

# Can Pretext-Based Self-Supervised Learning Be Boosted by Downstream Data? A Theoretical Analysis

Jiaye Teng\*

IIIS, Tsinghua University  
tjy20@mails.tsinghua.edu.cn

Weiran Huang\*

Huawei Noah's Ark Lab  
weiran.huang@outlook.com

## Abstract

Pretext-based self-supervised learning aims to learn the semantic representation via a hand-crafted pretext task over unlabeled data and then use the learned representation for downstream prediction tasks. Lee et al. (2020) prove that pretext-based self-supervised learning can effectively reduce the sample complexity of downstream tasks under Conditional Independence (CI) between the components of the pretext task conditional on the downstream label. However, the CI condition rarely holds in practice, and the downstream sample complexity will get much worse if the CI condition does not hold. In this paper, we explore the idea of applying a learnable function to the input to make the CI condition hold. In particular, we first rigorously formulate the criteria that the function needs to satisfy. We then design an ingenious loss function for learning such a function and prove that the function minimizing the proposed loss satisfies the above criteria. We theoretically study the number of labeled data required, and give a model-free lower bound showing that taking limited downstream data will hurt the performance of self-supervised learning. Furthermore, we take the model structure into account and give a model-dependent lower bound, which gets higher when the model capacity gets larger. Moreover, we conduct several numerical experiments to verify our theoretical results.

## 1. Introduction

Data representations used to be learned in a supervised or semi-supervised learning way (e.g., He et al., 2016; Lee et al., 2013). Recently, self-supervised learning has drawn massive attention for its fantastic data efficiency and generalization ability, with many state-of-the-art models following

this paradigm in computer vision (Chen et al., 2020; He et al., 2020), language modeling (Devlin et al., 2018; Radford et al., 2018), graph learning (Peng et al., 2020), etc. It learns data representations through self-supervised tasks, and then use the learned representations for downstream prediction tasks. Such paradigm involves a large number of unlabeled data that are easier to access, to reduce the sample complexity of labeled data.

The recent renaissance of self-supervised learning began with artificially designed *pretext tasks* as self-supervised tasks, such as image colorization (Zhang et al., 2016), inpainting (Pathak et al., 2016), solving jigsaw puzzles (Noroozi & Favaro, 2016), predicting image rotations (Gidaris et al., 2018), etc. Recently, such pretext-based methods achieve the state-of-the-art performance in natural language processing tasks, e.g., BERT (Devlin et al., 2018) and GPT (Radford et al., 2018). As an example of pretext-based self-supervised learning, let us consider predicting image rotations (Gidaris et al., 2018) as the pretext task. In this case, we rotate an image by  $k$  radians to get a sample  $x$  and use  $k$  as its pretext label  $z$ . Then the representation is learned from the joint distribution of  $(x, z)$ .

Since the representations are learned via pretext tasks, the performance of a downstream task highly depends on the choice of the pretext task. For example, if we consider the downstream task of classifying desert, forest, and sea images, then a meaningful pretext task can be predicting the background color of images (i.e., image colorization (Zhang et al., 2016)), rather than predicting image rotations (Gidaris et al., 2018). Moreover, recent experiments show that the self-supervised learned representations sometimes fail to transfer to other tasks (Yamaguchi et al., 2019; Zoph et al., 2020). Thus, quantitatively studying the connection between the pretext task and the downstream task is crucial to pretext-based self-supervised learning.

The theoretical study of pretext-based self-supervised learning is still at an early stage. A recent work (Lee et al., 2020) demonstrates that a reduced downstream sample complexity can be achieved by pretext-based self-supervised

\*Equal contribution. Author ordering determined by coin flip.

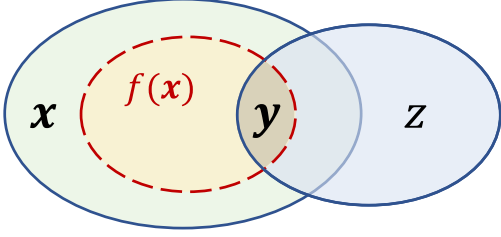


Figure 1. The situation that the CI condition  $x \perp z | y$  does not hold. To eliminate the information related to  $z$  in  $x$  conditional on  $y$ , we apply a function  $f$  to  $x$ , such that  $f(x) \perp z | y$  holds.

learning. In particular, it shows that for Gaussian variables, the sample complexity of downstream tasks can be reduced to  $\tilde{O}(\dim(y))$ <sup>1</sup> when conditional independence  $x \perp z | y$  holds, where  $x, z, y$  indicates the input variable, pretext label, and downstream label, respectively. As a comparison, the sample complexity of directly using  $x$  to predict  $y$  is  $\tilde{O}(\dim(x))$ , where the dimension of  $x$  is supposed to be much larger than the dimension of  $y$ . However, if the Conditional Independence (CI) does not hold, the downstream sample complexity will increase to  $\tilde{O}(\dim(z))$ .

In this paper, we further investigate the situation that the CI condition does not hold. We explore an intuitive idea of applying a function  $f$  to the input variable  $x$  to make conditional independence  $f(x) \perp z | y$  hold (See Figure 1). If such function  $f$  is known, then we can simply use  $(f(x), z)$  to learn the representation as the usual pretext task does. In this way, we only need  $\tilde{O}(\dim(y))$  samples for the downstream task using the learned representation. In particular, we rigorously formulate the criteria that the function  $f$  needs to satisfy in Section 3. Then we design an ingenious loss function to learn such function  $f$  in Section 4. We demonstrate the rationality of our loss by proving that any function  $f$  minimizing the proposed loss satisfies the criteria. Based on our loss, we theoretically study the number of labeled data required. We derive a model-free lower bound in Section 5.1 and show that taking limited downstream data will hurt the performance of pretext-based self-supervised learning. We also provide another lower bound based on a more general loss function. Furthermore, we take the model capacity of function class into account and give a model-dependent lower bound in Section 5.2. We show that the lower bound gets higher when the model capacity gets larger, indicating that the number of downstream samples needs to match the model capacity. To verify the above theoretical results, we conduct several numerical experiments in Section 6.

In summary, we list our contributions as follows:

- **Criterion Formulation:** We rigorously formulate the

criteria that function  $f$  should satisfy, and then design an ingenious loss function which is proved to be minimized when function  $f$  satisfying the proposed criteria.

- **Model-Free Lower Bound:** We provide a model-free lower bound of downstream sample size based on the proposed loss and a general loss, which indicates that taking insufficient downstream samples to the representation learning phase will hurt the performance of pretext-based self-supervised learning.
- **Model-Dependent Lower Bound:** We also give a model-dependent lower bound of downstream sample size, and show that the lower bound increases when the model capacity grows.

## 2. Related Works

**Self-Supervised Methods in Practice.** There are three common approaches for Self-Supervised Learning (SSL): generative model based, contrastive learning based, and pretext based. Generative model based SSL (Donahue et al., 2017; Dumoulin et al., 2017; Donahue & Simonyan, 2019) learns a bijective mapping between input and representation. Contrastive learning based SSL learns representations by maximizing the mutual information between the global and local features (Hjelm et al., 2018; Oord et al., 2018; Bachman et al., 2019) or between the features of positive samples (Tian et al., 2019; He et al., 2020; Chen et al., 2020). Pretext based SSL learns representations via handcrafted pretext tasks (Zhang et al., 2016; Pathak et al., 2016; Noroozi & Favaro, 2016; Gidaris et al., 2018). These three approaches are quite different in technique, and we focus on studying the pretext based SSL in this paper.

**Theory for Self-Supervised Learning.** Although there are a number of great empirical works for SSL, the theoretical study of SSL is still at an early stage. The most related work to ours is given by Lee et al. (2020). They are the first to give a formulation of pretext-based SSL, and show that it can reduce the sample complexity of downstream tasks compared with supervised learning. They also point out that when the CI condition holds, the downstream sample complexity achieves the optimal, and it gets worse when the CI condition does not hold. In this paper, we further study the situation that the CI condition does not hold, and explore the idea of applying a learnable function to the input to make the CI condition hold. Other works (Saunshi et al., 2019; Tosh et al., 2020) study the generalization error of contrastive learning based SSL, whose setting is different from our paper. Most recently, Bansal et al. (2020) analyze the generalization gap for most SSL methods. However, the rationality gap, which is a part of the generalization gap, cannot be theoretically bounded.

**Semi-Supervised Learning.** There are several key differ-

<sup>1</sup>We use notation  $\tilde{O}$  to hide log factors in this paper.

ences between self-supervised learning and semi-supervised learning (e.g., Wei et al. 2020): 1) Labeled data and unlabeled data are usually assumed to be drawn from the same distribution in semi-supervised learning, while they can be different in self-supervised learning; 2) Semi-supervised learning usually utilizes unlabeled data by giving them pseudo labels according to the labeled data, while self-supervised learning gives unlabeled data labels by human knowledge, which does not depend on the labeled data.

### 3. Notations and Problem Formulation

Denote  $\mathbf{x} \in \mathcal{X} \subset \mathbb{R}^{d_x}$ ,  $\mathbf{z} \in \mathcal{Z} \subset \mathbb{R}^{d_z}$ ,  $\mathbf{y} \in \mathcal{Y} \subset \mathbb{R}^{d_y}$  as the input variable, pretext label, and downstream label, respectively, where  $d_x \geq d_z \geq d_y$ . We use bold fonts to denote random variables in this paper. For simplicity, we assume that  $\mathbb{E}[\mathbf{x}] = \mathbb{E}[\mathbf{y}] = \mathbb{E}[\mathbf{z}] = 0$ . Unless stated otherwise, we consider the  $L_2$ -loss function in the following content. When subscript is omitted, notation  $\|\cdot\|$  stands for  $L_2$ -norm or Frobenius norm for vectors and matrices. With the above assumptions, the best representation  $\psi^*: \mathbb{R}^{d_x} \rightarrow \mathbb{R}^{d_z}$  learned from pretext task can be formulated by

$$\psi^* \triangleq \arg \min_{\psi \in \mathcal{H}} \mathbb{E} \|\psi(\mathbf{x}) - \mathbf{z}\|^2,$$

where  $\mathcal{H}$  is the candidate function class (e.g., the class of multi-layer neural network functions). For the downstream task, we directly use the learned representation  $\psi^*(\mathbf{x})$  followed by a linear layer  $W \in \mathbb{R}^{d_y \times d_z}$  to predict downstream label  $\mathbf{y}$ . The best  $W$  minimizes the downstream loss, i.e.,

$$W^* \triangleq \arg \min_{W \in \mathbb{R}^{d_y \times d_z}} \mathbb{E} \|W\psi^*(\mathbf{x}) - \mathbf{y}\|^2.$$

In this way, we have the best predictor  $g(\cdot) = W^*\psi^*(\cdot)$  for the downstream task.

In practice, we can only get finite samples both for pretext and downstream tasks. We assume that we have  $n_1$  pretext task samples  $(x_{pre}^1, z_{pre}^1), \dots, (x_{pre}^{n_1}, z_{pre}^{n_1})$  and  $n_2$  downstream task samples  $(x_{down}^1, y_{down}^1), \dots, (x_{down}^{n_2}, y_{down}^{n_2})$ , which are i.i.d. drawn from the joint distribution of  $(\mathbf{x}, \mathbf{z})$  and  $(\mathbf{x}, \mathbf{y})$ , respectively. For convenience, we express these samples as matrix pairs  $(X_{pre}, Z_{pre}) \in \mathbb{R}^{d_x \times n_1} \times \mathbb{R}^{d_z \times n_1}$  and  $(X_{down}, Y_{down}) \in \mathbb{R}^{d_x \times n_2} \times \mathbb{R}^{d_y \times n_2}$ . The empirical representation is trained by

$$\hat{\psi} = \arg \min_{\psi \in \mathcal{H}} \frac{1}{n_1} \|\psi(X_{pre}) - Z_{pre}\|^2,$$

where  $\psi(X_{pre})$  denotes  $[\psi(x_{pre}^1), \dots, \psi(x_{pre}^{n_1})] \in \mathbb{R}^{d_z \times n_1}$ . Then the empirical linear layer  $\hat{W}$  is learned by

$$\hat{W} = \arg \min_{W \in \mathbb{R}^{d_y \times d_z}} \frac{1}{n_2} \|W\hat{\psi}(X_{down}) - Y_{down}\|^2,$$

---

#### Algorithm 1 Pretext-Based Self-Supervised Learning

---

**Input:** Pretext task data  $(X_{pre}, Z_{pre})$  and downstream task data  $(X_{down}, Y_{down})$ .

- 1: Train representation  $\hat{\psi}(\cdot)$  via pretext task using data  $(X_{pre}, Z_{pre})$ ;
  - 2: Train linear layer  $\hat{W}$  for downstream task using data  $(\hat{\psi}(X_{down}), Y_{down})$ ;
  - 3: **return** predictor  $\hat{g}(\cdot) = \hat{W}\hat{\psi}(\cdot)$ .
- 

---

#### Algorithm 2 Modified Self-Supervised Learning

---

**Input:** Pretext task data  $(X_{pre}, Z_{pre})$  and downstream task data  $(X_{down}, Y_{down})$ .

- 1: Split downstream data  $(X_{down}, Y_{down})$  into two part:  $(X_{down}^{(1)}, Y_{down}^{(1)})$  and  $(X_{down}^{(2)}, Y_{down}^{(2)})$ ;
  - 2: Train coarse representation extractor  $f$  using data  $(X_{pre}, Z_{pre})$  and  $(X_{down}^{(1)}, Y_{down}^{(1)})$ ;
  - 3: Train linear layer  $\hat{W}_1$  using data  $(f(X_{pre}), Z_{pre})$  and obtain the representation  $\hat{\psi}(\cdot) = \hat{W}_1 f(\cdot)$ ;
  - 4: Train linear layer  $\hat{W}_2$  for downstream task using data  $(\hat{W}_1 f(X_{down}^{(2)}), Y_{down}^{(2)})$ ;
  - 5: **return** predictor  $\hat{g}(\cdot) = \hat{W}_2 \hat{W}_1 f(\cdot)$ .
- 

and the final predictor  $\hat{g}(\cdot) = \hat{W}\hat{\psi}(\cdot)$ . We summarize the above procedure as Algorithm 1.

Lee et al. (2020) point out that under the conditional independence  $\mathbf{x} \perp \mathbf{z} \mid \mathbf{y}$ , self-supervised learning can largely reduce the downstream sample complexity. For instance, when  $\mathbf{x}, \mathbf{z}, \mathbf{y}$  are jointly Gaussian, the downstream sample complexity can be reduced from  $\tilde{O}(d_x)^2$  to  $\tilde{O}(d_y)^3$ . However, when the Conditional Independence (CI) does not hold, the downstream sample complexity gets worse to  $\tilde{O}(d_z)$ . This is because there is some redundant information (irrelevant to the downstream) involved in the representation during the representation learning via the pretext task. Then for the downstream task, the mapping from the redundant features to downstream labels will also be learned, which causes a larger sample complexity. In practice, the CI condition rarely holds, and thus self-supervised learning is hard to realize its full potential. Lee et al. (2020) suggest to apply PCA on  $\hat{\psi}$  to predict downstream task, such that the dimension of  $\hat{\psi}$  is forced to be  $d_y$ . However, PCA is done disregarding the information of downstream tasks, which may remove the downstream-related information, leading to a constant generalization error.

In this paper, we propose an intuitive and effective idea to make the CI condition hold, that is, we apply a function  $f$

---

<sup>2</sup>The sample complexity of directly supervised learning  $\mathbf{y}$  from  $\mathbf{x}$  is  $\tilde{O}(d_x)$ , where  $d_x$  is supposed to be much larger than  $d_y$ .

<sup>3</sup>This statement ignores the  $\text{tr}(\Sigma_{\mathbf{y}|\mathbf{x}})$  factor.

to the input variable  $\mathbf{x}$  such that  $f(\mathbf{x}) \perp \mathbf{z} \mid \mathbf{y}$ . If such function  $f$  is known, then we can simply use  $(f(\mathbf{x}), \mathbf{z})$  to learn the representation via pretext task. Hence we call  $f(\mathbf{x})$  the coarse representation of  $\mathbf{x}$ . For simplicity, we assume that  $\mathbb{E}[f(\mathbf{x})] = 0$ . Since the CI condition is related to the downstream label, we need to involve part of downstream data to guide the coarse representation learning. Thus, we split the downstream data into two parts: one is for the coarse representation learning, and the other is for the downstream training. Since function  $f$  is allowed to be non-linear, for simplicity, we use a linear layer  $\widehat{W}_1$  following the coarse representation  $f(\mathbf{x})$  as the representation  $\widehat{\psi}(\mathbf{x})$ . We summarize the modified algorithm as Algorithm 2. We remark that step 2 and 3 of Algorithm 2 can be jointly done in a single pretext training step (See Section 4.1).

For better understanding, let us consider a simple case. Suppose  $\mathbf{x}$  is a  $d_x$ -dimensional random vector. Let  $\mathbf{z}$  and  $\mathbf{y}$  consist of the first  $d_z$  and  $d_y$  elements of  $\mathbf{x}$ , respectively, i.e.,  $\mathbf{z} = \mathbf{x}[0 : d_z - 1]$  and  $\mathbf{y} = \mathbf{x}[0 : d_y - 1]$ . When we execute the standard self-supervised learning (Algorithm 1, with linear representation  $\widehat{\psi}$ ), the learned representation in step 1 has rank  $d_z$ , consisting of  $d_y$ -dimensional useful information and  $(d_z - d_y)$ -dimensional redundancy information that we cannot eliminate (See Figure 2). Therefore, we need  $\tilde{O}(d_z)$  samples to train the linear layer  $\widehat{W}$  in the downstream tasks. In comparison, Algorithm 2 uses a function  $f$  (learned by step 2) to eliminate the redundancy information at the beginning. Hopefully,  $f(\mathbf{x})$  will consist of all the useful information and some information of  $\mathbf{x}[d_z : d_x - 1]$ , without any information of  $\mathbf{x}[d_y : d_z - 1]$ . In step 3, the information of  $\mathbf{x}[d_z : d_x - 1]$  is further removed from  $f(\mathbf{x})$ . Thus, the rank of representation  $\widehat{\psi}(\mathbf{x})$  could be reduced to  $d_y$ . In this way, we will only need  $\tilde{O}(d_y)$  samples for the downstream task.

We remark that setting  $f(\mathbf{x}) \perp \mathbf{z} \mid \mathbf{y}$  as the only objective may lead to non-informative coarse representations. For example, if  $f(\mathbf{x})$  is an independent random noise, criterion  $f(\mathbf{x}) \perp \mathbf{z} \mid \mathbf{y}$  naturally holds, but such  $f(\mathbf{x})$  does not carry any information of  $\mathbf{y}$  and thus it cannot be used to predict  $\mathbf{y}$ . Therefore, we need to ensure that applying function  $f$  to input variable  $\mathbf{x}$  will not lose the information for predicting  $\mathbf{y}$ . It can be formulated as

$$\mathbb{E} \left\| \mathbf{y} - W_{\mathbf{y}, f(\mathbf{x})}^* f(\mathbf{x}) \right\|^2 = \min_{f'} \mathbb{E} \left\| \mathbf{y} - W_{\mathbf{y}, f'(\mathbf{x})}^* f'(\mathbf{x}) \right\|^2,$$

where  $W_{\mathbf{y}, f(\mathbf{x})}^* = \arg \min_{W \in \mathbb{R}^{d_y \times d_f}} \mathbb{E} \left\| \mathbf{y} - W f(\mathbf{x}) \right\|^2$  is the best linear predictor of  $\mathbf{y}$  on  $f(\mathbf{x})$  when given  $f$ . Based on the discussions above, we can summarize the two criteria that coarse representation  $f$  needs to satisfy as follows.

$$f(\mathbf{x}) \perp \mathbf{z} \mid \mathbf{y} \iff \Sigma_{f(\mathbf{x}), \mathbf{z} \mid \mathbf{y}} = 0, \quad (\text{C1})$$

$$\mathbb{E} \left\| \mathbf{y} - W_{\mathbf{y}, f(\mathbf{x})}^* f(\mathbf{x}) \right\|^2 = \min_{f'} \mathbb{E} \left\| \mathbf{y} - W_{\mathbf{y}, f'(\mathbf{x})}^* f'(\mathbf{x}) \right\|^2. \quad (\text{C2})$$

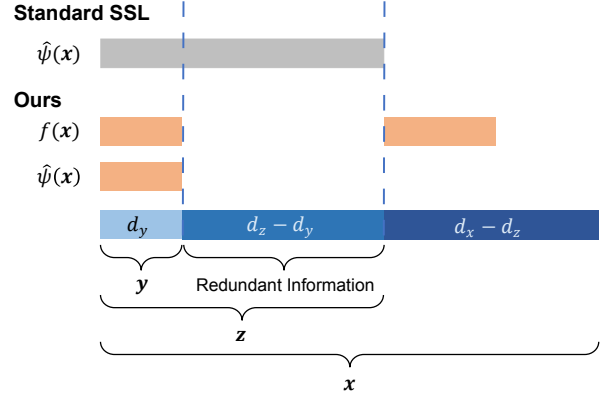


Figure 2. A simple case for understanding Algorithm 2. Here  $\mathbf{y}$  is the prefix of  $\mathbf{z}$ , and  $\mathbf{z}$  is the prefix of  $\mathbf{x}$ . For Algorithm 1, the learned representation  $\widehat{\psi}(\mathbf{x})$  consists of not only the information of  $\mathbf{y}$  but also the redundant information in  $\mathbf{z}$ . For Algorithm 2, the learned coarse representation  $f(\mathbf{x})$  contains the information of  $\mathbf{y}$  and some information of  $\mathbf{x}$ , and does not include any redundant information in  $\mathbf{z}$ . Then the final learned representation  $\widehat{\psi}(\mathbf{x})$  will only contain the information of  $\mathbf{y}$ , hopefully.

The above criteria can also be interpreted as that the coarse representation  $f(\mathbf{x})$  is the ones satisfying the CI condition (C1) among the global optima of  $\mathbb{E} \left\| \mathbf{y} - W_{\mathbf{y}, f(\mathbf{x})}^* f(\mathbf{x}) \right\|^2$ .

After rigorously formulating the criteria that function  $f$  requires to satisfy, the following questions arise: *How can we learn such  $f$ ? How many downstream data are needed to learn  $f$ ?* In Section 4, we will give a provably rational loss for training  $f$ , and in Section 5, we will show the lower bounds of downstream sample size needed.

## 4. Learning Function $f$

In this section, we focus on learning the coarse representation, which satisfies both Criterion (C1) and (C2), illustrated in Section 3. We first propose a carefully designed loss in Section 4.1, and then give proof of its rationality in Section 4.2.

### 4.1. Loss Design

Recall that Criterion (C1) requires variable  $f(\mathbf{x})$  and  $\mathbf{z}$  conditional independent. This indicates that  $\mathbf{z}$  cannot be fitted by  $f(\mathbf{x})$ . It can be measured by the  $L_2$  loss

$$\mathcal{L}_1 = - \mathbb{E}_{\mathbf{x}, \mathbf{z}} \left\| \mathbf{z} - W_{\mathbf{z}, f(\mathbf{x})}^* f(\mathbf{x}) \right\|^2,$$

where  $W^*$  represents the best linear projection from the second subscript variable to the first subscript variable, i.e.,

$$W_{\mathbf{b}, \mathbf{a}}^* = \arg \min_W \mathbb{E} \left\| \mathbf{b} - W \mathbf{a} \right\|^2.$$



If  $f(\mathbf{x})$  can fit  $\mathbf{z}$  well, then  $\mathcal{L}_1$  will be large. Thus, we can minimize  $\mathcal{L}_1$  to force  $\mathbf{z}$  not be fitted by  $f(\mathbf{x})$ .

For Criterion (C2), we need to guarantee that applying  $f$  to  $\mathbf{x}$  does not lose the information for predicting  $\mathbf{y}$ , indicating that  $f(\mathbf{x})$  can still fit  $\mathbf{y}$  well. It can be formulated as

$$\mathcal{L}_2 = \mathbb{E}_{\mathbf{x}, \mathbf{y}} \left\| \mathbf{y} - W_{\mathbf{y}, f(\mathbf{x})}^* f(\mathbf{x}) \right\|^2.$$

The smaller  $\mathcal{L}_2$  indicates the less loss of the information for predicting  $\mathbf{y}$ .

To minimize  $\mathcal{L}_1$  and  $\mathcal{L}_2$  simultaneously, we define the population loss as  $\lambda \mathcal{L}_1 + \mathcal{L}_2$ , where  $\lambda > 0$  is the penalty coefficient, i.e.,

$$\mathcal{L}(f) \triangleq \mathbb{E}_{\mathbf{x}, \mathbf{z}, \mathbf{y}} \left[ \left\| \mathbf{y} - W_{\mathbf{y}, f(\mathbf{x})}^* f(\mathbf{x}) \right\|^2 - \lambda \left\| \mathbf{z} - W_{\mathbf{z}, f(\mathbf{x})}^* f(\mathbf{x}) \right\|^2 \right]. \quad (1)$$

The intuition of such loss design is to extract the useful information for predicting  $\mathbf{y}$ , as well as, remove the redundancy information. In the following Section 4.2, we will give strict proof of its rationality.

The corresponding training loss (for step 2 of Algorithm 2) can be defined as

$$\begin{aligned} \mathcal{L}_{n_1, n'_2}(f) \triangleq & \frac{1}{n'_2} \left\| Y_{down}^{(1)} - \widetilde{W}_2 f(X_{down}^{(1)}) \right\|^2 \\ & - \frac{\lambda}{n_1} \left\| Z_{pre} - \widetilde{W}_1 f(X_{pre}) \right\|^2, \end{aligned} \quad (2)$$

where  $\widetilde{W}_1, \widetilde{W}_2$  represent respectively the best empirical linear predictors for the pretext data and downstream data, and  $n_1, n'_2$  are respectively the sample sizes of  $(X_{pre}, Z_{pre})$  and  $(X_{down}^{(1)}, Y_{down}^{(1)})$ . It is easy to verify that the training loss is an unbiased estimate of the population loss. Therefore, when the sample size goes to infinity, the training loss  $\mathcal{L}_{n_1, n'_2}(f)$  converges to the population loss  $\mathcal{L}(f)$  almost surely, according to the law of large numbers. We remark that  $\widetilde{W}_1$  in Equation 2 can be directly used as  $\widehat{W}_1$  in step 3 of Algorithm 2, thus step 2 and 3 of Algorithm 2 can be done in a single step.

## 4.2. Rationality of Loss

The loss function (Equation (1)) consists of two terms, and we want to maximize the term  $-\mathcal{L}_1$  for predicting  $\mathbf{z}$  while minimizing the term  $\mathcal{L}_2$  for predicting  $\mathbf{y}$ . Since  $\mathbf{z}$  contains the information of  $\mathbf{y}$ , one may wonder if maximizing  $-\mathcal{L}_1$  will make  $\mathcal{L}_2$  minimization failure, leading to the failure to meet Criterion (C1) and (C2). In fact, we provide the following Theorem 1 to show the rationality of loss  $\mathcal{L}(f)$ , which states that under some mild assumptions, with a small penalty coefficient  $\lambda$ , the function  $f$  which minimizes the population loss  $\mathcal{L}(f)$ , satisfies Criterion (C1) and (C2).

Since self-supervised learning uses a simple linear layer for the downstream task, to enable the theoretical analysis, unless stated otherwise, we assume a linear relationship between  $\mathbf{y}$  and  $\mathbf{z}$  in the following content, i.e.,  $\mathbf{y} = B\mathbf{z}$  where  $B$  is an unknown matrix.

**Theorem 1** (Rationality of Loss). *Assume that there exists a ground truth  $f^*$  satisfying both Criterion (C1) and (C2), and matrix  $\mathbb{E}[f(\mathbf{x})f^\top(\mathbf{x})]$  is nonsingular. Let the singular values of matrix  $B$  be  $\sigma_1, \dots, \sigma_{d_y}$ . If  $\sigma_1 = \dots = \sigma_{d_y} = \sigma$ , the penalty coefficient  $\lambda < \sigma^2$  and  $\mathbb{E}[\mathbf{z}\mathbf{z}^\top] = I$ , then for each  $f$  minimizing the loss  $\mathcal{L}(f)$  in Equation 1 satisfies both Criterion (C1) and (C2).*

**Remark 4.1.** When the sample size goes to infinity, the training loss converges to the population loss. Thus, the above theorem can also be regarded as the result of training with infinite samples.

**Remark 4.2.** We assume that  $\mathbb{E}[\mathbf{z}\mathbf{z}^\top] = I$  and  $\sigma_i = \sigma$  for each  $i \in [d_y]$  in the above theorem. They can be achieved by simply normalizing  $\mathbf{y}$  and  $\mathbf{z}$ .

Before giving the proof of Theorem 1, we first take a closer look at the loss function  $\mathcal{L}(f) = \mathcal{L}_2 + \lambda \mathcal{L}_1$ . We can rewrite it as  $\mathcal{L}(f) = (1 - \frac{\lambda}{\sigma^2})\mathcal{L}_2 + \lambda(\mathcal{L}_1 + \frac{1}{\sigma^2}\mathcal{L}_2)$ . The first term is  $\mathcal{L}_2$  multiplied by a coefficient, which still captures the information of  $\mathbf{y}$  as  $\mathcal{L}_2$  does. The second term intuitively captures the redundant information of  $\mathbf{z}$  (see supplementary materials for details), and the following lemma builds a connection between the second term and the conditional independence.

**Lemma 4.1.** *Under the assumptions of Theorem 1,  $\mathcal{L}_1 + \frac{1}{\sigma^2}\mathcal{L}_2$  is minimized if and only if the conditional independence criterion (C1) holds, i.e.,  $f(\mathbf{x}) \perp \mathbf{z} \mid \mathbf{y}$ .*

The advantage of decomposing  $\mathcal{L}(f)$  into  $(1 - \frac{\lambda}{\sigma^2})\mathcal{L}_2$  and  $\mathcal{L}_1 + \frac{1}{\sigma^2}\mathcal{L}_2$  is that these two terms can be optimized individually. In other words, when  $\mathcal{L}(f)$  is minimized, the above two terms are also minimized. We state it formally as the following lemma.

**Lemma 4.2.** *Under the assumptions of Theorem 1, for each function  $f$ , if  $\mathcal{L}(f)$  is minimized, then  $(1 - \frac{\lambda}{\sigma^2})\mathcal{L}_2$  and  $\mathcal{L}_1 + \frac{1}{\sigma^2}\mathcal{L}_2$  are both minimized.*

The full proofs of the above lemmas can be found in the supplementary materials. Equipped with the above two lemmas, we are now ready to prove Theorem 1.

*Proof of Theorem 1.* According to Lemma 4.2, for each function  $f$  that minimizes the loss, it also minimizes  $(1 - \frac{\lambda}{\sigma^2})\mathcal{L}_2$  and  $\mathcal{L}_1 + \frac{1}{\sigma^2}\mathcal{L}_2$  at the same time. For the first term, since  $1 - \frac{\lambda}{\sigma^2}$  is a positive coefficient,  $\mathcal{L}_2$  is minimized. Therefore, Criterion (C2) holds. Since the second term is minimized, Criterion (C1) holds by Lemma 4.1. Thus, such  $f$  satisfies both Criterion (C1) and (C2).  $\square$

## 5. Theoretical Bounds

Theorem 1 has shown that with unlimited data, we can learn  $f$  by minimizing the empirical loss, such that  $f$  satisfies both Criterion (C1) and (C2). However, there are only a finite number of training data in practice. In this section, we theoretically study the sample size needed for learning coarse representation  $f(\mathbf{x})$ . Compared with labeled data, unlabeled data are usually much easier to access. Thus, in the following contents, we consider the situation with unlimited unlabeled pretext data and study the sample size of labeled downstream data.

We will first provide a model-free lower bound in Section 5.1 as a warm-up. Furthermore, we take the structure of function class into account and give a model-dependent lower bound, which is tighter.

### 5.1. Warm-Up: Model-Free Results

When unlimited unlabeled pretext data are available, we can replace the second term associated with pretext data in the training loss (Equation (2)) by its expectation, since the loss term converges to its expectation when the sample size of unlabeled pretext data goes to infinity. Namely, we rewrite Equation (2) as

$$\mathcal{L}_{\infty, n'_2} \triangleq \frac{1}{n'_2} \left\| Y_{\text{down}}^{(1)} - \widetilde{W}_2 f(X_{\text{down}}^{(1)}) \right\|^2 - \lambda \mathbb{E}_{\mathbf{x}, \mathbf{z}} \left\| \mathbf{z} - W_{\mathbf{z}, f(\mathbf{x})}^* f(\mathbf{x}) \right\|^2. \quad (3)$$

Based on the above loss, we can now state our model-free lower bound of downstream sample size  $n'_2$  as the following theorem.

**Theorem 2 (Model-Free Lower Bound).** *Assume that there exist function  $f_1$  and  $f_2$  in the function class such that covariance  $\text{Cov}[f_1(\mathbf{x}), \mathbf{y}] \neq 0$  and  $\text{Cov}[f_2(\mathbf{x}), \mathbf{z}] = 0$ , and matrix  $\mathbb{E}[f(\mathbf{x})f^\top(\mathbf{x})]$  is nonsingular. If we train the model with  $n'_2 = o(d_f)$  labeled data and infinite unlabeled data, where  $d_f$  is the dimension of  $f(\mathbf{x})$ , then any function  $f$  that minimizes the loss  $\mathcal{L}_{\infty, n'_2}$  in Equation (3) cannot meet Criterion (C1) and (C2) simultaneously.*

Theorem 2 indicates that under the mild assumption, to train a function  $f$  satisfying both Criterion (C1) and (C2), we need at least  $\Omega(d_f)$  labeled downstream samples, although unlimited unlabeled data can be accessed. One can expect the larger downstream sample size with finite unlabeled data. Here  $\text{Cov}[f_1(\mathbf{x}), \mathbf{y}] \neq 0$  assumes that the function class of  $f$  is meaningful such that  $f$  has ability to predict  $\mathbf{y}$ , while  $\text{Cov}[f_2(\mathbf{x}), \mathbf{z}] = 0$  could be easily satisfied when the function class is chosen independently with the pretext task. We give a proof sketch of Theorem 2 as follows and defer the full proof in the supplementary materials.

*Proof Sketch of Theorem 2.* When we have only  $o(d_f)$  samples, the best empirical predictor  $\widetilde{W}_2$  guarantees the first term in the RHS of Equation (3) to be zero. Therefore, minimizing loss  $\mathcal{L}_{\infty, n'_2}$  leads to maximizing the second term of loss. In other words,  $f(\mathbf{x})$  is encouraged to not fit  $\mathbf{z}$ . Since  $\mathbf{z}$  contains all the information of  $\mathbf{y}$ , the learned  $f(\mathbf{x})$  fails to predict  $\mathbf{y}$ , which leads to a contradiction of Criterion (C2).  $\square$

**Remark 5.1.** If the learned  $f(\mathbf{x})$  does not satisfy Criterion (C2), then it contains less information of  $\mathbf{y}$  than the raw data  $\mathbf{x}$ , leading to the failure of self-supervised learning.

**Remark 5.2.** Recall that, with unlimited labeled data and unlimited unlabeled data, the learned  $f$  meets the criteria (Theorem 1 and Remark 4.1), while with limited labeled data and unlimited unlabeled data, the learned  $f(\mathbf{x})$  breaks the criteria (Theorem 2). Therefore, one can conclude that the failure is due to the lack of labeled data.

### Model-Free Results for General Loss

Theorem 2 is derived based on the  $L_2$  loss that we design in Section 4.1. One may concern the generality. In fact, we can extend Theorem 2 based on a more general loss. Since unlabeled data and labeled data come from different datasets, there should be two terms in the loss: one for unlabeled data and the other for labeled data. A rational loss should ensure that  $f(\mathbf{x})$  can fit the labeled downstream data well and does not tend to fit the redundant information in the unlabeled pretext data simultaneously. Therefore, we can define the loss as

$$\begin{aligned} \bar{\mathcal{L}}(f, \lambda; g_2, \rho_2, g_1, \rho_1) &= \mathbb{E}_{\mathbf{x}, \mathbf{y}} g_2(\rho_2(\mathbf{y}, W_{\mathbf{y}, f(\mathbf{x})}^* f(\mathbf{x}))) \\ &\quad - \lambda \mathbb{E}_{\mathbf{x}, \mathbf{z}} g_1(\rho_1(\mathbf{z}, W_{\mathbf{z}, f(\mathbf{x})}^* f(\mathbf{x}))). \end{aligned}$$

Here  $g_1$  and  $g_2$  are strictly increasing functions over  $[0, \infty)$ ,  $\lambda$  is a positive penalty coefficient. Mappings  $\rho_1: \mathbb{R}^{d_z} \times \mathbb{R}^{d_z} \rightarrow \mathbb{R}_{\geq 0}$  and  $\rho_2: \mathbb{R}^{d_y} \times \mathbb{R}^{d_y} \rightarrow \mathbb{R}_{\geq 0}$  are distance metrics, and  $\rho_1$  is consistent with  $\rho_2$  in subspace  $\mathbb{R}^{d_y} \times \mathbb{R}^{d_y}$ , i.e., for all  $a, b \in \mathbb{R}^{d_y}$ ,  $\rho_1((a, \vec{0}_{d_z-d_y}), (b, \vec{0}_{d_z-d_y})) = \rho_2(a, b)$  holds. We abuse  $W_{\mathbf{y}, f(\mathbf{x})}^*$  and  $W_{\mathbf{z}, f(\mathbf{x})}^*$  as the best linear predictors which minimize the corresponding loss  $\mathbb{E} g_2(\rho_2(\mathbf{y}, W f(\mathbf{x})))$  and  $\mathbb{E} g_1(\rho_1(\mathbf{z}, W f(\mathbf{x})))$ , respectively. We assume there exists a function  $f$  such that  $f(\mathbf{x}) \perp \mathbf{z}$  holds, and  $f(\mathbf{x}) \perp \mathbf{z}$  holds if and only if  $\mathbb{E} g_1(\rho_1(\mathbf{z}, W_{\mathbf{z}, f(\mathbf{x})}^* f(\mathbf{x})))$  reaches the maximum. Namely,  $f(\mathbf{x})$  can not predict  $\mathbf{z}$  if and only if the loss term  $\mathbb{E} g_1(\rho_1(\mathbf{z}, W_{\mathbf{z}, f(\mathbf{x})}^* f(\mathbf{x})))$  is maximized.

It is not hard to verify that the loss  $\mathcal{L}(f)$  defined in Section 4.1 is a special case of the above general loss, by setting  $\rho_2, \rho_1$  to be Euclidean distances and  $g_2, g_1$  the square functions.

Based on the general loss  $\bar{\mathcal{L}}$  and mild assumptions, we still have the lower bound  $\Omega(d_f)$  of downstream sample size. Denote  $\bar{\mathbf{y}} = (\mathbf{y}, \vec{0}_{d_z-d_y})$  as the augmented vector of  $\mathbf{y}$  from dimension in  $\mathbb{R}^{d_z}$ , and  $\hat{f}$  as the best linear predictor that minimizes the training loss with infinity unlabeled data

$$\begin{aligned} \bar{\mathcal{L}}_{\infty, n'_2}(f) &= \frac{1}{n'_2} \sum_{i \in [n'_2]} g_2(\rho_2(y_{down}^i, \widetilde{W}_1 f(x_{down}^i))) \\ &\quad - \lambda \mathbb{E}_{\mathbf{z}, \mathbf{z}} g_1(\rho_1(\mathbf{z}, W_{\mathbf{z}, f(\mathbf{x})}^* f(\mathbf{x}))), \end{aligned}$$

where  $\widetilde{W}_1$  is the best empirical linear predictor for the downstream data.

**Theorem 3** (Lower Bound for General Loss). *Let  $\hat{f}$  be a global minimum of loss  $\bar{\mathcal{L}}_{\infty, n'_2}(\hat{f})$ . Assume that  $g_2 g_1^{-1}$  is convex and non-decreasing, and there exists a linear transformation  $V \in \mathbb{R}^{d_z \times d_z}$  such that*

$$\begin{aligned} \mathbb{E}[\rho_1(\bar{\mathbf{y}}, V\mathbf{z})] &< \frac{1}{M} g_2 g_1^{-1} \max_f \mathbb{E}[g_1 \rho_1(W_{V\mathbf{z}, f(\mathbf{x})}^* f(\mathbf{x}), V\mathbf{z})] \\ &\quad - \frac{1}{M} \min_f \mathbb{E}[g_2 \rho_2(W_{\mathbf{y}, f(\mathbf{x})}^* f(\mathbf{x}), \mathbf{y})], \end{aligned}$$

where  $M$  is the upper bound of  $g_2$ 's derivative function over  $[0, \sup_{\mathbf{x}, \mathbf{y}} \rho_2(W_{\mathbf{y}, \hat{f}(\mathbf{x})}^* \hat{f}(\mathbf{x}), \mathbf{y}) + \sup_{\mathbf{y}, \mathbf{z}} \rho_1(\bar{\mathbf{y}}, V\mathbf{z})]$ , i.e.,

$$g_2'(x) \leq M, \forall x \in [0, \sup_{\mathbf{x}, \mathbf{y}} \rho_2(W_{\mathbf{y}, \hat{f}(\mathbf{x})}^* \hat{f}(\mathbf{x}), \mathbf{y}) + \sup_{\mathbf{y}, \mathbf{z}} \rho_1(\bar{\mathbf{y}}, V\mathbf{z})].$$

If the downstream sample size  $n'_2 = o(d_f)$ , then any function  $\hat{f}$  that minimizes the loss  $\bar{\mathcal{L}}_{\infty, n'_2}(\hat{f})$  cannot meet Criterion (C1) and (C2) simultaneously.

We leave the proof of Theorem 3 in the supplementary materials. Here we give some remarks on the theorem.

**Remark 5.3** (Compare with Theorem 2). Consider  $\mathcal{L}(f)$  defined in Eq. (1) as a special case. Notice that  $\mathbf{y} = B\mathbf{z}$ . If we let  $V = [B; 0_{(d_z-d_y) \times d_z}]$ , then  $\mathbb{E}[\rho_1(\bar{\mathbf{y}}, V\mathbf{z})] = 0$ . When  $\mathbb{E}[f(\mathbf{x}) f^\top(\mathbf{x})]$  is nonsingular and there exists  $f_1$  such that  $\text{Cov}[f_1(\mathbf{x}), \mathbf{y}] \neq 0$ ,  $\max_f \mathbb{E} \|W_{V\mathbf{z}, f(\mathbf{x})}^* f(\mathbf{x}) - V\mathbf{z}\|^2 - \min_f \mathbb{E} \|W_{\mathbf{y}, f(\mathbf{x})}^* f(\mathbf{x}) - \mathbf{y}\|^2 > 0 = \mathbb{E}[\rho_1(\bar{\mathbf{y}}, V\mathbf{z})]$ . The condition in Theorem 3 holds, and thus Theorem 2 can be derived from Theorem 3 under the condition that there exists  $f_2$  such that  $f_2(\mathbf{x}) \perp \mathbf{z}$ , which is slightly stronger than the condition  $\text{Cov}[f_2(\mathbf{x}), \mathbf{z}] = 0$  in Theorem 2.

**Remark 5.4.** At first glance, one might think that seeing labeled downstream data ahead in the self-supervised representation learning phase would always boost the downstream performance. However, Theorem 3 tell us that in most cases, taking *limited* downstream data into the representation learning phase will hurt the final performance instead. For example, for the few-shot downstream task, self-supervised pre-training and then fine-tuning is better than joint training with both labeled and unlabeled data.

## 5.2. Model-Dependent Results

The results stated in Section 5.1 hold without any assumption of model structures (i.e., the function class of  $f$ ). One may wonder whether we can take the model structures into account to get the fine-grained result. The answer is yes. Before providing the main theorem, we introduce a measure of *model capacity* at first, which is defined as follows.

**Definition 1** (Model Capacity). *Define the model capacity  $\mathcal{M}(\mathcal{F}, \mathcal{L})$  of function class  $\mathcal{F}$  with respect to loss function  $\mathcal{L}$  as*

$$\mathcal{M}(\mathcal{F}, \mathcal{L}) = \sup \left\{ n : \forall \mathcal{D}, \inf_{f \in \mathcal{F}} \sup_{(X, Y) \in \mathcal{D}^n} \mathcal{L}(f(X), Y) = 0 \right\},$$

where  $\mathcal{D}$  is the data distribution, and  $X, Y$  are data vectors, each of which consists of  $n$  samples.

Intuitively, model capacity measures how well the function class fits the noise under finite samples. It is similar to the VC dimension under regression settings. For instance, consider the case when  $\mathcal{F}$  is the linear function class mapping from  $\mathbb{R}^{d_x}$  to  $\mathbb{R}^{d_y}$  and  $\mathcal{L}$  is the  $L_2$  loss, then  $\mathcal{M}(\mathcal{F}, \mathcal{L}) = d_x$ . That is because linear models can interpolate the dataset when the sample size is no greater than the dimension of the covariate, namely,  $n \leq d_x$ , leading to a zero training loss.

**Theorem 4** (Model-Dependent Lower Bound). *Assume that the function class of the coarse representation can be decomposed as  $\mathcal{F} = \mathcal{F}_1 \times \mathcal{F}_2$ , where there exists a function  $f_0 \in \mathcal{F}$  such that  $\text{Cov}[f_0(\mathbf{x}), \mathbf{y}] \neq 0$  and a function  $f_2 \in \mathcal{F}_2$  such that  $f_2(\mathbf{x}) \perp \mathbf{z}$ . And assume matrix  $\mathbb{E}[f(\mathbf{x}) f^\top(\mathbf{x})]$  is nonsingular for all  $f \in \mathcal{F}$ . If we train the model with  $n'_2 = o(\mathcal{M}(\mathcal{F}_1, \mathcal{L}))$  labeled samples and infinite unlabeled data, then any function  $f$  that minimizes the loss  $\mathcal{L}_{\infty, n'_2}$  in Equation (3) cannot meet Criterion (C1) and (C2) simultaneously.*

The lower bound  $\Omega(\mathcal{M}(\mathcal{F}_1, \mathcal{L}))$  in Theorem 4 is closely related to the model capacity. We emphasize that when  $\mathcal{F}_1$  is the class of neural networks, a large model capacity is expected (even infinite large since overparameterized neural networks can fit any noise model). Thus, given a fixed number of labeled downstream data, more complex neural networks perform worse.

*Proof Sketch of Theorem 4.* On the one hand, notice that  $f_2(\mathbf{x}) \perp \mathbf{z}$ . It leads to  $\bar{f}(f_2(\mathbf{x})) \perp \mathbf{z}$  for any function  $\bar{f} \in \mathcal{F}_1$ . On the other hand, by the definition of model capacity, when  $n'_2 = o(\mathcal{M}(\mathcal{F}, \mathcal{L}))$ , there exists  $f_1 \in \mathcal{F}_1$  such that  $f_1(f_2(x_{down}^i)) = y_{down}^i$  for all downstream samples  $(x_{down}^i, y_{down}^i)$  used in the training process. Therefore, since the trained classifier  $\hat{f}$  minimize the training loss, it should at least minimize  $\mathcal{L}_1$ , which eliminates the information of  $\mathbf{z}$ . Furthermore, since  $\mathbf{z}$  contains all the information

of  $\mathbf{y}$ , the classifier  $\hat{f}$  loses the ability to predict  $\mathbf{y}$ . Therefore, any trained classifier  $\hat{f}$  that minimize the training loss cannot meet Criterion (C1) and (C2) simultaneously.  $\square$

## 6. Experiments

In this section, we conduct several numerical experiments on simulated data to verify some key arguments: (1) With enough downstream samples, the proposed algorithm works well as demonstrated in Theorem 1, meaning a small error on the downstream tasks; (2) A large penalty coefficient  $\lambda$  harms the performance of the proposed algorithm, as demonstrated in Theorem 1; (3) A large dimension  $d_f$  harms the performance of the proposed algorithm, as demonstrated in Theorem 2 and Theorem 4.

### 6.1. Setup

**Datasets and Environment.** The dataset is simulated as follows. Firstly, we randomly generate  $n_1$  pretext samples  $\{(x_{pre}^i, z_{pre}^i)\}_{i \in [n_1]}$ , where  $x_{pre}^i \in \mathbb{R}^{d_x}$  is generated from Gaussian distribution  $\mathcal{N}(0, I_{d_x})$ , and  $z_{pre}^i \in \mathbb{R}^{d_z}$  is a perturbation of the first  $d_z$  elements of  $x_{pre}^i$ , namely,  $z_{pre}^i = x_{pre}^i[0 : d_z] + 0.01 \cdot \varepsilon_1^i$ , where  $\varepsilon_1^i \sim \mathcal{N}(0, I_{d_z})$ . Similarly, we generate  $n_2$  downstream samples  $\{(x_{down}^i, y_{down}^i)\}_{i \in [n_2]}$ , where  $x_{down}^i \in \mathbb{R}^{d_x}$  is generated from Gaussian distribution  $\mathcal{N}(0, I_{d_x})$  and  $y_{down}^i \in \mathbb{R}^{d_y}$  is a perturbation of the first  $d_y$  elements of  $x_{down}^i$ , i.e.,  $y_{down}^i = x_{down}^i[0 : d_y] + 0.01 \cdot \varepsilon_2^i$ , where  $\varepsilon_2^i \sim \mathcal{N}(0, I_{d_y})$ . We generate  $n_t$  test samples under the same procedure.

In each run of our proposed algorithm, we split one-half of the downstream samples to train the coarse representation and leave the other half data for downstream task training. We run each experiment 5 times repeatedly and calculate its mean and standard deviation. We plot their 95% confidence bands in all the figures (light blue and light yellow).

**Algorithm and Metrics.** We run our newly proposed algorithm (labeled as Ours), which first train coarse representation and then do pretext and downstream tasks. We denote  $d_f$  as the coarse representation dimension and  $\lambda$  as the penalty coefficient in the loss (See Equation 1). We also run the standard self-supervised learning of Algorithm 1 (labeled as SSL). In terms of metrics, we calculate the MSE on test downstream samples as the model performance. Intuitively, MSE is small when the features are learned well.

### 6.2. Analysis

We plot the results in Figure 3 and defer the specific statistics in the supplementary materials due to space limitations.

**The rationality of the loss  $\mathcal{L}(f)$ .** Set  $d_x = 100$ ,  $d_z = 80$ ,  $d_y = 5$ ,  $d_f = 5$ ,  $n_1 = 20000$  and  $\lambda = 0.1$ . We plot the MSE as  $n_2$  varies from 30 to 140 in Figure 3(a). With large

$n_2$ , the downstream tasks benefit from small MSE, showing that the algorithm indeed finds the proper coarse representations. Theorem 1 demonstrates this phenomenon, which claims that with sufficient downstream task samples, the proposed algorithm finds the proper coarse representations.

We further remark that in the case  $n_2 = 80$ , the SSL suffers from a large MSE. It is related to the “double decent” phenomenon in the downstream tasks, and MSE reaches the maximum when  $n_2 = d_z$ . We finally remark that when  $n_2 \in [40, 80]$ , the proposed algorithm outperforms SSL in terms of MSE.

**Large coefficient  $\lambda$  harms the performance.** Set  $d_x = 100$ ,  $d_z = 80$ ,  $d_y = 5$ ,  $d_f = 5$ ,  $n_1 = 20000$  and  $n_2 = 30$ . We test the case  $\lambda$  varies from 0.1 to 1.5, as plotted in Figure 3(b). Figure 3(b) illustrates that when  $\lambda$  is large, the algorithm suffers from unsatisfying performances. Theorem 1 demonstrates the phenomenon. Intuitively, that is because  $z$  contains the information of  $\mathbf{y}$ . With a large coefficient  $\lambda$ , the trained classifier tends to eliminate the information of  $y$ , leading to a lousy model performance (MSE).

**Large dimension  $d_f$  harms the performance.** Set  $d_x = 100$ ,  $d_z = 80$ ,  $d_y = 5$ ,  $n_1 = 20000$ ,  $n_2 = 30$ , and  $\lambda = 0.1$ . We test the case  $d_f$  varies from 1 to 12, and plot them in Figure 3(c). Firstly, notice that MSE reaches the minimum when  $d_f = d_y$ , that is because we force the coarse representation to have no redundancy information by limiting its dimension. Secondly, notice that the figure shows a Basin-type phenomenon and reaches the minimum when  $d_f = d_y$ . On the one hand, when  $d_f < d_y$ , the coarse representation is forced to lose the information of  $y$ , leading to the underfitting phenomenon. On the other hand, when  $d_f > d_y$ , the performance of coarse representation is limited by the downstream sample numbers, as illustrated in Theorem 2 and Theorem 4.

## 7. Conclusion and Future Work

In this work, we explore the idea of applying a learnable function  $f$  to the input to make the conditional independence  $f(\mathbf{x}) \perp \mathbf{z} \mid \mathbf{y}$  hold. We rigorously formulate the criteria that function  $f$  needs to satisfy. We design an ingenious loss function and prove that a function  $f$  minimizing the proposed loss satisfies the above criteria. We then theoretically study the number of labeled data required, by giving both model-free and model-dependent lower bounds of downstream sample size for learning the function  $f$ . We show that taking limited downstream data will hurt the performance of self-supervised learning. Numerical experiments validate our theoretical results.

The experiments have shown that under some proper hyperparameters, taking downstream data to the representation learning phase can significantly boost the pretext-based



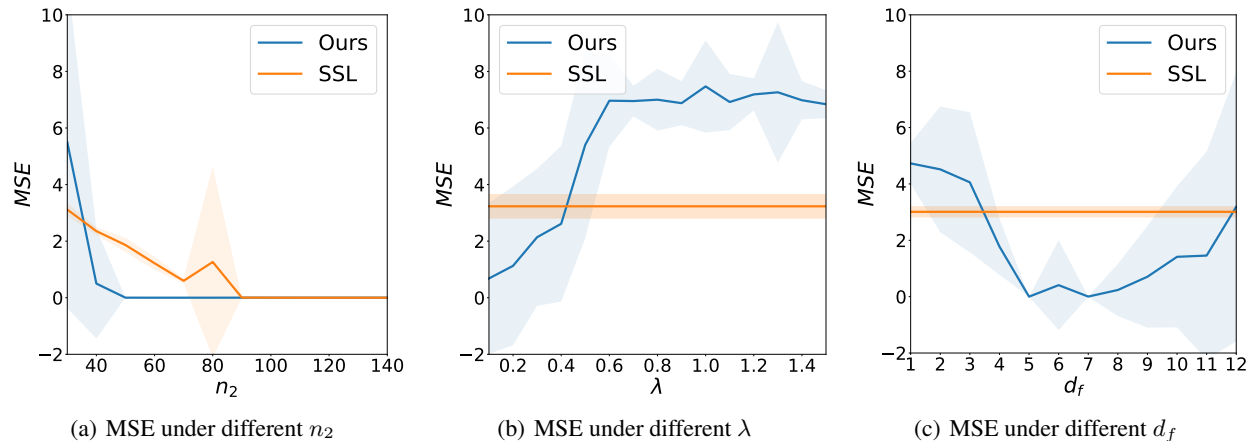


Figure 3. **Algorithm performance under different hyperparameters.** (a) Our algorithm learns better with sufficient downstream samples (large  $n_2$ ). MSE decreases to zero as downstream samples increase. (b) A large penalty forces the coarse representation to abandon the  $y$  information, leading to a large MSE. (c) When  $d_f$  is small, the model underfits; when  $d_f$  is large, the model suffers from a limited number of downstream samples.

self-supervised learning. Thus, one possible future work is to analyze the upper bound of downstream sample size for learning the function  $f$ . Another possible direction could be exploring the idea in this paper for other self-supervised learning approaches, i.e., contrastive learning based approach.

## References

- Bachman, P., Hjelm, R. D., and Buchwalter, W. Learning representations by maximizing mutual information across views. In *Advances in Neural Information Processing Systems*, pp. 15535–15545, 2019.
- Bansal, Y., Kaplun, G., and Barak, B. For self-supervised learning, rationality implies generalization, provably. *arXiv preprint arXiv:2010.08508*, 2020.
- Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020.
- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- Donahue, J. and Simonyan, K. Large scale adversarial representation learning. In *Advances in Neural Information Processing Systems*, pp. 10542–10552, 2019.
- Donahue, J., Krähenbühl, P., and Darrell, T. Adversarial feature learning. In *International Conference on Learning Representations (ICLR)*, 2017.
- Dumoulin, V., Belghazi, I., Poole, B., Mastropietro, O., Lamb, A., Arjovsky, M., and Courville, A. Adversarially learned inference. In *International Conference on Learning Representations (ICLR)*, 2017.
- Gidaris, S., Singh, P., and Komodakis, N. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*, 2018.
- He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.
- He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9729–9738, 2020.
- Hjelm, R. D., Fedorov, A., Lavoie-Marchildon, S., Grewal, K., Bachman, P., Trischler, A., and Bengio, Y. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*, 2018.
- Lee, D.-H. et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, 2013.
- Lee, J. D., Lei, Q., Saunshi, N., and Zhuo, J. Predicting what you already know helps: Provable self-supervised learning. *arXiv preprint arXiv:2008.01064*, 2020.

- Noroozi, M. and Favaro, P. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European Conference on Computer Vision*, pp. 69–84. Springer, 2016.
- Oord, A. v. d., Li, Y., and Vinyals, O. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.
- Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., and Efros, A. A. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2536–2544, 2016.
- Peng, Z., Dong, Y., Luo, M., Wu, X.-M., and Zheng, Q. Self-supervised graph representation learning via global context prediction. *arXiv preprint arXiv:2003.01604*, 2020.
- Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. Improving language understanding by generative pre-training, 2018.
- Saunshi, N., Plevrakis, O., Arora, S., Khodak, M., and Khandeparkar, H. A theoretical analysis of contrastive unsupervised representation learning. In *International Conference on Machine Learning*, pp. 5628–5637. PMLR, 2019.
- Tian, Y., Krishnan, D., and Isola, P. Contrastive multiview coding. *arXiv preprint arXiv:1906.05849*, 2019.
- Tosh, C., Krishnamurthy, A., and Hsu, D. Contrastive learning, multi-view redundancy, and linear models. *arXiv preprint arXiv:2008.10150*, 2020.
- Wei, C., Shen, K., Chen, Y., and Ma, T. Theoretical analysis of self-training with deep networks on unlabeled data. *arXiv preprint arXiv:2010.03622*, 2020.
- Yamaguchi, S., Kanai, S., Shioda, T., and Takeda, S. Multiple pretext-task for self-supervised learning via mixing multiple image transformations. *arXiv preprint arXiv:1912.11603*, 2019.
- Zhang, R., Isola, P., and Efros, A. A. Colorful image colorization. In *European conference on computer vision*, pp. 649–666. Springer, 2016.
- Zoph, B., Ghiasi, G., Lin, T.-Y., Cui, Y., Liu, H., Cubuk, E. D., and Le, Q. Rethinking pre-training and self-training. *Advances in Neural Information Processing Systems*, 33, 2020.

## Supplementary Materials

We give all the proofs of lemmas and theorems here organized by theorems, i.e., Section A for Theorem 1 and related lemmas, Section B for Theorem 2, Section C for Theorem 3, and Section D for Theorem 4.

### A. Proof of Theorem 1

In this section, we first prove Lemma 4.1 and Lemma 4.2 in Section A.1 and Section A.2, respectively, and then prove the Theorem 1 in Section A.3.

#### A.1. Proof of Lemma 4.1

**Lemma 4.1.** *Under the assumptions of Theorem 1,  $\mathcal{L}_1 + \frac{1}{\sigma^2} \mathcal{L}_2$  is minimized if and only if the conditional independence criterion (C1) holds, i.e.,  $f(\mathbf{x}) \perp \mathbf{z} \mid \mathbf{y}$ .*

*Proof.* Criterion (C1) is equivalent to

$$\Sigma_{f(\mathbf{x}), \mathbf{z} \mid \mathbf{y}} = \Sigma_{\mathbf{z} f(\mathbf{x})} - \Sigma_{\mathbf{y} f(\mathbf{x})} \Sigma_{\mathbf{y} \mathbf{y}}^{-1} \Sigma_{\mathbf{y} \mathbf{z}} = 0.$$

Notice that  $\mathbb{E}[\mathbf{y}] = \mathbb{E}[\mathbf{z}] = \mathbb{E}[f(\mathbf{x})] = 0$ , thus the above equation can be rewritten as

$$\mathbb{E}[f(\mathbf{x}) \mathbf{z}^\top] = \mathbb{E}[f(\mathbf{x}) \mathbf{y}^\top] (\mathbb{E}[\mathbf{y} \mathbf{y}^\top])^{-1} \mathbb{E}[\mathbf{y} \mathbf{z}^\top]. \quad (\text{A.1})$$

On the other hand, we express the term  $\mathcal{L}_1 + \frac{1}{\sigma^2} \mathcal{L}_2$  as

$$\begin{aligned} & \mathcal{L}_1 + \frac{1}{\sigma^2} \mathcal{L}_2 \\ &= \frac{1}{\sigma^2} \mathbb{E} \left\| \mathbf{y} - W_{\mathbf{y}, f(\mathbf{x})}^* f(\mathbf{x}) \right\|^2 - \mathbb{E} \left\| \mathbf{z} - W_{\mathbf{z}, f(\mathbf{x})}^* f(\mathbf{x}) \right\|^2 \\ &= \text{tr} \left[ \frac{1}{\sigma^2} \mathbb{E} \left[ \left( \mathbf{y} - W_{\mathbf{y}, f(\mathbf{x})}^* f(\mathbf{x}) \right) \left( \mathbf{y} - W_{\mathbf{y}, f(\mathbf{x})}^* f(\mathbf{x}) \right)^\top \right] - \mathbb{E} \left[ \left( \mathbf{z} - W_{\mathbf{z}, f(\mathbf{x})}^* f(\mathbf{x}) \right) \left( \mathbf{z} - W_{\mathbf{z}, f(\mathbf{x})}^* f(\mathbf{x}) \right)^\top \right] \right] \\ &= \text{tr} \left[ \frac{1}{\sigma^2} \left( \mathbb{E}[\mathbf{y} \mathbf{y}^\top] - \mathbb{E}[\mathbf{y} f^\top(\mathbf{x})] (\mathbb{E}[f(\mathbf{x}) f^\top(\mathbf{x})])^{-1} \mathbb{E}[f(\mathbf{x}) \mathbf{y}^\top] \right) - \left( \mathbb{E}[\mathbf{z} \mathbf{z}^\top] - \mathbb{E}[\mathbf{z} f^\top(\mathbf{x})] (\mathbb{E}[f(\mathbf{x}) f^\top(\mathbf{x})])^{-1} \mathbb{E}[f(\mathbf{x}) \mathbf{z}^\top] \right) \right] \\ &= \text{tr} \left[ \mathbb{E}[\mathbf{z} f^\top(\mathbf{x})] (\mathbb{E}[f(\mathbf{x}) f^\top(\mathbf{x})])^{-1} \mathbb{E}[f(\mathbf{x}) \mathbf{z}^\top] - \frac{1}{\sigma^2} \mathbb{E}[\mathbf{y} f^\top(\mathbf{x})] (\mathbb{E}[f(\mathbf{x}) f^\top(\mathbf{x})])^{-1} \mathbb{E}[f(\mathbf{x}) \mathbf{y}^\top] \right] + \left( \frac{1}{\sigma^2} \mathbb{E}[\mathbf{y}^\top \mathbf{y}] - \mathbb{E}[\mathbf{z}^\top \mathbf{z}] \right) \\ &= \text{tr} \left[ \left( I - \frac{1}{\sigma^2} B^\top B \right) \mathbb{E}[\mathbf{z} f^\top(\mathbf{x})] (\mathbb{E}[f(\mathbf{x}) f^\top(\mathbf{x})])^{-1} \mathbb{E}[f(\mathbf{x}) \mathbf{z}^\top] \right] + \left( \frac{1}{\sigma^2} \mathbb{E}[\mathbf{y}^\top \mathbf{y}] - \mathbb{E}[\mathbf{z}^\top \mathbf{z}] \right), \end{aligned}$$

where the third equation holds because  $W_{\mathbf{y}, f(\mathbf{x})}^* = \mathbb{E}[\mathbf{y} f^\top(\mathbf{x})] (\mathbb{E}[f(\mathbf{x}) f^\top(\mathbf{x})])^{-1}$  and the last equation follows the assumption that  $\mathbf{y} = B\mathbf{z}$ . Notice that the second term  $\frac{1}{\sigma^2} \mathbb{E}[\mathbf{y}^\top \mathbf{y}] - \mathbb{E}[\mathbf{z}^\top \mathbf{z}]$  is unrelated to  $f$ . Therefore, minimizing  $\mathcal{L}_1 + \frac{1}{\sigma^2} \mathcal{L}_2$  is equivalent to minimizing

$$T := \text{tr} \left[ \left( I - \frac{1}{\sigma^2} B^\top B \right) \mathbb{E}[\mathbf{z} f^\top(\mathbf{x})] (\mathbb{E}[f(\mathbf{x}) f^\top(\mathbf{x})])^{-1} \mathbb{E}[f(\mathbf{x}) \mathbf{z}^\top] \right]. \quad (\text{A.2})$$

(a) We first prove the **sufficient condition**, namely, the conditional independence criterion (C1) leads to minimizing  $\mathcal{L}_1 + \frac{1}{\sigma^2} \mathcal{L}_2$ .

Plugging Equation (A.1) to Equation (A.2), we have

$$\begin{aligned} T &= \text{tr} \left[ \left( I - \frac{1}{\sigma^2} B^\top B \right) \mathbb{E}[\mathbf{z} f^\top(\mathbf{x})] (\mathbb{E}[f(\mathbf{x}) f^\top(\mathbf{x})])^{-1} \mathbb{E}[f(\mathbf{x}) \mathbf{z}^\top] \right] \\ &= \text{tr} \left[ \mathbb{E}[f(\mathbf{x}) \mathbf{z}^\top] \left( I - \frac{1}{\sigma^2} B^\top B \right) \mathbb{E}[\mathbf{z} f^\top(\mathbf{x})] (\mathbb{E}[f(\mathbf{x}) f^\top(\mathbf{x})])^{-1} \right] \\ &= \text{tr} \left[ \mathbb{E}[f(\mathbf{x}) \mathbf{y}^\top] (\mathbb{E}[\mathbf{y} \mathbf{y}^\top])^{-1} \mathbb{E}[\mathbf{y} \mathbf{z}^\top] \left( I - \frac{1}{\sigma^2} B^\top B \right) \left( \mathbb{E}[f(\mathbf{x}) \mathbf{y}^\top] (\mathbb{E}[\mathbf{y} \mathbf{y}^\top])^{-1} \mathbb{E}[\mathbf{y} \mathbf{z}^\top] \right)^\top (\mathbb{E}[f(\mathbf{x}) f^\top(\mathbf{x})])^{-1} \right]. \end{aligned}$$

By assumption that  $\mathbf{y} = B\mathbf{z}$  and  $\mathbb{E}[\mathbf{z}\mathbf{z}^\top] = I$ , we have

$$\begin{aligned}
 & \mathbb{E} [f(\mathbf{x}) \mathbf{y}^\top] (\mathbb{E}[\mathbf{y}\mathbf{y}^\top])^{-1} \mathbb{E}[\mathbf{y}\mathbf{z}^\top] \left( I - \frac{1}{\sigma^2} B^\top B \right) \left( \mathbb{E} [f(\mathbf{x}) \mathbf{y}^\top] (\mathbb{E}[\mathbf{y}\mathbf{y}^\top])^{-1} \mathbb{E}[\mathbf{y}\mathbf{z}^\top] \right)^\top \\
 &= \mathbb{E} [f(\mathbf{x}) \mathbf{z}^\top] B^\top (\mathbb{E}[B\mathbf{z}\mathbf{z}^\top B^\top])^{-1} B \mathbb{E}[\mathbf{z}\mathbf{z}^\top] \left( I - \frac{1}{\sigma^2} B^\top B \right) \left( \mathbb{E} [f(\mathbf{x}) \mathbf{z}^\top] B^\top (\mathbb{E}[B\mathbf{z}\mathbf{z}^\top B^\top])^{-1} B \mathbb{E}[\mathbf{z}\mathbf{z}^\top] \right)^\top \\
 &= \mathbb{E} [f(\mathbf{x}) \mathbf{z}^\top] B^\top [BB^\top]^{-1} B \left( I - \frac{1}{\sigma^2} B^\top B \right) (\mathbb{E} [f(\mathbf{x}) \mathbf{z}^\top] B^\top [BB^\top]^{-1} B)^\top \\
 &= \mathbb{E} [f(\mathbf{x}) \mathbf{z}^\top] B^\top [BB^\top]^{-1} \left[ BB^\top - \frac{1}{\sigma^2} BB^\top BB^\top \right] [BB^\top]^{-1} B \mathbb{E} [\mathbf{z} f(\mathbf{x})^\top] \\
 &= \mathbb{E} [f(\mathbf{x}) \mathbf{z}^\top] B^\top [BB^\top]^{-1} \cdot 0_{d_y \times d_y} \cdot [BB^\top]^{-1} B \mathbb{E} [\mathbf{z} f(\mathbf{x})^\top] \\
 &= 0.
 \end{aligned}$$

Furthermore, notice that the eigenvalues of  $I - \frac{1}{\sigma^2} B^\top B$  is no less than zero by the definition of  $\sigma$ . And notice that the eigenvalues of  $\mathbb{E} [\mathbf{z} f^\top(\mathbf{x})] (\mathbb{E} [f(\mathbf{x}) f^\top(\mathbf{x})])^{-1} \mathbb{E} [f(\mathbf{x}) \mathbf{z}^\top]$  is also no less than zero. Therefore,  $T$  reaches minimum when it reaches zero. To conclude, the sufficient condition holds.

(b) We next prove the **necessary condition**, namely, minimizing  $\mathcal{L}_1 + \frac{1}{\sigma^2} \mathcal{L}_2$  leads to the conditional independence.

Under the assumption that there exists a ground truth  $f^*$  satisfying Criterion (C1) and (C2), we see from the sufficient condition that when  $T$  reaches the minimum,  $T$  must be equal to zero:

$$T = \text{tr} \left[ \left( I - \frac{1}{\sigma^2} B^\top B \right) \mathbb{E} [\mathbf{z} f^\top(\mathbf{x})] (\mathbb{E} [f(\mathbf{x}) f^\top(\mathbf{x})])^{-1} \mathbb{E} [f(\mathbf{x}) \mathbf{z}^\top] \right] = 0.$$

Since the matrix is semi-definite, we can omit the trace term and rewrite it as:

$$\mathbb{E} [f(\mathbf{x}) \mathbf{z}^\top] \left( I - \frac{1}{\sigma^2} B^\top B \right) \mathbb{E} [\mathbf{z} f^\top(\mathbf{x})] (\mathbb{E} [f(\mathbf{x}) f^\top(\mathbf{x})])^{-1} = 0.$$

Besides, since  $\mathbb{E} [f(\mathbf{x}) f^\top(\mathbf{x})]$  is non-singular, we have

$$\mathbb{E} [f(\mathbf{x}) \mathbf{z}^\top] \left( I - \frac{1}{\sigma^2} B^\top B \right) \mathbb{E} [\mathbf{z} f^\top(\mathbf{x})] = 0.$$

On the other hand, we represent the covariance as the follows based on the assumption  $\mathbb{E} [\mathbf{z}\mathbf{z}^\top] = I$ ,

$$\Sigma_{f(\mathbf{x}), \mathbf{z}|\mathbf{y}} = \mathbb{E} [f(\mathbf{x}) \mathbf{z}^\top] - \mathbb{E} [f(\mathbf{x}) \mathbf{y}^\top] (\mathbb{E} [\mathbf{y}\mathbf{y}^\top])^{-1} \mathbb{E} [\mathbf{y}\mathbf{z}^\top] = \mathbb{E} [f(\mathbf{x}) \mathbf{z}^\top] \left[ I - \frac{1}{\sigma^2} B^\top B \right].$$

Finally, the covariance meets

$$\Sigma_{f(\mathbf{x}), \mathbf{z}|\mathbf{y}} \Sigma_{f(\mathbf{x}), \mathbf{z}|\mathbf{y}}^\top = \mathbb{E} [f(\mathbf{x}) \mathbf{z}^\top] \left[ I - \frac{1}{\sigma^2} B^\top B \right] \mathbb{E} [\mathbf{z} f(\mathbf{x})^\top] = 0.$$

This leads to the conclusion that  $\Sigma_{f(\mathbf{x}), \mathbf{z}|\mathbf{y}} = 0$ .

Combining (a) and (b), we finish the proof.  $\square$

## A.2. Proof of Lemma 4.2

**Lemma 4.2.** *Under the assumptions of Theorem 1, for each function  $f$ , if  $\mathcal{L}(f)$  is minimized, then  $(1 - \frac{\lambda}{\sigma^2})\mathcal{L}_2$  and  $\mathcal{L}_1 + \frac{1}{\sigma^2}\mathcal{L}_2$  are both minimized.*



*Proof.* Notice that there exists a ground truth  $f^*$  that satisfies both Criterion (C1) and Criterion (C2). By definition of Criterion (C2), the ground truth  $f^*$  should minimize  $\mathcal{L}_2$ . On the other hand, we have proved in Section A.1 that satisfying Criterion (C2) leads to minimizing  $\mathcal{L}_1 + \frac{1}{\sigma^2} \mathcal{L}_2$ .

Notice that

$$\mathcal{L}(f) = \lambda \mathcal{L}_1 + \mathcal{L}_2 = \left(1 - \frac{\lambda}{\sigma^2}\right) \mathcal{L}_2 + \lambda \left(\mathcal{L}_1 + \frac{1}{\sigma^2} \mathcal{L}_2\right),$$

with  $\lambda > 0$ . Therefore, the ground truth  $f^*$  minimizes  $\left(1 - \frac{\lambda}{\sigma^2}\right) \mathcal{L}_2$  and  $\mathcal{L}_1 + \frac{1}{\sigma^2} \mathcal{L}_2$  at the same time. Thus, any function  $f$  minimizing the loss must minimize both  $\left(1 - \frac{\lambda}{\sigma^2}\right) \mathcal{L}_2$  and  $\mathcal{L}_1 + \frac{1}{\sigma^2} \mathcal{L}_2$ , or it would be larger than  $\mathcal{L}(f^*)$ .  $\square$

### A.3. Proof of Theorem 1

**Theorem 1** (Rationality of Loss). *Assume that there exists a ground truth  $f^*$  satisfying both Criterion (C1) and (C2), and matrix  $\mathbb{E}[f(\mathbf{x})f^\top(\mathbf{x})]$  is nonsingular. Let the singular values of matrix  $B$  be  $\sigma_1, \dots, \sigma_{d_y}$ . If  $\sigma_1 = \dots = \sigma_{d_y} = \sigma$ , the penalty coefficient  $\lambda < \sigma^2$  and  $\mathbb{E}[\mathbf{z}\mathbf{z}^\top] = I$ , then for each  $f$  minimizing the loss  $\mathcal{L}(f)$  in Equation 1 satisfies both Criterion (C1) and (C2).*

*Proof.* According to Lemma 4.2, for each function  $f$  that minimizes the loss, it also minimizes  $\left(1 - \frac{\lambda}{\sigma^2}\right) \mathcal{L}_2$  and  $\mathcal{L}_1 + \frac{1}{\sigma^2} \mathcal{L}_2$  at the same time. For the first term, since  $1 - \frac{\lambda}{\sigma^2}$  is a positive coefficient,  $\mathcal{L}_2$  is minimized. Therefore, Criterion (C2) holds. Since the second term is minimized, Criterion (C1) holds by Lemma 4.1. Thus, such  $f$  satisfies both Criterion (C1) and (C2). This finishes the proof.  $\square$

## B. Proof of Theorem 2

**Theorem 2** (Model-Free Lower Bound). *Assume that there exist function  $f_1$  and  $f_2$  in the function class such that covariance  $\text{Cov}[f_1(\mathbf{x}), \mathbf{y}] \neq 0$  and  $\text{Cov}[f_2(\mathbf{x}), \mathbf{z}] = 0$ , and matrix  $\mathbb{E}[f(\mathbf{x})f^\top(\mathbf{x})]$  is nonsingular. If we train the model with  $n'_2 = o(d_f)$  labeled data and infinite unlabeled data, where  $d_f$  is the dimension of  $f(\mathbf{x})$ , then any function  $f$  that minimizes the loss  $\mathcal{L}_{\infty, n'_2}$  in Equation (3) cannot meet Criterion (C1) and (C2) simultaneously.*

*Proof.* Let us first consider the training process. We first claim that when  $n'_2 = o(d_f)$ , for any  $f$ , the first term of  $\mathcal{L}_{\infty, n'_2}$  can be trained to zero, i.e.,

$$\frac{1}{n'_2} \left\| Y_{\text{down}}^{(1)} - \widetilde{W}_2 f \left( X_{\text{down}}^{(1)} \right) \right\|^2 = 0,$$

by taking

$$\widetilde{W}_2 = Y_{\text{down}}^{(1)} f^\top \left( X_{\text{down}}^{(1)} \right) \left[ f \left( X_{\text{down}}^{(1)} \right) f^\top \left( X_{\text{down}}^{(1)} \right) \right]^{-1},$$

where we abuse  $f \left( X_{\text{down}}^{(1)} \right) \in \mathbb{R}^{d_f \times n'_2}$ .

Therefore, to minimize the loss  $\mathcal{L}_{\infty, n'_2}$  only needs to minimize the second term of  $\mathcal{L}_{\infty, n'_2}$ :

$$\hat{f} \in \arg \min_f \left[ - \min_W \mathbb{E}_{\mathbf{x}, \mathbf{z}} \left\| \mathbf{z} - W f(\mathbf{x}) \right\|^2 \right].$$

Note that for any  $f$ ,

$$\min_W \mathbb{E}_{\mathbf{x}, \mathbf{z}} \left\| \mathbf{z} - W f(\mathbf{x}) \right\|^2 = \mathbb{E} \left\| \mathbf{z} \right\|^2 - \text{tr} \left[ \mathbb{E} \left[ \mathbf{z} f^\top(\mathbf{x}) \right] \left( \mathbb{E} \left[ f(\mathbf{x}) f^\top(\mathbf{x}) \right] \right)^{-1} \mathbb{E} \left[ f(\mathbf{x}) \mathbf{z}^\top \right] \right] \leq \mathbb{E} \left\| \mathbf{z} \right\|^2.$$

When  $f(\mathbf{x})$  is uncorrelated to  $\mathbf{z}$  (i.e.,  $f = f_2$ ), the equality holds, i.e.,

$$\min_W \mathbb{E}_{\mathbf{x}, \mathbf{z}} \left\| \mathbf{z} - W f_2(\mathbf{x}) \right\|^2 = \mathbb{E} \left\| \mathbf{z} \right\|^2.$$

Therefore, any function  $\hat{f}$  that minimizes the loss satisfies

$$\min_W \mathbb{E}_{\mathbf{x}, \mathbf{z}} \left\| \mathbf{z} - W \hat{f}(\mathbf{x}) \right\|^2 = \mathbb{E} \|\mathbf{z}\|^2.$$

We next prove that any  $\hat{f}$  that minimizes training loss  $\mathcal{L}_{\infty, n'_2}$  could not contain any information of  $\mathbf{y}$ . Under the condition that  $\mathbb{E} \left[ \hat{f}(\mathbf{x}) \hat{f}^\top(\mathbf{x}) \right]$  is invertible, we derive that

$$\begin{aligned} \min_W \mathbb{E}_{\mathbf{x}, \mathbf{y}} \left\| \mathbf{y} - W \hat{f}(\mathbf{x}) \right\|^2 &= \mathbb{E} \|\mathbf{y}\|^2 - \text{tr} \left[ \mathbb{E} \left[ \mathbf{y} \hat{f}^\top(\mathbf{x}) \right] \left( \mathbb{E} \left[ \hat{f}(\mathbf{x}) \hat{f}^\top(\mathbf{x}) \right] \right)^{-1} \mathbb{E} \left[ \hat{f}(\mathbf{x}) \mathbf{y}^\top \right] \right] \\ &= \mathbb{E} \|\mathbf{y}\|^2 - \text{tr} \left[ B \mathbb{E} \left[ \mathbf{z} \hat{f}^\top(\mathbf{x}) \right] \left( \mathbb{E} \left[ \hat{f}(\mathbf{x}) \hat{f}^\top(\mathbf{x}) \right] \right)^{-1} \mathbb{E} \left[ \hat{f}(\mathbf{x}) \mathbf{z}^\top \right] B^\top \right] \\ &= \mathbb{E} \|\mathbf{y}\|^2 - \text{tr} \left[ B^\top B \mathbb{E} \left[ \mathbf{z} \hat{f}^\top(\mathbf{x}) \right] \left( \mathbb{E} \left[ \hat{f}(\mathbf{x}) \hat{f}^\top(\mathbf{x}) \right] \right)^{-1} \mathbb{E} \left[ \hat{f}(\mathbf{x}) \mathbf{z}^\top \right] \right] \\ &= \mathbb{E} \|\mathbf{y}\|^2 - \text{tr} \left[ B^\top B \right] \left( \mathbb{E} \|\mathbf{z}\|^2 - \min_W \mathbb{E} \left\| \mathbf{z} - W \hat{f}(\mathbf{x}) \right\|^2 \right) \\ &= \mathbb{E} \|\mathbf{y}\|^2. \end{aligned}$$

However, since there exists  $f_1$  such that  $\text{Cov}(f_1(\mathbf{x}), \mathbf{y}) \neq 0$ , leading to

$$\min_W \mathbb{E}_{\mathbf{x}, \mathbf{y}} \left\| \mathbf{y} - W \hat{f}(\mathbf{x}) \right\|^2 < \min_W \mathbb{E}_{\mathbf{x}, \mathbf{y}} \left\| \mathbf{y} - W f_1(\mathbf{x}) \right\|^2$$

Therefore,  $\hat{f}$  violates Criterion (C2). The proof is done.  $\square$

### C. The proof of Theorem 3

**Theorem 3** (Lower Bound for General Loss). *Let  $\hat{f}$  be a global minimum of loss  $\bar{\mathcal{L}}_{\infty, n'_2}(\hat{f})$ . Assume that  $g_2 g_1^{-1}$  is convex and non-decreasing, and there exists a linear transformation  $V \in \mathbb{R}^{d_z \times d_z}$  such that*

$$\begin{aligned} \mathbb{E}[\rho_1(\bar{\mathbf{y}}, V\mathbf{z})] &< \frac{1}{M} g_2 g_1^{-1} \max_f \mathbb{E}[g_1 \rho_1(W_{V\mathbf{z}, f(\mathbf{x})}^* f(\mathbf{x}), V\mathbf{z})] \\ &\quad - \frac{1}{M} \min_f \mathbb{E}[g_2 \rho_2(W_{\mathbf{y}, f(\mathbf{x})}^* f(\mathbf{x}), \mathbf{y})], \end{aligned}$$

where  $M$  is the upper bound of  $g_2$ 's derivative function over  $[0, \sup_{\mathbf{x}, \mathbf{y}} \rho_2(W_{\mathbf{y}, \hat{f}(\mathbf{x})}^* \hat{f}(\mathbf{x}), \mathbf{y}) + \sup_{\mathbf{y}, \mathbf{z}} \rho_1(\bar{\mathbf{y}}, V\mathbf{z})]$ , i.e.,

$$g_2'(x) \leq M, \forall x \in [0, \sup_{\mathbf{x}, \mathbf{y}} \rho_2(W_{\mathbf{y}, \hat{f}(\mathbf{x})}^* \hat{f}(\mathbf{x}), \mathbf{y}) + \sup_{\mathbf{y}, \mathbf{z}} \rho_1(\bar{\mathbf{y}}, V\mathbf{z})].$$

If the downstream sample size  $n'_2 = o(d_f)$ , then any function  $\hat{f}$  that minimizes the loss  $\bar{\mathcal{L}}_{\infty, n'_2}(\hat{f})$  cannot meet Criterion (C1) and (C2) simultaneously.

*Proof.* Denote the trained predictor as  $\hat{f}$ . Notice that with  $n = o(d_f)$ ,  $\mathbf{y}_i = \widehat{W}_2 f(\mathbf{x}_i)$  by setting

$$\widehat{W}_2 = \left( Y_{\text{down}}^{(1)} \right)^\top \left[ f \left( X_{\text{down}}^{(1)} \right) f \left( X_{\text{down}}^{(1)} \right)^\top \right]^{-1} f \left( X_{\text{down}}^{(1)} \right),$$

where we abuse the notation  $f(X) \in \mathbb{R}^{n \times d_f}$ . This leads to

$$\min_W g_2 \rho_2 \left( Y_{\text{down}}^{(1)}, W \hat{f} \left( X_{\text{down}}^{(1)} \right) \right) = 0,$$

on the training set.

Therefore,  $\hat{f}$  is trained to minimize the loss

$$\mathcal{L}_1 = -\min_W \mathbb{E} g_1 \rho_1(z, W f(x)),$$

leading to the independence between  $f(x)$  and  $z$ .

By the definition of  $\rho_2$  and  $\rho_1$ , we have for any  $V > 0$ ,

$$\min_{W_1} \mathbb{E} g_2 \rho_2(\mathbf{y}, W_1 f(x)) \triangleq \mathbb{E} g_2 \rho_2(\mathbf{y}, \widehat{W}_1 \hat{f}(x)) = \mathbb{E} g_2 \rho_1(\bar{\mathbf{y}}, \widehat{W}_1 \hat{f}(x)), \quad (\text{A.3})$$

where  $\bar{\mathbf{y}}, \widehat{W}_1$  are the augmented version filled with zero. By the upper bound of  $g'(\cdot)$ , we have

$$\begin{aligned} & g_2 \rho_1(\bar{\mathbf{y}}, \widehat{W}_1 \hat{f}(x)) - g_2 \rho_1(Vz, \widehat{W}_1 \hat{f}(x)) \\ &= g_2'(\xi) \left[ \rho_1(\bar{\mathbf{y}}, \widehat{W}_1 \hat{f}(x)) - \rho_1(Vz, \widehat{W}_1 \hat{f}(x)) \right] \\ &\geq g_2'(\xi) [-\rho_1(\bar{\mathbf{y}}, Vz)] \\ &\geq -M \rho_1(\bar{\mathbf{y}}, Vz). \end{aligned} \quad (\text{A.4})$$

where the first equation is due to mean value theorem of integrals with point  $\xi$ . The second equality is due to  $g_2'(\xi) \geq 0$  and triangle inequality. And the third inequality is due to the condition  $g_2'(\xi) \leq M$ , and  $\xi \leq \max\{\rho_1(\bar{\mathbf{y}}, \widehat{W}_1 \hat{f}(x)), \rho_1(Vz, \widehat{W}_1 \hat{f}(x))\} \leq Q_1 + Q_2$ .

Besides, notice that

$$\begin{aligned} & \mathbb{E} g_2 \rho_1(Vz, \widehat{W}_1 \hat{f}(x)) \\ &= \mathbb{E} g_2 g_1^{-1} g_1(\rho_1(Vz, \widehat{W}_1 \hat{f}(x))) \\ &\stackrel{(i)}{\geq} g_2 g_1^{-1} \mathbb{E} g_1(\rho_1(Vz, \widehat{W}_1 \hat{f}(x))) \\ &\stackrel{(ii)}{\geq} g_2 g_1^{-1} \min_W \mathbb{E} g_1(\rho_1(Vz, W \hat{f}(x))) \\ &= g_2 g_1^{-1} \max_f \min_W \mathbb{E} g_1(\rho_1(Vz, W f(x))), \end{aligned} \quad (\text{A.5})$$

where inequality (i) follows Jensen's inequality with condition  $g_2 g_1^{-1}$  is convex, and inequality (ii) follows that  $g_2 g_1^{-1}$  is increasing. The final equation is due to the independence between  $f(x)$  and  $z$ .

Combining Equation (A.3), Equation (A.4), and Equation (A.5), leads to

$$\begin{aligned} & \min_{W_1} \mathbb{E} g_2 \rho_2(\mathbf{y}, W_1 f(x)) \\ &= \mathbb{E} g_2 \rho_1(\bar{\mathbf{y}}, \widehat{W}_1 \hat{f}(x)) \\ &\geq \mathbb{E} g_2(\rho_1(Vz, \widehat{W}_1 \hat{f}(x))) - M \mathbb{E} \rho_1(\bar{\mathbf{y}}, Vz) \\ &\geq g_2 g_1^{-1} \max_f \min_W \mathbb{E} g_1(\rho_1(Vz, W \hat{f}(x))) - M \mathbb{E} \rho_1(\bar{\mathbf{y}}, Vz) \\ &> \min_f \min_W \mathbb{E} g_2 \rho_2(\mathbf{y}, W_1 f(x)). \end{aligned}$$

This contradicts with Criterion (C2). The proof is done.  $\square$

## D. Proof of Theorem 4

**Theorem 4** (Model-Dependent Lower Bound). *Assume that the function class of the coarse representation can be decomposed as  $\mathcal{F} = \mathcal{F}_1 \times \mathcal{F}_2$ , where there exists a function  $f_0 \in \mathcal{F}$  such that  $\text{Cov}[f_0(x), \mathbf{y}] \neq 0$  and a function  $f_2 \in \mathcal{F}_2$  such that*

$f_2(\mathbf{x}) \perp \mathbf{z}$ . And assume matrix  $\mathbb{E}[f(\mathbf{x})f^\top(\mathbf{x})]$  is nonsingular for all  $f \in \mathcal{F}$ . If we train the model with  $n'_2 = o(\mathcal{M}(\mathcal{F}_1, \mathcal{L}))$  labeled samples and infinite unlabeled data, then any function  $f$  that minimizes the loss  $\mathcal{L}_{\infty, n'_2}$  in Equation (3) cannot meet Criterion (C1) and (C2) simultaneously.

*Proof.* Firstly, consider the training process. Under the assumptions on  $\mathcal{F}_1 \times \mathcal{F}_2$ , there exists a  $f_q(f_2(\cdot))$  such that:

(a) on the one hand,  $f_q(f_2(\cdot))$  minimize  $\mathcal{L}_1$ .

$$f_q(f_2(\cdot)) \in \arg \min_f \mathcal{L}_1(f).$$

Note that  $f_q f_2(\mathbf{x}) \perp \mathbf{z}$  follows  $f_2(\mathbf{x}) \perp \mathbf{z}$ . As a result, due to the assumption  $\mathbb{E} \mathbf{z} = 0$ , we have

$$\arg \min_W \|\mathbf{z} - W f_q f_2(\mathbf{x})\| = 0, \quad \min_W \|\mathbf{z} - W f_q f_2(\mathbf{x})\|^2 = \|\mathbf{z}\|^2.$$

And further notice that for any predictor  $f$ , we have

$$\min_W \|\mathbf{z} - W f(\mathbf{x})\|^2 \leq \|\mathbf{z} - 0 f(\mathbf{x})\|^2 = \|\mathbf{z}\|^2,$$

therefore,  $f_q(f_2(\cdot))$  minimize  $\mathcal{L}_1$ .

(b) on the other hand,  $f_q(f_2(\cdot))$  makes  $\mathcal{L}_2$  equal to zero on the training set. Due to the definition of model capacity, when  $n < o(\mathcal{M}(\mathcal{F}_1, \mathcal{L}))$ , by fixing  $f_2$ , there always exists  $f_q$  such that it can fit any  $n$  samples:

$$\min_W \left\| Y_{down}^{(1)} - W f_q(f_2(\mathbf{x}_i)) \right\|^2 = 0,$$

As a result,  $f_q(f_2(\cdot))$  can minimize the training loss since it minimize both  $\mathcal{L}_1$  and  $\mathcal{L}_2$ . Therefore, any predictor  $\hat{f}(\cdot)$  that minimize the training loss must minimize  $\mathcal{L}_1$ , leading to

$$\min_W \mathbb{E} \|\mathbf{z} - W \hat{f}(\mathbf{x})\|^2 = \mathbb{E} \|\mathbf{z}\|^2.$$

Next we consider the predictor  $\hat{f}(\cdot)$ .

$$\begin{aligned} & \min_W \mathbb{E}_{\mathbf{x}, \mathbf{y}} \left\| \mathbf{y} - W \hat{f}(\mathbf{x}) \right\|^2 \\ &= \mathbb{E}(\|\mathbf{y}\|^2) - \text{tr} \left[ \mathbb{E}(\mathbf{y} \hat{f}^\top(\mathbf{x})) \left( \mathbb{E}[\hat{f}(\mathbf{x}) \hat{f}^\top(\mathbf{x})] \right)^{-1} \mathbb{E}(\hat{f}(\mathbf{x}) \mathbf{y}^\top) \right] \\ &= \mathbb{E}(\|\mathbf{y}\|^2) - \text{tr} \left[ B \mathbb{E}(\mathbf{z} \hat{f}^\top(\mathbf{x})) \left( \mathbb{E}[\hat{f}(\mathbf{x}) \hat{f}^\top(\mathbf{x})] \right)^{-1} \mathbb{E}(\hat{f}(\mathbf{x}) \mathbf{z}^\top) B^\top \right] \\ &= \mathbb{E}(\|\mathbf{y}\|^2) - \text{tr} \left[ B^\top B \mathbb{E}(\mathbf{z} \hat{f}^\top(\mathbf{x})) \left( \mathbb{E}[\hat{f}(\mathbf{x}) \hat{f}^\top(\mathbf{x})] \right)^{-1} \mathbb{E}(\hat{f}(\mathbf{x}) \mathbf{z}^\top) \right] \\ &= \mathbb{E}(\|\mathbf{y}\|^2) - \text{tr} \left[ B^\top B \left( \min_W \|\mathbf{z} - W \hat{f}(\mathbf{x})\|^2 - \|\mathbf{z}\|^2 \right) \right] \\ &= \mathbb{E}(\|\mathbf{y}\|^2). \end{aligned}$$

Notice that  $\min_W \mathbb{E}_{\mathbf{x}, \mathbf{y}} \|\mathbf{y} - W f_1(\mathbf{x})\|^2 < \mathbb{E}(\|\mathbf{y}\|^2)$  since  $\text{Cov}[f_1(\mathbf{x}), \mathbf{y}] \neq 0$ . Therefore,

$$\mathbb{E}_{\mathbf{x}, \mathbf{y}} \left\| \mathbf{y} - W_{\mathbf{y}, \hat{f}(\mathbf{x})}^* \hat{f}(\mathbf{x}) \right\|^2 > \min_f \mathbb{E}_{\mathbf{x}, \mathbf{y}} \left\| \mathbf{y} - W_{\mathbf{y}, f(\mathbf{x})}^* f(\mathbf{x}) \right\|^2$$

which contradicts to the Criterion (C2).  $\square$