Predicting What You Already Know Helps: Provable Self-Supervised Learning

Jason D. Lee[1], Qi Lei[1], Nikunj Saunshi[1], Jiacheng Zhuo[2]

[1] Princeton University [2] University of Texas at Austin
{jasonlee@,qilei@,nsaunshi@cs}.princeton.edu, jzhuo@utexas.edu

Abstract

Self-supervised representation learning solves auxiliary prediction tasks (known as pretext tasks) without requiring labeled data to learn useful semantic representations. These pretext tasks are created solely using the input features, such as predicting a missing image patch, recovering the color channels of an image from context, or predicting missing words in text; yet predicting this *known* information helps in learning representations effective for downstream prediction tasks.
We posit a mechanism exploiting the statistical connections between certain *reconstruction-based* pretext tasks that guarantee to learn a good representation. Formally, we quantify how the approximate independence between the components of the pretext task (conditional on the label and latent variables) allows us to learn representations that can solve the downstream task by just training a linear layer on top of the learned representation. We prove the linear layer yields small approximation error even for complex ground truth function class and will drastically reduce labeled sample complexity.

1 Introduction

Self-supervised learning revitalizes machine learning models in computer vision, NLP, and control problems (see reference therein [36, 38, 15, 63, 35]). Training a model with auxiliary tasks based only on input features reduces the extensive costs of data collection and semantic annotations for downstream tasks. It is also known to improve the adversarial robustness of models [29, 11, 12]. Self-supervised learning creates pseudo labels solely based on input features, and solves auxiliary prediction tasks (or pretext tasks) in a supervised manner. However, the underlying principles of self-supervised learning are mysterious since it is a-priori unclear why predicting what we already know should help. We thus raise the following question:

What conceptual connection between pretext and downstream tasks ensures good representations? What is a good way to quantify this?

As a thought experiment, consider a simple downstream task of classifying desert, forest, and sea images. A meaningful pretext task is to predict the background color of images (known as image colorization [66]). Denote X_1, X_2, Y to be the input image, color channel, and the downstream label respectively. Given knowledge of the label Y, one can possibly predict the background X_2 without knowing much about X_1. In other words, X_2 is approximately independent of X_1 conditional on the label Y. Consider another task of inpainting [48] the front of a building (X_2) from the rest (X_1). While knowing the label "building" (Y) is not sufficient for successful inpainting, adding additional latent variables Z such as architectural style, location, window positions, etc. will ensure that variation in X_2 given Y, Z is small. We can mathematically interpret this as X_1 being approximate conditionally independent of X_2 given Y, Z.

The main insight that we exploit in this work is that with approximate conditional independence (as in the above examples), a method that predicts X_2 from X_1 will inadvertently implicitly encode and

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

learn to predict Y (and Z) from X_1 as an intermediate step, and then predict X_2 from Y[1]. Building upon this insight, we make the following contributions.

Contributions. The goal of this paper, as in statistical learning theory, is to investigate the *statistical connections* between the random variables of input features (in this paper (X_1, X_2)) and downstream labels Y, and show how specific connections can guarantee a successful learning procedure. For self-supervised learning (SSL), success is measured using the following 2 notions, 1) expressivity, i.e. does the learned representation from SSL have the ability to express the ground truth prediction function for labels Y, and 2) sample complexity, i.e. can it do so with way fewer labeled samples than what would be required without SSL.

In this work, we show such guarantees for a class of *reconstruction-based* SSL methods under a statistical assumption of *approximate conditional independence (ACI)*. In particular we show that under such an assumption, the learned representation from SSL will end up having the following properties, 1) it can express the ground truth label as a *linear function*, thus guaranteeing expressivity, and 2) will also end up being low-rank (or low-dimensional), thus guaranteeing smaller labeled sample complexity. Note that such an expressive and sample efficient (summarized as *good*) representation is often not a-priori available. For instance, the original input features themselves may not be able to express the ground truth function linearly, while kernel methods with a fixed kernel, while expressive, may not be sample efficient for many problems of interest. The strategy in modern machine learning is to find such a *good* representation as the output of a complicated neural network. The benefit of SSL, as we formally show here, is that the complicated but *good* representation function can be learned using just *unlabeled data*, so that labeled data is just needed to learn a linear function.

The *reconstruction-based* SSL method (differentiated from other SSL methods in Section 1.1) we consider is strongly motivated by empirical works [66, 48, 15, 25], but is a simplification that captures the essence of the problem and is amenable to a precise theoretical analysis. We consider a two-staged pipeline, where we first learn a representation function ψ (e.g. output of a neural network) from input X_1 and pretext target X_2 using *unlabeled data* by minimizing $\mathbb{E}_{(X_1,X_2)}[\|X_2 - \psi(X_1)\|^2]$. In the second stage of downstream task, we learn a *linear layer* on top of representation ψ using labeled samples (X_1, Y), thus restricting to learning from a significantly smaller hypothesis class of $\mathcal{H}_\psi = \{f : X_1 \to Y | f \text{ is linear in } \psi\}$. The key non-trivial question of expressivity is now whether the ground truth predictor $f^* \equiv \mathbb{E}[Y|X_1]$ can be approximated well by this class $\mathcal{H}_\psi$, and the question of sample complexity reduces to the understanding the sample complexity of learning $\mathcal{H}_\psi$. Under appropriate statistical connections[2] between input data X_1, X_2 and target Y, we prove both the desired properties, expressivity and low sample complexity, for the aforementioned SSL method.

Our statistical assumption based on approximate conditional independence (ACI) helps us demonstrate how solving pretext tasks created from *known information* can learn useful representations. Specifically, we show that once the complicated representation function ψ is learned using an abundance of *unlabeled data* in the SSL stage, not only is ψ expressive enough, but it will also require only $\tilde{\mathcal{O}}(k)$ *labeled samples* to solve a k-way supervised learning task under exact conditional independence (CI). In contrast, solving the downstream task without any pretraining will require a lot of labeled data to learn the representation function from scratch. Since the strong exact conditional independence assumption will likely not be satisfied in practice, our main contribution is to derive similar risk bounds when only approximate CI (ACI) is satisfied. We quantify the notion of ACI using the norm of a certain partial covariance matrix (Definition 4.1) and our risk bound scales linearly with it. We verify this and other aspects of our main Theorem 4.2 using simulations and also find that pretext task helps when CI is approximately enforced in text domain. We further demonstrate on a real-world image dataset that a pretext task-based linear model performs at least as well as many baselines.

1.1 Related work

Self-supervised learning (SSL) methods in practice: There has been a flurry of self-supervised methods lately. One class of methods reconstruct images from corrupted or incomplete versions of it, like denoising auto-encoders [61], image inpainting [48], and split-brain autoencoder [67]. Pretext

[1]This is formally demonstrated in the proof sketch of Lemma 3.1.

[2]We note that since the representation ψ is the result of the SSL method and not something we have access to a-priory, we cannot make any direct assumptions on it.

tasks are also created using visual common sense, including predicting rotation angle [22], relative patch position [16], recovering color channels [66], solving jigsaw puzzle games [45], and discriminating images created from distortion [17]. We refer to the above procedures as reconstruction-based SSL. Another popular paradigm is contrastive learning [13, 14]. The idea is to learn representations that bring similar data points closer while pushing randomly selected points further away [63, 39, 5] or to maximize a contrastive-based mutual information lower bound between different views [30, 46, 54]. A popular approach for text domain is based on language modeling where models like BERT and GPT create auxiliary tasks for next word predictions [15, 49]. The natural ordering or topology of data is also exploited in video-based [64, 43, 19], graph-based [65, 33] or map-based [68] SSL. For instance, the pretext task is to determine the correct temporal order for video frames as in [43].

Theory for SSL: While we theoretically study reconstruction-based SSL, prior work has different flavors of theoretical results for different kinds of SSL methods. Most relevant are the guarantees for representation learning using SSL methods on downstream tasks that just learn a linear classifier on top of the learned representations. [5] shows guarantees for representations from a contrastive learning objective: $L_1^{cont}(\psi) = \mathbb{E}_{(X_1,X_2),X_2'}[\log(1 + e^{-\psi(X_1)^\top \psi(X_2) + \psi(X_1)^\top \psi(X_2')})]$. Under a class conditional independence assumption, i.e. $X_1 \perp X_2 \mid Y$, they show that representation ψ that does well on contrastive objective, i.e. $L_1^{cont}(\psi) \leq \epsilon$, will have $\mathcal{O}(\epsilon)$ linear classification loss on the average binary task involving pairs of classes (y_1, y_2). However, their analysis cannot handle the general case of approximate conditional independence. Recently, Tosh *et al.* [56] show that contrastive learning representations can *linearly* recover continuous functions of the underlying topic posterior under a topic modeling assumption for text. While their assumption bears similarity to ours, the assumption of independent sampling of words is strong and does not generalizable to other domains like images. Most relevant is a concurrent work [57] that shows guarantees for a contrastive learning objective that looks like $L_2^{cont}(\psi, \eta) = \mathbb{E}_{(X_1,X_2),X_2'}\left[\log(1 + e^{-\psi(X_1)^\top \eta(X_2)}) + \log(1 + e^{\psi(X_1)^\top \eta(X_2')})\right]$, with a multi-view redundancy assumptions that is very similar to our ACI assumption. We take a closer look at their assumption in Section F.2. All the above objectives are different from the simple reconstruction-based objective we consider: $L(\psi) = \mathbb{E}_{(X_1,X_2)}\left[\|X_2 - \psi(X_1)\|^2\right]$. Saunshi *et al.* [51] show guarantees for representations learned using language modeling on sentence classification tasks. Some more recent work [58, 44, 55, 62] provide theoretical understanding on SSL respectively based on causality, mutual information, gradient-descent dynamics, and alignment/uniformity of representations, without explicit risk bounds for downstream tasks. There is a mutual information maximization view of contrastive learning, but [59] points out issues with it. Previous attempts to explain negative sampling [42] based methods use the theory of noise contrastive estimation [27, 40] to show asymptotic guarantees, without explicit connections to downstream tasks. CI is also used in sufficient dimension reduction [21, 20], while CI and redundancy assumptions on multiple views [37, 2] are used to analyze a canonical-correlation based dimension reduction algorithm and also for self-supervised learning algorithms like co-training [10]. Finally, [1, 60] provide a theoretical analysis for denoising auto-encoder.

1.2 Overview of results:

Section 2 introduces notation, setup, and the self-supervised learning procedure considered in this work. In Section 3, we analyze downstream sample complexity under exact CI and unlimited labeled data to highlight the key ideas. Section 4 presents our main result with relaxed conditions: under ACI with latent variables, and assuming finite samples in both pretext and downstream tasks, for various function classes, and both regression and classification tasks. Experiments verifying our theoretical findings are in Section 6. Proofs of most results are in the Appendix.

2 Preliminary

2.1 Notation

We use lower case symbols (x) to denote scalar quantities, bold lower case symbols ($\boldsymbol{x}$) for vector values, capital letters (X) for random variables, and capital and bold letters $\boldsymbol{X}$ for matrices. P_X denotes the probability law of random variable X, and the space of square-integrable functions with probability P is denoted by $L^2(P)$. We use standard $\mathcal{O}$ notation to hide universal factors and $\tilde{\mathcal{O}}$ to hide log factors. $\|\cdot\|$ stands for ℓ_2-norm for vectors or Frobenius norm for matrices.

Linear conditional expectation. $\mathbb{E}^L[Y|X]$ denotes the prediction of Y with linear regression:

$$\mathbb{E}^L[Y|X=\boldsymbol{x}] := \boldsymbol{W}^*\boldsymbol{x}+\boldsymbol{b}^*, \quad \text{where } \boldsymbol{W}^*, \boldsymbol{b}^* := \arg\min_{\boldsymbol{W},\boldsymbol{b}} \mathbb{E}[\|Y-\boldsymbol{W}X-\boldsymbol{b}\|^2].$$

In other words, $\mathbb{E}^L[Y|X]$ denotes the best linear predictor of Y given X. We also note that $\mathbb{E}[Y|X] \equiv \arg\min_f \mathbb{E}[\|Y-f(X)\|^2]$ is the best predictor of Y given X.

(Partial) covariance matrix. For random variables X, Y, we denote $\boldsymbol{\Sigma}_{XY}$ to be covariance matrix of X and Y. For simplicity in most cases, we assume $\mathbb{E}[X]=0$ and $\mathbb{E}[Y]=0$; thus we do not distinguish $\mathbb{E}[XY]$ and $\boldsymbol{\Sigma}_{XY}$. The partial covariance matrix between X and Y given Z is:

$$\boldsymbol{\Sigma}_{XY|Z} := \text{cov}\{X-\mathbb{E}^L[X|Z], Y-\mathbb{E}^L[Y|Z]\} \equiv \boldsymbol{\Sigma}_{XY}-\boldsymbol{\Sigma}_{XZ}\boldsymbol{\Sigma}_{ZZ}^{-1}\boldsymbol{\Sigma}_{ZY}, \tag{1}$$

which captures the correlation between X and Y setting aside the effect of Z.

Sub-gaussian random vectors. $X \in \mathbb{R}^d$ is ρ^2-sub-gaussian if for every fixed unit vector $\boldsymbol{v} \in \mathbb{R}^d$, the variable $\boldsymbol{v}^\top X$ is ρ^2-sub-gaussian, i.e., $\mathbb{E}[e^{s\cdot\boldsymbol{v}^\top(X-\mathbb{E}[X])}] \le e^{s^2\rho^2/2}$ $(\forall s \in \mathbb{R})$.

2.2 Setup and methodology

We denote by X_1 the input variable, X_2 the target random variable for the pretext task, and Y the label for the downstream task, with $X_1 \in \mathcal{X}_1 \subset \mathbb{R}^{d_1}, X_2 \in \mathcal{X}_2 \subset \mathbb{R}^{d_2}$ and $Y \in \mathcal{Y} \subset \mathbb{R}^k$. If $\mathcal{Y}$ is finite with $|\mathcal{Y}|=k$, we assume $\mathcal{Y} \subset \mathbb{R}^k$ is the one-hot encoding of the labels. $P_{X_1X_2Y}$ denotes the joint distribution over $\mathcal{X}_1 \times \mathcal{X}_2 \times \mathcal{Y}$. P_{X_1Y}, P_{X_1} denote the corresponding marginal distributions. Our proposed self-supervised learning aims to fulfill the following two steps:

Step 1 (pretext task): Learn a representation $\psi(\boldsymbol{x}_1)$ close to $\psi^* := \arg\min_{g\in\mathcal{H}} \mathbb{E}\|X_2-g(X_1)\|^2$, where $\mathcal{H}$ can vary for different settings that we will specify and discuss later.

Step 2 (downstream task): Perform linear regression on Y with $\psi(X_1)$, i.e. $f(\boldsymbol{x}_1) := (\boldsymbol{W}^*)^\top\psi(\boldsymbol{x}_1)$, where $\boldsymbol{W}^* \leftarrow \arg\min_{\boldsymbol{W}} \mathbb{E}_{X_1,Y}[\|Y-\boldsymbol{W}^\top\psi(X_1)\|^2]$. Namely we learn $f(\cdot)=\mathbb{E}^L[Y|\psi(\cdot)]$.

We study this simplified version in the main text, where in practice, the SSL procedure may utilize an encoder-decoder structure, while the downstream task uses both X_1 and X_2 to predict Y. We incorporate these extensions in Appendix C.3 and G.

With finite samples, performance of a learned representation ψ on the downstream task depends on the following quantities that capture expressivity and sample complexity respectively:

Approximation error indicates whether Y is *linearly separable* by the learned representation ψ, thus measuring expressivity. We measure this by comparing $\boldsymbol{W}\psi(X_1)$ to the optimal predictor $f^* := \mathbb{E}[Y|X_1=\boldsymbol{x}_1]$. Denote $e_{\text{apx}}(\psi) = \min_{\boldsymbol{W}} \mathbb{E}[\|f^*(X_1)-\boldsymbol{W}\psi(X_1)\|^2]$. This gives a measure of how well ψ can linearly predict Y when given infinite samples for the task.

Estimation error measure sample complexity of ψ on the downstream task and assume access to n_2 i.i.d. samples $(\boldsymbol{x}_1^{(1)}, \boldsymbol{y}^{(1)}), \cdots, (\boldsymbol{x}_1^{(n_2)}, \boldsymbol{y}^{(n_2)})$ drawn from P_{X_1Y}. We express the n_2 samples collectively as $\boldsymbol{X}_1^{\text{down}} \in \mathbb{R}^{n_2\times d_1}$, $\boldsymbol{Y} \in \mathbb{R}^{n_2\times k}$ and overload notation to say $\psi(\boldsymbol{X}_1^{\text{down}}) = \left[\psi(\boldsymbol{x}_1^{(1)})|\psi(\boldsymbol{x}_1^{(2)})\cdots|\psi(\boldsymbol{x}_1^{(n_2)})\right]^\top \in \mathbb{R}^{n_2\times d_2}$. We perform linear regression on the learned representation ψ and measure excess risk, that incorporates both approximation and estimation errors.

$$\hat{\boldsymbol{W}} \leftarrow \arg\min_{\boldsymbol{W}} \frac{1}{2n_2}\|\boldsymbol{Y}-\psi(\boldsymbol{X}_1)\boldsymbol{W}\|_F^2; \;\; \text{ER}_\psi(\hat{\boldsymbol{W}}) := \frac{1}{2}\mathbb{E}\|f^*(X_1)-\hat{\boldsymbol{W}}^\top\psi(X_1)\|_2^2.$$

3 Guaranteed recovery with conditional independence

In this section, we focus on the case where the input X_1 and pretext target X_2 are conditionally independent (CI) given the downstream label Y. While this is a strong assumption that is rarely satisfied in practice, it helps us understand the role of CI with clean results and builds up to our main results with ACI with latent variables in Section 4. As a warm-up, we show how CI helps when (X_1, X_2, Y) are jointly Gaussian to give us a flavor for the results to follow in Appendix B. We then analyze it for general random variables under two settings: (a) when the function class used for ψ is universal, (b) when ψ is restricted to be a linear function of given features. For now we assume

access to a large amount of unlabeled data so as to learn the optimal ψ^* perfectly and this will be relaxed later in Section 4. The general recipe for the results is as follows:

1. Find a closed-form expression for the optimal solution ψ^* for the pretext task.
2. Use conditional independence to show that optimal f^* is linear in ψ^*, i.e., $e_{\text{apx}}(\psi^*)$ is small.
3. Exploit the low rank structure of ψ^* to show small estimation error on downstream tasks.

Data assumption. Suppose $Y = f^*(X_1) + N$, where $f^* = \mathbb{E}[Y|X_1]$ and $\mathbb{E}[N] = 0$. We assume N is σ^2-subgaussian. For simplicity, we assume non-degeneracy: $\mathbf{\Sigma}_{X_iX_i}, \mathbf{\Sigma}_{YY}$ are full rank.

Assumption 3.1. *Let $X_1 \in \mathbb{R}^{d_1}, X_2 \in \mathbb{R}^{d_2}$ be random variables from some unknown distribution. Let label $Y \in \mathcal{Y}$ be a discrete random variable with $k = |\mathcal{Y}| < d_2$. We assume conditional independence: $X_1 \perp X_2 | Y$.*

Here Y can be interpreted as the multi-class labels where k is the number of classes. For regression problems, one can think about Y as the discretized values of continuous labels. We do not specify the dimension for Y since Y could be arbitrarily encoded but the results only depend on k and the variance of Y (conditional on the input X_1).

3.1 Universal function class.

Suppose we learn the optimal ψ^* among all measurable functions The optimal function ψ^* in this case is naturally given by conditional expectation: $\psi^*(\boldsymbol{x}_1) = \mathbb{E}[X_2|X_1 = \boldsymbol{x}_1]$. We show that CI implies that ψ^* is good for downstream tasks, which is not apriori clear.

Lemma 3.1 (Approximation error)**.** *If random variables X_1, X_2, Y satisfy Assumption 3.1, and $\boldsymbol{A} \in \mathbb{R}^{\mathcal{Y} \times d_2}$ with $\boldsymbol{A}_{y,:} := \mathbb{E}[X_2|Y = \boldsymbol{y}]$ has rank $k = |\mathcal{Y}|$. Then $f^* \equiv \boldsymbol{W}^*\psi^*$, i.e., $e_{apx}(\psi^*) = 0$.*

This tells us that although f^* could be nonlinear in $\boldsymbol{x}_1$, it is guaranteed to be linear in $\psi^*(\boldsymbol{x}_1)$.

Proof Sketch of Lemma 3.1. Lemma is proved by law of total expectation:

$$\begin{aligned}\psi^*(\cdot) := \mathbb{E}[X_2|X_1] =& \mathbb{E}[\mathbb{E}[X_2|X_1, Y]|X_1] = \mathbb{E}[\mathbb{E}[X_2|Y]|X_1] \qquad \text{(uses CI)}\\ =& \sum_y P(Y = y|X_1)\mathbb{E}[X_2|Y = y] =: f(X_1)^\top \boldsymbol{A},\end{aligned}$$

where $f(x_1)_y = P(Y = y|X_1 = x_1)$, and $\boldsymbol{A} \in \mathbb{R}^{\mathcal{Y} \times d_2}$ satisfies $\boldsymbol{A}_{y,:} = \mathbb{E}[X_2|Y = y]$. One could see that through predicting X_2, due to the CI assumption, ψ^* has implicitly encoded the information of $Y|X_1$. Finally due to the fact that matrix $\boldsymbol{A}$ is full rank, we get that f^* is linear in ψ^* as well. □

We see that besides CI, another important property is $\mathbb{E}[X_2|Y]$ being rank k. This means X_2 is correlated with every instance of Y, and thus captures information of every prediction class. This is naturally a necessary assumption for X_2 to be a reasonable pretext task for predicting Y. Note that this assumption does not trivialize the problem and that even though ψ is designed to predict X_2, it can still be a better representation than X_2 for downstream tasks. Note that Y does not have to be linear in X_2 but is proven to be linear in ψ, since ψ learns to ignore some information in X_2 that is irrelevant to Y. We provide this simple example for better understanding:

Example 3.1. *Let $Y \in \{-1, 1\}$ be binary labels, and X_1, X_2 be $2-$mixture Gaussian random variables with $X_1 \sim \mathcal{N}(Y\boldsymbol{\mu}_1, \mathbf{I}), X_2 \sim \mathcal{N}(Y\boldsymbol{\mu}_2, \mathbf{I})$. In this example, $X_1 \perp X_2|Y$. Although $\mathbb{E}[Y|X_2]$ and $\mathbb{E}[Y|X_1]$ are not linear, $\mathbb{E}[Y|\psi]$ is linear: $\psi(\boldsymbol{x}_1) = P(Y = 1|X_1 = \boldsymbol{x}_1)\boldsymbol{\mu}_2 - P(Y = -1|X_1 = \boldsymbol{x}_1)\boldsymbol{\mu}_2$ and $f^*(\boldsymbol{x}_1) = P(Y = 1|X_1 = \boldsymbol{x}_1) - P(Y = -1|X_1 = \boldsymbol{x}_1) \equiv \boldsymbol{\mu}_2^T\psi(\boldsymbol{x}_1)/\|\boldsymbol{\mu}_2\|^2$.*

Given that ψ^* is good for downstream, we now care about the sample complexity. We will need to assume that the representation has some nice concentration properties. We make an assumption about the whitened data $\psi^*(X_1)$ to ignore scaling factors.

Assumption 3.2. *We assume the whitened feature variable $U := \mathbf{\Sigma}_\psi^{-1/2}\psi(X_1)$ is a ρ^2-subgaussian random variable, where $\mathbf{\Sigma}_\psi = \mathbb{E}[\psi(X_1)\psi(X_1)^\top]$.*

We note that all bounded random variables satisfy sub-gaussian property.

Theorem 3.2 (General conditional independence)**.** *Fix a failure probability $\delta \in (0, 1)$, under the same assumption as Lemma 3.1 and Assumption 3.2 for ψ^*, if additionally $n \gg \rho^4(k + \log(1/\delta))$, then the excess risk of the learned predictor $\boldsymbol{x}_1 \to \hat{\boldsymbol{W}}^\top\psi^*(\boldsymbol{x}_1)$ on the downstream task satsifies*

$$\mathrm{ER}_{\psi^*}[\hat{\boldsymbol{W}}] \leq \tilde{\mathcal{O}}\left(\frac{k}{n_2}\sigma^2\right)^{3}$$

Remark 3.1. *This analysis assumes we could perfectly learn $\psi^* = \mathbb{E}[X_2|X_1]$ disregarding the number of samples in the SSL phase (unlabeled data is cheap to obtain). Here by sample complexity we refer to the labeled data (X_1, Y). We defer the effect of imprecise representation ψ in Section 4.*

3.2 Function class induced by feature maps.

Given feature map $\phi_1 : \mathcal{X}_1 \to \mathbb{R}^{D_1}$, we consider the function class $\mathcal{H}_1 = \{\psi : \mathcal{X}_1 \to \mathbb{R}^{d_2} | \exists \boldsymbol{B} \in \mathbb{R}^{d_2 \times D_1}, \psi(\boldsymbol{x}_1) = \boldsymbol{B}\phi_1(\boldsymbol{x}_1)\}$.

Claim 3.3 (Closed form solution)**.** *The optimal function in $\mathcal{H}$ is $\psi^*(\boldsymbol{x}_1) = \boldsymbol{\Sigma}_{X_2\phi_1}\boldsymbol{\Sigma}_{\phi_1\phi_1}^{-1}\phi_1(\boldsymbol{x}_1)$, where $\boldsymbol{\Sigma}_{X_2\phi_1} := \boldsymbol{\Sigma}_{X_2\phi_1(X_1)}$ and $\boldsymbol{\Sigma}_{\phi_1\phi_1} := \boldsymbol{\Sigma}_{\phi_1(X_1)\phi_1(X_1)}$.*

We again show the benefit of CI, but only comparing the performance of ψ^* to the original features ϕ_1. Since ψ^* is linear in ϕ_1, it cannot have smaller approximation error than ϕ_1. However CI will ensure that ψ^* has the same approximation error as ϕ_1 and enjoys better sample complexity.

Lemma 3.4 (Approximation error)**.** *If Assumption 3.1 is satisfied, and if the matrix $\boldsymbol{A} \in \mathbb{R}^{\mathcal{Y}\times d_2}$ with $\boldsymbol{A}_{y,:} := \mathbb{E}[X_2|Y = \boldsymbol{y}]$ is of rank $k = |\mathcal{Y}|$. Then $e_{apx}(\psi^*) = e_{apx}(\phi_1)$.*

We additionally need an assumption on the residual $a(\boldsymbol{x}_1) := \mathbb{E}[Y|X_1 = \boldsymbol{x}_1] - \mathbb{E}^L[Y|\phi_1(\boldsymbol{x}_1)]$.

Assumption 3.3. *(Bounded approx. error; Condition 3 in [32])) We have almost surely*

$$\|\boldsymbol{\Sigma}_{\phi_1\phi_1}^{-1/2}\phi_1(X_1)a(X_1)^\top\|_F \leq b_0\sqrt{k}$$

Theorem 3.5. *(CI with approximation error) Fix a failure probability $\delta \in (0, 1)$, under the same assumption as Lemma 3.4, Assumption 3.2 for ψ^* and Assumption 3.3, if $n_2 \gg \rho^4(k + \log(1/\delta))$, then the excess risk of the learned predictor $\boldsymbol{x}_1 \to \hat{\boldsymbol{W}}^\top\psi^*(\boldsymbol{x}_1)$ on the downstream task satisfies:*

$$\mathrm{ER}_{\psi^*}[\hat{\boldsymbol{W}}] \leq e_{apx}(\phi_1) + \tilde{\mathcal{O}}\left(\frac{k}{n_2}\sigma^2\right).$$

Thus with SSL, the requirement of labels is reduced from complexity for D_1 to $\mathcal{O}(k)$.

4 Beyond conditional independence

In the previous section, we focused on the case where we have exact CI. A weaker but more realistic assumption is that Y captures some portion of the dependence between X_1 and X_2 but not all. We quantify this notion of approximate ACI through a quantity ϵ^2_{CI} (Definition 4.1), and show excess risk bounds for the representation learned from SSL[4]. In particular, the excess risk will have the form $\tilde{\mathcal{O}}\left(\frac{d_2}{n_2} + \epsilon^2_{\text{CI}} + \epsilon^2_{\text{pre}}\right)$, which suggests that only $n_2 = \mathcal{O}(d_2)$ labeled samples will be required to get small error on downstream task, as long as approximate CI is satisfied (ϵ^2_{CI} is small) and the pretext task is solved well enough (ϵ^2_{pre} is small). This is in contrast to not doing SSL, where many more labeled samples will be required to learn a solve the downstream task that learns a complicated representation function from scratch. We now describe the SSL method on finite samples, followed by the definition of ACI which we use to discuss the main excess risk bound and its consequences.

SSL with finite samples and general function space: Let $\boldsymbol{X}_1^{\text{pre}} = [\boldsymbol{x}_1^{(1,\text{pre})}, \cdots, \boldsymbol{x}_1^{(n_1,\text{pre})}]^\top \in \mathbb{R}^{n_1\times d_1}$ and $\boldsymbol{X}_2 = [\boldsymbol{x}_2^{(1)}, \cdots, \boldsymbol{x}_2^{(n_1)}]^\top \in \mathbb{R}^{n_1\times d_2}$ be n_1 training samples for pretext task, where $(\boldsymbol{x}_1^{(i,\text{pre})}, \boldsymbol{x}_2^{(i)})$ is sampled from $P_{X_1X_2}$. The n_2 labeled samples for the downstream task are defined as $\boldsymbol{X}_1^{\text{down}} \in \mathbb{R}^{n_2\times d_1}$, $\boldsymbol{Y} \in \mathbb{R}^{n_2\times d_3}$[5]. Given a representation function space $\mathcal{H} : \mathcal{X}_1 \to \mathbb{R}^{d_2}$, we learn $\tilde{\psi}$ from $\mathcal{H}$ using the n_1 unlabeled samples and then use the n_2 labeled samples to learn a linear classifier on the learned representation $\tilde{\psi}(\boldsymbol{X}_1^{\text{down}})$ to fit $\boldsymbol{Y}$. This process is summarized below.

$$1)\ \tilde{\psi} := \arg\min_{f\in\mathcal{H}} \frac{1}{n_1}\|\boldsymbol{X}_2 - f(\boldsymbol{X}_1^{\text{pre}})\|_F^2,\ 2)\ \hat{\boldsymbol{W}} \leftarrow \arg\min_{\boldsymbol{W}} \frac{1}{2n_2}\|\boldsymbol{Y} - \tilde{\psi}(\boldsymbol{X}_1^{\text{down}})\boldsymbol{W}\|_F^2. \quad (2)$$

[3] We will use $\tilde{O}$ to hide log factor $\log(k/\delta)$ or $\log(d_2/\delta)$.

[4] Results for jointly-Gaussian variables is in Appendix D.1; ACI is quantified by the partial covariance matrix.

[5] $d_3 = k$ and $Y \equiv \phi_y(Y)$ (one-hot encoding) refers multi-class classification task, $d_3 = 1$ refers to regression.

In our main results, we consider two types of function spaces: $\mathcal{H} \in \{\mathcal{H}_1, \mathcal{H}_u\}$. Recall that $\mathcal{H}_1 = \{\psi(\cdot) = \boldsymbol{B}\phi_1(\cdot); \boldsymbol{B} \in \mathbb{R}^{d_2 \times D_1}\}$ is a class of *linear representations* induced by feature map $\phi_1 : \mathcal{X}_1 \to \mathbb{R}^{D_1}$. We use $\mathcal{H}_u$ to denote a function space with universal approximation power (e.g. deep networks) that ensures $\psi^* = \mathbb{E}[X_2|X_1] \in \mathcal{H}_u$. We define the optimal predictor in each case as $f^*_{\mathcal{H}}(X_1) = \mathbb{E}^L[Y|\phi_1(X_1)]$ when $\mathcal{H} = \mathcal{H}_1$, $f^*_{\mathcal{H}} = f^*$ for $\mathcal{H} = \mathcal{H}_u$, we define excess risk as

$$\mathrm{ER}_{\tilde{\psi}}(\hat{\boldsymbol{W}}) := \mathbb{E}_{X_1}\left[\|f^*_{\mathcal{H}}(X_1) - \hat{\boldsymbol{W}}^\top \tilde{\psi}(X_1)\|_2^2\right].$$

Approximate conditional independence: Our new assumption will generalize Assumption 3.1 in two ways, 1) we allow for additional latent variables Z that together with Y could potentially make X_1 and X_2 independent, and 2) we allow this conditional independence to be approximate. Note that allowing for extra latent variable can trivially make X_1 and X_2 to be conditionally independent by picking a large enough Z (e.g. $Z = (X_1, X_2)$). However the following assumption, that needs the pretext target X_2 to correlate with all instances of variable $\bar{Y} = [Y, Z]$ (analogous to Lemma 3.1), will impose this restriction on how large Z can be.

Assumption 4.1 (Correlation between X_2 and Y, Z)**.** *Suppose there exists latent variable $Z \in \mathcal{Z}, |\mathcal{Z}| = m$ that ensures $\boldsymbol{\Sigma}_{\phi_{\bar{y}} X_2}$ is full column rank and $\|\boldsymbol{\Sigma}_{Y\phi_{\bar{y}}}\boldsymbol{\Sigma}^\dagger_{X_2\phi_{\bar{y}}}\|_2 = 1/\beta$, where $A^\dagger$ is pseudo-inverse, and $\phi_{\bar{y}}$ is the one-hot embedding for $\bar{Y} = [Y, Z]$.*

Just as in Section 3, this assumption will not assume away the problem (Example 3.1 can be suitably extended). The additional term $1/\beta$ here captures both the "scale" of X_2 and also the strength of correlation between X_2 and $[Y, Z]$ that was discussed after Lemma 3.1. For $\boldsymbol{\Sigma}_{\phi_{\bar{y}} X_2}$ to be full column rank, it is essential that $d_2 \geq km$, and this already gives an upper bound on the size of Z. Given this restriction on Z (and thus $\bar{Y}$), we define the notion of approximate conditional independence.

Definition 4.1 (Approximate conditional independence with function space $\mathcal{H}$)**.** *For $\bar{Y} = [Y, Z]$,*
1. For $\mathcal{H} = \mathcal{H}_1$, define $\epsilon_{CI} := \|\boldsymbol{\Sigma}^{-1/2}_{\phi_1\phi_1}\boldsymbol{\Sigma}_{\phi_1 X_2|\phi_{\bar{y}}}\|_F$.
2. For $\mathcal{H} = \mathcal{H}_u$, define $\epsilon^2_{CI} := \mathbb{E}_{X_1}[\|\mathbb{E}[X_2|X_1] - \mathbb{E}_{\bar{Y}}[\mathbb{E}[X_2|\bar{Y}]|X_1]\|^2]$.

Firstly we note that this is indeed an extension of exact CI, since exact CI in both cases will imply that $\epsilon_{\text{CI}} = 0$. We present a unified analysis in the appendix that shows the ϵ_{CI} for the second case is same as the first case, with covariance operators instead of matrices (A direct derivation is in Claim D.7). We also present more relaxed and general form of the above assumptions in Appendix F.1. With this assumption, we are ready to present our main bound.

Bound on excess risk: Recall that we assume that the residual term $N := Y - \mathbb{E}[Y|X_1]$ is mean zero and σ^2-subgaussian. Before showing our main result, analogous to Assumption 3.3, for the class $\mathcal{H}_1$ with non-universal features ϕ_1, we will need an assumption[6] on the residual $a := f^* - f^*_{\mathcal{H}_1} = \mathbb{E}[Y|X_1] - \mathbb{E}^L[Y|\phi_1(X_1)]$:

Assumption 4.2. *(Bounded approximation error on pretext phase [32]) There exists a universal constant b_0, such that $\|\boldsymbol{\Sigma}^{-1/2}_{\phi_1\phi_1}\phi_1(X_1)a(X_1)^\top\|_F \leq b_0\sqrt{d_2}$ almost surely.*

Theorem 4.2. *For a fixed $\delta \in (0, 1)$, under Assumptions 4.1,3.2 for $\tilde{\psi}$ and ψ^* and 4.2 for non-universal feature maps, if $n_1, n_2 \gg \rho^4(d_2 + \log 1/\delta)$, and we learn the pretext tasks such that: $\mathbb{E}\|\tilde{\psi}(X_1) - \psi^*(X_1)\|_F^2 \leq \epsilon^2_{pre}$. Then the generalization error for downstream task w.p. $1 - \delta$ is:*

$$\mathrm{ER}_{\tilde{\psi}}(\hat{\boldsymbol{W}}) \leq \tilde{\mathcal{O}}\left(\underbrace{\sigma^2\frac{d_2}{n_2}}_{\textit{estimation error}} + \underbrace{\frac{\epsilon^2_{CI}}{\beta^2} + \frac{\epsilon^2_{pre}}{\beta^2}}_{\textit{approximation error}}\right) \tag{3}$$

We defer the proof to the appendix. The proof technique is similar to that of Section 3. The difference is that now $\tilde{\psi}(\boldsymbol{X}^{(\text{down})}) \in \mathbb{R}^{n_2 \times d_2}$ will be an approximately low rank matrix, where the low rank part is the high-signal features that implicitly comes from Y, Z that can linearly learn downstream task. The remaining part comes from ϵ_{CI} and ϵ_{pre} and causes the approximation error. Again by selecting the top km (dimension of $\phi_{\bar{y}}$) features we could further improve the bound:

[6] This rules out the failure if one chooses a very simple function class to learn $\mathbb{E}[X_2|X_1]$. In practice we usually use neural networks (with universal approximation power) and this bound should be very small.

Remark 4.1. *By applying PCA on $\tilde{\psi}(\boldsymbol{X}_1^{down})$ and keeping the top km principal components only, we can improve the bound in Theorem 4.2 to* $\mathrm{ER}_{\tilde{\psi}}(\hat{\boldsymbol{W}}) \leq \tilde{\mathcal{O}}\left(\sigma^2 \frac{km}{n_2} + \frac{\epsilon_{CI}^2}{\beta^2} + \frac{\epsilon_{pre}^2}{\beta^2}\right)$.

We take a closer look at the different sources of errors in Lemma 4.1: 1) The first term is estimation error on learning with finite samples n_2 with noise level σ^2 in $Y - f^*(X_1)$; 2) ϵ_{CI} measures the approximate CI; and 3) ϵ_{pre} is the error from not learning the pretext task exactly. The first term is optimal ignoring log factors as we do linear regression on mk-dimensional features. The second and third term together form approximation error. They are non-reducible due to the fact that f^* is not exactly linear in ψ and we use it as a fixed representation. Fine-tuning the representations might be necessary to get rid of these terms when we have sufficient downstream labeled data. We leave this exploring this as future work. Compared to traditional supervised learning, learning $f^*_{\mathcal{H}}$ requires sample complexity scaling with the (Rademacher/Gaussian) complexity of $\mathcal{H}$ (see e.g. [8, 52]), which is very large for complicated models such as deep networks. Thus SSL can significantly reduce the labeled sample complexity down from this complexity measure of $\mathcal{H}$ to $\tilde{\mathcal{O}}(km)$, demonstrating the power of predicting what you already know using unlabeled data. In Section H, we consider a similar result for classification.

5 Example: Topic Modeling

In this section, we will demonstrate how our framework can be instantiated for standard data model like topic modeling. Topic modeling for text that has a rich literature [47, 31, 9, 4, 3] and is used for analyzing and designing algorithms for information retrieval, dimensionality reduction and data analysis for large text corpora. We describe the basic setup below, followed by how our results for reconstruction-based SSL can be instantiated to learn such models.

For a set S, let Δ_S denote the set of all distributions on S. In the topic modeling framework, generation of a text document with a vocabulary set $[V] = \{1, \ldots, V\}$ is governed by certain latent topics from the set $[k]$, where k is the total number of topics. Each topic $i \in [k]$ is associated with a distribution over the vocabulary $[V]$ that is denoted by vector $A_i \in \Delta_{[V]}$; stack these vectors into the columns of a matrix $A \in \mathbb{R}^{V \times k}$. A document $X = (x_1, \ldots, x_n) \in [V]^N$ of length N is then sampled from a mixture of the k topics $\mu \in \Delta_{[k]}$. The generative process is described below:

1. Sample a topic mixture $\mu \sim \tau$. τ is some underlying distribution over Δ_k, i.e. $\tau \in \Delta_{\Delta_{[k]}}$
2. For each $i \in [N]$, sample a topic $t_i \sim \mu$ and sample a word $x_i \sim A_{t_i}$ from the topic

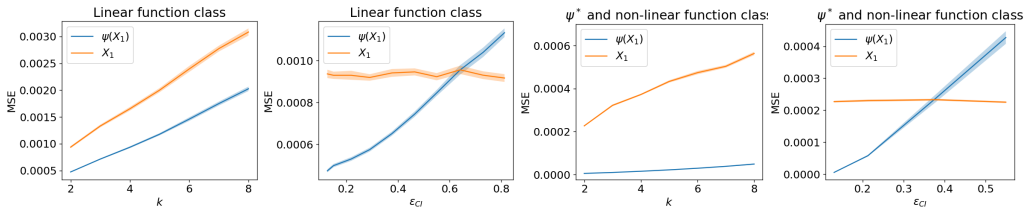
For the reconstruction SSL task, we evenly split the document as $X = (\bar{X}_1, \bar{X}_2)$, where $\bar{X}_1$ and $\bar{X}_2$ denote the first and second halves of the document; note that $\bar{X}_1, \bar{X}_2 \in [V]^{N/2}$. We let X_1 and X_2 be the multiset of words in the two halves by using the normalized bag-of-words representation, i.e. $X_i = \frac{2}{N}\text{bag-of-words}(\bar{X}_i) \in \mathbb{R}^V,\ i \in \{1, 2\}$[7]. The downstream task is chosen to be a linear function of the topic posterior distribution μ for a given document X, i.e. $Y = w^\top \mathbb{E}[\mu|X] + N$, where N is 0 mean and σ^2-subgaussian. The error of a predictor $f : [V]^N \to \mathbb{R}$ is measured as $\mathbb{E}_{\mu,X}\left[\left(f(X) - \mu^\top w\right)^2\right]$, the optimal predictor being $f^*(X) = \mathbb{E}\left[Y \mid X\right]$.

A crucial property of topic model described above is that words in the document are sampled independently given the topic mixture μ, thus giving us the property: $X_1 \perp X_2 \mid \mu$. Although the cardinality of $\mu \in \Delta_{[k]}$ (that implicitly shows up in Theorem 4.2) is infinite, we can still show the benefit of SSL using our theoretical framework. We will show appropriate bounds for ϵ_{CI} and β, that show up in Theorem 4.2, using the topic model generative process. We make the following standard assumptions about the topic modeling distribution, motivated by prior work [4, 3].

Assumption 5.1. *Let $A \in \mathbb{R}^{V \times k}$ be the word-topic matrix and $\Gamma = \mathbb{E}_{\mu \sim \tau}\left[\mu\mu^\top\right]$ be the topic covariance matrix, then the following hold*

- *(Anchor word) The word-topic matrix A is p-separable for $p > 0$, i.e. for every topic $i \in [k]$ there is a word j such that $A_i(j) \geq p$ and $A_{i'}(j) = 0$ when $i' \neq i$*
- *Γ is full rank, so condition number $\kappa = \frac{\lambda_{\max}(\Gamma)}{\lambda_{\min}(\Gamma)} < \infty$*

[7] We only need X_2 to be the bag-of-word representation, X_1 can be an ordered sentence.



(a) (b) (c) (d)

Figure 1: **Left two**: how MSE scales with k (the dimension of Y) and ϵ_{CI} (Definition 4.1) with the linear function class. **Right two**: how MSE scales with k and ϵ with ψ^* and non-linear function class. Mean of 30 trials are shown in solid line and one standard error is shown by shadow.

Theorem 5.1. *Let ϵ_{CI} be the definition (2) from Definition 4.1 and β as defined in Assumption 4.1 and suppose the topic model satisfies Assumption 5.1, then there exists a latent variable $\bar{Y} \in \bar{\mathcal{Y}}$ such that the following hold*

1. *$\bar{Y}$ takes k distinct values, i.e. $|\bar{\mathcal{Y}}| = k$*
2. *X_1 and X_1 are uncorrelated given $\bar{Y}$, which implies $\boldsymbol{\epsilon_{CI} = 0}$.*
3. *$\mathbb{E}[Y|X_1]$ is a linear function of $\mathbb{E}[\bar{Y}|X_1]$*
4. $\boldsymbol{\beta^{-1} \leq \kappa \|w\|_2 / \lambda_{\min}(A)} \leq \kappa \|w\|_2 / p$

The proof for this is presented in Section E.1. Note that the p-separability is not necessarily needed, and the bound with $\lambda_{\min}(A)$ can be invoked instead. Thus the upper bound from Theorem 4.2) will look like $\tilde{\mathcal{O}}\left(\sigma^2 \frac{k}{n_2} + \epsilon_{\text{pre}}^2 \frac{\kappa \|w\|_2}{p}\right)$, thus requiring only $\mathcal{O}(k)$ samples for the downstream task.

6 Experiments

In this section, we empirically verify our claim that SSL performs well when ACI is satisfied. More details for experiments can be found in Section J, including experiments in the text domain.

Simulations. With synthetic data, we verify how excess risk (ER) scales with the cardinality/feature dimension of $\mathcal{Y}$ (k), and ACI (ϵ_{CI} in Definition 4.1). We consider a mixture of Gaussian data and conduct experiments with both linear function space ($\mathcal{H}_1$ with ϕ_1 as identity map) and universal function space $\mathcal{H}_u$. We sample the label Y uniformly from $\{1, ..., k\}$. For i-th class, the centers $\mu_{1i} \in \mathbb{R}^{d_1}$ and $\mu_{2i} \in \mathbb{R}^{d_2}$ are uniformly sampled from $[0, 10)$. Given $Y = i$, $\alpha \in [0, 1]$, let $X_1 \sim \mathcal{N}(\mu_{1i}, \mathbf{I})$, $\hat{X}_2 \sim \mathcal{N}(\mu_{2i}, \mathbf{I})$, and $X_2 = (1 - \alpha)\hat{X}_2 + \alpha X_1$. Therefore α is a correlation coefficient: $\alpha = 0$ ensures X_2 being CI with X_1 given Y and when $\alpha = 1$, X_2 fully depends on X_1. (if $d_1 \neq d_2$, we append zeros or truncate to fit accordingly).

We first conduct experiments with linear function class. We learn a linear representation ψ with n_1 samples and the linear prediction of Y from ψ with n_2 samples. We set $d_1 = 50$, $d_2 = 40$, $n_1 = 4000$, $n_2 = 1000$ and ER is measured with Mean Squared Error (MSE). As shown in Figure 1(a)(b), the MSE of learning with $\psi(X_1)$ scales linearly with k as indicated in Theorem 3.5, and scales linearly with ϵ_{CI} associated with linear function class as indicated in Theorem 4.2. Next we move on to general function class, i.e., $\psi^* = \mathbb{E}[Y|X_1]$ with a closed form solution (see example 3.1). We use the same parameter settings as above. For baseline method, we use kernel linear regression to predict Y using X_1 (we use RBF kernel which also has universal approximation power). As shown in Figure 1(c)(d), the phenomenon is the same as what we observe in the linear function class setting, and hence they respectively verify Theorem 3.2 and Theorem 4.2 with $\mathcal{H}_u$.

Computer Vision Task. We verify if learning from ψ is more effective than learning directly from X_1, in a realistic setting (without enforcing conditional independence). Specifically, we test on the Yearbook dataset [23], and try to predict the date when the portraits are taken (denoted as Y_D), which ranges from 1905 to 2013. We resize all the portraits to be 128 by 128. We crop out the center 64 by 64 pixels (the face), and treat it as X_2, and treat the outer rim as X_1 as shown in Figure 2. Our task is to predict Y_D, which is the year when the portraits are taken, and the year ranges from 1905 to 2013. For ψ, we learn X_2 from X_1 with standard image inpainting techniques [48], and full set of training data (without labels). After that we fix the learned ψ and learn a linear model to predict Y_D from ψ using a smaller set of data (with labels). Besides linear model on X_1, another strong baseline that

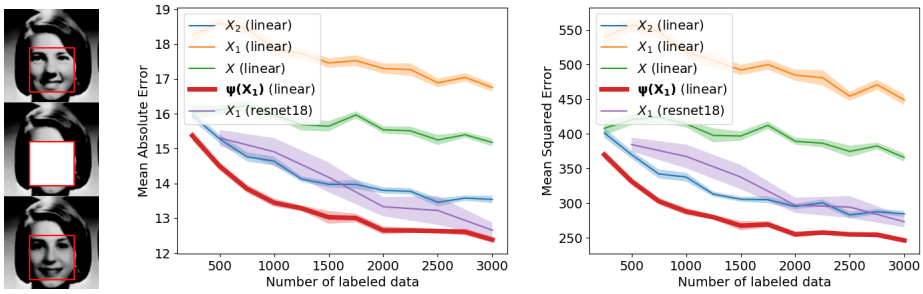


Figure 2: **Left**: Example of the X_2 (in the red box of the 1st row), the X_1 (out of the red box of the 1st row), the input to the inpainting task (the second row), $\psi(X_1)$ (the 3 row in the red box), and in this example $Y = 1967$. **Middle**: Mean Squared Error comparison of yearbook regression predicting dates. **Right**: Mean Absolute Error comparison of yearbook regression predicting dates. Experiments are repeated 10 times, with mean shown in solid line and one standard deviation in shadow.

we compare with is using ResNet18 [28] to predict Y_D from X_1. With the full set of training data, this model is able to achieve a Mean Absolute Difference of 6.89, close to what state-of-the-art can achieve [23]. ResNet18 has similar amount of parameters as our generator, and hence roughly in the same function class. We show the MSE result as in Figure 2. Learning from ψ is more effective than learning from X_1 or X_2 directly, with linear model as well as with ResNet18. Practitioner usually fine-tune ψ with the downstream task, which leads to more competitive performance [48].

7 Conclusion

In this work we theoretically quantify how an approximate conditional independence assumption that connects pretext and downstream task data distributions can give sample complexity benefits of self-supervised learning on downstream tasks. Our theoretical findings are also supported by experiments on simulated data and also on real CV and NLP tasks. We would like to note that approximate CI is only a sufficient condition for a useful pretext task. We leave it for future work to investigate other mechanisms by which pretext tasks help with downstream tasks.

Acknowledgment

JDL acknowledges support of the ARO under MURI Award W911NF-11-1-0304, the Sloan Research Fellowship, NSF CCF 2002272, NSF IIS 2107304, and an ONR Young Investigator Award. QL was supported by NSF #2030859 and the Computing Research Association for the CIFellows Project. NS is supported by NSF, ONR, Simons Foundation, DARPA and SRC. JZ acknowledges sponsorship of Army Research Office and support of Cooperative Agreement Number W911NF-19-2-0333 and NSF grant 2019844. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Office or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

References

[1] Guillaume Alain and Yoshua Bengio. What regularized auto-encoders learn from the data-generating distribution. *The Journal of Machine Learning Research*, 15(1):3563–3593, 2014.

[2] Rie Kubota Ando and Tong Zhang. Two-view feature generation model for semi-supervised learning. In *Proceedings of the 24th international conference on Machine learning*, pages 25–32, 2007.

[3] Sanjeev Arora, Rong Ge, Yonatan Halpern, David Mimno, Ankur Moitra, David Sontag, Yichen Wu, and Michael Zhu. A practical algorithm for topic modeling with provable guarantees. In *International conference on machine learning*. PMLR, 2013.

[4] Sanjeev Arora, Rong Ge, and Ankur Moitra. Learning topic models–going beyond svd. In *2012 IEEE 53rd annual symposium on foundations of computer science*. IEEE, 2012.

[5] Sanjeev Arora, Hrishikesh Khandeparkar, Mikhail Khodak, Orestis Plevrakis, and Nikunj Saunshi. A theoretical analysis of contrastive unsupervised representation learning. In *Proceedings of the 36th International Conference on Machine Learning*, 2019.

[6] Charles R Baker. Joint measures and cross-covariance operators. *Transactions of the American Mathematical Society*, 186:273–289, 1973.

[7] Andrew R Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information theory*, 39(3):930–945, 1993.

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

[9] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 2003.

[10] Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory*, 1998.

[11] Yair Carmon, Aditi Raghunathan, Ludwig Schmidt, John C Duchi, and Percy S Liang. Unlabeled data improves adversarial robustness. In *Advances in Neural Information Processing Systems*, pages 11190–11201, 2019.

[12] Tianlong Chen, Sijia Liu, Shiyu Chang, Yu Cheng, Lisa Amini, and Zhangyang Wang. Adversarial robustness: From self-supervised pre-training to fine-tuning. *arXiv preprint arXiv:2003.12862*, 2020.

[13] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020.

[14] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. Big self-supervised models are strong semi-supervised learners. *arXiv preprint arXiv:2006.10029*, 2020.

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[16] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1422–1430, 2015.

[17] Alexey Dosovitskiy, Philipp Fischer, Jost Tobias Springenberg, Martin Riedmiller, and Thomas Brox. Discriminative unsupervised feature learning with exemplar convolutional neural networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(9):1734–1747, 2015.

[18] Simon S Du, Wei Hu, Sham M Kakade, Jason D Lee, and Qi Lei. Few-shot learning via learning the representation, provably. *arXiv preprint arXiv:2002.09434*, 2020.

[19] Basura Fernando, Hakan Bilen, Efstratios Gavves, and Stephen Gould. Self-supervised video representation learning with odd-one-out networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3636–3645, 2017.

[20] Kenji Fukumizu, Francis R Bach, and Michael I Jordan. Dimensionality reduction for supervised learning with reproducing kernel hilbert spaces. *Journal of Machine Learning Research*, 5(Jan):73–99, 2004.

[21] Kenji Fukumizu, Francis R Bach, Michael I Jordan, et al. Kernel dimension reduction in regression. *The Annals of Statistics*, 37(4):1871–1905, 2009.

[22] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*, 2018.

[23] Shiry Ginosar, Kate Rakelly, Sarah Sachs, Brian Yin, and Alexei A Efros. A century of portraits: A visual historical record of american high school yearbooks. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 1–7, 2015.

[24] Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In *International conference on algorithmic learning theory*, pages 63–77. Springer, 2005.

[25] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint arXiv:2006.07733*, 2020.

[26] David Gross. Recovering low-rank matrices from few coefficients in any basis. *IEEE Transactions on Information Theory*, 57(3):1548–1566, 2011.

[27] Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, 2010.

[28] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[29] Dan Hendrycks, Mantas Mazeika, Saurav Kadavath, and Dawn Song. Using self-supervised learning can improve model robustness and uncertainty. In *Advances in Neural Information Processing Systems*, pages 15637–15648, 2019.

[30] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*, 2018.

[31] Thomas Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, 1999.

[32] Daniel Hsu, Sham M Kakade, and Tong Zhang. Random design analysis of ridge regression. In *Conference on learning theory*, pages 9–1, 2012.

[33] Weihua Hu, Bowen Liu, Joseph Gomes, Marinka Zitnik, Percy Liang, Vijay Pande, and Jure Leskovec. Strategies for pre-training graph neural networks. *arXiv preprint arXiv:1905.12265*, 2019.

[34] Tzee-Ming Huang. Testing conditional independence using maximal nonlinear conditional correlation. *The Annals of Statistics*, 38(4):2047–2091, 2010.

[35] Eric Jang, Coline Devin, Vincent Vanhoucke, and Sergey Levine. Grasp2vec: Learning object representations from self-supervised grasping. *arXiv preprint arXiv:1811.06964*, 2018.

[36] Longlong Jing and Yingli Tian. Self-supervised visual feature learning with deep neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[37] Sham M Kakade and Dean P Foster. Multi-view regression via canonical correlation analysis. In *International Conference on Computational Learning Theory*, pages 82–96. Springer, 2007.

[38] Alexander Kolesnikov, Xiaohua Zhai, and Lucas Beyer. Revisiting self-supervised visual representation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1920–1929, 2019.

[39] Lajanugen Logeswaran and Honglak Lee. An efficient framework for learning sentence representations. In *Proceedings of the International Conference on Learning Representations*, 2018.

[40] Zhuang Ma and Michael Collins. Noise contrastive estimation and negative sampling for conditional models: Consistency and statistical efficiency. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018.

[41] Charles A Micchelli, Yuesheng Xu, and Haizhang Zhang. Universal kernels. *Journal of Machine Learning Research*, 7(Dec):2651–2667, 2006.

[42] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, 2013.

[43] Ishan Misra, C Lawrence Zitnick, and Martial Hebert. Shuffle and learn: unsupervised learning using temporal order verification. In *European Conference on Computer Vision*, pages 527–544. Springer, 2016.

[44] Jovana Mitrovic, Brian McWilliams, Jacob Walker, Lars Buesing, and Charles Blundell. Representation learning via invariant causal mechanisms. *arXiv preprint arXiv:2010.07922*, 2020.

[45] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European Conference on Computer Vision*, pages 69–84. Springer, 2016.

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

[47] Christos H Papadimitriou, Prabhakar Raghavan, Hisao Tamaki, and Santosh Vempala. Latent semantic indexing: A probabilistic analysis. *Journal of Computer and System Sciences*, 2000.

[48] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2536–2544, 2016.

[49] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. *URL https://s3-us-west-2. amazonaws. com/openai-assets/researchcovers/languageunsupervised/language understanding paper. pdf*, 2018.

[50] Michael Reed. *Methods of modern mathematical physics: Functional analysis*. Elsevier, 2012.

[51] Nikunj Saunshi, Sadhika Malladi, and Sanjeev Arora. A mathematical exploration of why language models help solve downstream tasks. *arXiv preprint arXiv:2010.03648*, 2020.

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

[53] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, 2013.

[54] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. *arXiv preprint arXiv:1906.05849*, 2019.

[55] Yuandong Tian, Lantao Yu, Xinlei Chen, and Surya Ganguli. Understanding self-supervised learning with dual deep networks. *arXiv preprint arXiv:2010.00578*, 2020.

[56] Christopher Tosh, Akshay Krishnamurthy, and Daniel Hsu. Contrastive estimation reveals topic posterior information to linear models. *arXiv preprint arXiv:2003.02234*, 2020.

[57] Christopher Tosh, Akshay Krishnamurthy, and Daniel Hsu. Contrastive learning, multi-view redundancy, and linear models. *arXiv preprint arXiv:2008.10150*, 2020.

[58] Yao-Hung Hubert Tsai, Yue Wu, Ruslan Salakhutdinov, and Louis-Philippe Morency. Demystifying self-supervised learning: An information-theoretical framework. *arXiv preprint arXiv:2006.05576*, 2020.

[59] Michael Tschannen, Josip Djolonga, Paul K Rubenstein, Sylvain Gelly, and Mario Lucic. On mutual information maximization for representation learning. *arXiv preprint arXiv:1907.13625*, 2019.

[60] Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural computation*, 23(7):1661–1674, 2011.

[61] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pages 1096–1103, 2008.

[62] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. *arXiv preprint arXiv:2005.10242*, 2020.

[63] Xiaolong Wang and Abhinav Gupta. Unsupervised learning of visual representations using videos. In *Proceedings of the IEEE International Conference on Computer Vision*, 2015.

[64] Donglai Wei, Joseph J Lim, Andrew Zisserman, and William T Freeman. Learning and using the arrow of time. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8052–8060, 2018.

[65] Han Yang, Xiao Yan, Xinyan Dai, and James Cheng. Self-enhanced gnn: Improving graph neural networks using model outputs. *arXiv preprint arXiv:2002.07518*, 2020.

[66] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *European conference on computer vision*, pages 649–666. Springer, 2016.

[67] Richard Zhang, Phillip Isola, and Alexei A Efros. Split-brain autoencoders: Unsupervised learning by cross-channel prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1058–1067, 2017.

[68] Zaiwei Zhang, Zhenxiao Liang, Lemeng Wu, Xiaowei Zhou, and Qixing Huang. Path-invariant map networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11084–11094, 2019.

Checklist

The checklist follows the references. Please read the checklist guidelines carefully for information on how to answer these questions. For each question, change the default **[TODO]** to [Yes] , [No] , or [N/A] . You are strongly encouraged to include a **justification to your answer**, either by referencing the appropriate section of your paper or providing a brief inline description. For example:

- Did you include the license to the code and datasets? [Yes] See Section **??**.
- Did you include the license to the code and datasets? [No] The code and the data are proprietary.
- Did you include the license to the code and datasets? [N/A]

Please do not modify the questions and only use the provided macros for your answers. Note that the Checklist section does not count towards the page limit. In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] See Section 1, especially the contribution paragraph.
   (b) Did you describe the limitations of your work? [Yes] See the Conclusion section and line 314-326.
   (c) Did you discuss any potential negative societal impacts of your work? [No] This is mainly a theoretical work with simulations and should not cause much social impacts.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]

2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] Every assumption is listed as a standalone block. Every theorem or lemma is self-contained.
   (b) Did you include complete proofs of all theoretical results? [Yes] For every theorem, lemma, claim, and some technical remarks, we indicate where their proof is located in the Appendix.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [No] The codes will be made public after this work is accepted for publish.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] The details are provided in the sections 6 and J
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] See Figure 1, Figure 5 and Figure 2.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] The details are provided in the Appendix
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] See Section 6.
   (b) Did you mention the license of the assets? [Yes] The details are provided in the Appendix
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A] All the portrait in the Yearbook dataset are in public domain. Please find the detailed discussion in Section 3 of the Yearbook dataset paper [23]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Some Useful Facts

A.1 Relation of Inverse Covariance Matrix and Partial Correlation

For a covariance matrix of joint distribution for variables X, Y, the covariance matrix is

$$\begin{bmatrix} \mathbf{\Sigma}_{XX} & \mathbf{\Sigma}_{XY} \\ \mathbf{\Sigma}_{YX} & \mathbf{\Sigma}_{YY} \end{bmatrix} = \begin{bmatrix} \mathbf{\Sigma}_{X_1X_1} & \mathbf{\Sigma}_{X_1X_2} & \mathbf{\Sigma}_{X_1Y} \\ \mathbf{\Sigma}_{X_2X_1} & \mathbf{\Sigma}_{X_2X_2} & \mathbf{\Sigma}_{X_2Y} \\ \mathbf{\Sigma}_{YX_1} & \mathbf{\Sigma}_{X_2Y} & \mathbf{\Sigma}_{YY} \end{bmatrix}.$$

Its inverse matrix $\mathbf{\Sigma}^{-1}$ satisfies

$$\mathbf{\Sigma}^{-1} = \begin{bmatrix} \boldsymbol{A} & \rho \\ \rho^\top & \boldsymbol{B} \end{bmatrix}.$$

Here $\boldsymbol{A}^{-1} = \mathbf{\Sigma}_{XX} - \mathbf{\Sigma}_{XY}\mathbf{\Sigma}_{YY}^{-1}\mathbf{\Sigma}_{YX} \equiv \mathrm{cov}(X - \mathbb{E}^L[X|Y], X - \mathbb{E}^L[X|Y]) := \mathbf{\Sigma}_{XX\cdot Y}$, the partial covariance matrix of X given Y.

A.2 Relation to Conditional Independence

Proof of Lemma D.4.

Fact A.1. *When $X_1 \perp X_2 | Y$, the partial covariance between X_1, X_2 given Y is 0:*

$$\begin{aligned} \mathbf{\Sigma}_{X_1X_2\cdot Y} :=& \mathrm{cov}(X_1 - \mathbb{E}^L[X_1|Y], X_2 - \mathbb{E}^L[X_2|Y]) \\ \equiv& \mathbf{\Sigma}_{X_1X_2} - \mathbf{\Sigma}_{X_1Y}\mathbf{\Sigma}_{YY}^{-1}\mathbf{\Sigma}_{YX_2} = 0. \end{aligned}$$

The derivation comes from the following:

Lemma A.1 (Conditional independence (Adapted from [34]))**.** *For random variables X_1, X_2 and a random variable Y with finite values, conditional independence $X_1 \perp X_2 | Y$ is equivalent to:*

$$\sup_{f \in N_1, g \in N_2} \mathbb{E}[f(X_1)g(X_2)|Y] = 0. \tag{4}$$

Here $N_i = \{f : \mathbb{R}^{d_i} \to R : E[f(X_i)|Y] = 0\}$, $i = 1, 2$.

Notice for arbitrary function f, $\mathbb{E}[f(X)|Y] = \mathbb{E}^L[f(X)|\phi_y(Y)]$ with one-hot encoding of discrete variable Y. Therefore for any feature map we can also get that conditional independence ensures:

$$\begin{aligned} \mathbf{\Sigma}_{\phi_1(X_1)\phi_2(X_2)|Y} :=& \mathrm{cov}(\phi_1(X_1) - \mathbb{E}^L[\phi_1(X_1)|\phi_y(Y)], \phi_2(X_2) - \mathbb{E}^L[\phi_2(X_2)|\phi_y(Y)]) \\ =& \mathbb{E}[\bar{\phi}_1(X_1)\bar{\phi}_2(X_2)^\top] = 0. \end{aligned}$$

Here $\bar{\phi}_1(X_1) = \phi_1(X_1) - \mathbb{E}[\phi_1(X_1)|\phi_y(Y)]$ is mean zero given Y, and vice versa for $\bar{\phi}_2(X_2)$. This thus finishes the proof for Lemma D.4. □

A.3 Technical Facts for Matrix Concentration

We include this covariance concentration result that is adapted from Claim A.2 in [18]:

Claim A.2 (covariance concentration for gaussian variables)**.** *Let $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots \boldsymbol{x}_n]^\top \in \mathbb{R}^{n\times d}$ where each $x_i \sim \mathcal{N}(0, \mathbf{\Sigma}_X)$. Suppose $n \gg k + \log(1/\delta)$ for $\delta \in (0, 1)$. Then for any given matrix $B \in \mathbb{R}^{d\times m}$ that is of rank k and is independent of $\boldsymbol{X}$, with probability at least $1 - \frac{\delta}{10}$ over $\boldsymbol{X}$ we have*

$$0.9\boldsymbol{B}^\top\mathbf{\Sigma}_X\boldsymbol{B} \preceq \frac{1}{n}\boldsymbol{B}^\top\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{B} \preceq 1.1\boldsymbol{B}^\top\mathbf{\Sigma}_X\boldsymbol{B}. \tag{5}$$

And we will also use Claim A.2 from [18] for concentrating subgaussian random variable.

Claim A.3 (covariance concentration for subgaussian variables)**.** *Let $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots \boldsymbol{x}_n]^\top \in \mathbb{R}^{n\times d}$ where each $\boldsymbol{x}_i$ is ρ^2-sub-gaussian. Suppose $n \gg \rho^4(k + \log(1/\delta))$ for $\delta \in (0, 1)$. Then for any given matrix $B \in \mathbb{R}^{d\times m}$ that is of rank k and is independent of $\boldsymbol{X}$, with probability at least $1 - \frac{\delta}{10}$ over $\boldsymbol{X}$ we have*

$$0.9\boldsymbol{B}^\top\mathbf{\Sigma}_X\boldsymbol{B} \preceq \frac{1}{n}\boldsymbol{B}^\top\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{B} \preceq 1.1\boldsymbol{B}^\top\mathbf{\Sigma}_X\boldsymbol{B}. \tag{6}$$

Claim A.4. *Let $\boldsymbol{Z} \in \mathbb{R}^{n\times k}$ be a matrix with row vectors sampled from i.i.d Gaussian distribution $\mathcal{N}(0, \boldsymbol{\Sigma}_Z)$. Let $\mathbf{P} \in \mathbb{R}^{n\times n}$ be a fixed projection onto a space of dimension d. Then with a fixed $\delta \in (0,1)$, we have:*

$$\|\mathbf{P}\boldsymbol{Z}\|_F^2 \lesssim \mathrm{Tr}(\boldsymbol{\Sigma}_Z)(d + \log(k/\delta)),$$

with probability at least $1-\delta$.

Claim A.4. Each t-th column of Z is an n-dim vector that is i.i.d sampled from Gaussian distribution $\mathcal{N}(0, \boldsymbol{\Sigma}_{tt})$.

$$\begin{aligned}\|\mathbf{P}\boldsymbol{Z}\|_F^2 &= \sum_{t=1}^{k} \|\mathbf{P}\boldsymbol{z}_t\|^2 \\ &= \sum_{t=1}^{k} \boldsymbol{z}_t^\top \mathbf{P}\boldsymbol{z}_t.\end{aligned}$$

Each term satisfy $\boldsymbol{\Sigma}_{kk}^{-1}\|\mathbf{P}\boldsymbol{z}_t\|^2 \sim \chi^2(d)$, and therefore with probability at least $1-\delta'$ over $\boldsymbol{z}_t$,

$$\boldsymbol{\Sigma}_{kk}^{-1}\|\mathbf{P}\boldsymbol{z}_t\|^2 \lesssim d + \log(1/\delta').$$

Using union bound, take $\delta' = \delta/k$ and summing over $t \in [k]$ we get:

$$\|\mathbf{P}\boldsymbol{Z}\|_F^2 \lesssim \mathrm{Tr}(\boldsymbol{\Sigma}_Z)(d + \log(k/\delta)).$$

□

Theorem A.5 (Vector Bernstein Inequality (Theorem 12 in [26])). *Let $X_1, \cdots, X_m$ be independent zero-mean vector-valued random variables. Let*

$$N = \|\sum_{i=1}^{m} X_i\|_2.$$

Then

$$\mathbb{P}[N \geq \sqrt{V} + t] \leq \exp\left(\frac{-t^2}{4V}\right),$$

where $V = \sum_i \mathbb{E}\|X_i\|_2^2$ and $t \leq V/(\max \|X_i\|_2)$.

Lemma A.6. *Let $\boldsymbol{Z} \in \mathbb{R}^{n\times d}$ be a matrix whose row vectors are n independent mean-zero (conditional on $\mathbf{P}$ being a rank-d projection matrix) σ-sub-Gaussian random vectors. With probability $1-\delta$:*

$$\|\mathbf{P}\boldsymbol{Z}\|_F^2 \lesssim \sigma^2(d + \log(d/\delta)).$$

Proof of Lemma A.6. Write $\mathbf{P} = \boldsymbol{U}\boldsymbol{U}^\top = [\boldsymbol{u}_1, \cdots, \boldsymbol{u}_d]$ where $\boldsymbol{U}$ is orthogonal matrix in $\mathbb{R}^{n\times d}$ where $\boldsymbol{U}^\top\boldsymbol{U} = I$. Notice $\|\boldsymbol{U}\boldsymbol{U}^\top\boldsymbol{Z}\|_F^2 = \mathrm{Tr}(\boldsymbol{Z}^\top UU^\top UU^\top \boldsymbol{Z}) = \mathrm{Tr}(\boldsymbol{Z}^\top UU^\top \boldsymbol{Z})$. Therefore:

$$\begin{aligned}\|\mathbf{P}\boldsymbol{Z}\|_F^2 &= \|\boldsymbol{U}^\top\boldsymbol{Z}\|_F^2 \\ &= \sum_{j=1}^{d} \|\boldsymbol{u}_j^\top \boldsymbol{Z}\|^2 \\ &= \sum_{j=1}^{d} \|\sum_{i=1}^{n} \boldsymbol{u}_{ji}\boldsymbol{z}_i\|^2,\end{aligned}$$

where each $\boldsymbol{z}_i \in \mathbb{R}^k$ being the i-th row of $\boldsymbol{Z}$ is a centered independent σ sub-Gaussian random vectors. To use vector Bernstein inequality, we let $X := \sum_{i=1}^{n} X_i$ with X_i taking the value of $\boldsymbol{u}_{ji}\boldsymbol{z}_i$.

We have X_i is zero mean: $\mathbb{E}[X_i] = \mathbb{E}[\boldsymbol{u}_{ji}\mathbb{E}[\boldsymbol{z}_i|\boldsymbol{u}_{ji}]] = \mathbb{E}[\boldsymbol{u}_{ji} \cdot 0] = 0$.

$$\begin{aligned} V :=& \sum_i \mathbb{E}\|X_i\|_2^2 \\ =& \sum_i \mathbb{E}[\boldsymbol{u}_{ji}^2 \boldsymbol{z}_i^\top \boldsymbol{z}_i] \\ =& \sum_i \mathbb{E}_{\boldsymbol{u}_{ji}}[\boldsymbol{u}_{ji}^2 \mathbb{E}[\|\boldsymbol{z}_i\|_2^2|\boldsymbol{u}_{ji}]] \\ \leq& \sigma^2 \sum_i \mathbb{E}_{\boldsymbol{u}_{ji}}[\boldsymbol{u}_{ji}^2] \\ =& \sigma^2. \end{aligned}$$

Therefore by vector Bernstein Inequality, with probability at least $1-\delta/d$, $\|X\| \leq \sigma(1+\sqrt{\log(d/\delta)})$. Then by taking union bound, we get that $\|\mathbf{P}\boldsymbol{Z}\|^2 = \sum_{j=1}^d \|\boldsymbol{u}_j^\top \boldsymbol{Z}\|^2 \lesssim \sigma^2 d(1+\log(d/\delta))$ with probability $1-\delta$.

□

B Warm-up: jointly Gaussian variables

We assume X_1, X_2, Y are jointly Gaussian, and so the optimal regression functions are all linear, i.e., $\mathbb{E}[Y|X_1] = \mathbb{E}^L[Y|X_1]$. We also assume data is centered: $\mathbb{E}[X_i] = 0$ and $\mathbb{E}[Y] = 0$. Non-centered data can easily be handled by learning an intercept. All relationships between random variables can then be captured by the (partial) covariance matrix. Therefore it is easy to quantify the CI property and establish the necessary and sufficient conditions that make X_2 a reasonable pretext task.

Assumption B.1 (Jointly Gaussian)**.** *X_1, X_2, Y are jointly Gaussian.*

Assumption B.2 (Conditional independence)**.** *$X_1 \perp X_2 | Y$.*

Claim B.1 (Closed-form solution)**.** *Under Assumption B.1, the representation function and optimal prediction that minimize the population risk can be expressed as follows:*

$$\psi^*(\boldsymbol{x}_1) := \mathbb{E}^L[X_2|X_1 = \boldsymbol{x}_1] = \boldsymbol{\Sigma}_{X_2X_1}\boldsymbol{\Sigma}_{X_1X_1}^{-1}\boldsymbol{x}_1 \tag{7}$$

$$\textit{Our target } f^*(\boldsymbol{x}_1) := \mathbb{E}^L[Y|X_1 = \boldsymbol{x}_1] = \boldsymbol{\Sigma}_{YX_1}\boldsymbol{\Sigma}_{X_1X_1}^{-1}\boldsymbol{x}_1. \tag{8}$$

Our prediction for downstream task with representation ψ^* will be: $g(\cdot) := \mathbb{E}^L[Y|\psi^*(X_1)]$. Recall from Equation (1) that the partial covariance matrix between X_1 and X_2 given Y is $\boldsymbol{\Sigma}_{X_1X_2|Y} \equiv \boldsymbol{\Sigma}_{X_1X_2} - \boldsymbol{\Sigma}_{X_1Y}\boldsymbol{\Sigma}_{YY}^{-1}\boldsymbol{\Sigma}_{YX_2}$. This partial covariance matrix captures the correlation between X_1 and X_2 given Y. For jointly Gaussian random variables, CI is equivalent to $\boldsymbol{\Sigma}_{X_1X_2|Y} = 0$. We first analyze the approximation error based on the property of this partial covariance matrix.

Lemma B.2 (Approximation error)**.** *Under Assumption B.1, B.2, if $\boldsymbol{\Sigma}_{X_2Y}$ has rank k, we have $f^*(\boldsymbol{x}_1) \equiv \boldsymbol{W}^*\psi^*(\boldsymbol{x}_1)$, i.e., $e_{apx}(\psi^*) = 0$.*

Remark B.1. *$\boldsymbol{\Sigma}_{X_2Y}$ being full column rank implies that $\mathbb{E}[X_2|Y]$ has rank k, i.e., X_2 depends on all directions of Y and thus captures all directions of information of Y. This is a necessary assumption for X_2 to be a reasonable pretext task for predicting Y. $e_{apx}(\psi^*) = 0$ means f^* is linear in ψ^*. Therefore ψ^* selects d_2 out of d_1 features that are sufficient to predict Y.*

Next we consider the estimation error that characterizes the number of samples needed to learn a prediction function $f(\boldsymbol{x}_1) = \hat{\boldsymbol{W}}\psi^*(\boldsymbol{x}_1)$ that generalizes.

Theorem B.3 (Excess risk)**.** *Fix a failure probability $\delta \in (0,1)$. Under Assumption B.1,B.2, if $n_2 \gg k + \log(1/\delta)$, excess risk of the learned predictor $\boldsymbol{x}_1 \to \hat{\boldsymbol{W}}\psi^*(\boldsymbol{x}_1)$ on the target task satisfies*

$$\mathrm{ER}_{\psi^*}(\hat{\boldsymbol{W}}) \leq \mathcal{O}\left(\frac{\mathrm{Tr}(\boldsymbol{\Sigma}_{YY|X_1})(k+\log(k/\delta))}{n_2}\right),$$

with probability at least $1-\delta$.

Here $\mathbf{\Sigma}_{YY|X_1} \equiv \mathbf{\Sigma}_{YY} - \mathbf{\Sigma}_{YX_1}\mathbf{\Sigma}_{X_1X_1}^{-1}\mathbf{\Sigma}_{X_1Y}$ captures the noise level and is the covariance matrix of the residual term $Y - f^*(X_1) = Y - \mathbf{\Sigma}_{YX_1}\mathbf{\Sigma}_{X_1X_1}^{-1}X_1$. Compared to directly using X_1 to predict Y, self-supervised learning reduces the sample complexity from $\tilde{\mathcal{O}}(d_1)$ to $\tilde{\mathcal{O}}(k)$. We generalize these results even when only a weaker form of CI holds.

Assumption B.3 (Conditional independence given latent variables)**.** *There exists some latent variable $Z \in \mathbb{R}^m$ such that $X_1 \perp X_2 | \bar{Y}$, and $\mathbf{\Sigma}_{X_2\bar{Y}}$ is of rank $k+m$, where $\bar{Y} = [Y, Z]$.*

This assumption lets introduce some reasonable latent variables that capture the information between X_1 and X_2 apart from Y. $\mathbf{\Sigma}_{X_2\bar{Y}}$ being full rank says that all directions of $\bar{Y}$ are needed to predict X_2, and therefore Z is not redundant. For instance, when $Z = X_1$, the assumption is trivially true but Z is not the minimal latent information we want to add. Note it implicitly requires $d_2 \geq k+m$.

Corollary B.4. *Under Assumption B.1, B.3, we have $f^*(\boldsymbol{x}_1) \equiv \boldsymbol{W}^*\psi^*(\boldsymbol{x}_1)$, i.e., the approximation error $e_{apx}(\psi^*)$ is 0. We can also generalize Theorem B.3 by replacing k by $k+m$.*

C Omitted Proofs with Conditional Independence

Proof of Lemma B.2.

$$\mathrm{cov}(X_1|Y, X_2|Y) = \mathbf{\Sigma}_{X_1X_2} - \mathbf{\Sigma}_{X_1Y}\mathbf{\Sigma}_{YY}^{-1}\mathbf{\Sigma}_{YX_2} = 0.$$

By plugging it into the expression of $\mathbb{E}^L[X_2|X_1]$, we get that

$$\begin{aligned}\psi(x_1) := \mathbb{E}^L[X_2|X_1 = x_1] &= \mathbf{\Sigma}_{X_2X_1}\mathbf{\Sigma}_{X_1X_1}^{-1}x_1 \\ &= \mathbf{\Sigma}_{X_2Y}\mathbf{\Sigma}_{YY}^{-1}\mathbf{\Sigma}_{YX_1}\mathbf{\Sigma}_{X_1X_1}^{-1}x_1 \\ &= \mathbf{\Sigma}_{X_2Y}\mathbf{\Sigma}_{YY}^{-1}\mathbb{E}^L[Y|X_1].\end{aligned}$$

Therefore, as long as $\mathbf{\Sigma}_{X_2Y}$ is rank k, it has left inverse matrix and we get: $\mathbb{E}^L[Y|X_1 = x_1] = \mathbf{\Sigma}_{X_2Y}^{\dagger}\mathbf{\Sigma}_{YY}\psi(x_1)$. Therefore there's no approximation error in using ψ to predict Y.

□

Proof of Corollary B.4. Let selector operator $\boldsymbol{S}_y$ be the mapping such that $\boldsymbol{S}_y\bar{Y} = Y$, we overload it as the matrix that ensure $\boldsymbol{S}_y\mathbf{\Sigma}_{\bar{Y}X} = \mathbf{\Sigma}_{YX}$ for any random variable X as well.

From Lemma B.2 we get that there exists W such that $\mathbb{E}^L[\bar{Y}|X_1] = \boldsymbol{W}\mathbb{E}^L[X_2|X_1]$, just plugging in $\boldsymbol{S}_y$ we get that $\mathbb{E}^L[Y|X_1] = (\boldsymbol{S}_y\boldsymbol{W})\mathbb{E}^L[X_2|X_1]$.

□

Proof of Theorem B.3. Write $f^*(X_1) = \mathbb{E}[Y|X_1] = (\boldsymbol{A}^*)^\top X_1$. $\mathbb{E}^L[Y|X_1 = x_1] = \mathbf{\Sigma}_{X_2Y}^{\dagger}\mathbf{\Sigma}_{YY}\psi(x_1)$. Let $\boldsymbol{W}^* = \mathbf{\Sigma}_{YY}\mathbf{\Sigma}_{YX_2}^{\dagger}$. From Lemma B.2 we know $f^* = \boldsymbol{W}^*\psi$. Recall noise $N = Y - f^*(X_1)$ is mean zero conditional on X_1. We write $\boldsymbol{N} = \boldsymbol{Y} - f^*(\boldsymbol{X}_1)$.

First we have the basic inequality,

$$\begin{aligned}\frac{1}{2n_2}\|\boldsymbol{Y} - \psi(\boldsymbol{X}_1)\hat{\boldsymbol{W}}\|_F^2 \leq& \frac{1}{2n_2}\|\boldsymbol{Y} - \boldsymbol{X}_1A^*\|_F^2 \\ =& \frac{1}{2n_2}\|\boldsymbol{Y} - \psi(\boldsymbol{X}_1)\boldsymbol{W}^*\|_F^2 = \frac{1}{2n_2}\|\boldsymbol{N}\|_F^2.\end{aligned}$$

Therefore by rearranging both sides, we have:

$$\begin{aligned}\|\psi(\boldsymbol{X}_1)\boldsymbol{W}^* - \psi(\boldsymbol{X}_1)\hat{\boldsymbol{W}}\|^2 \leq& 2\langle \boldsymbol{N}, \psi(\boldsymbol{X}_1)\boldsymbol{W}^* - \psi(\boldsymbol{X}_1)\hat{\boldsymbol{W}}\rangle \\ =& 2\langle P_{\psi(\boldsymbol{X}_1)}\boldsymbol{N}, \psi(\boldsymbol{X}_1)\boldsymbol{W}^* - \psi(\boldsymbol{X}_1)\hat{W}\rangle \\ \leq& 2\|P_{\psi(\boldsymbol{X}_1)}\boldsymbol{N}\|_F\|\psi(\boldsymbol{X}_1)\boldsymbol{W}^* - \psi(\boldsymbol{X}_1)\hat{W}\|_F \\ \Rightarrow \|\psi(\boldsymbol{X}_1)\boldsymbol{W}^* - \psi(\boldsymbol{X}_1)\hat{\boldsymbol{W}}\| \leq& 2\|P_{\psi(\boldsymbol{X}_1)}\boldsymbol{N}\|_F \\ \lesssim& \sqrt{\mathrm{Tr}(\mathbf{\Sigma}_{YY|X_1})(k + \log k/\delta)}. \qquad \text{(from Claim A.4)}\end{aligned}$$

The last inequality is derived from Claim A.4 and the fact that each row of $\boldsymbol{N}$ follows gaussian distribution $\mathcal{N}(0, \boldsymbol{\Sigma}_{YY|X_1})$. Therefore

$$\frac{1}{n_2}\|\psi(\boldsymbol{X}_1)W^* - \psi(\boldsymbol{X}_1)\hat{W}\|_F^2 \lesssim \frac{\mathrm{Tr}(\boldsymbol{\Sigma}_{YY|X_1})(k + \log k/\delta)}{n_2}.$$

Next we need to concentrate $1/n\boldsymbol{X}_1^\top \boldsymbol{X}_1$ to $\boldsymbol{\Sigma}_X$. Suppose $\mathbb{E}^L[X_2|X_1] = \boldsymbol{B}^\top X_1$, i.e., $\psi(x_1) = \boldsymbol{B}^\top x_1$, and $\psi(\boldsymbol{X}_1) = \boldsymbol{X}_1\boldsymbol{B}$. With Claim A.2 we have $1/n\psi(\boldsymbol{X}_1)^\top\psi(\boldsymbol{X}_1) = 1/n\boldsymbol{B}^\top\boldsymbol{X}_1^\top\boldsymbol{X}_1\boldsymbol{B}$ satisfies:

$$0.9\boldsymbol{B}^\top\boldsymbol{\Sigma}_X\boldsymbol{B} \preceq 1/n_2\psi(\boldsymbol{X}_1)^\top\psi(\boldsymbol{X}_1) \preceq 1.1\boldsymbol{B}^\top\boldsymbol{\Sigma}_X\boldsymbol{B}$$

Therefore we also have:

$$\begin{aligned}
&\mathbb{E}[\|(\boldsymbol{W}^* - \hat{\boldsymbol{W}})^\top\psi(x_1)\|^2] \\
=&\|\boldsymbol{\Sigma}_X^{1/2}\boldsymbol{B}(\boldsymbol{W}^* - \hat{\boldsymbol{W}})\|_F^2 \\
\leq&\frac{1}{0.9n_2}\|\psi(\boldsymbol{X}_1)\boldsymbol{W}^* - \psi(\boldsymbol{X}_1)\hat{\boldsymbol{W}}\|_F^2 \lesssim \frac{\mathrm{Tr}(\boldsymbol{\Sigma}_{YY|X_1})(k + \log k/\delta)}{n_2}.
\end{aligned}$$

□

C.1 Omitted Proof for General Random Variables

Proof of Lemma 3.1. Let the representation function ψ be defined as:

$$\begin{aligned}
\psi(\cdot) := \mathbb{E}[X_2|X_1] =&\mathbb{E}[\mathbb{E}[X_2|X_1, Y]|X_1] \\
=&\mathbb{E}[\mathbb{E}[X_2|Y]|X_1] && \text{(uses CI)} \\
=&\sum_y P(Y = y|X_1)\mathbb{E}[X_2|Y = y] \\
=:&f(X_1)^\top\boldsymbol{A},
\end{aligned}$$

where $f : \mathbb{R}^{d_1} \to \Delta_{\mathcal{Y}}$ satisfies $f(x_1)_y = P(Y = y|X_1 = x_1)$, and $\boldsymbol{A} \in \mathbb{R}^{\mathcal{Y}\times d_2}$ satisfies $\boldsymbol{A}_{y,:} = \mathbb{E}[X_2|Y = y]$. Here Δ_d denotes simplex of dimension d, which represents the discrete probability density over support of size d.

Let $\boldsymbol{B} = \boldsymbol{A}^\dagger \in \mathbb{R}^{\mathcal{Y}\times d_2}$ be the pseudoinverse of matrix A, and we get $\boldsymbol{B}\boldsymbol{A} = \boldsymbol{I}$ from our assumption that $\boldsymbol{A}$ is of rank $|\mathcal{Y}|$. Therefore $f(x_1) = \boldsymbol{B}\psi(x_1), \forall x_1$. Next we have:

$$\begin{aligned}
\mathbb{E}[Y|X_1 = x_1] =&\sum_y P(Y = y|X_1 = x_1) \times y \\
=&\boldsymbol{Y}f(x_1) \\
=&(\boldsymbol{Y}\boldsymbol{B}) \cdot \psi(X_1).
\end{aligned}$$

Here we denote by $\boldsymbol{Y} \in \mathbb{R}^{k\times\mathcal{Y}}, \boldsymbol{Y}_{:,y} = y$ that spans the whole support $\mathcal{Y}$. Therefore let $\boldsymbol{W}^* = \boldsymbol{Y}\boldsymbol{B}$ will finish the proof.

□

Proof of Theorem 3.2. With Lemma 3.1 we know $e_{\text{apx}} = 0$, and therefore $\boldsymbol{W}^*\psi(X_1) \equiv f^*(X_1)$. Next from basic inequality and the same proof as in Theorem B.3 we have:

$$\|\psi(\boldsymbol{X}_1)\boldsymbol{W}^* - \psi(\boldsymbol{X}_1)\hat{\boldsymbol{W}}\| \leq 2\|\mathbf{P}_{\psi(\boldsymbol{X}_1)}\boldsymbol{N}\|_F$$

Notice $\mathcal{N}$ is a random noise matrix whose row vectors are independent samples from some centered distribution. Note we assumed $\mathbb{E}[\|N\|^2|\boldsymbol{X}_1] \leq \sigma^2$. $P_{\psi(\boldsymbol{X}_1)}$ is a projection to dimension k. From Lemma A.6 we have:

$$\|f^*(\boldsymbol{X}_1) - \psi(\boldsymbol{X}_1)\hat{\boldsymbol{W}}\| \leq \sigma\sqrt{k(1 + \log k/\delta)}.$$

Next, with Claim A.3 we have when $n \gg \rho^4(k+\log(1/\delta))$, since $\boldsymbol{W}^* - \hat{\boldsymbol{W}} \in \mathbb{R}^{d_2 \times k}$,

$$\begin{aligned}
&0.9(\boldsymbol{W}^* - \hat{\boldsymbol{W}})^\top \boldsymbol{\Sigma}_\psi (\boldsymbol{W}^* - \hat{\boldsymbol{W}}) \\
\preceq &\frac{1}{n_2}(\boldsymbol{W}^* - \hat{\boldsymbol{W}})^\top \sum_i \psi(x_1^{(i)})\psi(x_1^{(i)})^\top (\boldsymbol{W}^* - \hat{\boldsymbol{W}}) \preceq 1.1(\boldsymbol{W}^* - \hat{\boldsymbol{W}})^\top \boldsymbol{\Sigma}_\psi (\boldsymbol{W}^* - \hat{\boldsymbol{W}})
\end{aligned}$$

And therefore we could easily conclude that:

$$\mathbb{E}\|\hat{\boldsymbol{W}}^\top \psi(X_1) - f^*(X_1)\|^2 \lesssim \sigma^2 \frac{k(1+\log(k/\delta))}{n_2}.$$

□

C.2 Omitted proof of linear model with approximation error

Proof of Theorem 3.5. First we note that $Y = f^*(X_1) + N$, where $\mathbb{E}[N|X_1] = 0$ but $Y - (\boldsymbol{A}^*)^\top X_1$ is not necessarily mean zero, and this is where additional difficulty lies. Write approximation error term $a(X_1) := f^*(X_1) - (\boldsymbol{A}^*)^\top X_1$, namely $Y = a(X_1) + (\boldsymbol{A}^*)^\top X_1 + N$. Also, $(\boldsymbol{A}^*)^\top X_1 \equiv (\boldsymbol{W}^*)^\top \psi(X_1)$ with conditional independence.

Second, with KKT condition on the training data, we know that $\mathbb{E}[a(X_1)X_1^\top] = 0$.

Recall $\hat{\boldsymbol{W}} = \arg\min_{\boldsymbol{W}} \|\boldsymbol{Y} - \psi(\boldsymbol{X}_1)\boldsymbol{W}\|_F^2$. We have the basic inequality,

$$\begin{aligned}
\frac{1}{2n_2}\|\boldsymbol{Y} - \psi(\boldsymbol{X}_1)\hat{\boldsymbol{W}}\|_F^2 \leq &\frac{1}{2n_2}\|\boldsymbol{Y} - \boldsymbol{X}_1\boldsymbol{A}^*\|_F^2 \\
= &\frac{1}{2n_2}\|\boldsymbol{Y} - \psi(\boldsymbol{X}_1)\boldsymbol{W}^*\|_F^2. \\
\text{i.e., } \frac{1}{2n_2}\|\psi(\boldsymbol{X}_1)\boldsymbol{W}^* + a(\boldsymbol{X}_1) + \boldsymbol{N} - \psi(\boldsymbol{X}_1)\hat{\boldsymbol{W}}\|_F^2 \leq &\frac{1}{2n_2}\|a(\boldsymbol{X}_1) + \boldsymbol{N}\|_F^2.
\end{aligned}$$

Therefore

$$\begin{aligned}
&\frac{1}{2n_2}\|\psi(\boldsymbol{X}_1)\boldsymbol{W}^* - \psi(\boldsymbol{X}_1)\hat{\boldsymbol{W}}\|^2 \\
\leq &-\frac{1}{n_2}\langle a(\boldsymbol{X}_1) + \boldsymbol{N}, \psi(\boldsymbol{X}_1)\boldsymbol{W}^* - \psi(\boldsymbol{X}_1)\hat{\boldsymbol{W}}\rangle \\
= &-\frac{1}{n_2}\langle a(\boldsymbol{X}_1), \psi(\boldsymbol{X}_1)\boldsymbol{W}^* - \psi(\boldsymbol{X}_1)\hat{\boldsymbol{W}}\rangle - \langle \boldsymbol{N}, \psi(\boldsymbol{X}_1)\boldsymbol{W}^* - \psi(\boldsymbol{X}_1)\hat{\boldsymbol{W}}\rangle
\end{aligned} \tag{9}$$

With Assumption 3.3 and by concentration $0.9\frac{1}{n_2}\boldsymbol{X}_1\boldsymbol{X}_1^\top \preceq \boldsymbol{\Sigma}_{X_1} \preceq 1.1\frac{1}{n_2}\boldsymbol{X}_1\boldsymbol{X}_1^\top$, we have

$$\frac{1}{\sqrt{n_2}}\|a(\boldsymbol{X}_1)\boldsymbol{X}_1^\top \boldsymbol{\Sigma}_{X_1}^{-1/2}\|_F \leq 1.1 b_0\sqrt{k} \tag{10}$$

Denote $\psi(\boldsymbol{X}_1) = \boldsymbol{X}_1\boldsymbol{B}$, where $\boldsymbol{B} = \boldsymbol{\Sigma}_{X_1}^{-1}\boldsymbol{\Sigma}_{X_1X_2}$ is rank k under exact CI since $\boldsymbol{\Sigma}_{X_1X_2} = \boldsymbol{\Sigma}_{X_1Y}\boldsymbol{\Sigma}_Y^{-1}\boldsymbol{\Sigma}_{YX_2}$. We have

$$\begin{aligned}
&\frac{1}{n_2}\langle a(\boldsymbol{X}_1), \psi(\boldsymbol{X}_1)\boldsymbol{W}^* - \psi(\boldsymbol{X}_1)\hat{\boldsymbol{W}}\rangle \\
= &\frac{1}{n_2}\langle a(\boldsymbol{X}_1), \boldsymbol{X}_1\boldsymbol{B}\boldsymbol{W}^* - \boldsymbol{X}_1\boldsymbol{B}\hat{\boldsymbol{W}}\rangle \\
= &\frac{1}{n_2}\langle \boldsymbol{\Sigma}_{X_1}^{-1/2}\boldsymbol{X}_1^\top a(\boldsymbol{X}_1), \boldsymbol{\Sigma}_{X_1}^{1/2}(\boldsymbol{B}\boldsymbol{W}^* - \boldsymbol{B}\hat{\boldsymbol{W}})\rangle \\
\leq &1.1 b_0\sqrt{\frac{k}{n_2}}\|\boldsymbol{\Sigma}_{X_1}^{1/2}(\boldsymbol{B}\boldsymbol{W}^* - \boldsymbol{B}\hat{\boldsymbol{W}})\|_F \qquad \text{(from Ineq. (10))}
\end{aligned}$$

Back to Eqn. (9), we get

$$
\begin{aligned}
&\frac{1}{2n_2}\|\psi(\boldsymbol{X}_1)\boldsymbol{W}^* - \psi(\boldsymbol{X}_1)\hat{\boldsymbol{W}}\|_F^2 \\
\lesssim &\sqrt{\frac{k}{n_2}}\|\boldsymbol{\Sigma}_{X_1}^{1/2}(\boldsymbol{B}\boldsymbol{W}^* - \boldsymbol{B}\hat{\boldsymbol{W}})\|_F + \frac{1}{n_2}\|P_{\boldsymbol{X}_1}\boldsymbol{N}\|_F\|\boldsymbol{X}_1(\boldsymbol{B}\boldsymbol{W}^* - \boldsymbol{B}\hat{\boldsymbol{W}})\|_F \\
\lesssim &\left(\frac{\sqrt{k}}{n_2} + \frac{1}{n_2}\|P_{\boldsymbol{X}_1}\boldsymbol{N}\|_F\right)\|\boldsymbol{X}_1(\boldsymbol{B}\boldsymbol{W}^* - \boldsymbol{B}\hat{\boldsymbol{W}})\|_F \\
\Longrightarrow &\frac{1}{\sqrt{n_2}}\|\psi(\boldsymbol{X}_1)\boldsymbol{W}^* - \psi(\boldsymbol{X}_1)\hat{\boldsymbol{W}}\|_F \lesssim \sqrt{\frac{k(1+\log k/\delta)}{n_2}}. \qquad \text{(from Lemma A.6)}
\end{aligned}
$$

Finally, by concentration we transfer the result from empirical loss to excess risk and get:

$$
\mathbb{E}[\|\psi(X_1)\boldsymbol{W}^* - \psi(X_1)\hat{\boldsymbol{W}}\|^2] \lesssim \frac{k(1+\log(k/\delta))}{n_2}.
$$

□

C.3 Argument on Denoising Auto-encoder or Context Encoder

Remark C.1. *We note that since $X_1 \perp X_2 | Y$ ensures $X_1 \perp h(X_2)|Y$ for any deterministic function h, we could replace X_2 by $h(X_2)$ and all results hold. Therefore in practice, we could use $h(\psi(X_1))$ instead of $\psi(X_1)$ for downstream task. Specifically with denoising auto-encoder or context encoder, one could think about h as the inverse of decoder D ($h = D^{-1}$) and use $D^{-1}\psi \equiv E$ the encoder function as the representation for downstream tasks, which is more commonly used in practice.*

This section explains what we claim in Remark C.1. For context encoder, the reconstruction loss targets to find the encoder E^* and decoder D^* that achieve

$$
\min_E \min_D \mathbb{E}\|X_2 - D(E(X_1))\|_F^2, \tag{11}
$$

where X_2 is the masked part we want to recover and X_1 is the remainder.

If we naively apply our theorem we should use $D^*(E^*(\cdot))$ as the representation, while in practice we instead use only the encoder part $E^*(\cdot)$ as the learned representation. We argue that our theory also support this practical usage if we view the problem differently. Consider the pretext task to predict $(D^*)^{-1}(X_2)$ instead of X_2 directly, namely,

$$
\bar{E} \leftarrow \arg\min_E \mathbb{E}\|(D^*)^{-1}(X_2) - E(X_1)\|^2, \tag{12}
$$

and then we should indeed use $E(X_1)$ as the representation. On one hand, when $X_1 \perp X_2 | Y$, it also satisfies $X_1 \perp (D^*)^{-1}(X_2)|Y$ since $(D^*)^{-1}$ is a deterministic function of X_2 and all our theory applies. On the other hand, the optimization on (11) or (12) give us similar result. Let

$$
E^* = \arg\min_E \mathbb{E}[\|X_2 - D^*(E(X_1))\|^2],
$$

and $\mathbb{E}\|X_2 - D^*(E^*(X_1))\|^2 \le \epsilon$, then with pretext task as in (12) we have that:

$$
\begin{aligned}
\mathbb{E}\|(D^*)^{-1}(X_2) - E^*(X_1)\|^2 =& \mathbb{E}\|(D^*)^{-1}(X_2) - (D^*)^{-1} \circ D^*(E^*(X_1))\|^2 \\
\le& \|(D^*)^{-1}\|_{\text{Lip}}^2 \mathbb{E}\|X_2 - D^*(E^*(X_1))\|^2 \\
\le& L^2\epsilon,
\end{aligned}
$$

where $L := \|(D^*)^{-1}\|_{\text{Lip}}$ is the Lipschitz constant for function $(D^*)^{-1}$. This is to say, in practice, we optimize over (11), and achieves a good representation $E^*(X_1)$ such that $\epsilon_{\text{pre}} \le L\sqrt{\epsilon}$ and thus performs well for downstream tasks. (Recall ϵ_{pre} is defined in Theorem 4.2 that measures how well we have learned the pretext task.)

D Omitted Proofs Beyond Conditional Independence

D.1 Warm-up: Jointly Gaussian Variables

As before, for simplicity we assume all data is centered in this case.

Assumption D.1 (Approximate Conditional Independent Given Latent Variables)**.** *Assume there exists some latent variable $Z \in \mathbb{R}^m$ such that*

$$\|\boldsymbol{\Sigma}_{X_1}^{-1/2}\boldsymbol{\Sigma}_{X_1,X_2|\bar{Y}}\|_F \leq \epsilon_{CI},$$

$\sigma_{k+m}(\boldsymbol{\Sigma}_{Y\bar{Y}}^{\dagger}\boldsymbol{\Sigma}_{\bar{Y}X_2}) = \beta > 0$ [8] *and* $\boldsymbol{\Sigma}_{X_2,\bar{Y}}$ *is of rank* $k+m$*, where* $\bar{Y} = [Y, Z]$.

When X_1 is not exactly CI of X_2 given Y and Z, the approximation error depends on the norm of $\|\boldsymbol{\Sigma}_{X_1}^{-1/2}\boldsymbol{\Sigma}_{X_1,X_2|\bar{Y}}\|_2$. Let $\hat{\boldsymbol{W}}$ be the solution from Equation uses CI.

Theorem D.1. *Under Assumption D.1 with constant* ϵ_{CI} *and* β*, then the excess risk satisfies*

$$\mathrm{ER}_{\psi^*}[\hat{\boldsymbol{W}}] := \mathbb{E}[\|\hat{\boldsymbol{W}}^\top\psi^*(X_1) - f^*(X_1)\|_F^2] \lesssim \frac{\epsilon_{CI}^2}{\beta^2} + \mathrm{Tr}(\boldsymbol{\Sigma}_{YY|X_1})\frac{d_2 + \log(d_2/\delta)}{n_2}.$$

Proof of Theorem D.1. Let $\boldsymbol{V} := f^*(\boldsymbol{X}_1) \equiv \boldsymbol{X}_1\boldsymbol{\Sigma}_{X_1X_1}^{-1}\boldsymbol{\Sigma}_{1Y}$ be our target direction. Denote the optimal representation matrix by $\Psi := \psi(\boldsymbol{X}_1) \equiv \boldsymbol{X}_1\boldsymbol{A}$ (where $\boldsymbol{A} := \boldsymbol{\Sigma}_{X_1X_1}^{-1}\boldsymbol{\Sigma}_{X_1X_2}$).

Next we will make use of the conditional covariance matrix:

$$\boldsymbol{\Sigma}_{X_1X_2|\bar{Y}} := \boldsymbol{\Sigma}_{X_1X_2} - \boldsymbol{\Sigma}_{X_1\bar{Y}}\boldsymbol{\Sigma}_{\bar{Y}}^{-1}\boldsymbol{\Sigma}_{\bar{Y}X_2},$$

and plug it in into the definition of Ψ:

$$\begin{aligned}\Psi =& \boldsymbol{X}_1\boldsymbol{\Sigma}_{X_1X_1}^{-1}\boldsymbol{\Sigma}_{X_1\bar{Y}}\boldsymbol{\Sigma}_{\bar{Y}}^{-1}\boldsymbol{\Sigma}_{\bar{Y}X_2} + \boldsymbol{X}_1\boldsymbol{\Sigma}_{X_1X_1}^{-1}\boldsymbol{\Sigma}_{X_1X_2|\bar{Y}}\\ =:& \boldsymbol{L} + \boldsymbol{E},\end{aligned}$$

where $\boldsymbol{L} := \boldsymbol{X}_1\boldsymbol{\Sigma}_{X_1X_1}^{-1}\boldsymbol{\Sigma}_{X_1\bar{Y}}\boldsymbol{\Sigma}_{\bar{Y}}^{-1}\boldsymbol{\Sigma}_{\bar{Y}X_2}$ and $\boldsymbol{E} := \boldsymbol{X}_1\boldsymbol{\Sigma}_{X_1X_1}^{-1}\boldsymbol{\Sigma}_{X_1X_2|\bar{Y}}$. We analyze these two terms respectively.

For $\boldsymbol{L}$, we note that span($\boldsymbol{V}$) $\subseteq$span($\boldsymbol{L}$): $\boldsymbol{L}\boldsymbol{\Sigma}_{X_2\bar{Y}}^{\dagger}\boldsymbol{\Sigma}_{\bar{Y}} = \boldsymbol{X}_1\boldsymbol{\Sigma}_{X_1X_1}^{-1}\boldsymbol{\Sigma}_{X_1\bar{Y}}$. By right multiplying the selector matrix S_Y we have: $\boldsymbol{L}\boldsymbol{\Sigma}_{X_2\bar{Y}}^{\dagger}\boldsymbol{\Sigma}_{\bar{Y}Y} = \boldsymbol{X}_1\boldsymbol{\Sigma}_{X_1X_1}^{-1}\boldsymbol{\Sigma}_{X_1Y}$, i.e., $\boldsymbol{L}\bar{\boldsymbol{W}} = \boldsymbol{V}$, where $\bar{\boldsymbol{W}} := \boldsymbol{\Sigma}_{X_2\bar{Y}}^{\dagger}\boldsymbol{\Sigma}_{\bar{Y}Y}$. From our assumption that $\sigma_r(\boldsymbol{\Sigma}_{\bar{Y}Y}^{\dagger}\boldsymbol{\Sigma}_{\bar{Y}X_2}) = \beta$, we have $\|\bar{\boldsymbol{W}}\|_2 \leq \|\boldsymbol{\Sigma}_{X_2\bar{Y}}^{\dagger}\boldsymbol{\Sigma}_{\bar{Y}}\|_2 \leq 1/\beta$. (Or we could directly define β as $\sigma_k(\boldsymbol{\Sigma}_{Y\bar{Y}}^{\dagger}\boldsymbol{\Sigma}_{\bar{Y}X_2}) \equiv \|\bar{\boldsymbol{W}}\|_2$.)

By concentration, we have $\boldsymbol{E} = \boldsymbol{X}_1\boldsymbol{\Sigma}_{X_1X_1}^{-1}\boldsymbol{\Sigma}_{X_1X_2|\bar{Y}}$ converges to $\boldsymbol{\Sigma}_{X_1X_1}^{-1/2}\boldsymbol{\Sigma}_{X_1X_2|\bar{Y}}$. Specifically, when $n \gg k + \log 1/\delta$, $\|\boldsymbol{E}\|_F \leq 1.1\|\boldsymbol{\Sigma}_{X_1X_1}^{-1/2}\boldsymbol{\Sigma}_{X_1X_2|\bar{Y}}\|_F \leq 1.1\epsilon_{\mathrm{CI}}$ (by using Lemma A.2). Together we have $\|\boldsymbol{E}\bar{\boldsymbol{W}}\|_F \lesssim \epsilon_{\mathrm{CI}}/\beta$.

Let $\hat{\boldsymbol{W}} = \arg\min_{\boldsymbol{W}} \|\boldsymbol{Y} - \Psi\boldsymbol{W}\|^2$. We note that $\boldsymbol{Y} = \boldsymbol{N} + \boldsymbol{V} = \boldsymbol{N} + \Psi\bar{\boldsymbol{W}} - \boldsymbol{E}\bar{\boldsymbol{W}}$ where $\boldsymbol{V}$ is our target direction and $\boldsymbol{N}$ is random noise (each row of $\boldsymbol{N}$ has covariance matrix $\boldsymbol{\Sigma}_{YY|X_1}$).

From basic inequality, we have:

$$\begin{aligned}\|\Psi\hat{\boldsymbol{W}} - \boldsymbol{Y}\|_F^2 \leq& \|\Psi\bar{\boldsymbol{W}} - \boldsymbol{Y}\|_F^2 = \|\boldsymbol{N} - \boldsymbol{E}\bar{\boldsymbol{W}}\|_F^2.\\ \Longrightarrow \|\Psi\hat{\boldsymbol{W}} - \boldsymbol{V} - \boldsymbol{E}\bar{\boldsymbol{W}}\|^2 \leq& 2\langle\Psi\hat{\boldsymbol{W}} - \boldsymbol{V} - \boldsymbol{E}\bar{\boldsymbol{W}}, \boldsymbol{N} - \boldsymbol{E}\bar{\boldsymbol{W}}\rangle\\ \Longrightarrow \|\Psi\hat{\boldsymbol{W}} - \boldsymbol{V} - \boldsymbol{E}\bar{\boldsymbol{W}}\| \leq& \|P_{[\Psi,E,V]}\boldsymbol{N}\| + \|\boldsymbol{E}\bar{\boldsymbol{W}}\|\\ \Longrightarrow \|\Psi\hat{\boldsymbol{W}} - \boldsymbol{V}\| \lesssim& \|\boldsymbol{E}\|_F\|\bar{\boldsymbol{W}}\| + (\sqrt{d_2} + \sqrt{\log 1/\delta})\sqrt{\mathrm{Tr}(\boldsymbol{\Sigma}_{YY|X_1})}.\\ &\text{(from Lemma A.4)}\\ \leq& \sqrt{n_2}\frac{\epsilon_{\mathrm{CI}}}{\beta} + (\sqrt{d_2} + \sqrt{\log 1/\delta})\sqrt{\mathrm{Tr}(\boldsymbol{\Sigma}_{YY|X_1})}.\\ &\text{(from Assumption D.1)}\end{aligned}$$

[8] $\sigma_k(\boldsymbol{A})$ denotes k-th singular value of $\boldsymbol{A}$, and $\boldsymbol{A}^\dagger$ is the pseudo-inverse of $\boldsymbol{A}$.

Next, by the same procedure that concentrates $\frac{1}{n_2}\boldsymbol{X}_1^\top \boldsymbol{X}_1$ to $\boldsymbol{\Sigma}_{X_1X_1}$ with Claim A.2, we could easily get

$$\mathrm{ER}[\hat{\boldsymbol{W}}] := \mathbb{E}[\|\hat{\boldsymbol{W}}^\top \psi(X_1) - f^*(X_1)\|^2] \lesssim \frac{\epsilon_{\mathrm{CI}}^2}{\beta^2} + \mathrm{Tr}(\boldsymbol{\Sigma}_{YY|X_1})\frac{d_2 + \log 1/\delta}{n_2}.$$

□

D.2 Measuring conditional dependence with cross-covariance operator

$L^2(P_X)$ denotes the Hilbert space of square integrable function with respect to the measure P_X, the marginal distribution of X. We are interested in some function class $\mathcal{H}_x \subset L^2(P_X)$ that is induced from some feature maps:

Definition D.2 (General and Universal feature Map)**.** *We denote feature map $\phi : \mathcal{X} \to \mathcal{F}$ that maps from a compact input space $\mathcal{X}$ to the feature space $\mathcal{F}$. $\mathcal{F}$ is a Hilbert space associated with inner product: $\langle \phi(\boldsymbol{x}), \phi(\boldsymbol{x}')\rangle_{\mathcal{F}}$. The associated function class is: $\mathcal{H}_x = \{h : \mathcal{X} \to \mathbb{R} | \exists w \in \mathcal{F}, h(\boldsymbol{x}) = \langle w, \phi(\boldsymbol{x})\rangle_{\mathcal{F}}, \forall \boldsymbol{x} \in \mathcal{X}\}$. We call ϕ universal if the induced $\mathcal{H}_x$ is dense in $L^2(P_X)$.*

Linear model is a special case when feature map $\phi = Id$ is identity mapping and the inner product is over Euclidean space. A feature map with higher order polynomials correspondingly incorporate high order moments [20, 24]. For discrete variable Y we overload ϕ as the one-hot embedding.

Remark D.1. *For continuous data, any universal kernel like Gaussian kernel or RBF kernel induce the universal feature map that we require [41]. Two-layer neural network with infinite width also satisfy it, i.e., $\forall \boldsymbol{x} \in \mathcal{X} \subset \mathbb{R}^d, \phi_{NN}(\boldsymbol{x}) : \mathcal{S}^{d-1} \times \mathbb{R} \to \mathbb{R}, \phi_{NN}(\boldsymbol{x})[\boldsymbol{w}, b] = \sigma(\boldsymbol{w}^\top \boldsymbol{x} + b)$ [7].*

When there's no ambiguity, we overload ϕ_1 as the random variable $\phi_1(X_1)$ over domain $\mathcal{F}_1$, and $\mathcal{H}_1$ as the function class over X_1. Next we characterize CI using the cross-covariance operator.

Definition D.3 (Cross-covariance operator)**.** *For random variables $X \in \mathcal{X}, Y \in \mathcal{Y}$ with joint distribution $P : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, and associated feature maps ϕ_x and ϕ_y, we denote by $\mathcal{C}_{\phi_x\phi_y} = \mathbb{E}[\phi_x(X) \otimes \phi_y(Y)] = \int_{\mathcal{X}\times\mathcal{Y}} \phi_x(x) \otimes \phi_y(y) dP(x, y)$, the (un-centered) cross-covariance operator. Similarly we denote by $\mathcal{C}_{X\phi_y} = \mathbb{E}[X \otimes \phi_y(Y)] : \mathcal{F}_y \to \mathcal{X}$.*

To understand what $\mathcal{C}_{\phi_x\phi_y}$ is, we note it is of the same shape as $\phi_x(x) \otimes \phi_y(y)$ for each individual $x \in \mathcal{X}, y \in \mathcal{Y}$. It can be viewed as an operator: $\mathcal{C}_{\phi_x\phi_y} : \mathcal{F}_y \to \mathcal{F}_x$, $\mathcal{C}_{\phi_x\phi_y} f = \int_{\mathcal{X}\times\mathcal{Y}} \langle \phi_y(y), f\rangle \phi_x(x) dP(x, y), \forall f \in \mathcal{F}_y$. For any $f \in \mathcal{H}_x$ and $g \in \mathcal{H}_y$, it satisfies: $\langle f, \mathcal{C}_{\phi_x\phi_y} g\rangle_{\mathcal{H}_x} = \mathbb{E}_{XY}[f(X)g(Y)]$[6, 20]. CI ensures $\mathcal{C}_{\phi_1 X_2|\phi_y} = 0$ for arbitrary ϕ_1, ϕ_2:

Lemma D.4. *With one-hot encoding map ϕ_y and arbitrary ϕ_1, $X_1 \perp X_2 | Y$ ensures:*

$$\mathcal{C}_{\phi_1 X_2|\phi_y} := \mathcal{C}_{\phi_1 X_2} - \mathcal{C}_{\phi_1\phi_y}\mathcal{C}_{\phi_y\phi_y}^{-1}\mathcal{C}_{\phi_y X_2} = 0. \tag{13}$$

A more complete discussion of cross-covariance operator and CI can be found in [20]. Also, recall that an operator $\mathcal{C} : \mathcal{F}_y \to \mathcal{F}_x$ is Hilbert-Schmidt (HS) [50] if for complete orthonormal systems (CONSs) $\{\zeta_i\}$ of $\mathcal{F}_x$ and $\{\eta_i\}$ of $\mathcal{F}_y$, $\|\mathcal{C}\|_{\mathrm{HS}}^2 := \sum_{i,j} \langle \zeta_j, \mathcal{C}\eta_i\rangle_{\mathcal{F}_x}^2 < \infty$. The Hilbert-Schmidt norm generalizes the Frobenius norm from matrices to operators, and we will later use $\|\mathcal{C}_{\phi_1 X_2|\phi_y}\|$ to quantify approximate CI.

We note that covariance operators [21, 20, 6] are commonly used to capture conditional dependence of random variables. In this work, we utilize the covariance operator to quantify the performance of the algorithm even when the algorithm is *not a kernel method*.

D.3 Omitted Proof in General Setting

Claim D.5. *For feature maps ϕ_1 with universal property, we have:*

$$\begin{aligned}
\psi^*(X_1) :=& \mathbb{E}[X_2|X_1] = \mathbb{E}^L[X_2|\phi_1] \\
=& \mathcal{C}_{X_2\phi_1}\mathcal{C}_{\phi_1\phi_1}^{-1}\phi_1(X_1). \\
\textit{Our target } f^*(X_1) :=& \mathbb{E}[Y|X_1] = \mathbb{E}^L[Y|\phi_1] \\
=& \mathcal{C}_{Y\phi_1}\mathcal{C}_{\phi_1\phi_1}^{-1}\phi_1(X_1).
\end{aligned}$$

For general feature maps, we instead have:

$$\begin{aligned}\psi^*(X_1) :=& \arg\min_{f\in\mathcal{H}_1^{d_2}} \mathbb{E}_{X_1X_2}\|X_2 - f(X_1)\|_2^2\\ =&\mathcal{C}_{X_2\phi_1}\mathcal{C}_{\phi_1\phi_1}^{-1}\phi_1(X_1).\\ \textit{Our target } f^*(X_1) :=& \arg\min_{f\in\mathcal{H}_1^{k}} \mathbb{E}_{X_1Y}\|Y - f(X_1)\|_2^2\\ =&\mathcal{C}_{Y\phi_1}\mathcal{C}_{\phi_1\phi_1}^{-1}\phi_1(X_1).\end{aligned}$$

To prove Claim D.5, we show the following lemma:

Lemma D.6. *Let $\phi : \mathcal{X} \to \mathcal{F}_x$ be a universal feature map, then for random variable $Y \in \mathcal{Y}$ we have:*

$$\mathbb{E}[Y|X] = \mathbb{E}^L[Y|\phi(X)].$$

Proof of Lemma D.6. Denote by $\mathbb{E}[Y|X = x] =: f(x)$. Since ϕ is dense in $\mathcal{X}$, there exists a linear operator $a : \mathcal{X} \to \mathbb{R}$ such that $\int_{x\in\mathcal{X}} a(x)\phi(x)[\cdot]dx = f(\cdot)$ a.e. Therefore the result comes directly from the universal property of ϕ. □

Proof of Claim D.5. We want to show that for random variables Y, X, where X is associated with a universal feature map ϕ_x, we have $\mathbb{E}[Y|X] = \mathcal{C}_{Y\phi_x(X)}\mathcal{C}^{-1}_{\phi_x(X)\phi_x(X)}\phi_x(X)$.

First, from Lemma D.6, we have that $\mathbb{E}[Y|X] = \mathbb{E}^L[Y|\phi_x(X)]$. Next, write $A^* : \mathcal{F}_x \to \mathcal{Y}$ as the linear operator that satisfies

$$\begin{aligned}&\mathbb{E}[Y|X] = A^*\phi_x(X)\\ \text{s.t. } &A^* = \arg\min_A \mathbb{E}[\|Y - A\phi_x(X)\|^2].\end{aligned}$$

Therefore from the stationary condition we have $A^*\mathbb{E}_X[\phi_x(X) \otimes \phi_x(X)] = \mathbb{E}_{XY}[Y \otimes \phi_x(X)]$. Or namely we get $A^* = \mathcal{C}_{Y\phi_x}\mathcal{C}^{-1}_{\phi_x\phi_x}$ simply from the definition of the cross-covariance operator $\mathcal{C}$. □

Claim D.7. $\|\mathcal{C}^{-1/2}_{\phi_1\phi_1}\mathcal{C}_{\phi_1X_2|\phi_{\bar{y}}}\|^2_{HS} = \mathbb{E}_{X_1}[\|\mathbb{E}[X_2|X_1] - \mathbb{E}_{\bar{Y}}[\mathbb{E}[X_2|\bar{Y}]|X_1]\|^2] = \epsilon^2_{CI}$.

Proof.

$$\begin{aligned}&\|\mathcal{C}^{-1/2}_{\phi_1\phi_1}\mathcal{C}_{\phi_1X_2|\phi_{\bar{y}}}\|^2_{\text{HS}}\\ =&\int_{X_1}\left\|\int_{X_2}\left(\frac{p_{X_1X_2}(\boldsymbol{x}_1,\boldsymbol{x}_2)}{p_{X_1}(\boldsymbol{x}_1)} - \frac{p_{X_1\perp X_2|Y}(\boldsymbol{x}_1,\boldsymbol{x}_2)}{p_{X_1}(\boldsymbol{x}_1)}\right)X_2dp_{\boldsymbol{x}_2}\right\|^2 dp_{\boldsymbol{x}_1}\\ =&\mathbb{E}_{X_1}[\|\mathbb{E}[X_2|X_1] - \mathbb{E}_{\bar{Y}}[\mathbb{E}[X_2|\bar{Y}]|X_1]\|^2].\end{aligned}$$

□

D.4 Omitted Proof for Main Results

We first prove a simpler version without approximation error.

Theorem D.8. *For a fixed $\delta \in (0, 1)$, under Assumption 4.1, 3.2, if there is no approximation error, i.e., there exists a linear operator A such that $f^*(X_1) \equiv A\phi_1(X_1)$, if $n_1, n_2 \gg \rho^4(d_2 + \log 1/\delta)$, and we learn the pretext tasks such that:*

$$\mathbb{E}\|\tilde{\psi}(X_1) - \psi^*(X_1)\|_F^2 \le \epsilon^2_{pre}.$$

Then we are able to achieve generalization for downstream task with probability $1 - \delta$:

$$\mathbb{E}[\|f^*_{\mathcal{H}_1}(X_1) - \hat{\boldsymbol{W}}^\top\tilde{\psi}(X_1)\|^2] \le \tilde{\mathcal{O}}\{\sigma^2\frac{d_2}{n_2} + \frac{\epsilon^2_{CI}}{\beta^2} + \frac{\epsilon^2_{pre}}{\beta^2}\}. \tag{14}$$

Proof of Theorem D.8. We follow the similar procedure as Theorem D.1. For the setting of no approximation error, we have $f^* = f^*_{\mathcal{H}_1}$, and the residual term $N := Y - f^*(X_1)$ is a mean-zero random variable with $\mathbb{E}[\|N\|^2|X_1] \lesssim \sigma^2$ according to our data assumption in Section 3. $\boldsymbol{N} = \boldsymbol{Y} - f^*(\boldsymbol{X}_1^{\text{down}})$ is the collected n_2 samples of noise terms. We write $Y \in \mathbb{R}^{d_3}$. For classification task, we have $Y \in \{\boldsymbol{e}_i, i \in [k]\} \subset \mathbb{R}^k$ (i.e, $d_3 = k$) is one-hot encoded random variable. For regression problem, Y might be otherwise encoded. For instance, in the yearbook dataset, Y ranges from 1905 to 2013 and represents the years that the photos are taken. We want to note that our result is general for both cases: the bound doesn't depend on d_3, but only depends on the variance of N.

Let $\Psi^*, \boldsymbol{L}, \boldsymbol{E}, \boldsymbol{V}$ be defined as follows:

Let $\boldsymbol{V} = f^*(\boldsymbol{X}_1^{\text{down}}) \equiv f^*_{\mathcal{H}_1}(\boldsymbol{X}_1^{\text{down}}) \equiv \phi(\boldsymbol{X}_1^{\text{down}})\mathcal{C}^{-1}_{\phi_1}\mathcal{C}_{\phi_1 Y}$ be our target direction. Denote the optimal representation matrix by

$$\begin{aligned}\Psi^* :=& \psi^*(\boldsymbol{X}_1^{\text{down}})\\ =&\phi(\boldsymbol{X}_1^{\text{down}})\mathcal{C}^{-1}_{\phi_1\phi_1}\mathcal{C}_{\phi_1 X_2}\\ =&\phi(\boldsymbol{X}_1^{\text{down}})\mathcal{C}^{-1}_{\phi_1\phi_1}\mathcal{C}_{\phi_1\phi_{\bar{y}}}\mathcal{C}^{-1}_{\phi_{\bar{y}}}\boldsymbol{\Sigma}_{\phi_{\bar{y}}X_2} + \phi(\boldsymbol{X}_1^{\text{down}})\mathcal{C}^{-1}_{\phi_1\phi_1}\mathcal{C}_{\phi_1 X_2|\phi_{\bar{y}}}\\ =:&\boldsymbol{L} + \boldsymbol{E},\end{aligned}$$

where $\boldsymbol{L} = \phi(\boldsymbol{X}_1^{\text{down}})\mathcal{C}^{-1}_{\phi_1\phi_1}\mathcal{C}_{\phi_1\phi_{\bar{y}}}\mathcal{C}^{-1}_{\phi_{\bar{y}}}\mathcal{C}_{\phi_{\bar{y}}X_2}$ and $\boldsymbol{E} = \phi(\boldsymbol{X}_1^{\text{down}})\mathcal{C}^{-1}_{\phi_1\phi_1}\mathcal{C}_{\phi_1 X_2|\bar{Y}}$.

In this proof, we denote S_Y as the matrix such that $S_Y\phi_{\bar{y}} = Y$. Specifically, if Y is of dimension d_3, S_Y is of size $d_3 \times |\mathcal{Y}||\mathcal{Z}|$. Therefore $S_Y\boldsymbol{\Sigma}_{\phi_y A} = \boldsymbol{\Sigma}_{YA}$ for any random variable A.

Therefore, similarly we have:

$$\boldsymbol{L}\boldsymbol{\Sigma}^{\dagger}_{X_2\phi_{\bar{y}}}\boldsymbol{\Sigma}_{\phi_{\bar{y}}\phi_{\bar{y}}}S_Y^{\top} = \boldsymbol{L}\boldsymbol{\Sigma}^{\dagger}_{X_2\phi_{\bar{y}}}\boldsymbol{\Sigma}_{\phi_{\bar{y}}Y} = \boldsymbol{L}\bar{\boldsymbol{W}} = \boldsymbol{V}$$

where $\bar{\boldsymbol{W}} := \boldsymbol{\Sigma}^{\dagger}_{X_2\phi_{\bar{y}}}\boldsymbol{\Sigma}_{\phi_{\bar{y}}Y}$ satisfies $\|\bar{\boldsymbol{W}}\|_2 = 1/\beta$. Therefore span$(\boldsymbol{V}) \subseteq$span$(\boldsymbol{L})$ since we have assumed that $\boldsymbol{\Sigma}^{\dagger}_{X_2\phi_{\bar{y}}}\boldsymbol{\Sigma}_{\phi_{\bar{y}}Y}$ to be full rank.

On the other hand, $\boldsymbol{E} = \phi_1(\boldsymbol{X}_1^{\text{down}})\mathcal{C}^{-1}_{\phi_1\phi_1}\mathcal{C}_{\phi_1 X_2|\bar{Y}}$ concentrates to $\mathcal{C}^{-1/2}_{\phi_1\phi_1}\mathcal{C}_{\phi_1 X_2|\phi_{\bar{y}}}$. Specifically, when $n \gg c + \log 1/\delta$, $\frac{1}{n_2}\|\boldsymbol{E}\|_F^2 \leq 1.1\|\mathcal{C}^{-1/2}_{\phi_1\phi_1}\mathcal{C}_{\phi_1 X_2|\phi_{\bar{y}}}\|_F^2 \leq 1.1\epsilon_{\text{CI}}^2$ (by using Lemma A.3). Together we have $\|\boldsymbol{E}\bar{\boldsymbol{W}}\|_F \lesssim \epsilon_{\text{CI}}/\beta$.

We also introduce the error from not learning ψ^* exactly: $\boldsymbol{E}^{\text{pre}} = \Psi - \Psi^* := \tilde{\psi}(\boldsymbol{X}_1^{\text{down}}) - \psi^*(\boldsymbol{X}_1^{\text{down}})$. With proper concentration and our assumption, we have that $\mathbb{E}\|\psi(X_1) - \psi^*(X_1)\|^2 \leq \epsilon_{\text{pre}}$ and $\frac{1}{\sqrt{n_2}}\|\psi(\boldsymbol{X}_1^{\text{down}}) - \psi^*(\boldsymbol{X}_1^{\text{down}})\|^2 \leq 1.1\epsilon_{\text{pre}}$.

Also, the noise term after projection satisfies $\|P_{[\Psi,\boldsymbol{E},\boldsymbol{V}]}\boldsymbol{N}\| \lesssim \sqrt{d_2(1+\log d_2/\delta)}\sigma$ as using Corollary A.6. Therefore $\Psi = \Psi^* - \boldsymbol{E}^{\text{pre}} = \boldsymbol{L} + \boldsymbol{E} - \boldsymbol{E}^{\text{pre}}$.

Recall that $\hat{\boldsymbol{W}} = \arg\min_{\boldsymbol{W}} \|\psi(\boldsymbol{X}_1^{\text{down}})\boldsymbol{W} - \boldsymbol{Y}\|_F^2$. And with exactly the same procedure as Theorem D.1 we also get that:

$$\begin{aligned}\|\Psi\hat{\boldsymbol{W}} - \boldsymbol{V}\| \leq& 2\|\boldsymbol{E}\bar{\boldsymbol{W}}\| + 2\|\boldsymbol{E}^{\text{pre}}\bar{\boldsymbol{W}}\| + \|P_{[\Psi,\boldsymbol{E},\boldsymbol{V},\boldsymbol{E}^{\text{pre}}]}\boldsymbol{N}\|\\ \lesssim& \sqrt{n_2}\frac{\epsilon_{\text{CI}} + \epsilon_{\text{pre}}}{\beta} + \sigma\sqrt{d_2(1+\log(d_2/\delta))}.\end{aligned}$$

With the proper concentration we also get:

$$\mathbb{E}[\|\hat{\boldsymbol{W}}^{\top}\psi(X_1) - f^*_{\mathcal{H}_1}(X_1)\|^2] \lesssim \frac{\epsilon_{\text{CI}}^2 + \epsilon_{\text{pre}}^2}{\beta^2} + \sigma^2\frac{d_2(1+\log(d_2/\delta))}{n_2}.$$

□

Next we move on to the proof of our main result Theorem 4.2 where approximation error occurs.

Proof of Theorem 4.2. The proof is a combination of Theorem 3.5 and Theorem D.8. We follow the same notation as in Theorem D.8. Now the only difference is that an additional term $a(\boldsymbol{X}_1^{\text{down}})$ is included in $\boldsymbol{Y}$:

$$\begin{aligned}\boldsymbol{Y} =& \boldsymbol{N} + f^*(\boldsymbol{X}_1^{\text{down}})\\ =& \boldsymbol{N} + \Psi^*\bar{\boldsymbol{W}} + a(\boldsymbol{X}_1^{\text{down}})\\ =& \boldsymbol{N} + (\Psi + \boldsymbol{E}^{\text{pre}})\bar{\boldsymbol{W}} + a(\boldsymbol{X}_1^{\text{down}})\\ =& \Psi\bar{\boldsymbol{W}} + (\boldsymbol{N} + \boldsymbol{E}^{\text{pre}}\bar{\boldsymbol{W}} + a(\boldsymbol{X}_1^{\text{down}})).\end{aligned}$$

From re-arranging $\frac{1}{2n_2}\|\boldsymbol{Y} - \Psi\hat{\boldsymbol{W}}\|_F^2 \leq \frac{1}{2n_2}\|\boldsymbol{Y} - \Psi\bar{\boldsymbol{W}}\|_F^2$,

$$\frac{1}{2n_2}\|\Psi(\bar{\boldsymbol{W}} - \hat{\boldsymbol{W}}) + (\boldsymbol{N} + \boldsymbol{E}^{\text{pre}} + a(\boldsymbol{X}_1^{\text{down}}))\|_F^2 \leq \frac{1}{2n_2}\|\boldsymbol{N} + \boldsymbol{E}^{\text{pre}}\bar{\boldsymbol{W}} + a(\boldsymbol{X}_1^{\text{down}})\|_F^2 \tag{15}$$

$$\Rightarrow \frac{1}{2n_2}\|\Psi(\bar{\boldsymbol{W}} - \hat{\boldsymbol{W}})\|_F^2 \leq \frac{1}{n_2}\langle\Psi(\bar{\boldsymbol{W}} - \hat{\boldsymbol{W}}), \boldsymbol{N} + \boldsymbol{E}^{\text{pre}}\bar{\boldsymbol{W}} + a(\boldsymbol{X}_1^{\text{down}})\rangle. \tag{16}$$

Then with similar procedure as in the proof of Theorem 3.5, and write Ψ as $\phi(X_1^{\text{down}})\boldsymbol{B}$, we have:

$$\begin{aligned}&\frac{1}{n_2}\langle\Psi(\bar{\boldsymbol{W}} - \hat{\boldsymbol{W}}), a(\boldsymbol{X}_1^{\text{down}})\rangle\\ =&\frac{1}{n_2}\langle\boldsymbol{B}(\bar{\boldsymbol{W}} - \hat{\boldsymbol{W}}), \phi(\boldsymbol{X}_1^{\text{down}})^\top a(\boldsymbol{X}_1^{\text{down}})\rangle\\ =&\frac{1}{n_2}\langle\mathcal{C}_{\phi_1}^{1/2}\boldsymbol{B}(\bar{\boldsymbol{W}} - \hat{\boldsymbol{W}}), \mathcal{C}_{\phi_1}^{-1/2}\phi(\boldsymbol{X}_1^{\text{down}})^\top a(\boldsymbol{X}_1^{\text{down}})\rangle\\ \leq&\sqrt{\frac{d_2}{n_2}}\|\mathcal{C}_{\phi_1}^{1/2}\boldsymbol{B}(\bar{\boldsymbol{W}} - \hat{\boldsymbol{W}})\|_F\\ \leq&1.1\frac{1}{\sqrt{n_2}}\sqrt{\frac{d_2}{n_2}}\|\phi(\boldsymbol{X}_1^{\text{down}})\boldsymbol{B}(\bar{\boldsymbol{W}} - \hat{\boldsymbol{W}})\|_F\\ =&1.1\frac{\sqrt{d_2}}{n_2}\|\Psi(\bar{\boldsymbol{W}} - \hat{\boldsymbol{W}})\|_F.\end{aligned}$$

Therefore plugging back to (16) we get:

$$\begin{aligned}&\frac{1}{2n_2}\|\Psi(\bar{\boldsymbol{W}} - \hat{\boldsymbol{W}})\|_F^2 \leq \frac{1}{n_2}\langle\Psi(\bar{\boldsymbol{W}} - \hat{\boldsymbol{W}}), \boldsymbol{N} + \boldsymbol{E}^{\text{pre}}\bar{\boldsymbol{W}} + a(\boldsymbol{X}_1^{\text{down}})\rangle\\ &\Rightarrow \frac{1}{2n_2}\|\Psi(\bar{\boldsymbol{W}} - \hat{\boldsymbol{W}})\|_F \leq \frac{1}{2n_2}\|\boldsymbol{E}^{\text{pre}}\bar{\boldsymbol{W}}\|_F + \frac{1}{2n_2}\|P_\Psi\boldsymbol{N}\|_F + 1.1\frac{\sqrt{d_2}}{n_2}.\\ &\Rightarrow \frac{1}{2\sqrt{n_2}}\|\Psi\hat{\boldsymbol{W}} - f^*_{\mathcal{H}_1}(\boldsymbol{X}_1^{\text{down}})\|_F - \|\boldsymbol{E}\bar{\boldsymbol{W}}\|_F \leq \frac{1}{\sqrt{n_2}}(1.1\sqrt{d_2} + \|\boldsymbol{E}^{\text{pre}}\bar{\boldsymbol{W}}\| + \sqrt{d_2 + \log(d_2/\delta)})\\ &\Rightarrow \frac{1}{2\sqrt{n_2}}\|\Psi\hat{\boldsymbol{W}} - f^*_{\mathcal{H}_1}(\boldsymbol{X}_1^{\text{down}})\|_F \lesssim \sqrt{\frac{d_2(1+\log d_2/\delta)}{n_2}} + \frac{\epsilon_{\text{CI}} + \epsilon_{\text{pre}}}{\beta}.\end{aligned}$$

Finally by concentrating $\frac{1}{n_2}\Psi^\top\Psi$ to $\mathbb{E}[\tilde{\psi}(X_1)\tilde{\psi}(X_1)^\top]$ we get:

$$\mathbb{E}[\|\hat{\boldsymbol{W}}^\top\tilde{\psi}(X_1) - f^*_{\mathcal{H}_1}(X_1)\|_2^2] \lesssim \frac{d_2(1+\log d_2/\delta}{n_2}) + \frac{\epsilon_{\text{CI}}^2 + \epsilon_{\text{pre}}^2}{\beta^2},$$

with probability $1-\delta$. □

D.5 Principal Component Regression

Claim D.9 (Approximation Error of Principle Component Analysis)**.** *Let matrix $\boldsymbol{A} = \boldsymbol{L} + \boldsymbol{E} \in \mathbb{R}^{n\times d}$ where $\boldsymbol{L}$ has rank r <size of $\boldsymbol{A}$.Let $\boldsymbol{A}_r$ be the rank-r PCA of A. Then we have: $\|\boldsymbol{A}_r - \boldsymbol{L}\|_F \leq 2\|\boldsymbol{E}\|_F$, and $\|\boldsymbol{A}_r - \boldsymbol{L}\|_2 \leq 2\|\boldsymbol{E}\|_2$.*

Proof. Due to the property of PCA, $\|\boldsymbol{A}_r - \boldsymbol{A}\|_F \le \|\boldsymbol{E}\|_F$ and $\|\boldsymbol{A}_r - \boldsymbol{A}\|_2 \le \|\boldsymbol{E}\|_2$.

$$\begin{aligned}\|\boldsymbol{A}_r - \boldsymbol{L}\|_2 =& \|\boldsymbol{A}_r - \boldsymbol{A} + \boldsymbol{A} - \boldsymbol{L}\|_2 \\ \le& \|\boldsymbol{A}_r - \boldsymbol{A}\|_F + \|\boldsymbol{E}\|_F \\ \le& 2\|\boldsymbol{E}\|_2.\end{aligned}$$

Similarly we have $\|\boldsymbol{A}_r - \boldsymbol{L}\|_F \le 2\|\boldsymbol{E}\|_F$. □

This technical fact could be used to complete the proof for Remark 4.1.

Proof of Remark 4.1. We replace the key steps of D.8.

Recall $\Psi^*, \boldsymbol{L}, \boldsymbol{E}, \boldsymbol{V}$ are defined as follows:

$\Psi^* := \psi^*(\boldsymbol{X}_1^{\text{down}})$ is the optimal representation matrix. Ψ_r is the features obtained from r-PCA of Ψ^*. $\Psi^* = \boldsymbol{L} + \boldsymbol{E}$ which is low rank plus small norm. ($\boldsymbol{L} = \phi(\boldsymbol{X}_1^{\text{down}})\mathcal{C}_{\phi_1\phi_1}^{-1}\mathcal{C}_{\phi_1\phi_{\bar{y}}}\mathcal{C}_{\phi_{\bar{y}}}^{-1}\mathcal{C}_{\phi_{\bar{y}}X_2}$ and $\boldsymbol{E} = \phi(\boldsymbol{X}_1^{\text{down}})\mathcal{C}_{\phi_1\phi_1}^{-1}\mathcal{C}_{\phi_1 X_2|\bar{Y}}$. Suppose $r = |\mathcal{Y}||\mathcal{Z}|$.) Let $\boldsymbol{V} = f^*(\boldsymbol{X}_1^{\text{down}}) \equiv f^*_{\mathcal{H}_1}(\boldsymbol{X}_1^{\text{down}}) \equiv \phi(\boldsymbol{X}_1^{\text{down}})\mathcal{C}_{\phi_1}^{-1}\mathcal{C}_{\phi_1 Y} = \boldsymbol{L}\bar{\boldsymbol{W}}$ be our target direction, where $\bar{\boldsymbol{W}} := \boldsymbol{\Sigma}^{\dagger}_{X_2\phi_{\bar{y}}}\boldsymbol{\Sigma}_{\phi_{\bar{y}}Y}$.

Due to representation learning error (finite sample in the first stage) and approximate conditional independence, the target direction $\boldsymbol{V}$ is not perfectly linear in Ψ^* or its r-PCA features Ψ.

Now with PCR we learn the linear model with $\hat{\boldsymbol{W}} \leftarrow \arg\min_{\boldsymbol{W}} \|\Psi_r \boldsymbol{W} - \boldsymbol{Y}\|_F^2$. Together with D.9 and the same procedure as Theorem D.8 we also get that:

Let $\bar{\boldsymbol{E}} = \boldsymbol{L} - \Psi_r$ is of rank at most $2r$.

$$\begin{aligned}\|\Psi_r\hat{\boldsymbol{W}} - \boldsymbol{Y}\|_F^2 \le& \|\Psi_r\bar{\boldsymbol{W}} - \boldsymbol{Y}\|_F^2 = \|\boldsymbol{N} - \bar{\boldsymbol{E}}\bar{\boldsymbol{W}}\|_F^2. \\ \implies \|\Psi_r\hat{\boldsymbol{W}} - \boldsymbol{V} - \bar{\boldsymbol{E}}\bar{\boldsymbol{W}}\|^2 \le& 2\langle \Psi_r\hat{\boldsymbol{W}} - \boldsymbol{V} - \bar{\boldsymbol{E}}\bar{\boldsymbol{W}}, \boldsymbol{N} - \bar{\boldsymbol{E}}\bar{\boldsymbol{W}}\rangle \\ \implies \|\Psi_r\hat{\boldsymbol{W}} - \boldsymbol{V} - \bar{\boldsymbol{E}}\bar{\boldsymbol{W}}\| \le& \|P_{[\Psi_r,\boldsymbol{L}]}\boldsymbol{N}\| + \|\bar{\boldsymbol{E}}\bar{\boldsymbol{W}}\| \\ \implies \|\Psi_r\hat{\boldsymbol{W}} - \boldsymbol{V}\| \le& 2\|\bar{\boldsymbol{E}}\|_F\|\bar{\boldsymbol{W}}\| + \|P_{2r}\boldsymbol{N}\| \\ \lesssim& \|\boldsymbol{E}\|_F\|\bar{\boldsymbol{W}}\| + \sigma\sqrt{r}(1 + \sqrt{\log(r/\delta)}).\end{aligned}$$

With concentration on the downstream labeled samples we also get the result in Remark 4.1:

$$\mathbb{E}[\|\hat{\boldsymbol{W}}^\top\psi_r(X_1) - f^*_{\mathcal{H}_1}(X_1)\|^2] \lesssim \frac{\epsilon_{\text{CI}}^2 + \epsilon_{\text{pre}}^2}{\beta^2} + \sigma^2\frac{r(1+\log(r/\delta))}{n_2}.$$

Here $r = |\mathcal{Y}||\mathcal{Z}|$. □

E Omitted Proofs Beyond Conditional Independence

E.1 Proof for topic modeling example

Proof for Theorem 5.1. We will construct a latent variable $\bar{Y}$ such that $\epsilon_{\text{CI}} = 0$. We pick the domain of $\bar{Y}$ to be $[k]$ and the distribution $P(\bar{Y}|X_1)$ to be the distribution $\mathbb{E}[\mu|X_1] \in \Delta_{[k]}$, and define $P\left(X_2|\bar{Y} = i\right) = P\left(X_2|\mu = e_i\right)$. More specifically we have

$$\begin{aligned}P(\bar{Y} = i|X_1) = \mathbb{E}[\mu|X_1](i) = \mathbb{E}[\mu(i)|X_1] \text{ and thus } \mathbb{E}\left[\bar{Y}|X_1\right] = \mathbb{E}[\mu|X_1] \\ P\left(X_2|\bar{Y} = i\right) = P\left(X_2|\mu = e_i\right) \text{ and thus } \mathbb{E}\left[X_2|\bar{Y} = i\right] = \mathbb{E}[X_2|\mu = e_i]\end{aligned}$$

To show $\epsilon_{\text{CI}} = 0$, from Definition 4.1 we need to show $\mathbb{E}[X_2|X_1] = \mathbb{E}\left[\mathbb{E}\left[X_2|\bar{Y}\right]|X_1\right]$. Since X_2 is the bag of words representation, we know that $X_2 = \frac{2}{N}\sum_{i=N/2+1}^{N} e_{w_i}$. So for any $\mu \in \Delta_{[k]}$ we get

$$\mathbb{E}[X_2|\mu] =^{(a)} \frac{2}{N}\sum_{i=N/2+1}^{N}\mathbb{E}[e_{w_i}|\mu] =^{(b)} \frac{2}{N}\sum_{i=N/2+1}^{N} A\mu = A\mu$$

where (a) follows from linearity of expectation and (b) follows from the linearity of the probability distribution of each word given μ for topic models. Thus from the definition of $\bar{Y}$, $\mathbb{E}\left[X_2|\bar{Y}=i\right] = \mathbb{E}\left[X_2|\mu=e_i\right] = Ae_i$. To check if $\epsilon_{\text{CI}}=0$, we compute the following

$$\begin{aligned}\mathbb{E}\left[\mathbb{E}\left[X_2|\bar{Y}\right]|X_1\right] &= \sum_{i=1}^{k}\mathbb{E}\left[X_2|\bar{Y}=i\right]P(\bar{Y}=i|X_1)\\ &= \sum_{i=1}^{k}Ae_i\,\mathbb{E}\left[\mu(i)|X_1\right] = A\sum_{i=1}^{k}\mathbb{E}\left[\mu(i)e_i|X_1\right]\\ &= \mathbb{E}\left[A\mu|X_1\right] = \mathbb{E}\left[\mathbb{E}\left[X_2|\mu\right]|X_1\right]\end{aligned}$$

Due to the topic modeling assumption and the independent sampling of words given μ, we know that $X_1 \perp X_2|\mu$ and thus $\mathbb{E}\left[X_2|X_1\right] = \mathbb{E}\left[\mathbb{E}\left[X_2|\mu\right]|X_1\right]$. Combining with the above calculation, we get that $\mathbb{E}\left[\mathbb{E}\left[X_2|\bar{Y}\right]|X_1\right] = \mathbb{E}\left[X_2|X_1\right]$, thus giving $\epsilon_{\text{CI}}=0$. This proves points 1. and 2.

For point 3., note that $\mathbb{E}[Y|X_1] = \mathbb{E}[w^\top\mu|X_1] = w^\top\mathbb{E}[\mu|X_1] = w^\top\mathbb{E}[\bar{Y}|X_1]$.

Finally for point 4., we use the definition $1/\beta = \|\mathbf{\Sigma}_{Y\phi_{\bar{y}}}\mathbf{\Sigma}^{\dagger}_{X_2\phi_{\bar{y}}}\|_2$. For the first term, we note that $\mathbb{E}\left[\phi_{\bar{Y}}|\mu\right] = \mathbb{E}\left[\mathbb{E}\left[\phi_{\bar{Y}}|X_1\right]|\mu\right] = \mathbb{E}\left[\mathbb{E}\left[\bar{\mu}|X_1\right]|\mu\right] = \mu$

$$\begin{aligned}\mathbf{\Sigma}_{Y\phi_{\bar{y}}} &= \mathbb{E}_{\mu\sim\tau}\left[Y\phi_{\bar{Y}}^\top\right] = \mathbb{E}_{\mu\sim\tau}\left[w^\top\mu\phi_{\bar{Y}}^\top\right]\\ &= \mathbb{E}_{\mu\sim\tau}\left[w^\top\mu\mathbb{E}\left[\phi_{\bar{Y}}^\top|\mu\right]\right] = \mathbb{E}_{\mu\sim\tau}\left[w^\top\mu\mu^\top\right]\\ &= w^\top\Gamma\end{aligned}$$

where Γ was defined as the topic covariance $\Gamma = \mathbb{E}_{\mu\sim\tau}\left[\mu\mu^\top\right]$. The second term is

$$\mathbf{\Sigma}_{X_2\phi_{\bar{y}}} = \mathbb{E}_{\mu\sim\tau}\left[\mathbb{E}\left[X_2|\mu\right]\mathbb{E}\left[\phi_{\bar{Y}}^\top|\mu\right]\right] = \mathbb{E}_{\mu\sim\tau}\left[A\mu\mu^\top\right] = A\Gamma$$

The upper bound for $1/\beta$ can be computed as follows

$$\begin{aligned}1/\beta &= \left\|\mathbf{\Sigma}_{Y\phi_{\bar{y}}}\mathbf{\Sigma}^{\dagger}_{X_2\phi_{\bar{y}}}\right\|_2 = \left\|w^\top\Gamma\left(A\Gamma\right)^\dagger\right\|_2\\ &\le \|w\|_2\,\lambda_{\max}(\Gamma)\,\lambda_{\max}\left(\left(A\Gamma\right)^\dagger\right) = \|w\|_2\,\lambda_{\max}(\Gamma)\,\lambda_{\min}\left(A\Gamma\right)^{-1}\\ &\le \|w\|_2\,\lambda_{\max}(\Gamma)\,\lambda_{\min}\left(A\right)^{-1}\,\lambda_{\min}\left(\Gamma\right)^{-1}\\ &= \|w\|_2\,\frac{\lambda_{\max}(\Gamma)}{\lambda_{\min}\left(\Gamma\right)^{-1}}\,\lambda_{\min}\left(A\right)^{-1} = \frac{\kappa\|w\|_2}{\lambda_{\min}\left(A\right)}\end{aligned}$$

For $\lambda_{\min}\left(A\right)$, we need to lower bound $\min_{\|v\|_2=1}\|Av\|_2$ for which we will use the anchor word assumption. Let $\pi(i)$ be the anchor word for topic $i\in[k]$; so $A_i(\pi(i))\ge p$ and $A_{i'}(\pi(i))=0$ for $i'\ne i$. Using the assumption, for any $v\in\mathbb{R}^k$ with $\|v\|_2=1$, we get that $|(Av)(\pi(i))| = |v_i|\,A_i(\pi(i))\ge p|v_i|$ and thus $\|Av\|_2^2\ge\sum_i(Av)(\pi(i))^2\ge p^2\sum_i v_i^2 = p^2$. This shows that $\lambda_{\min}\left(A\right)\ge p$, thus combining with the above calculation to prove point 4. and completing the proof. □

F General Results and Comparison to [57]

We now show a more general form of our results and also connect the multi-view redundancy assumption from [57] to ours.

F.1 General Results

We first note that all our results hold for a generalized version of Assumption 4.1 and Definition 4.1 that we state below.

Assumption F.1. *Suppose $\bar{Y}$ with $|\bar{Y}|\le m$ is a discrete latent variable that satisfies*

1. *$\bar{Y}$ makes X_1 and X_2 approximately CI as in Definition 4.1, i.e.*

$$\epsilon_{CI}^2 := \mathbb{E}_{X_1}\left[\|\mathbb{E}[X_2|X_1] - \mathbb{E}_{\bar{Y}}[\mathbb{E}[X_2|\bar{Y}]|X_1]\|^2\right]$$

2. *$\bar{Y}$ also makes X_1 and Y approximately CI with*

$$\epsilon_{\bar{Y}}^2 := \mathbb{E}_{X_1}\left[\|\mathbb{E}[Y|X_1] - \mathbb{E}_{\bar{Y}}[\mathbb{E}[Y|\bar{Y}]|X_1]\|^2\right]$$

3. *$\mathbf{\Sigma}_{\phi_{\bar{y}} X_2}$ is full column rank and $\|\mathbf{\Sigma}_{Y\phi_{\bar{y}}}\mathbf{\Sigma}^{\dagger}_{X_2\phi_{\bar{y}}}\|_2 = 1/\beta$, where $A^{\dagger}$ is pseudo-inverse, and $\phi_{\bar{y}}$ is the one-hot embedding for $\bar{Y}$.*

Note that our assumptions from the main paper are a special case of Assumption F.1, with $\epsilon_{\bar{Y}} = 0$ being satisfied automatically as $\bar{Y} = [Y, Z]$ is explicitly defined to contain Y in it. Unlike Assumption 4.1, we do not need Y to be a discrete variable, but just need $\bar{Y}$ to be discrete. We state the generalization of Theorem 4.2 below

Theorem F.1. *For a fixed $\delta \in (0,1)$, under Assumptions F.1, 4.2 for $\tilde{\psi}$ and ψ^* and 3.2 for non-universal feature maps, if $n_1, n_2 \gg \rho^4(d_2 + \log 1/\delta)$, and we learn the pretext tasks such that: $\mathbb{E}\|\tilde{\psi}(X_1) - \psi^*(X_1)\|_F^2 \leq \epsilon_{pre}^2$. Then the generalization error for downstream task w.p. $1-\delta$ is:*

$$\mathbb{E}_{X_1}\left[\|\mathbb{E}[Y|X_1] - \hat{\boldsymbol{W}}^\top \tilde{\psi}(X_1)\|_2^2\right] \leq \tilde{\mathcal{O}}\left(\sigma^2 \frac{d_2}{n_2} + \frac{\epsilon_{CI}^2}{\beta^2} + \frac{\epsilon_{pre}^2}{\beta^2} + \epsilon_{\bar{Y}}^2\right) \tag{17}$$

The result is pretty much the same as Theorem 4.2, except for an additional term of $\epsilon_{\bar{Y}}^2$. The proof is also very similar, the difference being that $\mathbb{E}[\mathbb{E}[Y|\bar{Y}]|X_1]$ can now be expressed as a linear function of ψ^* instead of $\mathbb{E}[Y|X_1]$, and the additional error incurred during to the mismatch between $\mathbb{E}[Y|X_1]$ and $\mathbb{E}[\mathbb{E}[Y|\bar{Y}]|X_1]$ that is $\epsilon_{\bar{Y}}^2$ will be incurred.

F.2 Comparison to [57]

We show guarantees for our algorithm under the assumption from [57] in the following special case that satisfies: (1) X_1 and X_2 are *exactly* CI given $\bar{Y}$ (thus $\epsilon_{\text{CI}} = 0$), (2) the variation in the target Y is small given X_1 and X_2. The assumption from [57], in our setting, is equivalent to saying that ϵ_{X_1} and ϵ_{X_2} are small, where

$$\epsilon_{X_i}^2 = \mathbb{E}\left[\|\mathbb{E}[Y|X_i] - \mathbb{E}[Y|X_1, X_2]\|^2\right], \quad i \in \{1, 2\}$$

A similar assumption of multi-view redundancy also appears in [58]; however they state it in terms of information-theoretic quantities instead. We will show that these assumptions are also almost sufficient to show results in our setting. In particular we show that if $Y|X_1, X_2$ is almost deterministic (which makes sense for a many regression tasks) and if $\epsilon_{X_2}^2$ is small, then $\epsilon_{\bar{Y}}$ defined in the previous subsection will be small and thus we have meaningful guarantees.

Lemma F.2. *Let $\sigma_Y^2 = Var[Y|X_1, X_2]$ be the variance of Y. $\bar{Y}$ is as defined in Assumption F.1 with the extra condition that X_1 and X_2 are exactly CI given $\bar{Y}$. Then we have*

$$\epsilon_{\bar{Y}} \leq \sqrt{2}(\sigma_Y + \epsilon_{X_2})$$

Plugging this into Theorem F.1 will give us the desired result. Note however that we did not even use the fact that ϵ_{X_1} is small. Using this part of the assumption, we can get an even stronger result that shows that even though our learned representation will only X_1, if will still predict $Y|X_1, X_2$ well.

Corollary F.3. *For a fixed $\delta \in (0,1)$, under Assumptions F.1, 4.2 for $\tilde{\psi}$ and ψ^* and 3.2 for non-universal feature maps, if $n_1, n_2 \gg \rho^4(d_2 + \log 1/\delta)$, and we learn the pretext tasks such that: $\mathbb{E}\|\tilde{\psi}(X_1) - \psi^*(X_1)\|_F^2 \leq \epsilon_{pre}^2$. Then the generalization error for downstream task w.p. $1-\delta$ is:*

$$\mathbb{E}_{X_1, X_2}\left[\|\mathbb{E}[Y|X_1, X_2] - \hat{\boldsymbol{W}}^\top \tilde{\psi}(X_1)\|_2^2\right] \leq \tilde{\mathcal{O}}\left(\sigma^2 \frac{d_2}{n_2} + \frac{\epsilon_{pre}^2}{\beta^2} + \epsilon_{X_1}^2 + \epsilon_{X_2}^2 + \sigma_Y^2\right)$$

Thus we see that the assumption from [57] is strong enough for us to be able to show stronger results than just our assumption. We complete this section by proving Lemma F.2

Lemma F.2. We will also make use of the following lemma that is easily proved using Cauchy-Schwarz inequality

Lemma F.4. *For random variables $Z_1, \dots, Z_n$ for which $\mathbb{E}[\|Z_i\|^2] < \infty$ for every $i \in [n]$, we have*

$$\mathbb{E}[\|Z_1 + \cdots + Z_n\|^2] \leq \left(\sqrt{\mathbb{E}[\|Z_1\|^2]} + \cdots + \sqrt{\mathbb{E}[\|Z_n\|^2]}\right)^2$$

The proof follows from the following sequence of inequalities that uses Jensen's inequality, conditional independence of X_1 and X_2 and the above lemma. For simplicity we assume that Y is a scalar random variable, the proof is the same for vector values Y, except squared values will replaced by norm squared values.

$$\begin{aligned}
\epsilon_{\bar{Y}}^2 &= \mathbb{E}_{X_1}\left[(\mathbb{E}[Y|X_1] - \mathbb{E}_{\bar{Y}}[\mathbb{E}[Y|\bar{Y}]|X_1])^2\right] = \mathbb{E}_{X_1}\left[(\mathbb{E}_{\bar{Y}}[\mathbb{E}[Y|\bar{Y}, X_1]|X_1] - \mathbb{E}_{\bar{Y}}[\mathbb{E}[Y|\bar{Y}]|X_1])^2\right] \\
&\leq \mathbb{E}_{X_1,\bar{Y}}\left[(\mathbb{E}[Y|X_1, \bar{Y}] - \mathbb{E}[Y|\bar{Y}])^2\right] \\
&= \mathbb{E}_{\bar{Y}}\mathbb{E}_{X_1|\bar{Y}}\mathbb{E}_{X_1'|\bar{Y}}\left[(\mathbb{E}[Y|X_1, \bar{Y}] - \mathbb{E}[Y|X_1', \bar{Y}])^2\right] \\
&= \frac{1}{2}\mathbb{E}_{\bar{Y}}\mathbb{E}_{X_1|\bar{Y}}\mathbb{E}_{X_1'|\bar{Y}}\left[(\mathbb{E}_{X_2}[\mathbb{E}[Y|X_1, X_2, \bar{Y}]|\bar{Y}] - \mathbb{E}_{X_2}[\mathbb{E}[Y|X_1', X_2, \bar{Y}]|\bar{Y}])^2\right] \\
&\leq \frac{1}{2}\mathbb{E}_{\bar{Y}}\mathbb{E}_{X_1|\bar{Y}}\mathbb{E}_{X_1'|\bar{Y}}\mathbb{E}_{X_2|\bar{Y}}\left[(\mathbb{E}[Y|X_1, X_2, \bar{Y}] - \mathbb{E}[Y|X_1', X_2, \bar{Y}])^2\right] \\
&= \frac{1}{2}\mathbb{E}\left[(Z_1 + Z_2 + Z_3 + Z_4)^2\right]
\end{aligned}$$

where $Z_1 = \mathbb{E}[Y|X_1, X_2, \bar{Y}] - \mathbb{E}[Y|X_1, X_2]$, $Z_2 = -\mathbb{E}[Y|X_1', X_2, \bar{Y}] + \mathbb{E}[Y|X_1', X_2]$, $Z_3 = \mathbb{E}[Y|X_1, X_2] - \mathbb{E}[Y|X_2]$ and $Z_4 = -\mathbb{E}[Y|X_1', X_2] + \mathbb{E}[Y|X_2]$. The first and third inequality follow from Jensen's inequality, second inequality follows from $\mathbb{E}[(X - \mathbb{E}[X])^2] = \frac{1}{2}\mathbb{E}[(X - X')^2]$, and the third equality follows from the CI assumption.

We will bound $\mathbb{E}[Z_1^2] = \mathbb{E}[Z_2^2] \leq \mathbb{E}[(\mathbb{E}[Y|X_1, X_2, \bar{Y}] - \mathbb{E}[Y|X_1, X_2])^2] \leq \mathbb{E}[(Y - \mathbb{E}[Y|X_1, X_2])^2] = \sigma_Y^2$ again from Jensen's inequality. Z_3 and Z_4 can be handled by observing that $\mathbb{E}[Z_3^2] = \mathbb{E}[Z_4^2] = \mathbb{E}[(\mathbb{E}[Y|X_1, X_2] - \mathbb{E}[Y|X_2])^2] = \epsilon_{X_2}^2$.

Thus using the above lemma, we get the desired upper bound on $\epsilon_{\bar{Y}}$. □

G Showing $\mathbb{E}[Y|X_1] \approx \mathbb{E}[Y|X_1, X_2]$

Our main result Theorem 4.2 shows that self-supervised learning can help approximate $\mathbb{E}[Y|X_1]$ as a linear function of the learned features $\tilde{\psi}$. In practice, however, it is more common to predict the label Y using the entire input $X = (X_1, X_2)$ rather than just X_1. We show here that learning $\mathbb{E}[Y|X_1]$ is sufficient, under mild assumptions on the task being solved: the Bayes error of the classification task (X_1, Y) is low. We first upper bound the discrepancy between $\mathbb{E}[Y|X_1]$ and $\mathbb{E}[Y|X_1, X_2]$ based on the Bayes error rate.

Lemma G.1. *Suppose $\|Y\| \leq 1$ and $k = |\mathcal{Y}|$. Denote the Bayes error for distribution $P_{X_1,Y}$ to be Bayes-error$(P_{X_1,Y}) = \mathbb{E}_{X_1}\left[1 - \max_y P(y|X_1)\right]$[9]. Then we have*

$$\mathbb{E}_{X_1,X_2}\left[\|\mathbb{E}[Y|X_1] - \mathbb{E}[Y|X_1, X_2]\|^2\right] \leq 2k \text{ Bayes-error}(P_{X_1,Y})$$

We will show below (for $\mathcal{H} = \mathcal{H}_u$) that if $P_{X_1,Y}$ has low Bayes error, then predicting $\mathbb{E}[Y|X_1]$ is as good as predicting $\mathbb{E}[Y|X_1, X_2]$ up to this small additive error.

Theorem G.2. *Suppose $\epsilon_{Bayes} =$ Bayes-error$(P_{X_1,Y})$ and that $\tilde{\psi}$ is ϵ_{pre}^2-optimal on the SSL task (as in Theorem 4.2). Under the same conditions as Theorem 4.2, with probability $1 - \delta$ we have*

$$\mathbb{E}_{X_1,X_2}\left[\|\mathbb{E}[Y|X_1, X_2] - \hat{\boldsymbol{W}}^\top \tilde{\psi}(X_1)\|_2^2\right] \leq \tilde{\mathcal{O}}\left(\sigma^2 \frac{d_2}{n_2} + \frac{\epsilon_{CI}^2}{\beta^2} + \frac{\epsilon_{pre}^2}{\beta^2}\right) + 2\epsilon_{Bayes}$$

Proof. The law of total expectation gives $\mathbb{E}_{X_2}[\mathbb{E}[Y|X_1, X_2]|X_1] = \mathbb{E}[Y|X_1]$, thus it is easy to obtain the following decomposition

$$\begin{aligned}
\mathbb{E}_{X_1,X_2}\left[\|\mathbb{E}[Y|X_1, X_2] - \hat{\boldsymbol{W}}^\top \tilde{\psi}(X_1)\|_2^2\right] =& \mathbb{E}_{X_1}\left[\|\mathbb{E}[Y|X_1] - \hat{\boldsymbol{W}}^\top \tilde{\psi}(X_1)\|_2^2\right] \\
&+ \mathbb{E}_{X_1,X_2}\left[\|\mathbb{E}[Y|X_1] - \mathbb{E}[Y|X_1, X_2]\|_2^2\right]
\end{aligned}$$

[9] We abuse notation and use $P(y|X_1)$ instead of $P_{X_1,Y}(y|X_1)$.

The first term can be upper bounded using Theorem 4.2: $\mathbb{E}_{X_1}\left[\|\mathbb{E}[Y|X_1] - \hat{\boldsymbol{W}}^\top \tilde{\psi}(X_1)\|_2^2\right] = \mathrm{ER}_{\tilde{\psi}}(\hat{\boldsymbol{W}}) \leq \tilde{\mathcal{O}}\left(\sigma^2 \frac{d_2}{n_2} + \frac{\epsilon_{\mathrm{CI}}^2}{\beta^2} + \frac{\epsilon_{\mathrm{pre}}^2}{\beta^2}\right)$. The second term is upper bounded by $2\epsilon_{\text{Bayes}}$ by invoking Lemma G.1, and this completes the proof □

Proof of Lemma G.1. Notice the following inequality

$$\mathbb{E}_{X_1,X_2}\left[\|\mathbb{E}[Y|X_1] - \mathbb{E}[Y|X_1,X_2]\|^2\right] = \mathbb{E}_{X_1,X_2}\left[\left\|\sum_{y\in\mathcal{Y}} y\left(P(y|X_1) - P(y|X_1,X_2)\right)\right\|^2\right]$$
$$\leq |\mathcal{Y}|(\max_y \|y\|^2)\mathbb{E}_{X_1,X_2}\left[\sum_y \left(P(y|X_1) - P(y|X_1,X_2)\right)^2\right]$$
$$\leq k\mathbb{E}_{X_1}\left[\mathbb{E}_{X_2}\left[\sum_y \left(P(y|X_1) - P(y|X_1,X_2)\right)^2 \mid X_1\right]\right]$$

where the first inequality follows from Cauchy-Schwartz and second inequality follows from $\|Y\| \leq 1$. Thus the problem reduces to bounding the inner expectation for every X_1. We first note that for every X_1, y, we have $P(y|X_1) = \mathbb{E}_{X_2}[P(y|X_1,X_2)|X_1]$ from the law of total expectation. This gives

$$\mathbb{E}_{X_2}\left[\sum_y \left(P(y|X_1) - P(y|X_1,X_2)\right)^2 \mid X_1\right] = \sum_y \mathbb{E}_{X_2}\left[P(y|X_1,X_2)^2|X_1\right] - P(y|X_1)^2$$
$$\leq \sum_y \mathbb{E}_{X_2}\left[P(y|X_1,X_2)|X_1\right] - P(y|X_1)^2 = \mathbb{E}_{X_2}\left[\sum_y P(y|X_1,X_2)|X_1\right] - \sum_y P(y|X_1)^2$$
$$= 1 - \sum_y P(y|X_1)^2 \leq 1 - \max_y P(y|X_1)^2 \leq 2(1 - \max_y P(y|X_1))$$

where the first inequality follows because $P(y|X_1,X_2) \in [0,1]$ and second follows trivially and third follows from $1 - x^2 \leq 2(1-x)$ for $x \in [0,1]$. Combining everything, we get $\mathbb{E}_{X_1,X_2}\left[\|\mathbb{E}[Y|X_1] - \mathbb{E}[Y|X_1,X_2]\|^2\right] \leq 2k\mathbb{E}_{X_1}\left[1 - \max_y P(y|X_1)\right] = 2k$ Bayes-error$(P_{X_1,Y})$, thus proving the result. □

H Theoretical analysis for classification tasks

H.1 Classification tasks

We now consider the benefit of learning ψ from a class $\mathcal{H}_1$ on linear classification task for label set $\mathcal{Y} = [k]$. The performance of a classifier is measured using the standard logistic loss

Definition H.1. *For a task with $\mathcal{Y} = [k]$, classification loss for a predictor $f : \mathcal{X}_1 \to \mathbb{R}^k$ is*

$$\ell_{clf}(f) = \mathbb{E}[\ell_{log}(f(X_1), Y)]\text{, where } \ell_{log}(\hat{y}, y) = \left[-\log\left(\frac{e^{\hat{y}_y}}{\sum_{y'} e^{\hat{y}_{y'}}}\right)\right]$$

The loss for representation $\psi : \mathcal{X}_1 \to \mathbb{R}^{d_1}$ and linear classifier $\boldsymbol{W} \in \mathbb{R}^{k\times d_1}$ is denoted by $\ell_{clf}(\boldsymbol{W}\psi)$.

We note that the function $\ell_{\log}$ is 1-Lipschitz in the first argument. The result will also hold for the hinge loss $\ell_{\text{hinge}}(\hat{y}, y) = (1 - \hat{y}_y + \max_{y'\neq y} \hat{y}_{y'})_+$ which is also 1-Lipschitz, instead of $\ell_{\log}$.

We assume that the optimal regressor $f^*_{\mathcal{H}_1}$ for one-hot encoding also does well on linear classification.

Assumption H.1. *The best regressor for 1-hot encodings in $\mathcal{H}_1$ does well on classification, i.e. $\ell_{clf}(\gamma f^*_{\mathcal{H}_1}) \leq \epsilon_{\text{one-hot}}$ is small for some scalar γ.*

Remark H.1. *Note that if $\mathcal{H}_1$ is universal, then $f^*_{\mathcal{H}_1}(\boldsymbol{x}_1) = \mathbb{E}[Y|X_1 = \boldsymbol{x}_1]$ and we know that $f^*_{\mathcal{H}_1}$ is the Bayes-optimal predictor for binary classification. In general one can potentially predict the label by looking at $\arg\max_{i\in[k]} f^*_{\mathcal{H}_1}(\boldsymbol{x}_1)_i$. The scalar γ captures the margin in the predictor $f^*_{\mathcal{H}_1}$.*

We now show that using the classifier $\hat{\boldsymbol{W}}$ obtained from linear regression on one-hot encoding with learned representations $\tilde{\psi}$ will also be good on linear classification. The proof is in Section H

Theorem H.2. *For a fixed $\delta \in (0,1)$, under the same setting as Theorem 4.2 and Assumption H.1, we have:*

$$\ell_{clf}\left(\gamma \hat{\boldsymbol{W}} \tilde{\psi}\right) \leq \tilde{\mathcal{O}}\left(\gamma \sqrt{\sigma^2 \frac{d_2}{n_2} + \frac{\epsilon^2}{\beta^2} + \frac{\epsilon_{pre}^2}{\beta^2}}\right) + \epsilon_{one\text{-}hot},$$

with probability $1-\delta$.

Proof of Theorem H.2. We simply follow the following sequence of steps

$$\begin{aligned}
\ell_{\text{clf}}\left(\gamma \hat{\boldsymbol{W}} \tilde{\psi}\right) &= \mathbb{E}[\ell_{\log}\left(\gamma \hat{\boldsymbol{W}} \tilde{\psi}(X_1), Y\right)] \\
&\leq^{(a)} \mathbb{E}\left[\ell_{\log}\left(\gamma f^*_{\mathcal{H}_1}(X_1), Y\right) + \gamma \|\hat{\boldsymbol{W}} \tilde{\psi}(X_1) - f^*_{\mathcal{H}_1}(X_1)\|\right] \\
&\leq^{(b)} \epsilon_{\text{one-hot}} + \gamma \sqrt{\mathbb{E}\left[\|\hat{\boldsymbol{W}} \tilde{\psi}(X_1) - f^*_{\mathcal{H}_1}(X_1)\|^2\right]} \\
&= \epsilon_{\text{one-hot}} + \gamma \sqrt{\text{ER}_{\tilde{\psi}}[\hat{\boldsymbol{W}}]}
\end{aligned}$$

where (a) follows because $\ell_{\log}$ is 1-Lipschitz and (b) follows from Assumption H.1 and Jensen's inequality. Plugging in Theorem 4.2 completes the proof. □

I Four Different Ways to Use CI

In this section we propose four different ways to use conditional independence to prove zero approximation error, i.e.,

Claim I.1 (informal)**.** *When conditional independence is satisfied: $X_1 \perp X_2 | Y$, and some non-degeneracy is satisfied, there exists some matrix $\boldsymbol{W}$ such that $\mathbb{E}[Y|X_1] = \boldsymbol{W}\mathbb{E}[X_2|X_1]$.*

We note that for simplicity, most of the results are presented for the jointly Gaussian case, where everything could be captured by linear conditional expectation $\mathbb{E}^L[Y|X_1]$ or the covariance matrices. When generalizing the results for other random variables, we note just replace X_1, X_2, Y by $\phi_1(X_1), \phi_2(X_2), \phi_y(Y)$ will suffice the same arguments.

I.1 Inverse Covariance Matrix

Write $\boldsymbol{\Sigma}$ as the covariance matrix for the joint distribution $P_{X_1 X_2 Y}$.

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{XX} & \boldsymbol{\Sigma}_{XY} \\ \boldsymbol{\Sigma}_{YY}^\top & \boldsymbol{\Sigma}_{YY} \end{bmatrix}, \quad \boldsymbol{\Sigma}^{-1} = \begin{bmatrix} \boldsymbol{A} & \rho \\ \rho^\top & \boldsymbol{B} \end{bmatrix}$$

where $\boldsymbol{A} \in \mathbb{R}^{(d_1+d_2)\times(d_1+d_2)}, \rho \in \mathbb{R}^{(d_1+d_2)\times k}, \boldsymbol{B} \in \mathbb{R}^{k\times k}$. Furthermore

$$\rho = \begin{bmatrix} \rho_1 \\ \rho_2 \end{bmatrix}; \quad \boldsymbol{A} = \begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} \\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} \end{bmatrix}$$

for $\rho_i \in \mathbb{R}^{d_i \times k}, i = 1,2$ and $\boldsymbol{A}_{ij} \in \mathbb{R}^{d_i \times d_j}$ for $i,j \in \{1,2\}$.

Claim I.2. *When conditional independence is satisfied, $\boldsymbol{A}$ is block diagonal matrix, i.e., $\boldsymbol{A}_{12}$ and $\boldsymbol{A}_{21}$ are zero matrices.*

Lemma I.3. *We have the following*

$$\mathbb{E}[X_1|X_2] = (\boldsymbol{A}_{11} - \bar{\rho}_1\bar{\rho}_1^\top)^{-1}(\bar{\rho}_1\bar{\rho}_2{}^\top - \boldsymbol{A}_{12})X_2 \tag{18}$$

$$\mathbb{E}[X_2|X_1] = (\boldsymbol{A}_{22} - \bar{\rho}_2\bar{\rho}_2^\top)^{-1}(\bar{\rho}_2\bar{\rho}_1{}^\top - \boldsymbol{A}_{21})X_1 \tag{19}$$

$$\mathbb{E}[Y|X] = -B^{-\frac{1}{2}}(\bar{\rho}_1^\top X_1 + \bar{\rho}_2^\top X_2) \tag{20}$$

where $\bar{\rho}_i = \rho_i \boldsymbol{B}^{-\frac{1}{2}}$ for $i \in \{1,2\}$. Also,

$$(\boldsymbol{A}_{11} - \bar{\rho}_1\bar{\rho}_1^\top)^{-1}\bar{\rho}_1\bar{\rho}_2^\top = \frac{1}{1 - \bar{\rho}_1^\top \boldsymbol{A}_{11}^{-1}\bar{\rho}_1}\boldsymbol{A}_{11}^{-1}\bar{\rho}_1\bar{\rho}_2^\top$$

$$(\boldsymbol{A}_{22} - \bar{\rho}_2\bar{\rho}_2^\top)^{-1}\bar{\rho}_2\bar{\rho}_1^\top = \frac{1}{1 - \bar{\rho}_2^\top \boldsymbol{A}_{22}^{-1}\bar{\rho}_2}\boldsymbol{A}_{22}^{-1}\bar{\rho}_2\bar{\rho}_1^\top$$

Proof. We know that $\mathbb{E}[X_1|X_2] = \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}X_2$ and $\mathbb{E}[X_2|X_1] = \mathbf{\Sigma}_{21}\mathbf{\Sigma}_{11}^{-1}x_1$, where

$$\mathbf{\Sigma}_{XX} = \begin{bmatrix} \mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12} \\ \mathbf{\Sigma}_{21} & \mathbf{\Sigma}_{22} \end{bmatrix}$$

First using $\mathbf{\Sigma}\mathbf{\Sigma}^{-1} = I$, we get the following identities

$$\mathbf{\Sigma}_{XX}\boldsymbol{A} + \mathbf{\Sigma}_{XY}\rho^\top = \boldsymbol{I} \tag{21}$$

$$\mathbf{\Sigma}_{XY}^\top\boldsymbol{A} + \mathbf{\Sigma}_{YY}\rho^\top = 0 \tag{22}$$

$$\mathbf{\Sigma}_{XX}\rho + \mathbf{\Sigma}_{XY}\boldsymbol{B} = 0 \tag{23}$$

$$\mathbf{\Sigma}_{XY}^\top\rho + \mathbf{\Sigma}_{YY}\boldsymbol{B} = \boldsymbol{I} \tag{24}$$

From Equation (23) we get that $\mathbf{\Sigma}_{XY} = -\mathbf{\Sigma}_{XX}\rho\boldsymbol{B}^{-1}$ and plugging this into Equation (21) we get

$$\begin{aligned} &\mathbf{\Sigma}_{XX}\boldsymbol{A} - \mathbf{\Sigma}_{XX}\rho\boldsymbol{B}^{-1}\rho^\top = \boldsymbol{I} \\ \implies &\mathbf{\Sigma}_{XX} = (\boldsymbol{A} - \rho\boldsymbol{B}^{-1}\rho^\top)^{-1} = (\boldsymbol{A} - \bar{\rho}\bar{\rho}^\top)^{-1} \\ \implies &\begin{bmatrix} \mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12} \\ \mathbf{\Sigma}_{21} & \mathbf{\Sigma}_{22} \end{bmatrix} = \left(\begin{bmatrix} \boldsymbol{A}_{11} - \bar{\rho}_1\bar{\rho}_1^\top & \boldsymbol{A}_{12} - \bar{\rho}_1\bar{\rho}_2^\top \\ \boldsymbol{A}_{21} - \bar{\rho}_2\bar{\rho}_1^\top & \boldsymbol{A}_{22} - \bar{\rho}_2\bar{\rho}_2^\top \end{bmatrix} \right)^{-1} \end{aligned}$$

We now make use of the following expression for inverse of a matrix that uses Schur complement: $\boldsymbol{M}/\alpha = \delta - \gamma\alpha^{-1}\beta$ is the Schur complement of α for $\boldsymbol{M}$ defined below

$$\text{If } \boldsymbol{M} = \begin{bmatrix} \alpha & \beta \\ \gamma & \delta \end{bmatrix}, \text{ then, } \boldsymbol{M}^{-1} = \begin{bmatrix} \alpha^{-1} + \alpha^{-1}\beta(\boldsymbol{M}/\alpha)^{-1}\gamma\alpha^{-1} & -\alpha^{-1}\beta(\boldsymbol{M}/\alpha)^{-1} \\ -(\boldsymbol{M}/\alpha)^{-1}\gamma\alpha^{-1} & (\boldsymbol{M}/\alpha)^{-1} \end{bmatrix}$$

For $\boldsymbol{M} = (\boldsymbol{A} - \bar{\rho}\bar{\rho}^\top)$, we have that $\mathbf{\Sigma}_{XX} = \boldsymbol{M}^{-1}$ and thus

$$\begin{aligned} \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1} &= -\alpha^{-1}\beta(\boldsymbol{M}/\alpha)^{-1}((\boldsymbol{M}/\alpha)^{-1})^{-1} \\ &= -\alpha^{-1}\beta \\ &= (\boldsymbol{A}_{11} - \bar{\rho}_1\bar{\rho}_1^\top)^{-1}(\bar{\rho}_1\bar{\rho}_2^\top - \boldsymbol{A}_{12}) \end{aligned}$$

This proves Equation (18) and similarly Equation (19) can be proved.

For Equation (20), we know that $\mathbb{E}[Y|X = (X_1, X_2)] = \mathbf{\Sigma}_{YX}\mathbf{\Sigma}_{XX}^{-1}X = \mathbf{\Sigma}_{XY}^\top\mathbf{\Sigma}_{XX}^{-1}X$. By using Equation (23) we get $\mathbf{\Sigma}_{XY} = -\mathbf{\Sigma}_{XX}\rho\boldsymbol{B}^{-1}$ and thus

$$\begin{aligned} \mathbb{E}[Y|X = (X_1, X_2)] &= -\boldsymbol{B}^{-1}\rho^\top\mathbf{\Sigma}_{XX}\mathbf{\Sigma}_{XX}^{-1}X \\ &= -\boldsymbol{B}^{-1}\rho^\top X = \boldsymbol{B}^{-1}(\rho_1^\top X_1 + \rho_2^\top X_2) \\ &= -\boldsymbol{B}^{-\frac{1}{2}}(\bar{\rho}_1^\top X_1 + \bar{\rho}_2^\top X_2) \end{aligned}$$

For the second part, we will use the fact that $(\boldsymbol{I} - \boldsymbol{a}\boldsymbol{b}^\top)^{-1} = \boldsymbol{I} + \frac{1}{1-\boldsymbol{a}^\top\boldsymbol{b}}\boldsymbol{a}\boldsymbol{b}^\top$. Thus

$$\begin{aligned} (\boldsymbol{A}_{11} - \bar{\rho}_1\bar{\rho}_1^\top)^{-1}\bar{\rho}_1\bar{\rho}_2 &= (\boldsymbol{I} - \boldsymbol{A}_{11}^{-1}\bar{\rho}_1\bar{\rho}_1^\top)\boldsymbol{A}_{11}^{-1}\bar{\rho}_1\bar{\rho}_2^\top \\ &= (\boldsymbol{I} + \frac{1}{1 - \bar{\rho}_1^\top\boldsymbol{A}_{11}^{-1}\bar{\rho}_1}\boldsymbol{A}_{11}^{-1}\bar{\rho}_1\bar{\rho}_1)\boldsymbol{A}_{11}^{-1}\bar{\rho}_1\bar{\rho}_2^\top \\ &= \boldsymbol{A}_{11}^{-1}(I + \frac{1}{1 - \bar{\rho}_1^\top\boldsymbol{A}_{11}^{-1}\bar{\rho}_1}\bar{\rho}_1\bar{\rho}_1\boldsymbol{A}_{11}^{-1})\bar{\rho}_1\bar{\rho}_2^\top \\ &= \boldsymbol{A}_{11}^{-1}(\bar{\rho}_1\bar{\rho}_2^\top + \frac{\bar{\rho}_1\boldsymbol{A}_{11}^{-1}\bar{\rho}_1}{1 - \bar{\rho}_1^\top\boldsymbol{A}_{11}^{-1}\bar{\rho}_1}\bar{\rho}_1\bar{\rho}_2^\top) \\ &= \boldsymbol{A}_{11}^{-1}\bar{\rho}_1\bar{\rho}_2^\top(1 + \frac{\bar{\rho}_1\boldsymbol{A}_{11}^{-1}\bar{\rho}_1}{1 - \bar{\rho}_1^\top\boldsymbol{A}_{11}^{-1}\bar{\rho}_1}) \\ &= \frac{1}{1 - \bar{\rho}_1^\top\boldsymbol{A}_{11}^{-1}\bar{\rho}_1}A_{11}^{-1}\bar{\rho}_1\bar{\rho}_2^\top \end{aligned}$$

The other statement can be proved similarly. □

Claim I.4.

$\mathbb{E}[X_2|X_1] = (\boldsymbol{A}_{22} - \bar{\rho}_2\bar{\rho}_2^\top)^{-1}\bar{\rho}_2\bar{\rho}_1^\top X_1.\mathbb{E}[Y|X_1] = -\boldsymbol{B}^{-1/2}\bar{\rho}_1^\top X_1 - \boldsymbol{B}^{-1/2}\bar{\rho}_2^\top\mathbb{E}[X_2|X_1]$

Therefore $\mathbb{E}[Y|X_1]$ *is in the same direction as* $\mathbb{E}[X_2|X_1]$.

I.2 Closed form of Linear Conditional Expectation

Refer to Claim B.1 and proof of Lemma B.2. As this is the simplest proof we used in our paper.

I.3 From Law of Iterated Expectation

$$\begin{aligned}\mathbb{E}^L[X_2|X_1] =& \mathbb{E}^L[\mathbb{E}^L[X_2|X_1, Y]|X_1] \\ =& \mathbb{E}\left[[\mathbf{\Sigma}_{X_2X_1}, \mathbf{\Sigma}_{X_2Y}]\begin{bmatrix}\mathbf{\Sigma}_{X_1X_1} & \mathbf{\Sigma}_{X_1Y} \\ \mathbf{\Sigma}_{YX_1} & \mathbf{\Sigma}_{YY}\end{bmatrix}^{-1}\begin{bmatrix}X_1 \\ Y\end{bmatrix} \mid X_1\right] \\ =& \boldsymbol{A}X_1 + \boldsymbol{B}\mathbb{E}^L[Y|X_1].\end{aligned}$$

Using block matrix inverse,

$$\begin{aligned}\boldsymbol{A} &= (\mathbf{\Sigma}_{X_2X_1} - \mathbf{\Sigma}_{X_2Y}\mathbf{\Sigma}_{YY}^{-1}\mathbf{\Sigma}_{YX_1})(\mathbf{\Sigma}_{X_1X_1} - \mathbf{\Sigma}_{X_1Y}\mathbf{\Sigma}_{YY}^{-1}\mathbf{\Sigma}_{YX_1})^{-1} \in \mathbb{R}^{d_2\times d_1} \\ &= \mathbf{\Sigma}_{X_1X_2|Y}(\mathbf{\Sigma}_{X_1X_1|Y})^{-1} \\ \boldsymbol{B} &= \mathbf{\Sigma}_{X_2Y|X_1}(\mathbf{\Sigma}_{YY|X_1})^{-1} \in \mathbb{R}^{d_2\times\mathcal{Y}}.\end{aligned}$$

Therefore in general (without conditional independence assumption) our learned representation will be $\psi(x_1) = \boldsymbol{A}x_1 + \boldsymbol{B}f^*(x_1)$, where $f^*(\cdot) := \mathbb{E}^L[Y|X_1]$.

It's easy to see that to learn f^* from representation ψ, we need A to have some good property, such as light tail in eigenspace, and B needs to be full rank in its column space.

Notice in the case of conditional independence, $\mathbf{\Sigma}_{X_1X_2|Y} = 0$, and $A = 0$. Therefore we could easily learn f^* from ψ if X_2 has enough information of Y such that $\mathbf{\Sigma}_{X_2Y|X_1}$ is of the same rank as dimension of Y.

I.4 From $\mathbb{E}[X_2|X_1, Y] = \mathbb{E}[X_2|Y]$

Proof. Let the representation function ψ be defined as follows, and let we use law of iterated expectation:

$$\begin{aligned}\psi(\cdot) := \mathbb{E}[X_2|X_1] =& \mathbb{E}[\mathbb{E}[X_2|X_1, Y]|X_1] \\ =& \mathbb{E}[\mathbb{E}[X_2|Y]|X_1] && \text{(uses CI)} \\ =& \sum_y P(Y = y|X_1)\mathbb{E}[X_2|Y = y] \\ =:& f(X_1)^\top A,\end{aligned}$$

where $f : \mathbb{R}^{d_1} \to \Delta_{\mathcal{Y}}$ satisfies $f(x_1)_y = P(Y = y|X_1 = x_1)$, and $\boldsymbol{A} \in \mathbb{R}^{\mathcal{Y}\times d_2}$ satisfies $\boldsymbol{A}_{y,:} = \mathbb{E}[X_2|Y = y]$. Here Δ_d denotes simplex of dimension d, which represents the discrete probability density over support of size d.

Let $\boldsymbol{B} = \boldsymbol{A}^\dagger \in \mathbb{R}^{\mathcal{Y}\times d_2}$ be the pseudoinverse of matrix $\boldsymbol{A}$, and we get $\boldsymbol{BA} = \boldsymbol{I}$ from our assumption that A is of rank $|\mathcal{Y}|$. Therefore $f(\boldsymbol{x}_1) = \boldsymbol{B}\psi(\boldsymbol{x}_1), \forall x_1$. Next we have:

$$\begin{aligned}\mathbb{E}[Y|X_1 = \boldsymbol{x}_1] =& \sum_y P(Y = y|X_1 = \boldsymbol{x}_1) \times y \\ =& \hat{\boldsymbol{Y}}f(\boldsymbol{x}_1) \\ =& (\hat{\boldsymbol{Y}}\boldsymbol{B}) \cdot \psi(X_1).\end{aligned}$$

Here we denote by $\hat{\boldsymbol{Y}} \in \mathbb{R}^{k\times\mathcal{Y}}, \hat{\boldsymbol{Y}}_{:,y} = y$ that spans the whole support $\mathcal{Y}$. Therefore let $\boldsymbol{W}^* = \hat{\boldsymbol{Y}}\boldsymbol{B}$ will finish the proof.

□

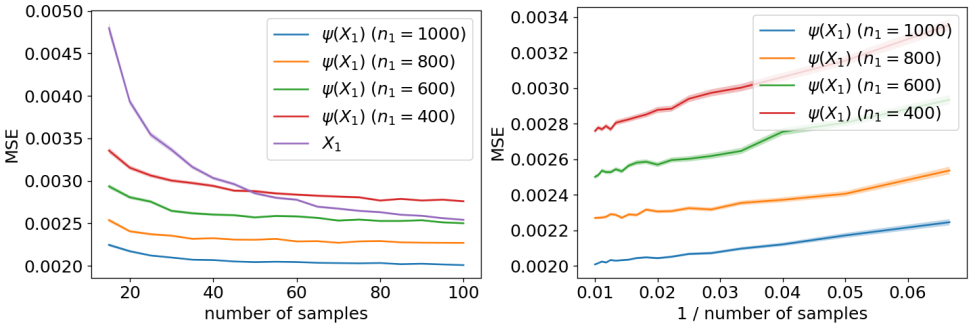


Figure 3: **Left**: MSE of using ψ to predict Y versus using X_1 directly to predict Y. Using ψ consistently outperforms using X_1. **Right**: MSE of ψ learned with different n_1. The MSE scale with $1/n_2$ as indicated by our analysis. Simulations are repeated 100 times, with the mean shown in solid line and one standard error shown in shadow.

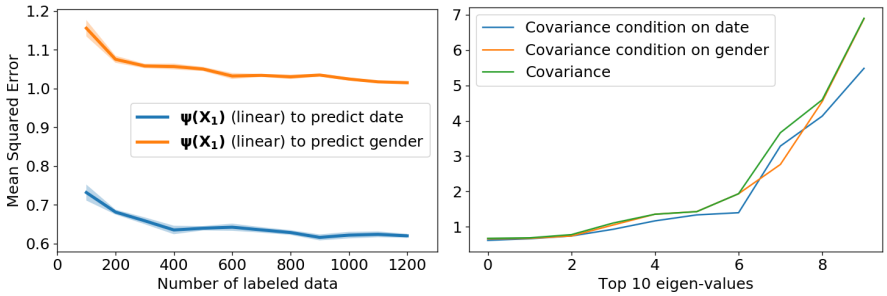


Figure 4: **Left**: Mean Squared Error comparison of predicting gender and predicting date. **Right**: the spectrum comparison of covariance condition on gender and condition on date.

J More on the experiments

In this section, we include more experiment setup and results.

Simulations. All the experiments are performed on a desktop computer with Intel i7-8700K, 16GB RAM.

Following Theorem 4.2, we know that the Excessive Risk (ER) is also controlled by (1) the number of samples for the pretext task (n_1), and (2) the number of samples for the downstream task (n_2), besides k and ϵ_{CI} as discussed in the main text. In this simulation, we enforce strict conditional independence, and explore how ER varies with n_1 and n_2. We generate the data the same way as in the main text, and keep $\alpha = 0, k = 2$, $d_1 = 50$ and $d_2 = 40$ We restrict the function class to linear model. Hence ψ is the linear model to predict X_2 from X_1 given the pretext dataset. We use Mean Squared Error (MSE) as the metric, since it is the empirical version of the ER. As shown in Figure 3, ψ consistently outperforms X_1 in predicting Y using a linear model learnt from the given downstream dataset, and ER does scale linearly with $1/n_2$, as indicated by our analysis.

Computer Vision Task. For the context encoder part, we use all the recommended hyperparameter as in the provided source codes. For the downstream resnet18 regression, we perform grid search over the hyperparameters to achieve best performance. Specifically, we set the batch size to be 24, and traing the resnet18 for 50 epoches. One pass of training (loops over all the settings with different number of labeled data) is finished within 6 hours. All the experiments are performed on a desktop computer with Intel i7-8700K, 16GB RAM, and NVIDIA Geforce 1080. Training of the context encoder is finished within 12 hours. The yearbook dataset is distributed under BSD license.

Following the same procedure, we try to predict the gender Y_G. We normalize the label (Y_G, Y_D) to unit variance, and confine ourself to linear function class. That is, instead of using a context encoder to impaint X_2 from X_1, we confine ψ to be a linear function. As shown on the left of Figure 4, the MSE

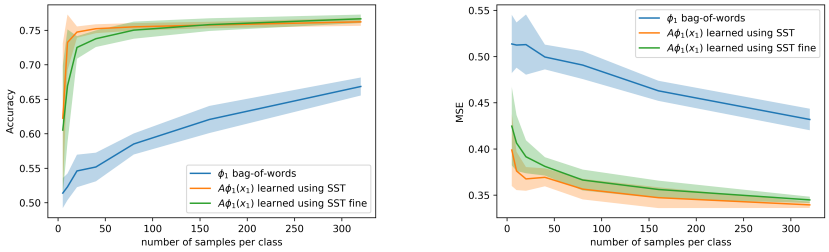


Figure 5: Performance on SST of baseline $\phi_1(\boldsymbol{x}_1)$, i.e. bag-of-words, and learned $\psi(\boldsymbol{x}_1)$ for the two settings. **Left:** Classification accuracy, **Right:** Regression MSE.

of predicting gender is higher than predicting dates. We find that $\|\boldsymbol{\Sigma}_{\boldsymbol{X}_1\boldsymbol{X}_1}^{-1/2}\boldsymbol{\Sigma}_{\boldsymbol{X}_1 X_2|Y_G}\|_F = 9.32$, while $\|\boldsymbol{\Sigma}_{\boldsymbol{X}_1\boldsymbol{X}_1}^{-1/2}\boldsymbol{\Sigma}_{\boldsymbol{X}_1 X_2|Y_D}\|_F = 8.15$. Moreover, as shown on the right of Figure 4, conditioning on Y_D cancels out more spectrum than conditioning on Y_G. In this case, we conjecture that, unlike Y_D, Y_G does not capture much dependence between X_1 and X_2. And as a result, ϵ_{CI} is larger, and the downstream performance is worse, as we expected.

NLP Task. We look at the setting where both $\mathcal{X}_1$ and $\mathcal{X}_2$ are the set of sentences and perform experiments by enforcing CI with and without latent variables. The downstream task is sentiment classification with the Stanford Sentiment Treebank (SST) dataset [53], where inputs are movie reviews and the label set $\mathcal{Y}$ is $\{\pm 1\}$. We learn a linear representation $\psi(X_1) = \boldsymbol{B}\phi(X_1)$ in the SSL phase as defined in Section 4. Here we X_1, we pick $\phi(X_1)$ to be the bag-of-words representations of the movie review X_1, which has a vocabulary size of 13848 For X_2 we use a $d_2 = 300$ dimensional embedding of the sentence, that is the mean of word vectors (random Gaussians) for the words in the review X_2. For SSL data we consider 2 settings, (a) enforce CI with the labels $\mathcal{Y}$, (b) enforce CI with extra latent variables, for which we use fine-grained version of SST with label set $\bar{\mathcal{Y}} = \{1, 2, 3, 4, 5\}$[10].. In this setting, for every label $y \in \mathcal{Y}$ (or $\bar{y} \in \bar{\mathcal{Y}}$), we independently sample movie reviews X_1 and X_2 from the class y (or $\bar{y}$), thus simulating the CI (or approximate CI) condition. We test the learned ψ on SST binary task with linear regression and linear classification; results are presented in Figure 5. We observe that in both settings ψ outperforms ϕ_1, especially in the small-sample-size regime. Exact CI is better than CI with latent variables, as suggested by theory.

The function ψ (or equivalently matrix $\boldsymbol{B} \in \mathbb{R}^{300\times 13848}$) is learnt by minimizing $\|X_2 - \boldsymbol{B}\phi(X_1)\|^2$ averaged over the SSL train data with an $\|\cdot\|_F^2$ penalty on the matrix $\boldsymbol{B}$. We use the scikit-learn RidgeRegressionCV[11] solver for this with regularizer parameters in the list $[0.001, 0.1, 10, 1000]$. Plotting Figure 5 took less than an hour when using 8 Intel(R) Xeon(R) Silver 4214 CPUs on a cluster.

[10]Ratings $\{1, 2\}$ correspond to $y = -1$ and $\{4, 5\}$ correspond to $y = 1$

[11]https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.RidgeCV.html