^{m_\ell}$ from a distribution $\mathcal{D}$ over the domains $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{X} \subseteq \mathbb{R}^d$ is the input feature space and $\mathcal{Y}$ is the label space. The goal is to learn a predictor $p : \mathcal{X} \to \mathcal{Y}$ that fits $\mathcal{D}$. This can be achieved by first learning a representation function $\phi = h(x) \in \mathbb{R}^r$ over the input and then learning a predictor $y{=}f(\phi){\in}\mathcal{Y}$ on the representation. Denote the hypothesis classes for h and f by $\mathcal{H}$ and $\mathcal{F}$ respectively, and the loss function by $\ell_c(f(h(x)), y)$. Without loss of generality, assume $\ell_c \in [0, 1]$. We are interested in representation learning approaches where $h(x)$ is learned with the help of an auxiliary task on unlabeled data $U = \{\tilde{x}_i\}_{i=1}^{m_u}$ from a distribution $\mathcal{U}_X$ (same or different from the marginal distribution $\mathcal{D}_X$ of $\mathcal{D}$).

Representation learning using auto-encoders is an example that fits this consideration, where given input x, the goal is to learn $\phi = h(x)$ s.t. x can be decoded back from $h(x)$. More precisely, the decoder d takes the representation ϕ and decodes it to $\hat{x}{=}g(\phi) \in \mathbb{R}^d$. h and g are learnt by minimizing the reconstruction error between $\hat{x}$ and x (e.g., $\|x - \hat{x}\|_2{=}\|x{-}g(h(x))\|_2$).

3 Functional Regularization: A Unified Perspective

We make a key observation that the auxiliary task in several representation learning approaches provides a regularization on the representation function via a learnable function. To better illustrate this viewpoint, consider the auxiliary task of an auto-encoder, where the decoder $g(\phi)$ can be viewed as such a learnable function, and the reconstruction error $\|x - g(h(x))\|_2$ can be viewed as a regularization penalty imposed on h through the decoder g for the data point x.

To formalize this notion, we consider learning representations via an auxiliary task which involves: a learnable function g, and a loss of the form $L_r(h, g; x)$ on the representation h via g for an input x. We refer to g as the regularization function and L_r as the regularization loss. Let $\mathcal{G}$ denote the hypothesis class for g. Without loss of generality we assume that $L_r \in [0, 1]$.

Definition 1. Given a loss function $L_r(h, g; x)$ for an input x involving a representation h and a regularization function g, the regularization loss of h and g on a distribution $\mathcal{U}_X$ over $\mathcal{X}$ is defined as $L_r(h, g\,; \mathcal{U}_X){=}\mathbb{E}_{x\sim\mathcal{U}_X}\,[L_r(h, g; x)]$. The regularization loss of a representation h on $\mathcal{U}_X$ is defined as $L_r(h\,; \mathcal{U}_X){=}\min_{g\in\mathcal{G}} L_r(h, g\,; \mathcal{U}_X)$.

We can similarly define $L_r(h, g\,; U)$ and $L_r(h\,; U)$ to denote the loss over a fixed set U of unlabeled data points, i.e., $L_r(h, g\,; U) := \frac{1}{|U|}\sum_{x\in U} L_r(h, g\,; x)$ and $L_r(h\,; U) := \min_{g\in\mathcal{G}} L_r(h\,; U)$.
Here, $L_r(h\,; \mathcal{U}_X)$ can be viewed as a notion of incompatibility of a representation function h on the data distribution $\mathcal{U}$. This formalizes the prior knowledge about the representation function and the data. For example, in auto-encoders $L_r(h\,; \mathcal{U}_X)$ measures how well the representation function h complies with the prior knowledge of the input being reconstructible from the representation.

We now introduce a notion for the subset of representation functions having a bounded regularization loss, which is crucial for our sample complexity analysis.

Definition 2. Given $\tau \in [0, 1]$, the τ-regularization-loss subset of representation hypotheses $\mathcal{H}$ is: $\mathcal{H}_{\mathcal{D}_X, L_r}(\tau){=}\{h \in \mathcal{H} : L_r(h\,; \mathcal{D}_X) \le \tau\}$.

We also define the prediction loss over the data distribution $\mathcal{D}$ for a prediction function f on top of h: $L_c(f, h\,; \mathcal{D}) := \mathbb{E}_{(x,y)\sim\mathcal{D}}\,[\ell_c(f(h(x)), y)]$, where ℓ_c is the loss function for prediction. Similarly, the empirical loss on the labeled data set S is $L_c(f, h\,; S) := \frac{1}{|S|}\sum_{(x,y)\in S} \ell_c(f(h(x)), y)$.
In summary, given hypothesis classes $\mathcal{H}, \mathcal{F}$, and $\mathcal{G}$, a labeled dataset S, an unlabeled dataset U, and a threshold $\tau > 0$ on the regularization loss, we consider the following learning problem: $\min_{f\in\mathcal{F}, h\in\mathcal{H}} L_c(f, h\,; S)$, s.t. $L_r(h\,; U) \le \tau$.

3.1 Sample Complexity Analysis

To analyze the sample complexity of these approaches using functional regularization, we generalize the analysis framework [3] which shows that in semi-supervised learning, unlabeled data can reduce the labeled sample complexity. We first enumerate the considerations on the data and the hypothesis classes: 1) the labeled and unlabeled data can either be from the same or different distributions (i.e., same domain or different domains); 2) the hypothesis classes can contain zero error hypothesis or not (i.e., being realizable or unrealizable); 3) the hypothesis classes can be finite or infinite in size. We perform the analysis for different combinations of these assumptions. Our proofs share a common high-level intuition across different settings. We now present sample complexity bounds for 2 interesting, characteristic settings. Due to limited space, we present bounds for several other settings, proofs of all the theorems, and additional remarks and discussions in Appendix C.

Same Domain, Realizable, Finite Hypothesis Classes. We begin with the simplest setting, where the unlabeled dataset U and the labeled dataset S are from the same distribution $\mathcal{D}_X$, and the hypothesis classes $\mathcal{F}, \mathcal{G}, \mathcal{H}$ contain functions f^*, g^*, h^* with a zero prediction and regularization loss. We further assume that the hypothesis classes are finite in size. We get the following result:

Theorem 1. *Suppose there exist $h^* \in \mathcal{H}, f^* \in \mathcal{F}, g^* \in \mathcal{G}$ such that $L_c(f^*, h^*; \mathcal{D}) = 0$ and $L_r(h^*, g^*; \mathcal{D}_X) = 0$. For any $\epsilon_0, \epsilon_1 \in (0, 1/2)$, a set U of m_u unlabeled examples and a set S of m_l labeled examples are sufficient to learn to an error ϵ_1 with probability $1 - \delta$, where $m_u \geq \frac{1}{\epsilon_0}\left[\ln|\mathcal{G}| + \ln|\mathcal{H}| + \ln\frac{2}{\delta}\right]$,$m_l \geq \frac{1}{\epsilon_1}\left[\ln|\mathcal{F}| + \ln|\mathcal{H}_{\mathcal{D}_X, L_r}(\epsilon_0)| + \ln\frac{2}{\delta}\right]$. In particular, with probability at least $1 - \delta$, all hypotheses $h \in \mathcal{H}, f \in \mathcal{F}$ with $L_c(f, h; S) = 0$ and $L_r(h; U) = 0$ will have $L_c(f, h; \mathcal{D}) \leq \epsilon_1$.*

Recall that standard analysis shows that without unlabeled data, $\frac{1}{\epsilon_1}\left[\ln|\mathcal{F}| + \ln|\mathcal{H}| + \ln\frac{2}{\delta}\right]$ labeled points are needed to get the same error guarantee. On comparing the bounds, Theorem 1 shows that the functional regularization can prune away some hypotheses in $\mathcal{H}$; thereby replacing the factor $\mathcal{H}$ with its subset $\mathcal{H}_{\mathcal{D}_X, L_r}(\epsilon_0)$ in the bound. Thus, the sample complexity bound is reduced by $\frac{1}{\epsilon_1}\left[\ln|\mathcal{H}| - \ln|\mathcal{H}_{\mathcal{D}_X, L_r}(\epsilon_0)|\right]$. So the auxiliary task is helpful for learning the predictor when $\mathcal{H}_{\mathcal{D}_X, L_r}(\epsilon_0)$ is significantly smaller than $\mathcal{H}$, avoiding the requirement of a large number of labeled points to find a good representation function among them.

Same Domain, Unrealizable, Infinite Hypothesis Classes. We now present the result for a more elaborate setting, where both the prediction and regularization losses are non-zero. We also relax the assumptions on the hypothesis classes being finite. We use metric entropy to measure the capacity of the hypothesis classes for demonstration here. Alternative capacity measures like VC-dimension or Rademacher complexity can also be used with essentially no change to the analysis. Assume that the parameter space of $\mathcal{H}$ is equipped with a norm and let $\mathcal{N}_{\mathcal{H}}(\epsilon)$ denote the ϵ-covering number of $\mathcal{H}$; similarly for $\mathcal{F}$ and $\mathcal{G}$. Let the Lipschitz constant of the losses w.r.t. these norms be bounded by L.

Theorem 2. *Suppose there exist $h^* \in \mathcal{H}, f^* \in \mathcal{F}, g^* \in \mathcal{G}$ such that $L_c(f^*, h^*; \mathcal{D}) \leq \epsilon_c$ and $L_r(h^*, g^*; \mathcal{D}_X) \leq \epsilon_r$. For any $\epsilon_0, \epsilon_1 \in (0, 1/2)$, a set U of m_u unlabeled examples and a set S of m_l labeled examples are sufficient to learn to an error $\epsilon_c + \epsilon_1$ with probability $1 - \delta$, where $m_u \geq \frac{C}{\epsilon_0^2}\ln\frac{1}{\delta}\left[\ln\mathcal{N}_{\mathcal{G}}\left(\frac{\epsilon_0}{4L}\right) + \ln\mathcal{N}_{\mathcal{H}}\left(\frac{\epsilon_0}{4L}\right)\right]$, $m_l \geq \frac{C}{\epsilon_1^2}\ln\frac{1}{\delta}\left[\ln\mathcal{N}_{\mathcal{F}}\left(\frac{\epsilon_1}{4L}\right) + \ln\mathcal{N}_{\mathcal{H}_{\mathcal{D}_X, L_r}(\epsilon_r + 2\epsilon_0)}\left(\frac{\epsilon_1}{4L}\right)\right]$ for some absolute constant C. In particular, with probability at least $1 - \delta$, the $h \in \mathcal{H}, f \in \mathcal{F}$ that optimize $L_c(f, h; S)$ subject to $L_r(h; U) \leq \epsilon_r + \epsilon_0$ have $L_c(f, h; \mathcal{D}) \leq L_c(f^*, h^*; \mathcal{D}) + \epsilon_1$.*

This leads to a reduction of $\frac{C}{\epsilon_1^2}\left[\ln\mathcal{N}_{\mathcal{H}}\left(\frac{\epsilon_1}{4L}\right) - \ln\mathcal{N}_{\mathcal{H}_{\mathcal{D}_X, L_r}(\epsilon_r + 2\epsilon_0)}\left(\frac{\epsilon_1}{4L}\right)\right]$ with the standard bound on m_l without unlabeled data. We present bounds for the setting where the unlabeled data is from a distribution $\mathcal{U}_X$ different from $\mathcal{D}_X$ in Appendix C.3.

We bring attention to some subtleties which are worth noting. Firstly, the regularization loss ϵ_r of g^*, h^* need not be optimal; there may be other g, h which get a smaller $L_r(h, g; \mathcal{D}_X)$ (even $\ll \epsilon_r$). Secondly, the prediction loss is bounded by $L_c(f^*, h^*; \mathcal{D}) + \epsilon_1$, which is independent of ϵ_r. Similarly, the bounds on m_u and m_ℓ mainly depend on ϵ_0 and ϵ_1 respectively, while only m_ℓ depends on ϵ_r through the $\mathcal{H}_{\mathcal{D}_X, L_r}(\epsilon_r + 2\epsilon_0)$ term. Thus, even when the regularization loss is large (e.g., the reconstruction of an auto-encoder is far from accurate), it is still possible to learn an accurate predictor with a significantly reduced labeled data size using the unlabeled data. This suggests that when designing an auxiliary task ($\mathcal{G}$ and L_r), it is *not* necessary to ensure that the "ground-truth" h^* has a small regularization loss. Rather, one should ensure that only a small fraction of $h \in \mathcal{H}$ have a smaller (or similar) regularization loss than h^* so as to reduce the label sample complexity.

This bound also shows that τ should be carefully chosen for the constraint $L_r(h;U) \leq \tau$. With a very small τ, the ground-truth h^* (or hypotheses of similar quality) may not satisfy the constraint and become infeasible for learning. With a very large τ, the auxiliary task may not reduce the labeled sample complexity. Practical learning algorithms typically turn this constrain into a regularization like term, i.e., by optimizing $L_c(f,h;S) + \lambda L_r(h;U)$. For such objectives, the requirement on τ translates to carefully choosing λ. When λ is very large, this leads to a small $L_r(h;U)$ but a large $L_c(f,h;S)$, while when λ is very small, this may not reduce the labeled sample complexity.

4 Applying the Theoretical Framework to Concrete Examples

The analysis in Section 3 shows that the sample complexity bound reduction depends on the notion of the pruned subset $\mathcal{H}_{\mathcal{D}_X,L_r}$, which captures the effect of the regularization function and the property of the unlabeled data distribution. Our generic framework can be applied to various concrete configurations of the hypothesis classes and data distributions. This way we can quantify the reduction more explicitly by investigating $\mathcal{H}_{\mathcal{D}_X,L_r}$. We provide 2 such examples: one using auto-encoders (here in the main paper) and the other using a masked self-supervision (in Appendix D.2). We outline the sample complexity bounds for these examples, and present the complete proofs in Appendix D.

4.1 An Example of Functional Regularization via Auto-encoder

Learning Without Functional Regularization. Consider $\mathcal{H}$ to be the class of linear functions from $\mathbb{R}^d$ to $\mathbb{R}^r$ where $r < d/2$, and $\mathcal{F}$ to be the class of linear functions over some activations. That is, $\phi{=}h_W(x){=}Wx$, $y{=}f_a(\phi){=}\sum_{i=1}^r a_i\sigma(\phi_i)$, where $W \in \mathbb{R}^{r\times d}$, $a \in \mathbb{R}^r$ Here $\sigma(t)$ is an activation function (e.g., $\sigma(t){=}t^2$), the rows of W and a have ℓ_2 norms bounded by 1. We consider the Mean Square Error prediction loss, i.e., $L_c(f,h;x){=}\|y{-}f(h(x))\|_2^2$. Without prior knowledge on data, no functional regularization corresponds to end-to-end training on $\mathcal{F}{\circ}\mathcal{H}$.

Data Property. We consider a setting where the data has properties which allows functional regularization. We assume that the data consists of a signal and noise, where the signal lies in a certain r-dimensional subspace. Formally, let columns of $B \in \mathbb{R}^{d\times d}$ be eigenvectors of $\Sigma{:=}\mathbb{E}[xx^\top]$, then the prediction labels are largely determined by the signal in the first r directions: $y{=}\sum_{i=1}^r a_i^*\sigma(\phi_i^*){+}\nu$ and $\phi^*{=}B_{1:r}^\top x$, where $a^* \in \mathbb{R}^r$ is a ground-truth parameter with $\|a^*\|_2{\leq}1$, $B_{1:r}$ is the set of first r eigenvectors of Σ, and ν is a small Gaussian noise. We assume a difference in the r^{th} and $r{+}1^{\text{th}}$ eigenvalues of Σ to distinguish the corresponding eigenvectors. Let ϵ_r denote $\mathbb{E}\|x{-}B_{1:r}B_{1:r}^\top x\|_2^2$.

Learning With Functional Regularization. Knowing that the signal lies in an r-dimensional subspace, we can perform auto-encoder functional regularization. Let $\mathcal{G}$ be a class of linear functions from $\mathbb{R}^r$ to $\mathbb{R}^d$, i.e., $\hat{x}{=}g_V(\phi){=}V\phi$ where $V{\in}\mathcal{R}^{d\times r}$ has orthonormal columns. The regularization loss $L_r(h,g;x){=}\|x{-}g(h(x))\|_2^2$. For simplicity, we assume access to infinite unlabeled data.

Without regularization, the standard ϵ-covering argument shows that the labeled sample complexity, for an error ϵ close to the optimal, is $\frac{C}{\epsilon^2}\left[\ln\mathcal{N}_{\mathcal{F}}\left(\frac{\epsilon}{4L}\right) + \ln\mathcal{N}_{\mathcal{H}}\left(\frac{\epsilon}{4L}\right)\right]$ for some absolute constant C. Applying our framework when using regularization with $\tau = \epsilon_r$, the sample complexity is bounded by $\frac{C}{\epsilon^2}\left[\ln\mathcal{N}_{\mathcal{F}}\left(\frac{\epsilon}{4L}\right) + \ln\mathcal{N}_{\mathcal{H}_{\mathcal{D}_X,L_r(\epsilon_r)}}\left(\frac{\epsilon}{4L}\right)\right]$. Then we show that $\mathcal{N}_{\mathcal{H}}\left(\frac{\epsilon}{4L}\right) \geq \binom{d-r}{r}\mathcal{N}_{\mathcal{H}_{\mathcal{D}_X,L_r(\epsilon_r)}}\left(\frac{\epsilon}{4L}\right)$ (Proof in Appendix D.1) since $\mathcal{H}_{\mathcal{D}_X,L_r}(\epsilon_r){=}\{h_W(x) : W{=}OB_{1:r}^\top, O \in \mathbb{R}^{r\times r}, O \text{ is orthonormal }\}$, $\mathcal{H} \supseteq \{h_W(x) : W = OB_S^\top, O \in \mathbb{R}^{r\times r}, O \text{ is orthonormal}, S \subseteq \{r+1,\ldots,d\}, |S|{=}r\}$, where B_S refers to the sub-matrix of columns in B having indices in S. Therefore, the label sample complexity bound is reduced by $\frac{C}{\epsilon^2}\ln\binom{d-r}{r}$, i.e., the error bound is reduced by $\frac{C}{\sqrt{m_\ell}}\ln\binom{d-r}{r}$ when using m_ℓ labeled points. Note that $\ln\binom{d-r}{r} = \Theta(r\ln(d/k))$ when r is small, and thus the reduction is roughly linear initially and then grows slower with r. Interestingly, the reduction depends on the hidden dimension r but has little dependence on the input dimension d.

Due to limited space, experiments on synthetic and real data are presented in Appendix E and F.

5 Conclusion

In this paper we have presented a unified discriminative framework for analyzing many representation learning approaches using unlabeled data, by viewing them as imposing a regularization on the representation via a learnable function. We have derived sample complexity bounds under various assumptions on the hypothesis classes and data, and shown that the functional regularization can be used to prune the hypothesis class and reduce the labeled sample complexity. We have also applied our framework to two concrete examples. An interesting future work direction is to investigate the effect of such functional regularization on the optimization of the learning methods.

References

[1] P. Agrawal, A. Nair, P. Abbeel, J. Malik, and S. Levine. Learning to poke by poking: Experiential learning of intuitive physics. *Conference on Neural Information Processing Systems*, 06 2016.

[2] P. Bachman, R. D. Hjelm, and W. Buchwalter. Learning representations by maximizing mutual information across views. In *Advances in Neural Information Processing Systems*, pages 15535–15545, 2019.

[3] M.-F. Balcan and A. Blum. A discriminative model for semi-supervised learning. *J. ACM*, 57(3), Mar. 2010.

[4] P. Baldi. Autoencoders, unsupervised learning and deep architectures. In *Proceedings of the 2011 International Conference on Unsupervised and Transfer Learning Workshop - Volume 27*, UTLW'11, page 37–50. JMLR.org, 2011.

[5] N. Bansal, X. Chen, and Z. Wang. Can we gain more from orthogonality regularizations in training deep cnns? In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 4266–4276, Red Hook, NY, USA, 2018. Curran Associates Inc.

[6] J. Baxter. A model of inductive bias learning. *J. Artif. Int. Res.*, 12(1):149–198, Mar. 2000.

[7] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan. A theory of learning from different domains. *Machine learning*, 79(1-2):151–175, 2010.

[8] S. Ben-David and R. Urner. On the hardness of domain adaptation and the utility of unlabeled target samples. In *Proceedings of the 23rd International Conference on Algorithmic Learning Theory*, ALT'12, page 139–153, Berlin, Heidelberg, 2012. Springer-Verlag.

[9] Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives, 2012.

[10] A. Brock, T. Lim, J. M. Ritchie, and N. Weston. Neural photo editing with introspective adversarial networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, 2017.

[11] L. Cayton. Algorithms for manifold learning. *Univ. of California at San Diego Tech. Rep*, 12(1-17):1, 2005.

[12] O. Chapelle, B. Schölkopf, and A. Zien. *Semi-Supervised Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2006.

[13] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[14] C. Doersch, A. Gupta, and A. A. Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, ICCV '15, page 1422–1430, USA, 2015. IEEE Computer Society.

[15] W. B. Dolan and C. Brockett. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*, 2005.

[16] A. Dosovitskiy, J. T. Springenberg, M. Riedmiller, and T. Brox. Discriminative unsupervised feature learning with convolutional neural networks. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 766–774. Curran Associates, Inc., 2014.

[17] S. S. Du, W. Hu, S. M. Kakade, J. D. Lee, and Q. Lei. Few-shot learning via learning the representation, provably. *arXiv preprint arXiv:2002.09434*, 2020.

[18] F. Ebert, C. Finn, A. X. Lee, and S. Levine. Self-supervised visual planning with temporal skip connections. In *CoRL*, 2017.

[19] D. Erhan, Y. Bengio, A. Courville, P.-A. Manzagol, P. Vincent, and S. Bengio. Why does unsupervised pre-training help deep learning? *J. Mach. Learn. Res.*, 11:625–660, Mar. 2010.

[20] S. Gidaris, P. Singh, and N. Komodakis. Unsupervised representation learning by predicting image rotations. In *International Conference on Learning Representations*, 2018.

[21] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[22] S. Hanneke and S. Kpotufe. A no-free-lunch theorem for multitask learning. *arXiv preprint arXiv:2006.15785*, 2020.

[23] E. Hazan and T. Ma. A non-generative framework and convex relaxations for unsupervised learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 3314–3322, Red Hook, NY, USA, 2016. Curran Associates Inc.

[24] G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507, 2006.

[25] E. Hoffer and N. Ailon. Deep metric learning using triplet network. In Y. Bengio and Y. LeCun, editors, *ICLR (Workshop)*, 2015.

[26] O. J. Hénaff, A. Srinivas, J. D. Fauw, A. Razavi, C. Doersch, S. M. A. Eslami, and A. van den Oord. Data-efficient image recognition with contrastive predictive coding, 2019.

[27] E. Jang, C. Devin, V. Vanhoucke, and S. Levine. Grasp2vec: Learning object representations from self-supervised grasping. In A. Billard, A. Dragan, J. Peters, and J. Morimoto, editors, *Proceedings of The 2nd Conference on Robot Learning*, volume 87 of *Proceedings of Machine Learning Research*, pages 99–112. PMLR, 29–31 Oct 2018.

[28] O. Kovaleva, A. Romanov, A. Rogers, and A. Rumshisky. Revealing the dark secrets of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4365–4374, Hong Kong, China, Nov. 2019. Association for Computational Linguistics.

[29] H. Larochelle and Y. Bengio. Classification using discriminative restricted boltzmann machines. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, page 536–543, New York, NY, USA, 2008. Association for Computing Machinery.

[30] H. Larochelle, M. Mandel, R. Pascanu, and Y. Bengio. Learning algorithms for the classification restricted boltzmann machine. *J. Mach. Learn. Res.*, 13(1):643–669, Mar. 2012.

[31] L. Le, A. Patterson, and M. White. Supervised autoencoders: Improving generalization performance with unsupervised regularizers. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 107–117. Curran Associates, Inc., 2018.

[32] Y. Li, T. Ma, and H. Zhang. Algorithmic regularization in over-parameterized matrix sensing and neural networks with quadratic activations. In *Conference On Learning Theory*, pages 2–47. PMLR, 2018.

[33] T. Liu, D. Tao, M. Song, and S. J. Maybank. Algorithm-dependent generalization bounds for multi-task learning. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(2):227–241, Feb. 2017.

[34] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019.

[35] J. Mairal, J. Ponce, G. Sapiro, A. Zisserman, and F. R. Bach. Supervised dictionary learning. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 1033–1040. Curran Associates, Inc., 2009.

[36] M. Mohri, A. Rostamizadeh, and A. Talwalkar. *Foundations of machine learning*. MIT press, 2018.

[37] V. Nagarajan and J. Z. Kolter. Uniform convergence may be unable to explain generalization in deep learning. In *Advances in Neural Information Processing Systems*, pages 11615–11626, 2019.

[38] B. Neyshabur, R. Tomioka, and N. Srebro. In search of the real inductive bias: On the role of implicit regularization in deep learning. *arXiv preprint arXiv:1412.6614*, 2014.

[39] M. Noroozi and P. Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *ECCV*, 2016.

[40] K. Nozawa, P. Germain, and B. Guedj. Pac-bayesian contrastive unsupervised representation learning, 2019.

[41] D. Pathak, P. Krähenbühl, J. Donahue, T. Darrell, and A. A. Efros. Context encoders: Feature learning by inpainting. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2536–2544, 2016.

[42] L. Pinto and A. Gupta. Supersizing self-supervision: Learning to grasp from 50k tries and 700 robot hours. *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3406–3413, 2015.

[43] S. Rifai, Y. N. Dauphin, P. Vincent, Y. Bengio, and X. Muller. The manifold tangent classifier. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 2294–2302. Curran Associates, Inc., 2011.

[44] F. Rodriguez and G. Sapiro. Sparse representations for image classification: Learning discriminative and reconstructive non-parametric dictionaries. Technical report, MINNESOTA UNIV MINNEAPOLIS, 2008.

[45] D. E. Rumelhart and J. L. McClelland. *Learning Internal Representations by Error Propagation*, pages 318–362. MIT Press, 1987.

[46] M. Sadeghi, M. Babaie-Zadeh, and C. Jutten. Dictionary learning for sparse representation: A novel approach. *IEEE Signal Processing Letters*, 20(12):1195–1198, 2013.

[47] N. Saunshi, O. Plevrakis, S. Arora, M. Khodak, and H. Khandeparkar. A theoretical analysis of contrastive unsupervised representation learning. In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5628–5637, Long Beach, California, USA, 09–15 Jun 2019. PMLR.

[48] P. Smolensky. Information processing in dynamical systems: foundations of harmony theory. In *Parallel distributed processing: explorations in the microstructure of cognition, vol. 1: foundations*, pages 194–281. Bradford Books, MIT Press, 1986.

[49] B. Sofman, E. Lin, J. Bagnell, J. Cole, N. Vandapel, and A. Stentz. Improving robot navigation through self-supervised online learning. *J. Field Robotics*, 23:1059–1075, 11 2006.

[50] J. B. Tenenbaum, V. d. Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

[51] N. Tripuraneni, M. I. Jordan, and C. Jin. On the theory of transfer learning: The importance of task diversity. *arXiv preprint arXiv:2006.11650*, 2020.

[52] M. Tschannen, J. Djolonga, P. K. Rubenstein, S. Gelly, and M. Lucic. On mutual information maximization for representation learning. In *International Conference on Learning Representations*, 2020.

[53] L. G. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.

[54] L. van der Maaten and G. Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

[55] R. Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

[56] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, page 1096–1103, New York, NY, USA, 2008. Association for Computing Machinery.

[57] H. Xiao, K. Rasul, and R. Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *CoRR*, abs/1708.07747, 2017.

[58] Z. Yang, Z. Hu, R. Salakhutdinov, and T. Berg-Kirkpatrick. Improved variational autoencoders for text modeling using dilated convolutions. In D. Precup and Y. W. Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3881–3890, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR.

[59] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.

[60] R. Zhang, P. Isola, and A. Efros. Colorful image colorization. In *European Conference on Computer Vision*, volume 9907, pages 649–666, 10 2016.

[61] R. Zhang, P. Isola, and A. Efros. Split-brain autoencoders: Unsupervised learning by cross-channel prediction. In *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, pages 645–654, United States, 11 2017. Institute of Electrical and Electronics Engineers Inc.

Appendix

A Related Work

Self-supervised learning approaches for images have been extensively used in computer vision through auxiliary tasks such as masked image patch prediction [16], image rotations [20], pixel colorization [60], context prediction of image patches [14, 41, 39, 26], etc. Additionally, variants of these approaches find practical use in the field of robotics [49, 42, 1, 18, 27]. Masked self-supervision (a type of denoising auto-encoder), where representations are learnt by hiding a portion of the input and then reconstructing it, has lead to powerful language models like BERT [13] and RoBERTa [34] in natural language processing. There have also been numerous studies on other representation learning approaches such as RBMs [29, 30], dictionary learning [44, 35] and manifold learning [43]; [9] presents an extensive review of multiple representation learning approaches.

On the theoretical front, [3] presents a discriminative framework for analyzing semi-supervised learning showing that unlabeled data can reduce the labeled sample complexity. Our framework in this paper is inspired from [3], and generalizes their analysis for utilizing unlabeled data through a *learnable* regularization function. This allows a unified theoretical framework to study multiple representation learning approaches. In addition to [3], [12] also studies the benefits of using unlabeled data, but by restricting that the unlabeled data be utilized through a fixed function. Some other works [7, 8] have explored the benefits of unlabeled data for domain adaptation. Our setting differs from this since our goal is to learn a prediction function on the labeled data, rather than for a change in the domain of labeled data from source to target. Another line of related work considers multi-task learning, such as [6, 33]. These works show that multiple supervised learning tasks on different, but related, data distributions can help generalization. Our work differs from these since we focus on learning a supervised task using auxiliary unsupervised tasks on unlabeled data.

[19] presents a comprehensive empirical study on the benefits of unsupervised pre-training for image-classification tasks in computer vision. Our analysis in this paper is motivated by their empirical results showing that pre-training shrinks the hypothesis space searched during learning. There have also been theoretical studies on several representation approaches individually, without providing a holistic perspective. [47] presents a theoretical framework to analyse unsupervised representation learning techniques that can be posed as a contrastive learning problem, with their results later improved by [40]. [23] provide a theoretical analysis of unsupervised learning from an optimization viewpoint, with applications to dictionary learning and spectral auto-encoders. [31] prove uniform stability generalization bounds for linear auto-encoders and empirically demonstrate the benefits of using supervised auto-encoders. Additionally, there are some studies on learning transferable representations using multiple tasks [17, 51, 22]. Another line of related work includes approaches [2, 52] that analyze representation learning from the perspective of maximizing the mutual information between the data and the representation. Connecting these mutual information approaches with our framework is left as future work.

B Instantiations of Functional Regularization

Some popular examples of the auxiliary task used for representation learning are auto-encoders [45, 4], sparse dictionaries [46], masked self-supervision [13], manifold learning [11], and others [9]. These approaches have been extensively used in applications in various domains, such as computer vision (e.g., [56, 61, 16]) and natural language processing (e.g., [58, 13, 34]), and have achieved impressive empirical performance. Here we show that several unsupervised (self-supervised) representation learning strategies can be viewed as imposing a learnable function to regularize the representations being learned. We note that the class $\mathcal{G}$ can be an index set instead of a class of functions; our framework applies as long as the loss $L_r(h, g; x)$ is well defined (see the manifold learning example). $\mathcal{G}$ can also only have a single g, corresponding to the special case of a fixed regularizer (see the ℓ_p norm penalty example).

Auto-encoder. Auto-encoders use an encoder function h to map the input x to a lower dimensional space ϕ and a decoder network d to reconstruct the input back from ϕ using a MSE loss $\|x - d(h(x))\|^2$. One can view d as a regularizer on the feature representation $\phi = h(x)$ through the regularization

loss $L_r(h, g\,; x) = \|x - d(h(x))\|^2$. $\mathcal{H}_{\mathcal{D}_X, L_r}(\tau)$ is the subset of representation functions with at most τ reconstruction error using the best decoder in $\mathcal{G}$.

Variants of standard auto-encoders like noisy auto-encoders or sparse auto-encoders can be formulated similarly as a functional regularization on the representation being learnt.

Masked Self-supervised Learning. Masked self-supervision techniques, in abstract terms, cover a portion of the input and then predict the masked input portion [13]. More concretely, say the input $x = [x_1, x_2, \ldots, x_d]$ is masked as $x' = [x_1, \ldots, x_i, 0, \ldots, 0, x_j, \ldots, x_d]$ and a function g is learned to predict the masked input $[x_{i+1}, \ldots, x_{j-1}]$ over an input representation $h(x)$. This function g used to reconstruct x, can be viewed as imposing a regularization on h through a MSE regularization loss given by $\|x_{[i+1:j-1]} - g(h(x'))\|^2$. $\mathcal{H}_{\mathcal{D}_X, L_r}(\tau)$ is the subset of $\mathcal{H}$ which have at most τ MSE on predicting $x_{[i+1:j-1]}$ using the best function $g \in G$.

Variational Auto-encoder. VAEs encode the input x as a distribution $q_\phi(z|x)$ over a parametric latent space z instead of a single point, and sample from it to reconstruct x using a decoder $p_\theta(x|z)$. The encoder $q_\phi(z|x)$ is used to model the underlying mean μ_z and co-variance matrix σ_z of the distribution over z. VAEs are trained by minimising a loss

$$\mathcal{L}_x(\theta, \phi) = -\mathbb{E}_{z \sim q_\phi(z|x)}[\log\ p_\theta(x|z)] + \mathbb{KL}(q_\phi(z|x)||p(z))$$

where $p(z)$ is specified as the prior distribution over z (e.g., $\mathcal{N}(0, 1)$). The encoder $q_\phi(z|x)$ can be viewed as the representation function h, the decoder $p_\theta(x|z)$ as the learnable regularization function g, and the loss $\mathcal{L}_x(\theta, \phi)$ as the regularization loss $L_r(h, g; x)$ in our framework. Then $\mathcal{H}_{\mathcal{D}_X, L_r}(\tau)$ is the subset of encoders q_ϕ which have at most τ VAE loss when using the best decoder p_θ for it.

Manifold Learning through the Triplet Loss. Learning manifold representations through metric learning is a popular technique used in computer vision applications [25]. A triplet loss formulation is used to learn a distance metric for the representations, by trying to minimise this metric between a baseline and positive sample and maximising the metric between the baseline and a negative sample. This is achieved by learning a representation function h for an input x. Considering a triple of input samples $\bar{x} = (x_b, x_p, x_n)$ corresponding to a baseline, positive and negative sample, we use a loss $L_{\text{Triplet}}(\bar{x}) = \max(\|h(x_b) - h(x_p)\|_2^2 - \|h(x_b) - h(x_n)\|_2^2, 0)$ to learn h. This is a special instantiation of our framework using a dummy $\mathcal{G}$ having a single function g, where the regularization loss $L_r(h, g\,; \bar{x}) = L_{\text{Triplet}}(\bar{x})$ is computed over a triple of input samples.

Further, one can also consider some variants of the standard triplet loss formulation under our functional regularization perspective. For example, let the triplet loss be $L^{(\alpha)}_{\text{Triplet}}(\bar{x}) = \max(\|h(x_b) - h(x_p)\|_2^2 - \|h(x_b) - h(x_n)\|_2^2 + \alpha, 0)$ where $\alpha \in \mathbb{R}$ is a margin between the positive and negative pairs. When α is learnable, this corresponds to a functional regularization where $\mathcal{G} = \{\alpha : \alpha \in \mathbb{R}\}$, and the regularization loss is $L_r(h, g; \bar{x}) = L^{(\alpha)}_{\text{Triplet}}(\bar{x})$. In this case, the class $\mathcal{G}$ is not defined on top of the representation $h(x)$. However, our framework and the sample complexity analysis can still be applied through the definition of $L_r(h, g; \bar{x})$.

Sparse Dictionary Learning. Sparse dictionary learning is an unsupervised learning approach to obtain a sparse low-dimensional representation of the input data. Here we consider a distributional view of sparse dictionary learning. Give a distribution $\mathcal{D}_X$ over unlabeled data $x \in \mathbb{R}^d$ and a hyper-parameter $\lambda > 0$, we want to find a dictionary matrix $D \in \mathbb{R}^{d \times K}$ and a sparse representation $z \in \mathbb{R}^K$ for each x, so as to minimize the error $\mathbb{E}[L_D(x)]$, where $L_D(x)$ is the error on one point x defined as $L_D(x) := \|x - Dz\|_2^2 + \lambda\|z\|_0$, subject to the constraint that each column of D has ℓ_2 norm bounded by 1. The learned representations z can then be used for a target prediction task. Under our framework, we can view the representation function corresponding to $z = h_D(x) = \arg\min_{z \in \mathbb{R}^K} \|x - Dz\|_2^2 + \lambda\|z\|_0$, and D is the parameter of the representation function. The regularization function class $\mathcal{G}$ has a single g, and the regularization loss is $L_r(h_D, g; x) = L_D(x)$.

Our framework also captures an interesting variant of dictionary learning. Consider another dictionary matrix F and a hyper-parameter $\eta > 0$. The representation function still corresponds to $z = h_D(x) = \arg\min_{z \in \mathbb{R}^K} \|x - Dz\|_2^2 + \lambda\|z\|_0$, with D as the parameter. The regularization function class is now given by $\mathcal{G} = \{g_F(z) = Fz : F \in \mathbb{R}^{d \times K}\}$, and the regularization loss $L_r(h_D, g_F; x)$ is defined as $\|x - g_F(h_D(x))\|_2^2 + \lambda\|z\|_0 + \eta\|D - F\|_F^2$. This special case of dictionary learning allows the

encoding and decoding steps to use two different dictionaries D and E but constraining the difference between them. When $\eta \to +\infty$, this variant reduces to the original version described earlier.

Explicit ℓ_p Norm Penalty. Techniques imposing explicit regularizations on the representation h being learned, often use an ℓ_p norm penalty on $h(x)$ i.e, $\|h(x)\|_p^p$ to the prediction loss while jointly training f and h. This can be viewed as a special case of our framework using a fixed regularization function $g(h(x)) = \|h(x)\|_p^p$.

Restricted Boltzmann Machines. Restricted Boltzmann Machines (RBM) [48, 24] generate hidden representations for an input through unsupervised learning on unlabeled data. RBMs are characterized by a joint distribution over the input $x \in \{0,1\}^d$ and the representation $z \in \{0,1\}^r$: $P(x,z) = \frac{1}{Z}e^{-E(x,z)}$, where Z is the partition function and $E(x,z)$ is the energy function defined as: $E(x,z) = -a^\top x - b^\top z - x^\top W z$, where $a \in \mathbb{R}^d, b \in \mathbb{R}^r, W \in \mathbb{R}^{d\times r}$ are parameters to be learned.

Then $P(z|x)$, for a fixed x, is a distribution parameterized by b and W; which can be denoted as $q_{W,b}(z|x)$. Similarly, $P(x|z)$ is parameterized by a and W and thus can be denoted as $p_{W,a}(x|z)$. Given $x \sim \mathcal{D}_X$, the objective of the RBM is to minimize $-\mathbb{E}_{x\sim\mathcal{D}_X}[\log P(x)]$.

While the standard RBM objective does not have a direct analogy under our functional regularization framework, a heuristic variant can be formulated under our framework. If we use $\mathbb{E}_{P(x)}$ to denote the expectation over the marginal distribution of x in the RBM, $\mathbb{E}_{P(z)}$ to denote the expectation over the marginal distribution of z, and $\mathbb{E}_{\mathcal{D}_X}$ to denote the expectation over $x \sim \mathcal{D}_X$. Then the following hold for the standard RBM:

$$P(z) = \mathbb{E}_{P(x)}[P(z|x)] = \mathbb{E}_{P(x)}[q_{W,b}(z|x)] \tag{1}$$

$$P(x) = \mathbb{E}_{P(z)}[P(x|z)] = \mathbb{E}_{P(z)}[p_{W,a}(x|z)] \tag{2}$$

In the heuristic variant, we replace $P(x)$ with $\mathcal{D}_X$ in Equation (1):

$$\hat{P}(z) = \mathbb{E}_{\mathcal{D}_X}[P(z|x)] = \mathbb{E}_{\mathcal{D}_X}[q_{W,b}(z|x)], \quad \hat{P}(x) = \mathbb{E}_{\hat{P}(z)}[P(x|z)] = \mathbb{E}_{\hat{P}(z)}[p_{W,a}(x|z)], \tag{3}$$

and train using the loss:

$$L(W,a,b;x) := -\log \hat{P}(x) = -\log \mathbb{E}_{\hat{P}(z)}\{p_{W,a}(x|z)\} = -\log \mathbb{E}_{\mathbb{E}_{\mathcal{D}_X}[q_{W,b}(z|x)]}\{p_{W,a}(x|z)\}. \tag{4}$$

Furthermore, on introducing another weight matrix $F \in \mathbb{R}^{d\times r}$ for $P(x|z)$ and a hyper-parameter $\eta > 0$, we can train the RBM using the loss:

$$L_\eta(W,a,b;x) := -\log \mathbb{E}_{\mathbb{E}_{\mathcal{D}_X}[q_{W,b}(z|x)]}\{p_{F,a}(x|z)\} + \eta\|W - F\|_F^2. \tag{5}$$

When $\eta \to +\infty$, this loss function reduces to the loss $L(W,a,b;x)$. Here $q_{W,b}(z|x)$ can be viewed as the representation function h of our framework, $p_{F,a}(x|z)$ as the regularization function g, and $L_\eta(W,a,b;x)$ as the regularization loss $L_r(h,g;x)$.

Comparison to GANs. Finally, we would like to comment on Generative Adversarial Networks (GANs) [21]. While both functional regularization and GANs use auxiliary tasks having a function class, the goal of GANs is to learn a generative model using an auxiliary task through a discriminative function (the discriminator), while the goal of functional regularization is to learn a discriminative model using an auxiliary task which is usually (though not always) through a generative function (e.g., the decoder in auto-encoders).

C Sample Complexity Bounds

Consider a simple example of a regression problem where $y = \sum_{i=1}^d x_i$ and we use masked self-supervision to learn x_1 from x'. If each $x_i \sim \mathcal{N}(0,1)$ *i.i.d.*, then h will not be able to learn a meaningful representation of x for predicting y, since x_1 is independent of all other coordinates of x. On the other extreme, if all x_i's are equal, h can learn the perfect underlying representation for predicting y, which corresponds to a single coordinate of x. This shows two contrasting abstractions of the inherent structure in the data and how the benefits of using a specific auxiliary task may vary. Our framework aims at analyzing the sample complexity of labeled data and clarifying these subtleties on the benefits of the auxiliary task depending on the data distribution.

We generalize the analysis presented by [3] which shows that in semi-supervised learning, unlabeled data can reduce the labeled sample complexity. Their analysis is for a fixed regularization function, and we observe that it can be generalized to our general setting of *learnable* regularization functions and thus allow a unified framework for functional regularization.

We consider different assumptions on the data distribution and the hypothesis classes: 1) the labeled data and unlabeled data can be from the same or different distributions (i.e., same domain v.s. different domains); 2) the hypothesis classes can contain zero error hypothesis or not (realizable v.s. unrealizable); 3) the hypothesis classes can be of finite or infinite sizes.

C.1 Same Domain, Realizable, Finite Hypothesis Classes

For simplicity, we begin with the realizable case, where the hypothesis classes contain functions g^*, h^*, f^* with a zero prediction and regularization loss. Here we consider that the unlabeled U and labeled S samples are from the same domain distribution $\mathcal{D}_X$. We derive the following Theorem.

Theorem 1. *Suppose there exist $h^* \in \mathcal{H}, f^* \in \mathcal{F}, g^* \in \mathcal{G}$ such that $L_c(f^*, h^*; \mathcal{D}) = 0$ and $L_r(h^*, g^*; \mathcal{D}_X) = 0$. For any $\epsilon_0, \epsilon_1 \in (0, 1/2)$, a set U of m_u unlabeled examples and a set S of m_l labeled examples are sufficient to learn to an error ϵ_1 with probability $1 - \delta$, where $m_u \geq \frac{1}{\epsilon_0}\left[\ln|\mathcal{G}| + \ln|\mathcal{H}| + \ln\frac{2}{\delta}\right]$,$m_l \geq \frac{1}{\epsilon_1}\left[\ln|\mathcal{F}| + \ln|\mathcal{H}_{\mathcal{D}_X, L_r}(\epsilon_0)| + \ln\frac{2}{\delta}\right]$. In particular, with probability at least $1 - \delta$, all hypotheses $h \in \mathcal{H}, f \in \mathcal{F}$ with $L_c(f, h; S) = 0$ and $L_r(h; U) = 0$ will have $L_c(f, h; \mathcal{D}) \leq \epsilon_1$.*

Proof. We first show that with high probability, only the hypotheses h in $\mathcal{H}_{\mathcal{D}_X, L_r}(\epsilon_0)$ have $L_r(h; U) = 0$. For a given pair g and h with $L_r(h, g; \mathcal{D}_X) \geq \epsilon_0$, the probability that $L_r(h, g; U) = 0$ is at most

$$\mathbb{P}[L_r(h, g; U) = 0] \leq (1 - \epsilon_0)^{m_u} \leq \frac{\delta}{2|\mathcal{H}||\mathcal{G}|} \tag{6}$$

for the given value of m_u. By the union bound, with probability at least $1 - \delta/2$, only those g and h with $L_r(h, g; \mathcal{D}_X) \leq \epsilon_0$ have $L_r(h, g; U) = 0$. Then only hypotheses $h \in \mathcal{H}_{\mathcal{D}_X, L_r}(\epsilon_0)$ have $L_r(h; U) = 0$.

Then we show that with high probability, for all $h \in \mathcal{H}_{\mathcal{D}_X, L_r}(\epsilon_0)$, only those f and h with $L_c(f, h; \mathcal{D}) \leq \epsilon_1$ can have $L_c(f, h; S) = 0$. Similarly as above, for a pair $f \in \mathcal{F}$ and $h \in \mathcal{H}_{\mathcal{D}_X, L_r}(\epsilon_0)$ with $L_c(f, h; \mathcal{D}) \geq \epsilon_1$, the probability that $L_c(f, h; S) = 0$ is at most

$$\mathbb{P}[L_c(f, h; S) = 0] \leq (1 - \epsilon_1)^{m_\ell} \leq \frac{\delta}{2|\mathcal{H}_{\mathcal{D}_X, L_r}(\epsilon_0)||\mathcal{G}|} \tag{7}$$

for the given value of m_ℓ. By the union bound, with probability $1 - \delta/2$, for $f \in \mathcal{F}$ and $h \in \mathcal{H}_{\mathcal{D}_X, L_r}(\epsilon_0)$, only those with $L_c(f, h; \mathcal{D}) \leq \epsilon_1$ can have $L_c(f, h; S) = 0$, proving the theorem. □

C.2 Same Domain, Unrealizable Case, Infinite Hypothesis Classes

When the hypothesis classes are of an infinite size, we use metric entropy to measure the capacity. Suppose $\mathcal{H}$ is indexed by parameter set Θ_H with norm $\|\cdot\|_H$, $\mathcal{G}$ by Θ_G with norm $\|\cdot\|_G$, and $\mathcal{F}$ by Θ_F with norm $\|\cdot\|_F$. Assume that the losses are L-Lipschitz with respect to the parameters. That is,

$$\begin{aligned}
|L_r(h_\theta, g; x) - L_r(h_{\theta'}, g; x)| &\leq L\|\theta - \theta'\|_H, \forall g \in \mathcal{G}, x \in \mathcal{X}, \\
|L_r(h, g_\theta; x) - L_r(h, g_{\theta'}; x)| &\leq L\|\theta - \theta'\|_G, \forall h \in \mathcal{H}, x \in \mathcal{X}, \\
|L_c(h_\theta, f; x) - L_c(h_{\theta'}, f; x)| &\leq L\|\theta - \theta'\|_H, \forall f \in \mathcal{F}, x \in \mathcal{X}, \\
|L_c(h, f_\theta; x) - L_c(h, f_{\theta'}; x)| &\leq L\|\theta - \theta'\|_G, \forall h \in \mathcal{H}, x \in \mathcal{X}.
\end{aligned}$$

Let $\mathcal{N}_{\mathcal{G}}(\epsilon)$ be the ϵ-covering number of $\mathcal{G}$ w.r.t. the associated norm. This is similarly defined for the other function classes.

The assumptions that the regularization and prediction losses are 0 are usually impractical due to noise in the data distribution. Realistically we may assume that there exist ground-truth functions that can make the regularization and prediction losses small. We begin by considering a setting where the prediction loss can be zero while the regularization loss is not.

Theorem 3. *Suppose there exist $h^* \in \mathcal{H}, f^* \in \mathcal{F}, g^* \in \mathcal{G}$ such that $L_c(f^*, h^*; \mathcal{D}) = 0$ and $L_r(h^*, g^*; \mathcal{D}_X) \leq \epsilon_r$. For any $\epsilon_0, \epsilon_1 \in (0, 1/2)$, a set U of m_u unlabeled examples and a set S of m_l labeled examples is sufficient to learn to an error ϵ_1 with probability $1-\delta$, where $m_u \geq \frac{C}{\epsilon_0^2}\ln\frac{1}{\delta}\left[\ln\mathcal{N}_{\mathcal{G}}\left(\frac{\epsilon_0}{4L}\right) + \ln\mathcal{N}_{\mathcal{H}}\left(\frac{\epsilon_0}{4L}\right)\right]$, $m_l \geq \frac{C}{\epsilon_1}\ln\frac{1}{\delta}\left[\ln\mathcal{N}_{\mathcal{F}}\left(\frac{\epsilon_1}{4L}\right) + \ln\mathcal{N}_{\mathcal{H}_{\mathcal{D}_X, L_r}(\epsilon_r+\epsilon_0)}\left(\frac{\epsilon_1}{4L}\right)\right]$ for some absolute constant C. In particular, with probability at least $1-\delta$, the hypotheses $f \in \mathcal{F}, h \in \mathcal{H}$ with $L_c(f, h; S) = 0$ and $L_r(h, g; U) \leq \epsilon_r + \epsilon_0$ for some $g \in \mathcal{G}$ satisfy $L_c(f, h; \mathcal{D}) \leq \epsilon_1$.*

Proof. First, we show that with m_u unlabeled examples, by a covering argument over $\mathcal{H}$ and $\mathcal{G}$ (see, e.g., [55]), it is guaranteed that with probability $1-\delta/2$, all $h \in \mathcal{H}$ and $g \in \mathcal{G}$ satisfy $|L_r(h, g; U) - L_r(h, g; \mathcal{D}_X)| \leq \epsilon_0$. More precisely, let $\mathcal{C}_{\mathcal{G}}\left(\frac{\epsilon_0}{4L}\right)$ be a $\frac{\epsilon_0}{4L}$-covering of $\mathcal{G}$, and $\mathcal{C}_{\mathcal{H}}\left(\frac{\epsilon_0}{4L}\right)$ be a $\frac{\epsilon_0}{4L}$-covering of $\mathcal{H}$. Then by the union bound, all $h' \in \mathcal{C}_{\mathcal{H}}\left(\frac{\epsilon_0}{4L}\right)$ and $g' \in \mathcal{C}_{\mathcal{G}}\left(\frac{\epsilon_0}{4L}\right)$ satisfy $|L_r(h', g'; U) - L_r(h', g'; \mathcal{D}_X)| \leq \epsilon_0/4$. Then the claim follows from the definition of the coverings and the Lipschitzness of the losses.

By the claim, we have $L_r(h^*, g^*; U) \leq L_r(h^*, g^*; \mathcal{D}_X) + \epsilon_0 \leq \epsilon_r + \epsilon_0$. So $h^* \in \mathcal{H}_{\mathcal{D}_X, L_r}(\epsilon_r + \epsilon_0)$, and thus the optimal value $L_c(f, h; S)$ subject to $L_r(h, g; U) \leq \epsilon_r + \epsilon_0$ for some $g \in \mathcal{G}$ is 0. On the other hand, again by a covering argument over $\mathcal{H}$ and $\mathcal{F}$, with probability at least $1-\delta/2$, for all $h \in \mathcal{H}_{\mathcal{D}_X, L_r}(\epsilon_r + \epsilon_0)$ and all $f \in \mathcal{F}$, only those with $L_c(f, h; \mathcal{D}) \leq \epsilon_1$ can have $L_c(f, h; S) = 0$. The theorem statement then follows. □

The theorem shows that when the optimal regularization loss is not zero but $\epsilon_r > 0$, one needs to do the learning subject to $L_r(h; U) \leq \epsilon_r + \epsilon_0$ and the unlabeled sample complexity has a dependence on ϵ_0 by $\frac{1}{\epsilon_0^2}$, instead of $\frac{1}{\epsilon_0}$.

We are now ready to present the result for the setting where both the optimal prediction and regularization losses are non-zero.

Theorem 2. *Suppose there exist $h^* \in \mathcal{H}, f^* \in \mathcal{F}, g^* \in \mathcal{G}$ such that $L_c(f^*, h^*; \mathcal{D}) \leq \epsilon_c$ and $L_r(h^*, g^*; \mathcal{D}_X) \leq \epsilon_r$. For any $\epsilon_0, \epsilon_1 \in (0, 1/2)$, a set U of m_u unlabeled examples and a set S of m_l labeled examples are sufficient to learn to an error $\epsilon_c + \epsilon_1$ with probability $1-\delta$, where $m_u \geq \frac{C}{\epsilon_0^2}\ln\frac{1}{\delta}\left[\ln\mathcal{N}_{\mathcal{G}}\left(\frac{\epsilon_0}{4L}\right) + \ln\mathcal{N}_{\mathcal{H}}\left(\frac{\epsilon_0}{4L}\right)\right]$, $m_l \geq \frac{C}{\epsilon_1^2}\ln\frac{1}{\delta}\left[\ln\mathcal{N}_{\mathcal{F}}\left(\frac{\epsilon_1}{4L}\right) + \ln\mathcal{N}_{\mathcal{H}_{\mathcal{D}_X, L_r}(\epsilon_r+2\epsilon_0)}\left(\frac{\epsilon_1}{4L}\right)\right]$ for some absolute constant C. In particular, with probability at least $1-\delta$, the $h \in \mathcal{H}, f \in \mathcal{F}$ that optimize $L_c(f, h; S)$ subject to $L_r(h; U) \leq \epsilon_r + \epsilon_0$ have $L_c(f, h; \mathcal{D}) \leq L_c(f^*, h^*; \mathcal{D}) + \epsilon_1$.*

Proof. With m_u unlabeled examples, by a standard covering argument, it is guaranteed that with probability $1-\delta/4$, all $h \in \mathcal{H}$ and $g \in \mathcal{G}$ satisfy $|L_r(h, g; U) - L_r(h, g; \mathcal{D}_X)| \leq \epsilon_0$. In particular, $L_r(h^*, g^*; U) \leq L_r(h^*, g^*; \mathcal{D}_X) + \epsilon_0 \leq \epsilon_r + \epsilon_0$. Then again by a covering argument, the labeled sample size m_l implies that with probability at least $1-\delta/2$, all hypotheses $h \in \mathcal{H}_{\mathcal{D}_X, L_r}(\epsilon_r + 2\epsilon_0)$ and all $f \in \mathcal{F}$ have $L_c(f, h; S) \leq L_c(f, h; \mathcal{D}) + \epsilon_1/2$. Finally, by using Hoeffding's bounds, with probability at least $1-\delta/4$, we have

$$L_c(f^*, h^*; S) \leq L_c(f^*, h^*; \mathcal{D}) + \mathcal{O}\left(\sqrt{\frac{1}{m_l}\ln\frac{1}{\delta}}\right) \leq L_c(f^*, h^*; \mathcal{D}) + \epsilon_1/2.$$

Therefore, with a probability of at least $1-\delta$, the hypotheses $f \in \mathcal{F}, h \in \mathcal{H}$ that optimizes $L_c(f, h; S)$ subject to $L_r(h, g; U) \leq \epsilon_r + \epsilon_0$ for some $g \in \mathcal{G}$ have the following guarantee. First, since $L_r(h, g; U) \leq \epsilon_r + \epsilon_0$, we have $L_r(h, g; \mathcal{D}_X) \leq \epsilon_r + 2\epsilon_0$, and thus $h \in \mathcal{H}_{\mathcal{D}_X, L_r}(\epsilon_r + 2\epsilon_0)$. Then we have

$$\begin{aligned}
L_c(f, h; \mathcal{D}) &\leq L_c(f, h; S) + \epsilon_1/2 && (8)\\
&\leq L_c(f^*, h^*; S) + \epsilon_1/2 && (9)\\
&\leq L_c(f^*, h^*; \mathcal{D}) + \mathcal{O}\left(\sqrt{\frac{1}{m_l}\ln\frac{1}{\delta}}\right) + \epsilon_1/2 && (10)\\
&\leq L_c(f^*, h^*; \mathcal{D}) + \epsilon_1. && (11)
\end{aligned}$$

This completes the proof of the theorem. □

The above analysis also holds with some other capacity measure of the hypothesis classes, like the VC-dimension or Rademacher complexity. We give an example for using the VC-dimension

(assuming the prediction task is a classification task). The proof follows similarly to Theorem 2, but using the VC-dimension bound instead of the ϵ-net argument.

Theorem 4. *Suppose there exist $h^* \in \mathcal{H}, f^* \in \mathcal{F}, g^* \in \mathcal{G}$ such that $L_c(f^*, h^*; \mathcal{D}) \leq \epsilon_c$ and $L_r(h^*, g^*; \mathcal{D}_X) \leq \epsilon_r$. For any $\epsilon_0, \epsilon_1 \in (0, 1/2)$, a set U of m_u unlabeled examples and a set S of m_l labeled examples are sufficient to learn to an error $\epsilon_c + \epsilon_1$ with probability $1-\delta$, where*

$$m_u \geq \frac{C}{\epsilon_0^2}\left[d(\mathcal{G} \circ \mathcal{H}) \ln \frac{1}{\epsilon_0} + \ln \frac{1}{\delta}\right], \quad m_l \geq \frac{C}{\epsilon_1^2}\left[d(\mathcal{F} \circ \mathcal{H}_{\mathcal{D}_X, L_r}(\epsilon_r + 2\epsilon_0)) \ln \frac{1}{\epsilon_1} + \ln \frac{1}{\delta}\right] \tag{12}$$

for some absolute constant C. In particular, with probability at least $1-\delta$, the hypotheses $f \in \mathcal{F}, h \in \mathcal{H}$ that optimize $L_c(f, h; S)$ subject to $L_r(h; U) \leq \epsilon_r + \epsilon_0$ satisfy $L_c(f, h; \mathcal{D}) \leq L_c(f^*, h^*; \mathcal{D}) + \epsilon_1$.

C.3 Different Domains, Unrealizable, Infinite Hypothesis Classes

In practice, it is often the case that the unlabeled data is from a different domain than the labeled data. For example, state-of-the-art NLP systems are often trained on a large general unlabeled corpus (e.g., the entire Wikipedia) and a small specific labeled corpus (e.g., a set of medical records). That is, the unlabeled data U is from a distribution $\mathcal{U}_X$ different from $\mathcal{D}_X$, the marginal distribution of x in the labeled data. In this setting, we show that our previous analysis still holds.

Proof. The proof follows that for the setting with the same distribution for input feature vectors in the labeled data and unlabeled data; here we only mention the proof steps involving $\mathcal{U}_X$.

Even when the unlabeled data is from a different distribution $\mathcal{U}_X$, we still have that with probability $1 - \delta/4$, all $h \in \mathcal{H}$ and $g \in \mathcal{G}$ satisfy $|L_r(h, g; U) - L_r(h, g; \mathcal{U}_X)| \leq \epsilon_0$ for the given value of m_u. In particular, $L_r(h^*, g^*; U) \leq L_r(h^*, g^*; \mathcal{U}_X) + \epsilon_0 \leq \epsilon_r + \epsilon_0$. Then the labeled sample size m_l implies that with probability at least $1 - \delta/2$, all hypotheses $h \in \mathcal{H}_{\mathcal{U}_X, L_r}(\epsilon_r + 2\epsilon_0)$ and all $f \in \mathcal{F}$ have $L_c(f, h; S) \leq L_c(f, h; \mathcal{D}) + \epsilon_1/2$. Also, for any h, g with $L_r(h, g; U) \leq \epsilon_r + \epsilon_0$, we have $L_r(h, g; \mathcal{U}_X) \leq \epsilon_r + 2\epsilon_0$, and thus $h \in \mathcal{H}_{\mathcal{D}_X, L_r}(\epsilon_r + 2\epsilon_0)$. The rest of the proof follows that of Theorem 2. □

The bound is similar to that in the setting of the domain distributions being same. It implies that unlabeled data from a different domain than the labeled data can help in learning the target task, as long as there exists a "ground-truth" representation function h^*, which is shared across the two domains, having a small prediction loss on the labeled data and a suitable regularization loss on the unlabeled data. The former (small prediction loss) is typically assumed according to domain knowledge, e.g., for image data, common visual perception features are believed to be shared across different types of images. The latter (suitable regularization loss) means only a small fraction of $h \in \mathcal{H}$ have a smaller (or similar) regularization loss than h^*, which requires a careful design of $\mathcal{G}$ and L_r.

Remarks We would like to briefly comment on interpreting the reduction in sample complexity of labeled data when using functional regularization in our bounds. The sample complexity bounds are *upper bounds* and are aimed at aiding quantitative analysis by bounding the actual sample size needed for learning (under assumptions on the data and the hypothesis class). However, there exist settings where these bounds are nearly tighter mathematically (e.g., the standard lower bound via VC-dimension). More precisely, there exist hypothesis classes, such that for any learning algorithm, there exists a data distribution and a target function such that a sample, equal in size to the upper bound up to logarithmic factors, is required for learning (a more precise statement can be found in [36]). Additionally, these bounds usually do not take into account the effect of optimization [59].

While these upper bounds are not an exact quantification, they usually align well with the sample size needed for learning in practice, thereby providing useful insights. The reduction in our bounds on using functional regularization can roughly estimate the actual reduction in practice. Further this can provide useful theoretical insights such as the regularization restricting the learning to a subset of the hypothesis class of representation functions. The findings from our experiments in Section E indeed correlate well with the theory. Similar to prior sample complexity studies, we believe our sample complexity bounds can prove to be a useful analysis tool.

C.4 Discussions

When is functional regularization not helpful? Our theorems and analysis also provide implications for when the auxiliary self-supervised task may *not* help the prediction task. First, the regularization may not significantly reduce the hypothesis class. For example, consider Theorem 1, if $\mathcal{H}_{\mathcal{D}_X,L_r}(\epsilon_0)$ is not significantly smaller than $\mathcal{H}$, then using unlabeled data will not reduce the sample size of the labeled data much compared to only using the labeled data for prediction. In fact, the size of the regularized hypothesis class $\mathcal{H}_{\mathcal{D}_X,L_r}(\epsilon_0)$ needs to be exponentially small compare to that of the whole class $\mathcal{H}$ to get significant gain; a polynomially smaller $\mathcal{H}_{\mathcal{D}_X,L_r}(\epsilon_0)$ only leads to minor logarithmic reduction in the sample complexity. Section 4 presents two concrete examples where the regularized hypothesis class is exponentially small and leads to significant gain, but it is also to find fail examples. Second, the regularization can fail if the regularization loss threshold (τ) is not set properly. For example, if τ is set too small, then the feasible set may contain no hypothesis with a small prediction loss. Finally, another possible reason that these representation learning approaches can fail is that the optimization does not provide a good solution. It is an interesting future direction to analyze the optimization in function regularization approaches.

Is uniform convergence suitable for the analysis? Our analysis is based on uniform convergence. Careful readers may question if uniform convergence will be suitable for analyzing the generalization, since there is evidence that naïvely applying uniform convergence bounds may not result in good generalization/sample bounds for deep learning (e.g., [59, 37]). However, these existing studies are not for the setting in this work. To the best of our knowledge, they are for supervised learning without the auxiliary representation learning tasks, while our setting is with the auxiliary tasks. The difference in the settings is the key in making uniform convergence bounds meaningful. More precisely, in supervised deep learning without auxiliary tasks, it is generally believed that the hypothesis class is larger than statistically necessary, and the optimization has an implicit regularization on the training, and hence uniform convergence fails to explain the generalization (e.g., [38, 32]). However, in our setting with the auxiliary tasks, functional regularization has a regularization effect of restricting the learning to a smaller subset of the hypothesis space, as shown by our analysis below and supported by empirical evidence in existing work (e.g., [19]) and our experiments in Section E and Section F. Once regularized to a smaller subset of hypotheses, the implicit regularization of the optimization is no longer significant, and thus the generalization can be explained by uniform convergence. More thorough investigation is left for future work.

D Applying the Theoretical Framework to Concrete Examples

D.1 Auto-encoder

We first recall the details of the example: $\mathcal{H}$ is the class of linear functions from $\mathbb{R}^d$ to $\mathbb{R}^r$ where $r < d/2$, and $\mathcal{F}$ to be the class of linear functions over some activations. That is,

$$z = h_W(x) = Wx,\ \ y = f_a(z) = \sum_{i=1}^{r} a_i \sigma(z_i)\ ,\ \text{where } W{\in}\mathbb{R}^{r\times d},\ \ a{\in}\mathbb{R}^r$$

Here $\sigma(t)$ is an activation function, the rows of W and a have ℓ_2 norm bounded by 1. We consider the Mean Square Error (MSE) prediction loss, i.e., $L_c(f,h;x){=}\|y{-}f(h(x))\|_2^2$.

Also recall that we assume the data distribution having the following property: let the columns of $B \in \mathbb{R}^{d\times d}$ be the eigenvectors of $\Sigma{:=}\mathbb{E}[xx^\top]$, then the labels are largely determined by the signal in the first r directions: $y{=}(a^*)^\top z^*{+}\nu$ and $z^* = B_{1:r}^\top x$, where a^* is a ground-truth parameter with $\|a^*\|_2{\le}1$, $B_{1:r}$ is the set of first r eigenvectors of Σ, and ν is a small Gaussian noise. We also assume that the r^{th} and $r{+}1^{\text{th}}$ eigenvalues of Σ are different so that the corresponding eigenvectors can be distinguished. Let ϵ_r denote $\mathbb{E}\|x - B_{1:r}B_{1:r}^\top x\|_2^2$.

Finally, we recall that the functional regularization $\mathcal{G}$ we used is given by the class of linear functions from $\mathbb{R}^r$ to $\mathbb{R}^d$, i.e., $\hat{x}{=}g_V(z){=}Vz$ where $V{\in}\mathcal{R}^{d\times r}$ with orthonormal columns. The regularization loss $L_r(h,g;x){=}\|x{-}g(h(x))\|_2^2$.

For simplicity of analysis, we assume access to infinite unlabeled data, and set the threshold $\tau = \epsilon_r$. Strictly speaking, we need to allow $L_r(h,g;\mathcal{D}_X) \le \epsilon_r + \epsilon$ for a small $\epsilon > 0$ due to finite unlabeled data. A similar but more complex argument holds for that case. Here we assume infinite unlabeled

data to simplify the presentation and better illustrate the intuition, since our focus is on the labeled data.

Formally, we calculate the sample complexity bounds in the limit $m_u \to +\infty$. Equivalently we consider the learning problem:

$$\min_{f\in\mathcal{F},h\in\mathcal{H}} L_c(f,h\,;S),\ \text{ s.t. } L_r(h\,;\mathcal{D}_X)\le\epsilon_r. \tag{13}$$

Let $\mathcal{N}_{\mathcal{C}}(\epsilon)$ denote the ϵ-covering number of a class $\mathcal{C}$ w.r.t. the ℓ_2 norm (i.e., Euclidean norm for the weight vector a, and Frobenius norm for the weight matrices W and V). Let L denote the Lipschitz constant of the losses (See Appendix C.2). Without regularization, the standard ϵ-net argument shows that the labeled sample complexity, for an error ϵ close to the optimal, is $\frac{C}{\epsilon^2}\left[\ln\mathcal{N}_{\mathcal{F}}\left(\frac{\epsilon}{4L}\right)+\ln\mathcal{N}_{\mathcal{H}}\left(\frac{\epsilon}{4L}\right)\right]$ for some absolute constant C. Applying our framework when using regularization, the sample complexity is bounded by $\frac{C}{\epsilon^2}\left[\ln\mathcal{N}_{\mathcal{F}}\left(\frac{\epsilon}{4L}\right)+\ln\mathcal{N}_{\mathcal{H}_{\mathcal{D}_X,L_r(\epsilon_r)}}\left(\frac{\epsilon}{4L}\right)\right]$. To quantify the reduction in the bound, we show the following lemma.

Lemma 5. *For $\epsilon/4L<1/2$,*

$$\mathcal{N}_{\mathcal{H}}\left(\frac{\epsilon}{4L}\right)\ge\binom{d-r}{r}\mathcal{N}_{\mathcal{H}_{\mathcal{D}_X,L_r(\epsilon_r)}}\left(\frac{\epsilon}{4L}\right). \tag{14}$$

Proof. First, recall that the regularization loss is

$$\begin{aligned} L_r(h,g;\mathcal{D}_X)&=\mathbb{E}_x\|x-g(h(x))\|_2^2\\ &=\mathbb{E}_x\|x-VWx\|_2^2 \end{aligned} \tag{15}$$

which is the r-rank approximation of the data. So in the optimal solution, the columns of V and the rows of W should span the subspace of the top r eigenvectors Σ. More precisely,

$$\begin{aligned} L_r(h,g;\mathcal{D}_X)&=\mathbb{E}_x[x^\top(I-VW)^\top(I-VW)x]\\ &=\mathbb{E}_x[\mathrm{trace}(x^\top(I-VW)^\top(I-VW)x)]\\ &=\mathbb{E}_x[\mathrm{trace}((I-VW)^\top(I-VW)xx^\top)]\\ &=\mathrm{trace}((I-VW)^\top(I-VW)\Sigma).\\ &=\mathrm{trace}((I-VW)\Sigma). \end{aligned} \tag{16}$$

Since V and W are orthonormal and have rank r, the optimal VW should span the subspace of the top r eigenvectors of Σ and the optimal loss is given by ϵ_r.[2] Since the r-th and $r+1$-th eigenvalues of Σ are different, the optimal VW is unique, and thus we have

$$\mathcal{H}_{\mathcal{D}_X,L_r}(\epsilon_r)=\{OB_{1:r}^\top : O\in\mathbb{R}^{r\times r}, O \text{ is orthonormal}\}.$$

On the other hand, if B_S refers to the sub-matrix of columns in B having indices in S, then clearly,

$$\mathcal{H}\supseteq\mathcal{H}_S:=\{OB_S^\top : O\in\mathbb{R}^{r\times r}, O \text{ is orthonormal}\},$$

for any $S\subseteq\{r+1,r+2,\ldots,d\}, |S|=r$. By symmetry, $\mathcal{N}_{\mathcal{H}_S}(\epsilon')=\mathcal{N}_{\mathcal{H}_{\mathcal{D}_X,L_r(\epsilon_r)}}(\epsilon')$ for any $\epsilon'>0$. Now it is sufficient to prove that $\mathcal{H}_S$ and $\mathcal{H}_{S'}$ are sufficiently far away for different S and S'. This is indeed the case, since $\|OB_S^\top-O'B_{S'}^\top\|_F^2>1$ for any orthonormal O and O':

$$\|OB_S^\top-O'B_{S'}^\top\|_F^2=\mathrm{trace}\left((OB_S^\top-O'B_{S'}^\top)^\top(OB_S^\top-O'B_{S'}^\top)\right) \tag{17}$$

$$\begin{aligned} &=\mathrm{trace}\left((OB_S^\top)^\top(OB_S^\top)\right)+\mathrm{trace}\left((O'B_{S'}^\top)^\top(O'B_{S'}^\top)\right)\\ &\quad-\mathrm{trace}\left((OB_S^\top)^\top(O'B_{S'}^\top)\right)-\mathrm{trace}\left((O'B_{S'}^\top)^\top(OB_S^\top)\right) \end{aligned} \tag{18}$$

$$\begin{aligned} &=\|OB_S^\top\|_F^2+\|O'B_{S'}^\top\|_F^2\\ &\quad-\mathrm{trace}\left((O'B_{S'}^\top)(OB_S^\top)^\top\right)-\mathrm{trace}\left((OB_S^\top)(O'B_{S'}^\top)^\top\right) \end{aligned} \tag{19}$$

$$=\|B_S^\top\|_F^2+\|B_{S'}\|_F^2-\mathrm{trace}\left(B_{S'}^\top B_S\right)-\mathrm{trace}\left(B_S^\top B_{S'}\right) \tag{20}$$

$$\ge r+r-(r-1)-(r-1)=2. \tag{21}$$

This completes the proof. □

[2]The optimal product of V and W should span the subspace of the top r eigenvectors of Σ. But note that there are different pairs of V and W which can achieve the same product.

D.2 Masked Self-supervision

Learning Without Functional Regularization. Let $\mathcal{H}$ be linear functions from $\mathbb{R}^d$ to $\mathbb{R}^r$ where $r < (d-1)/2$ followed by a quadratic activation, and $\mathcal{F}$ be linear functions from $\mathbb{R}^r$ to $\mathbb{R}$. That is,

$$\phi = h_W(x) = [\sigma(w_1^\top x), \ldots, \sigma(w_r^\top x)] \in \mathbb{R}^r \ , \ y = f_a(\phi) = a^\top \phi, \text{ where } w_i \in \mathbb{R}^d \ , a \in \mathbb{R}^r. \quad (22)$$

Here $\sigma(t)=t^2$ for $t \in \mathbb{R}$ is the quadratic activation function. W.l.o.g, we assume that w_i and a have ℓ_2 norm bounded by 1. Without prior knowledge on the data, no functional regularization corresponds to end-to-end training on $\mathcal{F} \circ \mathcal{H}$.

Data Property. We consider the setting where the data point x satisfies $x_1 = \sum_{i=1}^r ((u_i^*)^\top x_{2:d})^2$, where $x_{2:d} = [x_2, x_3, \ldots, x_d]$ and u_i^* is the i-th eigenvector of $\Sigma := \mathbb{E}[x_{2:d} x_{2:d}^\top]$. Furthermore, the label y is given by $y = \sum_{i=1}^r a_i^* \sigma((u_i^*)^\top x_{2:d}) + \nu$ for some $\|a^*\|_2 \leq 1$ and a small Gaussian noise ν. We also assume a difference in the r^{th} and $r{+}1^{\text{th}}$ eigenvalues of Σ.

Learning With Functional Regularization. Suppose we have prior knowledge that $x_1 = \sum_{i=1}^r (u_i^\top x_{2:d})^2$ and $y = \sum_{i=1}^r a_i \sigma(u_i^\top x_{2:d})$ for some vectors $u_i \in \mathbb{R}^{d-1}$ and an a with $\|a\|_2 \leq 1$. Based on this, we perform masked self-supervision by constraining the first coordinate of w_i to be 0 for h, and choosing the regularization function $g(\phi) = \sum_{i=1}^r \phi_i$ and the regularization loss $L_r(h, g; x) = (x_1 - g(h_W(x)))^2$. Again for simplicity, we assume access to infinite unlabeled data and set the regularization loss threshold $\tau = 0$.

On applying our framework, we get that functional regularization can reduce the labeled sample bound by $\frac{C}{\epsilon^2}\left[\ln \mathcal{N}_{\mathcal{H}}\left(\frac{\epsilon}{4L}\right) - \ln \mathcal{N}_{\mathcal{H}_{\mathcal{D}_X, L_r(0)}}\left(\frac{\epsilon}{4L}\right)\right]$ for some absolute constant C. We first present Lemma 6 along with its proof below.

Lemma 6. *For $\epsilon/4L < 1/2$,*

$$\mathcal{N}_{\mathcal{H}}\left(\frac{\epsilon}{4L}\right) \geq \binom{d-1-r}{r} \mathcal{N}_{\mathcal{H}_{\mathcal{D}_X, L_r(0)}}\left(\frac{\epsilon}{4L}\right). \quad (23)$$

Proof. By definition,

$$\mathbb{E}\left[L_r(h, g; x)\right] = \mathbb{E}\left[\sum_{i=1}^r u_i^\top x_{2:d} x_{2:d}^\top u_i - \sum_{i=1}^r (u_i^*)^\top x_{2:d} x_{2:d}^\top u_i^*\right]^2 \quad (24)$$

$$\geq \left(\mathbb{E}\left|\sum_{i=1}^r u_i^\top x_{2:d} x_{2:d}^\top u_i - \sum_{i=1}^r (u_i^*)^\top x_{2:d} x_{2:d}^\top u_i^*\right|\right)^2 \quad (25)$$

$$\geq \left|\mathbb{E}\sum_{i=1}^r u_i^\top x_{2:d} x_{2:d}^\top u_i - \mathbb{E}\sum_{i=1}^r (u_i^*)^\top x_{2:d} x_{2:d}^\top u_i^*\right|^2, \quad (26)$$

$$\geq \left|\sum_{i=1}^r u_i^\top \Sigma u_i - \sum_{i=1}^r (u_i^*)^\top \Sigma u_i^*\right|^2. \quad (27)$$

Therefore, $\mathbb{E}\left[L_r(h, g; x)\right] = 0$ if and only if $u_1, \ldots, u_r$ span the same subspace as $u_1^*, \ldots, u_r^*$, i.e.,

$$\mathcal{H}_{\mathcal{D}_X, L_r}(0) = \{h_W(x) : w_i = [0, u_i], [u_1, \ldots, u_r]^\top = O[u_1^*, \ldots, u_r^*]^\top, O \in \mathbb{R}^{r \times r}, O \text{ is orthonormal}\}.$$

On the other hand, if $u_1^*, u_2^*, \ldots, u_{d-1}^*$ are the eigenvectors of Σ, and $U_I^* := [u_{i_1}^*, \ldots, u_{i_r}^*]^\top$ for indices $I = \{i_1, i_2, \ldots, i_r\} \subseteq \{r+1, r+2, \ldots, d-1\}$, then clearly

$$\mathcal{H} \subseteq \mathcal{H}_I := \{h_W(x) : w_i = [0, u_i], [u_1, \ldots, u_r]^\top = O(U_I^*)^\top, O \in \mathbb{R}^{r \times r}, O \text{ is orthonormal}\}$$

for any $I = \{i_1, i_2, \ldots, i_r\} \subseteq \{r+1, r+2, \ldots, d-1\}$. By symmetry, $\mathcal{N}_{\mathcal{H}_I}(\epsilon') = \mathcal{N}_{\mathcal{H}_{\mathcal{D}_X, L_r(0)}}(\epsilon')$ for any $\epsilon' > 0$. Using an argument similar to Section D.1, we can show that for two different index sets I and I', any hypothesis in $\mathcal{H}_I$ and any hypothesis in $\mathcal{H}_{I'}$ cannot be covered by the same ball in any ϵ'-cover with $\epsilon' < 1/2$. This completes the proof. □

Using Lemma 6, we can derive the following using:

$$\mathcal{H}_{\mathcal{D}_X, L_r}(0) = \{h_W(x) : w_i = [0, u_i], [u_1, \ldots, u_r]^\top = O[u_1^*, \ldots, u_r^*]^\top, O \in \mathbb{R}^{r \times r}, O \text{ is orthonormal}\}$$

Using this we can show that the reduction of the sample bound is $\frac{C}{\epsilon^2}\ln\binom{d-1-r}{r}$, i.e., a reduction in the error bound by $\frac{C}{\sqrt{m_\ell}}\ln\binom{d-1-r}{r}$ when using m_ℓ labeled data. We also note that this reduction depends on r but has little dependence on d.

E Experiments on Concrete Functional Regularization Examples

While there is abundant empirical evidence on the benefits of auxiliary tasks in various applications, our framework allows mathematical analysis and we can get non-trivial implications from the two examples. Therefore, here we focus on experimentally verifying the following implications for the two examples: 1) the reduction in prediction error (between end-to-end training and functional regularization) using the same labeled data; 2) the reduction in prediction error on varying a property of the data and hypotheses (specifically, varying parameter r); 3) the reduction in prediction error is obtained due to pruning the hypothesis class.

E.1 Auto-Encoder

Data: We first generate d orthonormal vectors($\{u_i\}_{i=1}^{i=d}$) in $\mathbb{R}^d$. We then randomly generate means μ_i and variances σ_i corresponding to each principal component $i \in [1, d]$ such that $\sigma_1 > \cdots > \sigma_r \gg \sigma_{r+1} > \cdots > \sigma_d$. The μ_i's are randomly generated integers in $[0, 20]$ and the variances σ_i , $i \in [1, r]$ are each generated randomly from $[1, 10]$ and σ_i , $i \in [r+1, d]$ are each generated randomly from $[0, 0.1]$. We also generate a vector $a \in \mathbb{R}^r$ randomly such that $||a||_2 \leq 1$. To generate a data point (x, y), we sample $\alpha_i \sim \mathcal{N}(\mu_i, \sigma_i)\ \forall i \in [1, d]$ and set $x = \sum_{i=1}^{d} \alpha_i u_i$ and $y = \sum_{i=1}^{r} a_i {\alpha_i}^2 + \nu$ where $\nu \sim \mathcal{N}(0, 10^{-2})$. We use an unlabeled dataset of 10^4 points (when using the auto-encoder functional regularization), a labeled training set of 10^4 points and a labeled test set of 10^3 points.

Models: h_W corresponds to a fully connected NN, without any activation function, to transform $x \in \mathbb{R}^d$ to its representation $h(s) \in \mathbb{R}^r$. For prediction on the target task, we use a linear classifier after a quadratic activation on $h(x)$ to obtain a scalar output $\hat{y}$. For functional regularization g_V, we use a fully connected NN to transform the representation $h(s) \in \mathbb{R}^r$ to reconstruct the input back $\hat{x} \in \mathbb{R}^d$. Our example additionally constrains V, W to be orthonormal. For achieving this, we enforce an orthonormal regularization [10, 5] penalty for each V, W weighted by hyper-parameters λ_1 and λ_2 respectively during the auto-encoder reconstruction. For a matrix $M \in R^{a \times b}$, the orthonormal regularization penalty to ensure that the rows of M are orthonormal, is given by $\sum_{ij} |(MM^\top)_{ij} - I_{ij}|$ where $\sum_{ij}$ is summing over all the matrix elements, I is the identity matrix in $R^{a \times a}$.

Training Details: For end-to-end training, we train the predictor and h jointly using a cross entropy loss between the predicted target $\hat{y}$ and the true y on the labeled training data set. For functional regularization, we first train h and g using the MSE loss between the reconstructed input $\hat{x}$ and the original input x over the unlabeled data set. Here, we also add the orthonormal regularization penalties. We tune the weights λ_1 and λ_2 using grid search in $[10^{-3}, 10^3]$ in multiplicative steps of 10 to get the best reconstruction (least MSE) on the training data inputs. Now using h initialized from the auto-encoder, we use the labeled training data set to jointly learn the predictor and h using a cross entropy loss between the predicted target $\hat{y}$ and the true y. We report the MSE on the test set as the metric. For all optimization steps we use an SGD optimizer with momentum set to 0.9 where the learning rate is tuned using grid search in $[10^{-5}, 10^{-1}]$ in multiplicative steps of 10. We set the data dimension $d = 100$ and report the test MSE averaged over 10 runs.

t-SNE Plots of Functional Approximations To get a functional approximation from a model, we compute and concatenate the output predictions $\hat{y}$ from the model over the test data set. For every model, we obtain a $\mathbb{R}^{1000}$ vector corresponding to the size of the test set. We perform 1000 independent runs for each model (with and without functional regularization) obtaining $2,000$ functional approximation vectors in $\mathbb{R}^{1000}$. We visualise these vectors in 2D using the t-SNE [54] algorithm.

E.2 Masked Self-Supervision

Data: We first generate $d-1$ orthonormal vectors($\{u_i\}_{i=2}^{i=d}$) in $\mathbb{R}^{d-1}$. We then randomly generate means μ_i and variances σ_i corresponding to each principal component $i \in [2, d]$ such that $\sigma_2 > \cdots > \sigma_{r+1} \gg \sigma_{r+2} > \cdots > \sigma_d$. The μ_i's are randomly generated integers in $[0, 20]$ and the

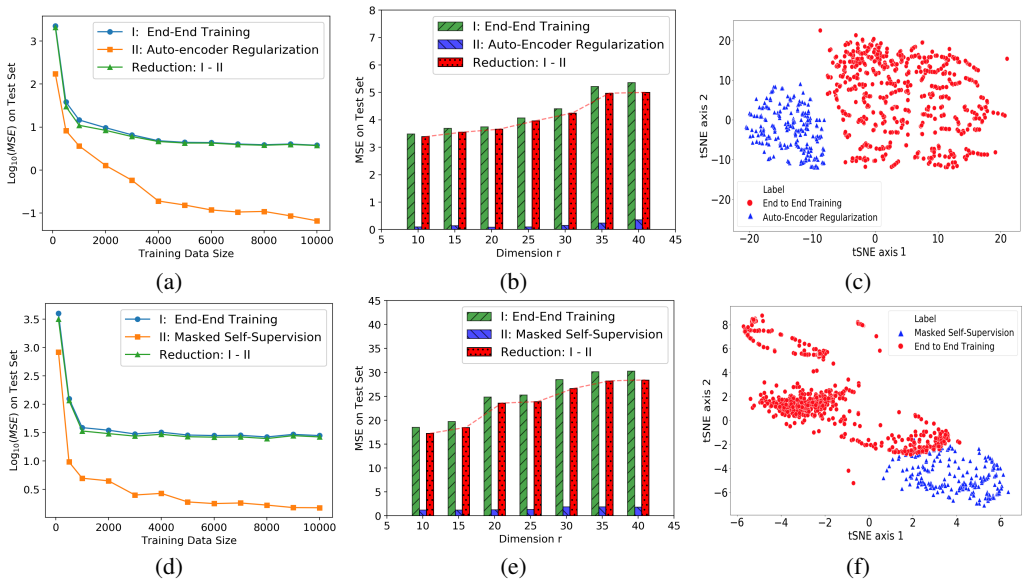


Figure 1: **Auto-Encoder:** 1(a) shows the variation of test MSE with labeled data size (here r=30), 1(b) shows this variation with the parameter r, and 1(c) shows the 2D visualization of the functional approximation using t-SNE. **Masked Self-Supervision:** 1(d), 1(e) and 1(f) show the same corresponding plots. Reduction refers to Test MSE of end-to-end training - Test MSE with regularization.

variances σ_i , $i \in [2, r+1]$ are each generated randomly from $[1, 10]$ and σ_i , $i \in [r+2, d]$ are each generated randomly from $[0, 0.1]$. We also generate a vector $a \in \mathbb{R}^r$ randomly such that $||a||_2 \leq 1$. To generate a data point (x, y), we sample $\alpha_i \sim \mathcal{N}(\mu_i, \sigma_i)$ $\forall i \in [2, d]$ and set $x_1 = \sum_{i=2}^{r+1} \alpha_i^2$, $x_{2:d} = \sum_{i=2}^{d} \alpha_i u_i$ and $y = \sum_{i=2}^{r+1} a_i \alpha_i^2 + \nu$ where $\nu \sim \mathcal{N}(0, 10^{-2})$. We use an unlabeled dataset of 10^4 points, a labeled training set of 10^4 points and a labeled test set of 10^3 points.

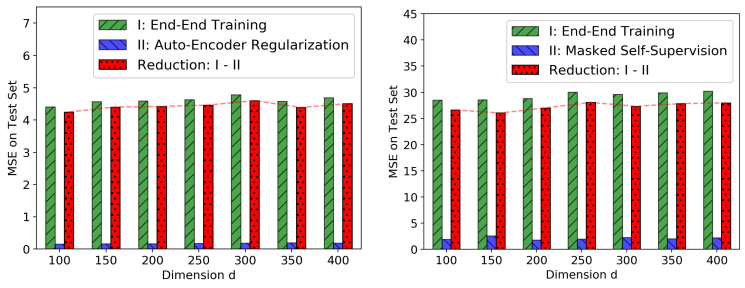
Models: h_W corresponds to a fully connected NN, using a quadratic activation function, to transform $x \in \mathbb{R}^d$ to its representation $h(s) \in \mathbb{R}^r$. For prediction on the target task we use a linear classifier to obtain the output $\hat{y}$ from the representation $h(x)$. For functional regularization, we sum the elements of the representation $h(x) \in \mathbb{R}^r$ to reconstruct the first input $\hat{x_1} \in \mathbb{R}$ back.

Training Details: For functional regularization, we mask the first dimension of the unlabeled data by setting it to 0 and train h using the MSE loss between the reconstructed $\hat{x_1}$ and the original input dimension x_1. Other experimental details remain similar to Section E.1.

Experimental details for the t-SNE plots of functional approximation remain similar to Section E.1.

E.3 Results

Figure 1(a) plots the Test MSE loss v.s. the size of the labeled data when $r = 30$. We observe that with the same labeled data size, functional regularization can significantly reduce the error compared to end-to-end training. Equivalently, it needs much fewer labeled samples to achieve the same error as end-to-end training (e.g., 500 v.s. 10,000 points). Also, the error without regularization does not decrease for sample sizes ≥ 2000 while it decreases with regularization, suggesting that the regularization can even help alleviate optimization difficulty. Figure 1(b) shows the effect of varying r (i.e., the dimension of the subspace containing signals for prediction). We observe that the reduction in the error increases roughly linearly with r and then grows slower, as predicted by our analysis. Figure 1(c) visualizes the prediction functions learned. It shows that when using the functional regularization, the learned functions stay in a small functional space, while they are scattered otherwise. This supports our intuition for the theoretical analysis. This result also interestingly suggests that pruning the representation hypothesis space via functional regularization translates to a compact functional space for the prediction, even through optimization. We make similar observations for the masked self-supervision example in Figure 1(d)-1(f), to support our analysis.



(a) Auto-Encoder (b) Masked Self-Supervision

Figure 2: Reduction in Test MSE (on using functional regularization with respect to end-to-end training) with dimension d. Here $r = 30$ and the Test MSE are normalized by the average test $\|x\|_2^2$.

Additionally, we plot the reduction in test MSE between end-to-end training and using functional regularization on varying d in Figure 2. Here we fix $r = 30$ and vary the data dimension d and present the test MSE scores normalized with the average norm $\|x\|_2^2$ over the test data. As per indications from our derived bounds, the reduction remains more or less constant on varying d.

F Additional Experiments on Functional Regularization

There have been several empirical studies verifying the benefits of functional regularization across different applications. Here we present empirical results showing the benefits of using functional regularization on a computer vision and natural language processing application.

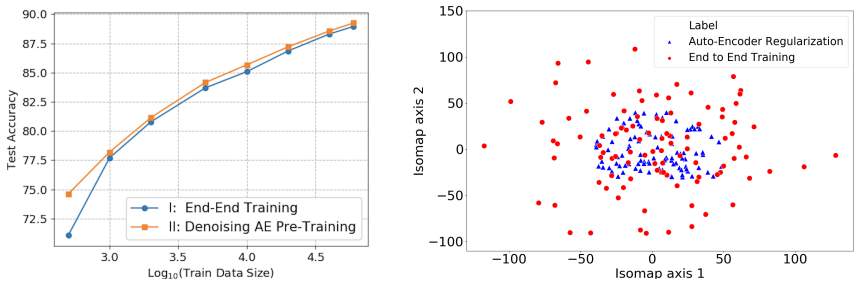
F.1 Image Classification

We consider the application of image classification using the Fashion MNIST dataset [57] which contains 28×28 gray-scale images of fashion products from 10 categories. This dataset has 60k images for training and 10k for testing. We consider a denoising auto-encoder functional regularization using unlabeled data and evaluate its benefits to supervised classification using labeled data.

Experimental Details We use a denoising auto-encoder as the functional regularization when learning from unlabeled data. The encoder consists of three fully connected layers with ReLU activations to obtain the input representation $h(x)$ of 1024 dimensions from an input x. The decoder consists of three fully connected layers with ReLU activations to reconstruct the 28×28 image $\hat{x}$ back from the 1024 dimensional representation $h(x)$. For training, the pixel values of x are normalized to $[0, 1]$ and independently corrupted by adding a Gaussian noise with mean 0 and standard deviation 0.2. The MSE loss between the x and $\hat{x}$ is used as the regularization loss L_r. Training is performed using the Adam optimizer with a learning rate of 3×10^{-4}. For classification, we use a simple linear layer which maps $h(x)$ to the class label $\hat{y}$. The classifier and the encoder are trained jointly using the cross entropy loss between $\hat{y}$ and the original label y. We compare the test set accuracy of 1) directly training the encoder and the target classifier using the labeled training data, and 2) pre-training the encoder using the de-noising auto-encoder functional regularization and then fine-tuning its weights along with the target classifier using the labeled training data. We vary the size of the labeled training data and plot the test accuracy averaged across 5 runs in Figure 3(a).

To visualize the impact of the denoising auto-encoder functional regularization, we follow the details in Appendix E.1 to get the functional approximation of the model. For each model, we obtain a $\mathbb{R}^{10000 \times 10}$ matrix with softmax values for 10 target classes for each of the 10000 test points. We perform 100 independent runs for each method (with and without the functional regularization) obtaining 200 functional approximation vectors in $\mathbb{R}^{100,000}$. We visualise these vectors in 2D using the Isomap [50] algorithm [3] in Figure 3(b).

[3] The t-SNE algorithm focuses more on neighbour distances by allowing large variance in large distances, while Isomap approximates geodesic distance via shortest paths thereby working well in practice with larger distances. Compared to the controlled data experiments where the functional approximation lies in $\mathbb{R}^{1000}$, the functional approximation for Fashion-MNIST lies in $\mathbb{R}^{100,000}$, thereby visualizing better via Isomap than t-SNE.



(a) (b)

Figure 3: Experimental results on **Fashion-MNIST**. (a) Test accuracy using de-noising auto-encoder functional regularization compared to end-to-end training on varying the size of labeled training data. (b) The 2D visualization of the functional approximation of 100 independent runs for each method.

Results From Figure 3(a), we observe that the test accuracy of end to end training is inferior to that of using functional regularization with unlabeled data across a variety of labeled data sizes. We observe that the difference in the test accuracy between the two methods is highest when the amount of labeled data available is small and the performance gap decreases as the amount of labeled data increases, as predicted by our theory.

Figure 3(b) visualizes the functional approximation learned by the model. It shows that when using the denoising auto-encoder functional regularization, the learned functions stay in a smaller functional space, while they are scattered when using end to end training. This is in line with our empirical observations on controlled data, and our intuition for the theoretical analysis: pruning the representation hypothesis space via functional regularization translates to a compact functional space.

F.2 Sentence Pair Classification

We consider the application of sentence pair classification using the Microsoft Research Paraphrase Corpus [15][4] which has sentence pairs with annotations of whether the two sentences are semantically equivalent. This dataset has approximately 3.7k and 1.7k sentence pairs in the train and test splits respectively. Here we specifically choose the MRPC dataset as it has a smaller size of labeled training data in comparison to most NLP datasets. To show the empirical benefits of using unlabeled data in addition to the limited train data available, we use a pre-trained BERT [13] language model. BERT, based on a transformer architecture, has been pre-trained using a masked token self-supervision task which involves masking a portion of the input sentence and using BERT to predict the masked tokens. This pre-training is done over a large text corpus ($\sim$ 2 billion words) and hence we can view the pre-trained BERT, under our framework, as having already pruned a large fraction of the hypothesis space of $\mathcal{H}$ for learning the representation on the input text.

Experimental Details We compare the performance of fine-tuning the pre-trained BERT with training a randomly initialised BERT from scratch. For the latter, we use three different loss formulations to further study the benefits of regularization on the text representation being learnt: (i) the Cross-Entropy loss $\mathcal{L}_{CE}$ on the predicted output (ii) $\mathcal{L}_{CE}$ along with a ℓ_1 norm penalty on the representation (i.e, the 768-dimensional representation from BERT corresponding to the [CLS] token) (iii) $\mathcal{L}_{CE}$ along with a ℓ_2 norm penalty on the representation. We refer to these three different loss formulations as Random, Random-ℓ_1 and Random-ℓ_2 respectively for notational simplicity. We want to study how varying the labeled data can impact the performance of different training methods. We present the results in Table 1. We use the 12-layer BERT Base uncased model for our experiments with an Adam optimizer having a learning rate $2e^{-5}$. We perform end to end training on the training data and tune the number of fine-tuning epochs. We report the accuracy and F1 scores as the metric on the test data averaged over 3 runs. When randomly initialising the weights of BERT, we use a standard normal distribution with mean 0 and standard deviation of 0.02 for the layer weights and set all the biases to zero vectors. We set the layer norms to have weights as a vector of ones with a zero vector as the bias. When adding the l_p penalty on the BERT representations on randomly initialising the weights, we choose an appropriate weighting function λ to make the training loss a sum of the

[4] `https://www.microsoft.com/en-us/download/details.aspx?id=52398`

Train Data Size	200	500	1000	2000	3668
BERT-FT	**68.1 / 80.6**	**71.0 / 80.6**	**72.7 / 81.8**	**74.9 / 82.4**	**80.3 / 85.7**
Random	64.1 / 74.8	64.7 / 75.66	67.0 / 80.1	68.9 / 79.0	68.9 / 79.3
Random-ℓ_1	54.7 / 65.1	62.6 / 75.5	63.6 / 76.7	63.4 / 76.6	66.3 / 79.6
Random-ℓ_2	65.3 / 78.6	66.4 / 79.7	65.3 / 78.6	65.0 / 78.4	66.5 / 79.9

Table 1: Performance of fine-tuning pre-trained BERT (BERT-FT) and end-to-end training of a randomly initialised BERT on varying the **MRPC** training dataset size. Metrics are reported in the format Accuracy/F1 scores on the test dataset. The training data size is 3668 sentence pairs.

cross entropy classification loss and λ times the l_p norm of the BERT representation. The λ is chosen $\in [10^{-3}, 10^3]$ by validation over a set of 300 data points randomly sampled from the training split. We use the huggingface transformers repository [5] for our experiments.

Results From the table, we observe that the performance of training BERT from pre-trained weights is better than the performance of training the BERT architecture from randomly initialised weights. When viewed under our framework, this empirically shows the benefits of using a learnable regularization function over fixed functions like the ℓ_1 or ℓ_2 norms of the representation.

On increasing the training data size, we observe that the performance of all the four training modes increases. However, we can see that the performance improvement of Random, Random-ℓ_1 and Random-ℓ_2 is marginal when compared to the improvement in BERT Fine-tuning. The latter can be attributed to the fact that the pre-trained weights of BERT are adjusted by specialising them towards the target data domain. To support this, in addition to Table 1, we also experimented by keeping the BERT weights fixed and only training the classifier. We observe that under such a setting, when we use a small training set, the model is unable to converge to a model different from the initialisation as similarly observed by [28]. This means that the learning indeed needs searching over a set of suitable hypotheses. Thus, we can conclude that unlabeled data helps in restricting the search space, and a small labeled data set can find a hypothesis suitable for the target domain data within the restricted search space, consistent with our analysis.

[5] `https://github.com/huggingface/transformers`