

A Statistical Theory of Contrastive Learning via Approximate Sufficient Statistics

Licong Lin* Song Mei*†

Abstract

Contrastive learning—a modern approach to extract useful representations from unlabeled data by training models to distinguish similar samples from dissimilar ones—has driven significant progress in foundation models. In this work, we develop a new theoretical framework for analyzing data augmentation-based contrastive learning, with a focus on SimCLR as a representative example. Our approach is based on the concept of *approximate sufficient statistics*, which we extend beyond its original definition in [OLCM25] for contrastive language-image pretraining (CLIP) using KL-divergence. We generalize it to equivalent forms and general f-divergences, and show that minimizing SimCLR and other contrastive losses yields encoders that are approximately sufficient. Furthermore, we demonstrate that these near-sufficient encoders can be effectively adapted to downstream regression and classification tasks, with performance depending on their sufficiency and the error induced by data augmentation in contrastive learning. Concrete examples in linear regression and topic classification are provided to illustrate the broad applicability of our results.

1 Introduction

Leveraging massive unlabeled data to learn useful representations has played a central role in recent advances in foundation models. A prominent approach of this kind is contrastive learning, which has driven significant progress in visual representation learning [CKNH20, HFW⁺20], large-scale speech processing [BZMA20], and multimodal AI [RKH⁺21, LLSH23].

In short, contrastive learning finds useful representations of the data by maximizing similarity between paired samples while minimizing it for non-paired samples. Consider SimCLR [CKNH20] for visual representation learning as an illustrative example. Given a dataset of images $\mathbf{x} \in \mathcal{X}$, SimCLR generates two augmented views $(\mathbf{z}^{(1)}, \mathbf{z}^{(2)}) \in \mathcal{X} \times \mathcal{X}$ for each image \mathbf{x} using random transformations (i.e., data augmentations) such as random cropping, random color distortions, and random Gaussian blur, etc. It then trains an encoder f that aligns the paired views and separates the non-paired views through minimizing the loss in Eq. (2). The learned representation $f(\mathbf{x})$ (or $f(\mathbf{z}^{(1)})$) can then be adapted to downstream tasks with few labeled samples and minimal fine-tuning.

Despite its remarkable empirical performance, the theoretical aspects of contrastive learning remain an active area of study [SPA⁺19, OLCM25]. In this work, we present a theoretical analysis of data augmentation-based contrastive learning, with a specific focus on the SimCLR framework [CKNH20] as a representative example. Notably, recent work by [OLCM25] has introduced new theoretical insights into contrastive language-image pretraining (CLIP). They first introduced the concept of approximate sufficient statistics, showing that the image and text encoders obtained from the empirical risk minimizer of CLIP are approximately sufficient. Additionally, under the joint graphical hierarchical model (JGHM) assumption for image and text data, they demonstrated that such encoders can be efficiently adapted to various downstream multimodal tasks.

Our work complements and extends the work by [OLCM25] in two key ways.

- (1) We extend the concept of approximate sufficient statistics, which was originally defined for CLIP in a specific form based on KL-divergence, to three equivalent forms and general f-divergences. Based on the

*Department of Statistics, UC Berkeley. Email: liconglin@berkeley.edu.

†Department of Statistics and Department of EECS, UC Berkeley. Email: songmei@berkeley.edu.

- equivalent forms of the definition, we establish that minimizing the contrastive loss (e.g., the InfoNCE loss [OLV18]) is essentially finding approximate sufficient statistics that are adaptable to downstream tasks.
- (2) We focus on data augmentation-based contrastive learning following the SimCLR framework. In contrast to CLIP, the random transformations in SimCLR introduce additional challenges for theoretical analysis. We show that the downstream performance of the learned encoder depends on its sufficiency and the error induced by the random transformations. Furthermore, motivated by the generalized definition of approximate sufficient statistics, we theoretically demonstrate that encoders trained using alternative contrastive losses can achieve similar downstream performance to those trained using standard SimCLR.

The remainder of this work is organized as follows. Section 2 reviews related literature. In Section 3, we introduce the concept of approximate sufficient statistics. Sections 4.1–4.2 present the setup of data augmentation-based contrastive learning and analyze the downstream performance of the SimCLR-trained encoder. In Section 4.3, we extend our analysis to general f-contrastive losses. Examples in linear regression and topic classification are discussed in Section 5. We also conduct synthetic experiments to compare contrastive learning losses in Section 6.

2 Related work

Self-supervised learning and contrastive learning. Self-supervised learning (SSL) dates back to the early work of [DS93], which leverages cross-modality information as a self-supervised substitute for labels to improve classification performance. In the past decade, SSL has been explored in image classification through various data augmentations, including rotation [GSK18], colorization [ZIE16], and Jigsaw puzzles [NF16]. More recently, contrastive learning based on paired and non-paired samples has emerged as a prominent approach in SSL [HFW⁺20, CKNH20, GSA⁺20, JYX⁺21, RKH⁺21]. Notably, SimCLR [CKNH20] learns image representations by minimizing the InfoNCE loss [OLV18] on randomly augmented views of images, while CLIP [RKH⁺21] does so on paired and non-paired image-text samples.

Choices of the loss function. Various loss functions have been used in contrastive learning, including NCE [GH10], InfoNCE [OLV18], Multi-class N-pair loss [Soh16], SigLIP [ZMKB23], f-MICL [LZS⁺24]. These losses utilize cross-entropy and its variants to distinguish paired from non-paired samples. Most relevant to our work is the InfoNCE loss [OLV18], which is derived based on the InfoMax principle [Lin88, HFLM⁺18].

Theoretical understanding of contrastive learning. Thus far, there is a rich body of literature on the theoretical understanding of self-supervised learning [SPA⁺19, POVDO⁺19, TKI20, WI20, VKSG⁺21, NS21, ZSS⁺21, AGKM21, TKH21a, TKH21b, HWGM21, HYJZ21, WL21, LLSZ21, WZW⁺22, DBP⁺23, SZZ⁺23, SCL⁺23, NGD⁺23, SZL24, VEG24, LZS⁺24, OLCM25]. Notably, early works [SPA⁺19, WI20, AGKM21] derived generalization error bounds for downstream classification tasks, using linear classifiers trained on representations learned by minimizing the InfoNCE loss. [WI20] explained contrastive learning through alignment (pulling paired samples together) and uniformity (separating non-paired samples). [ZSS⁺21] showed that InfoNCE minimization can implicitly learn the inverse of the data-generating function. [TKH21a] demonstrated that contrastive learning recovers document representations that reveal topic posterior information in a document classification problem. More recently, [VEG24] derived new PAC-Bayes bounds on the generalization error of SimCLR using bounded difference concentration and applied them to downstream linear classification. Compared with their results, our generalization error bound in Theorem 1 is independent of the batch size K and thus allows for large or full-batch learning. The most related work to ours is [OLCM25], which introduced the concept of approximate sufficiency to assess the quality of representations. They also demonstrated that the learned representation from CLIP [RKH⁺21] can be effectively adapted to several multimodal downstream tasks in a joint hierarchical graphical model.

Our work differs from existing theories of contrastive learning in several aspects: (1) Similar to [OLCM25], we derive more refined “excess risk bounds” instead of the “absolute risk bounds” established under structural conditions for downstream tasks in many prior works. (2) We derive novel unified risk bounds for downstream tasks that depend solely on the sufficiency of the encoder and the error induced by data augmentation. (3) We extend the concept of approximate sufficient statistics and theoretically analyze a broader class of contrastive losses.

3 Approximate sufficient statistics

Before diving into the analysis of contrastive learning, we first introduce the concept of approximate sufficient statistics, which provides a novel viewpoint for characterizing the quality of encoders f used in contrastive learning. Let $f : \mathbb{R}_+ \mapsto \mathbb{R}$ be a convex function such that $f(1) = 0$. For random variables (X, Y) on $\mathcal{X} \times \mathcal{Y}$ with joint density $\mathbb{P}(x, y)$ with respect to some measure μ ¹, we define the f-mutual information (f-MI) as

$$I_f(X, Y) = \int f\left(\frac{\mathbb{P}(x, y)}{\mathbb{P}(x)\mathbb{P}(y)}\right) \mathbb{P}(x)\mathbb{P}(y) d\mu.$$

Note that the f-MI is essentially the f-divergence between the joint distribution and the product of marginal distributions. It is non-negative and symmetric in X and Y . Moreover, provided that f is strictly convex, $I_f(X, Y) = 0$ if and only if X and Y are independent. Let (X, Y) be random variables that have the joint density $\mathbb{P}(X, Y)$ (Y could be thought as the parameter θ in Bayesian statistics). For any statistic $T : \mathcal{X} \mapsto T(\mathcal{X})$, to characterize the information loss of using $T(X)$ instead of X for predicting Y , we introduce the following definition of the sufficiency of $T(X)$.

Definition 1 (Approximate sufficiency). *Let $T : \mathcal{X} \rightarrow T(\mathcal{X})$ be a mapping (i.e., a statistic). We define three forms of sufficiency of T , which will be shown to be equivalent:*

- **Information Loss Sufficiency (ILS):** The information loss sufficiency of T is defined as

$$\text{Suff}_{\text{il}, f}(T) = I_f(X, Y) - I_f(T(X), Y).$$

- **Variational Form Sufficiency (VFS):** The variational form sufficiency of T is given by

$$\text{Suff}_{\text{vf}, f}(T) = \inf_{S: T(\mathcal{X}) \times \mathcal{Y} \mapsto \mathbb{R}} R_f(S \circ T) - \inf_{S: \mathcal{X} \times \mathcal{Y} \mapsto \mathbb{R}} R_f(S),$$

where $S \circ T(x, y) := S(T(x), y)$, and the f-contrastive loss

$$R_f(S) := \mathbb{E}_{\mathbb{P}(x, y)}[-S(x, y)] + \inf_{S_x: \mathcal{X} \mapsto \mathbb{R}} \mathbb{E}_{\mathbb{P}(x)\mathbb{P}(y)}[f^*(S(x, y) - S_x(x)) + S_x(x)], \quad (1)$$

where f^* is the Fenchel-dual of f .

- **Conditional Bregman Sufficiency (CBS):** The conditional Bregman sufficiency of T is defined as

$$\text{Suff}_{\text{cb}, f}(T) = \mathbb{E}_{\mathbb{P}(x) \times \mathbb{P}(y)} \left[B_f \left(\frac{\mathbb{P}(y|x)}{\mathbb{P}(y)}, \frac{\mathbb{P}(y|T(x))}{\mathbb{P}(y)} \right) \right],$$

where $B_f(a, b) := f(a) - f(b) - (a - b)f'(b)$ is the Bregman divergence of f .

Indeed, these definitions will be shown to be equivalent (Lemma 1), i.e.,

$$\text{Suff}_{\text{il}, f}(T) = \text{Suff}_{\text{vf}, f}(T) = \text{Suff}_{\text{cb}, f}(T) =: \text{Suff}_f(T).$$

We say $T(X)$ is an ε -approximate sufficient statistic if $\text{Suff}_f(T) \leq \varepsilon$.

The Information Loss Sufficiency (ILS) is closely linked to the InfoMax principle [Lin88, HFLM⁺18], which finds a statistic T that maximizes mutual information $I(T(X), Y)$ under certain constraints. The equivalence between ILS and CBS suggests that the loss in mutual information can be represented as a divergence between the conditional probabilities $\mathbb{P}(Y|X)$ and $\mathbb{P}(Y|T(X))$. This provides a concrete measure for interpreting the information loss.

In Variational Form Sufficiency (VFS), by definition, the excess risk $R_f(S \circ T) - \inf_{\tilde{S}} R_f(\tilde{S})$ serves as an upper bound on the sufficiency $\text{Suff}_f(T)$, and they are nearly equal when S is obtained by minimizing $R_f(S \circ T)$ over a sufficiently rich space \mathcal{S} . Consequently, VFS provides a loss minimization framework for finding T with low sufficiency by minimizing the f-contrastive loss $R_f(S)$ over S in some space \mathcal{S} and extracting T from S . Moreover, an extension of approximate sufficiency to similarity scores S is introduced in Appendix A.3.

¹For example, μ can be the Lebesgue measure on Euclidean spaces, or the counting measure on discrete spaces.

The concept of approximate sufficient statistics was first proposed in [OLCM25], but only in the CBS form for KL divergence (i.e., $f(x) = x \log x$). In this work, we extend the definition to general f -divergences and establish the equivalence among three forms of sufficiency. Notably, for f that is strictly convex, we have $\text{Suff}_f(T) = 0$ if and only if $Y \perp\!\!\!\perp X|T(X)$ from the CBS form, aligning with the classic definition of sufficient statistics (see e.g., [Kee10]). We will mainly consider two special cases of f : $f(x) = x \log x$ (KL-divergence) and $f(x) = (x - 1)^2/2$ (χ^2 -divergence), with the corresponding sufficiency denoted by Suff_{KL} and Suff_{χ^2} , respectively. For more examples and properties regarding approximate sufficient statistics, we refer the readers to Appendix A.

In the context of data augmentation-based contrastive learning, we may choose X and Y as two augmented views of the sample, and T as the encoder f . The sufficiency $\text{Suff}_f(f)$ then quantifies the loss of recovering augmented views from the encoder representation. We will show that the downstream performance of f can be controlled by its sufficiency (in the CBS form) and the error induced by data augmentation. Specifically, for any downstream task, a small risk can be achieved using f if it is near-sufficient and the random transformations in contrastive learning do not significantly change the downstream outcomes. As a preview of the results, we have

Theorem (Informal). *The risk on a downstream task using encoder f (denoted by $\mathcal{R}(f)$) satisfies*

$$\mathcal{R}(f) \leq c \cdot \left(\sqrt{\text{Suff}_f(f)} + \epsilon_{\mathcal{G}} \right)$$

for some constant $c > 0$, where $\text{Suff}_f(f)$ is the f -sufficiency of f and $\epsilon_{\mathcal{G}}$ denotes the error on the downstream task induced by data augmentation.

Contrastive learning with general f -divergence was also studied in [LZS⁺24, XZ24], but the loss functions considered in these works differ from the variational form in (1). In particular, while [LZS⁺24] considered a variational form similar to (1), they set $S_x = 0$ instead of taking the infimum over S_x .

4 Statistical properties of contrastive learning

In this section, we demonstrate that data augmentation-based contrastive learning can find near-sufficient encoders that are effectively adaptable to downstream tasks. We focus on the SimCLR framework in Section 4.1–4.2, and extend the results to general f -contrastive losses in Section 4.3.

4.1 Setup and the ERM estimator

Let $\mathbf{x} \in \mathcal{X}$ be a random sample drawn from a distribution $\mathbb{P}_{\mathcal{X}}$ on \mathcal{X} . Consider a set of transformations \mathcal{G} in which each transformation $g : \mathcal{X} \rightarrow \mathcal{X}$ maps \mathcal{X} to itself.² Let $\mathbb{P}_{\mathcal{G}}$ denote a distribution over the transformations in \mathcal{G} . Given a sample \mathbf{x} and two transformations $g^{(1)}, g^{(2)} \sim_{iid} \mathbb{P}_{\mathcal{G}}$, we generate two augmented views of \mathbf{x} , denoted as $\mathbf{z}^{(1)} = g^{(1)}(\mathbf{x})$ and $\mathbf{z}^{(2)} = g^{(2)}(\mathbf{x})$. The marginal distribution of $\mathbf{z}^{(1)}$ (or equivalently $\mathbf{z}^{(2)}$) is denoted by $\mathbb{P}_{\mathbf{z}}$. Often, we will omit the superscripts and let $\mathbf{z} = g(\mathbf{x})$ denote a single augmented view generated by a transformation $g \sim \mathbb{P}_{\mathcal{G}}$.

Throughout the remainder of this work, unless otherwise specified, we set $(X, Y) \stackrel{d}{=} (\mathbf{z}^{(1)}, \mathbf{z}^{(2)})$ in Definition 1, i.e., we define the sufficiency $\text{Suff}_f(T) = I_f(\mathbf{z}^{(1)}, \mathbf{z}^{(2)}) - I_f(T(\mathbf{z}^{(1)}), T(\mathbf{z}^{(2)}))$. For simplicity, we assume the joint distribution of $(\mathbf{x}, \mathbf{z}^{(1)}, \mathbf{z}^{(2)})$ is either discrete or has a continuous density w.r.t. some base measure on $\mathcal{X}^{\otimes 3}$. We abuse the notation $\mathbb{P}(\cdot)$ to refer to either discrete distributions or the density of continuous distributions, with the intended meaning clear from the context. Also, we occasionally omit the subscript KL when referring to KL-sufficiency.

SimCLR [CKNH20] learns a representation of the sample \mathbf{x} (i.e., $f(\mathbf{x})$ or $f(g(\mathbf{x}))$) through performing contrastive learning on the augmented views $(\mathbf{z}^{(1)}, \mathbf{z}^{(2)})$. Specifically, given a batch of K i.i.d. samples $\{\mathbf{x}_i\}_{i=1}^K$ from $\mathbb{P}_{\mathcal{X}}$, we generate K pairs of augmented views $\{(\mathbf{z}_i^{(1)}, \mathbf{z}_i^{(2)})\}_{i=1}^K$ using $2K$ i.i.d. transformations $\{(g_i^{(1)}, g_i^{(2)})\}_{i=1}^K$ from $\mathbb{P}_{\mathcal{G}}$. Let $f : \mathcal{X} \mapsto \mathbb{R}^p$ be an encoder function, potentially parametrized by neural

²More generally, we only need each transformation $g : \mathcal{X} \rightarrow \mathcal{Z}$ maps \mathcal{X} to a space \mathcal{Z} , which entails a natural injective map back to \mathcal{X} .

networks. The SimCLR risk function is defined as the expected InfoNCE loss [OLV18]:

$$\bar{R}_{\text{simclr}, K}(\mathbf{S}) := \frac{1}{2} \mathbb{E} \left[-\log \frac{\exp(S(\mathbf{z}_1^{(1)}, \mathbf{z}_1^{(2)}))}{\sum_{j \in [K]} \exp(S(\mathbf{z}_1^{(1)}, \mathbf{z}_j^{(2)}))} \right] + \frac{1}{2} \mathbb{E} \left[-\log \frac{\exp(S(\mathbf{z}_1^{(1)}, \mathbf{z}_1^{(2)}))}{\sum_{j \in [K]} \exp(S(\mathbf{z}_j^{(1)}, \mathbf{z}_1^{(2)}))} \right], \text{ and} \quad (2)$$

$R_{\text{simclr}, K}(f) := \bar{R}_{\text{simclr}, K}(\mathbf{S}_f)$, where $\mathbf{S}_f := \tau(\langle f(\mathbf{z}^{(1)}), f(\mathbf{z}^{(2)}) \rangle)$, $\tau : \mathbb{R} \mapsto \mathbb{R}$ is some simple link function.

Given a set of encoders denoted by \mathcal{F} and $n = n_1 K$ i.i.d. pairs of augmented views $\{(\mathbf{z}_i^{(1)}, \mathbf{z}_i^{(2)})\}_{i=1}^n$, SimCLR learns an encoder function $\hat{f} \in \mathcal{F}$ through empirical risk minimization (ERM), namely,

$$\begin{aligned} \hat{f} := \operatorname{argmin}_{f \in \mathcal{F}} \left\{ \hat{R}_{\text{simclr}, K}(\mathbf{S}_f) := \frac{1}{2n} \sum_{i=1}^{n_1} \left[\sum_{j=1}^K \left[-\log \frac{\exp(S_f(\mathbf{z}_{(i-1)K+j}^{(1)}, \mathbf{z}_{(i-1)K+j}^{(2)}))}{\sum_{l \in [K]} \exp(S_f(\mathbf{z}_{(i-1)K+j}^{(1)}, \mathbf{z}_{(i-1)K+l}^{(2)}))} \right] \right. \right. \\ \left. \left. + \left[-\log \frac{\exp(S_f(\mathbf{z}_{(i-1)K+j}^{(1)}, \mathbf{z}_{(i-1)K+j}^{(2)}))}{\sum_{l \in [K]} \exp(S_f(\mathbf{z}_{(i-1)K+l}^{(1)}, \mathbf{z}_{(i-1)K+j}^{(2)}))} \right] \right] \right\}. \end{aligned} \quad (3)$$

With the encoder $\hat{f}(\cdot)$ at hand, $\hat{f}(\mathbf{x})$ (or $\hat{f}(g(\mathbf{x}))$) serves as a representation for each $\mathbf{x} \in \mathcal{X}$, which can be used for downstream tasks.

We now show that the sufficiency of the ERM estimator \hat{f} can be properly controlled. We will demonstrate in Section 4.2 that the downstream performance of \hat{f} is closely tied to its sufficiency. First, we note that a global minimizer of the SimCLR risk is $S_\star(\mathbf{z}^{(1)}, \mathbf{z}^{(2)}) := \log \left[\frac{\mathbb{P}(\mathbf{z}^{(1)}, \mathbf{z}^{(2)})}{\mathbb{P}(\mathbf{z}^{(1)}) \cdot \mathbb{P}(\mathbf{z}^{(2)})} \right]$ (see Lemma 2 for the proof). To analyze the properties of the ERM estimator, we introduce the following boundedness assumption on the score function S and regularity assumption on τ .

Assumption 1 (Bounded score). *There exists a constant $B_S > 0$ such that for all pairs $(\mathbf{z}^{(1)}, \mathbf{z}^{(2)})$, we have $\exp(S_f(\mathbf{z}^{(1)}, \mathbf{z}^{(2)})) \in [1/B_S, B_S]$ for all $f \in \mathcal{F}$ and $\frac{\mathbb{P}(\mathbf{z}^{(1)}, \mathbf{z}^{(2)})}{\mathbb{P}(\mathbf{z}^{(1)}) \mathbb{P}(\mathbf{z}^{(2)})} \in [1/B_S, B_S]$.*

Assumption 2 (Simple link function). *The link function $\tau : \mathbb{R} \mapsto \mathbb{R}$ is invertible and there exists some constant $B_\tau > 0$ such that $|\tau(0)| \leq B_\tau$ and τ, τ^{-1} are B_τ -Lipschitz.*

Note that the first part of Assumption 1 is satisfied with $B_S = \exp(B_f^2)$ when $\|f(\mathbf{x})\|_2 \leq B_f$ for all $f \in \mathcal{F}, \mathbf{x} \in \mathcal{X}$ and τ is the identity function. Based on these assumptions, we have

Theorem 1 (Sufficiency bound for the ERM estimator). *Suppose Assumption 1 and 2 hold for some $B_S \geq 1, B_\tau > 0$. Let \hat{f} be the empirical risk minimizer defined in Eq. (3) and let S_\star be as defined in Section 4.1. Let $\text{supp}(\mathbf{z}^{(1)})$ be the support of $\mathbf{z}^{(1)}$ and $\mathcal{N}(u, \|\cdot\|_{2,\infty}, \mathcal{F})$ be the u -covering number of \mathcal{F} under the $(2, \infty)$ -norm $\|f\|_{2,\infty} := \sup_{x \in \text{supp}(\mathbf{z}^{(1)})} \|f(x)\|_2$. Then, with probability at least $1 - \delta$, we have*

$$\text{Suff}_{\text{kl}}(\hat{f}) \leq \left(1 + \frac{C}{K}\right) \cdot [\text{generalization error} + \text{approximation error}], \quad (4)$$

where

$$\text{generalization error} := \frac{C}{\sqrt{n}} \left[\sqrt{\log(1/\delta)} + B_\tau^2 \int_0^{2(\log B_S + B_\tau)} \sqrt{\log \mathcal{N}(u, \|\cdot\|_{2,\infty}, \mathcal{F})} du \right], \quad (5a)$$

$$\text{approximation error} := \inf_{f \in \mathcal{F}} \bar{R}_{\text{simclr}, K}(\mathbf{S}_f) - \bar{R}_{\text{simclr}, K}(\mathbf{S}_\star) \quad (5b)$$

for some constant $C > 0$ depending polynomially on B_S .

See the proof in Appendix B.2. In the decomposition on the R.H.S. of (4), the approximation error term represents the error incurred when approximating the optimal score S_\star within the function class \mathcal{F} . It is a property of the function class \mathcal{F} , and a richer class tends to have a smaller approximation error. The generalization error bound is derived using concentration properties of functions with bounded differences. Notably, it depends only on the total sample size $n = n_1 K$ rather than the batch size K or the number of batches n_1 . This allows our results to account for large or full-batch training, as used in SimCLR [CKNH20] and CLIP [RKH⁺21]. When $n \rightarrow \infty$, the generalization error vanishes while the approximation error remains constant.

Why does the SimCLR loss work? Intuitively, $\bar{R}_{\text{simclr}, K}(\mathbf{S})$ can be viewed as an approximation of the KL-contrastive loss $R_{\text{kl}}(\mathbf{S})$ in Eq. (1) using a finite batch size K . Namely,

$$R_{\text{kl}}(\mathbf{S}) = -\mathbb{E}[\mathbf{S}(\mathbf{z}^{(1)}, \mathbf{z}^{(2)})] + \mathbb{E}_{\mathbf{z}_1^{(1)}} \left[\log \mathbb{E}_{\mathbf{z}_2^{(2)}} [\exp(\mathbf{S}(\mathbf{z}_1^{(1)}, \mathbf{z}_2^{(2)}))] \right] = \lim_{K \rightarrow \infty} \bar{R}_{\text{simclr}, K}(\mathbf{S}) - \log K. \quad (6)$$

See the proof in Appendix B.1. As a result, by the definition of VFS in Definition 1

$$\text{Suff}_{\text{kl}}(f) \leq R_{\text{kl}}(\mathbf{S}_f) - \inf_{\mathbf{S}} R_{\text{kl}}(\mathbf{S}) \approx \underbrace{\bar{R}_{\text{simclr}, K}(\mathbf{S}_f) - \inf_{\mathbf{S}} \bar{R}_{\text{simclr}, K}(\mathbf{S})}_{\text{Excess risk}},$$

and therefore minimizing the SimCLR loss $\hat{R}_{\text{simclr}, K}(\mathbf{S}_f)$ effectively controls the sufficiency $\text{Suff}_{\text{kl}}(f)$.

4.2 Using the encoder for downstream tasks

Given an encoder function $f : \mathcal{X} \rightarrow \mathbb{R}^p$, we are interested in applying it to downstream tasks. Specifically, the goal is to leverage the learned representation $f(\mathbf{x})$ (or $f(g(\mathbf{x}))$) to facilitate learning in downstream tasks, such as regression or classification. By mapping the raw sample \mathbf{x} to the feature space \mathbb{R}^p , the representation $f(\mathbf{x})$ (or $f(g(\mathbf{x}))$) is expected to capture the most salient information of \mathbf{x} , simplifying the downstream task while maintaining high performance. In this section, we demonstrate that the downstream performance of the encoder depends on its sufficiency $\text{Suff}_{\text{kl}}(f)$ and the robustness of the downstream task to the random transformation $g \sim \mathbb{P}_{\mathcal{G}}$.

Adaptation to downstream regression tasks. We first study regression tasks. Consider the task of learning an unknown target function $h_{\star} : \mathcal{X} \mapsto \mathbb{R}$. Given an encoder f , our objective is to find a function $h : \mathbb{R}^p \mapsto \mathbb{R}$ such that $h(f(\mathbf{x})) \approx h_{\star}(\mathbf{x})$ (or $h(f(g(\mathbf{x}))) \approx h_{\star}(\mathbf{x})$). The estimation error of h is measured by the risk

$$R_{\mathcal{G}}(h \circ f) := \mathbb{E}_{\mathbf{x} \sim \mathbb{P}_{\mathcal{X}}, g \sim \mathbb{P}_{\mathcal{G}}} [(h(f(g(\mathbf{x}))) - h_{\star}(\mathbf{x}))^2], \quad \text{or} \quad R(h \circ f) := \mathbb{E}_{\mathbf{x} \sim \mathbb{P}_{\mathcal{X}}} [(h(f(\mathbf{x})) - h_{\star}(\mathbf{x}))^2].$$

For example, in regression tasks where the goal is to predict the outcome \mathbf{y} based on the covariates \mathbf{x} , one can choose $h_{\star}(\mathbf{x}) = \mathbb{E}[\mathbf{y} | \mathbf{x}]$. The two risks $R_{\mathcal{G}}(\cdot)$, $R(\cdot)$ correspond to the cases where a random transformation g is (or is not) applied before passing the input to the encoder f , respectively. Theorem 2 illustrates how the downstream performance of the encoder f depends on its sufficiency.

Theorem 2 (Performance on downstream regression). *Suppose h_{\star} satisfies $|\mathbb{E}[h_{\star}(\mathbf{x}) | g(\mathbf{x})]| \leq B_{h_{\star}}$ almost surely. Given an encoder $f : \mathcal{X} \mapsto \mathbb{R}^p$, there exists a measurable function $h : \mathbb{R}^p \mapsto \mathbb{R}$ such that*

$$R_{\mathcal{G}}(h \circ f) \leq c(B_{h_{\star}}^2 \sqrt{\text{Suff}_{\text{kl}}(f)} + \epsilon_{\mathcal{G}}), \quad (7a)$$

where $c > 0$ is some absolute constant and $\epsilon_{\mathcal{G}} := \mathbb{E}_{\mathbf{x} \sim \mathbb{P}_{\mathcal{X}}, g \sim \mathbb{P}_{\mathcal{G}}} [(h_{\star}(g(\mathbf{x})) - h_{\star}(\mathbf{x}))^2]$. Moreover, if the augmented view has the same marginal distribution as the original sample, i.e., $\mathbf{z}^{(1)} \stackrel{d}{=} \mathbf{x}$, then

$$R(h \circ f) \leq c(B_{h_{\star}}^2 \sqrt{\text{Suff}_{\text{kl}}(f)} + \epsilon_{\mathcal{G}}) \quad (7b)$$

for some absolute constant $c > 0$.

The proof of Theorem 2 is contained in Appendix B.3. The term $\epsilon_{\mathcal{G}}$ characterizes the impact of a random transformation g on the value of the target function h_{\star} . In SimCLR, since the encoder f is trained only on the augmented views $(\mathbf{z}^{(1)}, \mathbf{z}^{(2)})$, the random transformation g needs to preserve sufficient information on h_{\star} (e.g., $\epsilon_{\mathcal{G}}$ is small) for f to be effective. This is often the case in practice: for example, random cropping (g) typically does not alter the class label (h_{\star}) of an image; similarly, rotations and scaling (g) should not affect the true age (h_{\star}) of a person in facial images. In addition, Eq. (7a) still holds when $\epsilon_{\mathcal{G}}$ is replaced by the minimum error $\tilde{\epsilon}_{\mathcal{G}} := \inf_h \mathbb{E}_{\mathbf{x} \sim \mathbb{P}_{\mathcal{X}}, g \sim \mathbb{P}_{\mathcal{G}}} [(h(g(\mathbf{x})) - h_{\star}(\mathbf{x}))^2] \leq \epsilon_{\mathcal{G}}$. We refer to the proof for more details.

Adaptation to downstream classification tasks. We next turn to classification tasks. Suppose in the downstream we are given samples (\mathbf{x}, \mathbf{y}) from some joint distribution \mathbb{P} on $\mathcal{X} \times [K]$, where $\mathbf{x} \sim \mathbb{P}_{\mathcal{X}}$ is the input and $\mathbf{y} \in [K]$ is the corresponding label. Note that for any \mathbf{x} , the label \mathbf{y} follows the conditional probability $\mathbb{P}(\mathbf{y}|\mathbf{x})$. Given an encoder f , for any function $h : \mathbb{R}^p \mapsto \Delta([K])$, we measure its classification error by

$$R_{\mathcal{G}}^{\text{cls}}(h \circ f) := \mathbb{E}_{(\mathbf{x}, \mathbf{y}) \sim \mathbb{P}, g} [\text{D}_{\text{KL}}(\mathbb{P}(\mathbf{y}|\mathbf{x}) || h(f(g(\mathbf{x}))))].$$

Theorem 3 (Performance on downstream classification). *Suppose $\inf_{y \in [K]} \mathbb{P}(y|g(\mathbf{x})) \geq \exp(-B)$ for some $B > 0$ on the support of $g(\mathbf{x})$. Given an encoder $f : \mathcal{X} \mapsto \mathbb{R}^p$, there exists a measurable function $h : \mathbb{R}^p \mapsto \Delta([K])$ such that*

$$R_{\mathcal{G}}^{\text{cls}}(h \circ f) \leq c \left(B \sqrt{\text{Suff}_{\text{kl}}(f)} + \epsilon_{\mathcal{G}}^{\text{cls}} \right), \quad (8)$$

where $\epsilon_{\mathcal{G}}^{\text{cls}} := \mathbb{E}_{\mathbf{x} \sim \mathbb{P}_{\mathcal{X}}, g \sim \mathbb{P}_{\mathcal{G}}} [\text{D}_2(\mathbb{P}(\mathbf{y}|\mathbf{x}) || \mathbb{P}(\mathbf{y}|z)) + \text{D}_2(\mathbb{P}(\mathbf{y}|z) || \mathbb{P}(\mathbf{y}|\mathbf{x}))]$ and $c > 0$ is some absolute constant. Here, D_2 denotes the 2-Rényi divergence.

The proof of Theorem 3 is contained in Appendix B.4. Similar to the regression case in Theorem 2, the downstream classification error is bounded by the sum of a sufficiency term and an error term that characterizes the change in label probabilities induced by the transformation g .

4.3 General f-contrastive learning

We generalize our theoretical framework to using general f-sufficiency as defined in Definition 1, which could be controlled by minimizing the f-contrastive learning risk. We discuss (1) how to find encoders f with low f-sufficiency $\text{Suff}_f(f)$ via data augmentation-based contrastive learning and (2) the implications of low f-sufficiency on downstream performance. Note that $f(x) = x \log x$ yields the standard SimCLR setup.

4.3.1 Finding encoders with low f-sufficiency

Recall the variational form sufficiency (VFS) in Definition 1. We see that for any f and encoder f

$$\text{Suff}_f(f) \leq \inf_{S: f(\mathcal{X}) \times \mathcal{X} \mapsto \mathbb{R}} R_f(S \circ f) - \underbrace{\inf_{S: \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}} R_f(S)}_{\text{Excess risk}} \leq R_f(S_f) - \underbrace{\inf_{S: \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}} R_f(S)}_{\text{Excess risk}}.$$

Thus, for any $\varepsilon > 0$, if there exists an encoder $\hat{f} \in \mathcal{F}$ such that the excess risk of $S_{\hat{f}}$ is less than ε , then the sufficiency $\text{Suff}_f(\hat{f}) \leq \varepsilon$. Consequently, given i.i.d. pairs of augmented views, we can obtain an encoder \hat{f} with low f-sufficiency by choosing \hat{f} as the empirical risk minimizer (ERM) of a finite-sample estimate $\hat{R}_f(S_f)$ of $R_f(S_f)$, provided that $\hat{R}_f(S_f) \approx R_f(S_f)$, the function class \mathcal{F} is sufficiently rich, and its $\|\cdot\|_{2,\infty}$ -covering number is well-controlled.

We focus on χ^2 -sufficiency (i.e., $f(x) = (x-1)^2/2$) in the following. For general f, the $S_x(x)$ that attains the infimum in Eq. (1) may not have a closed-form solution, and estimating $\hat{R}_f(S_f)$ requires solving estimating equations, adding complexity to the analysis. Thus, we leave a detailed investigation of the general f case for future work.

When $f(x) = (x-1)^2/2$, basic algebra shows that the χ^2 -contrastive loss (1) takes the form

$$R_{\chi^2}(S) = \mathbb{E}_{\mathbb{P}(x,y)}[-S(x,y)] + \mathbb{E}_{\mathbb{P}(x)\mathbb{P}(y)}[(S(x,y) - \mathbb{E}_{\mathbb{P}(y)}[S(x,y)])^2/2 + S(x,y)]. \quad (9)$$

Given $n = n_1 K$ i.i.d. pairs of augmented views $\{(\mathbf{z}_i^{(1)}, \mathbf{z}_i^{(2)})\}_{i=1}^n$, an unbiased finite-sample estimate of $R_{\chi^2}(S)$ gives

$$\begin{aligned} \hat{R}_{\text{chisq}, K}(S_f) &:= \frac{1}{n} \sum_{i=1}^{n_1} \sum_{j=1}^K \left[\frac{1}{4(K-1)(K-2)} \sum_{\substack{k,l \in [K] \\ j \neq k, k \neq l, l \neq j}} (S_f(\mathbf{z}_{ij}^{(1)}, \mathbf{z}_{ik}^{(2)}) - S_f(\mathbf{z}_{ij}^{(1)}, \mathbf{z}_{il}^{(2)}))^2 \right. \\ &\quad \left. + \frac{1}{K-1} \sum_{k \neq j} S_f(\mathbf{z}_{ij}^{(1)}, \mathbf{z}_{ik}^{(2)}) - S_f(\mathbf{z}_{ij}^{(1)}, \mathbf{z}_{ij}^{(2)}) \right], \quad S_f := \tau(\langle f(\mathbf{z}^{(1)}), f(\mathbf{z}^{(2)}) \rangle), \end{aligned} \quad (10)$$

where we adopt the shorthand $\mathbf{z}_{ab}^{(i)} = \mathbf{z}_{(a-1)K+b}^{(i)}$ for $i \in [2]$. Let $\hat{f} = \operatorname{argmin}_{f \in \mathcal{F}} \widehat{\mathsf{R}}_{\text{chisq}, K}(S_f)$ be the ERM estimator. Similar to Theorem 1, we have

Theorem 4 (χ^2 -sufficiency bound for the ERM estimator). *Suppose $S_f(\mathbf{z}^{(1)}, \mathbf{z}^{(2)}) \in [-\bar{B}_S, \bar{B}_S]$ for all $f \in \mathcal{F}$ and pairs $(\mathbf{z}^{(1)}, \mathbf{z}^{(2)})$, and that Assumption 2 holds for some $B_\tau > 0$. Let $S_*(\mathbf{z}^{(1)}, \mathbf{z}^{(2)}) := \frac{\mathbb{P}(\mathbf{z}^{(1)}, \mathbf{z}^{(2)})}{\mathbb{P}(\mathbf{z}^{(1)})\mathbb{P}(\mathbf{z}^{(2)})}$. For any $K \geq 3$, with probability at least $1 - \delta$, we have*

$$\operatorname{Suff}_{\chi^2}(\hat{f}) \leq \text{generalization error} + \text{approximation error}, \quad (11)$$

where

$$\begin{aligned} \text{generalization error} &:= \frac{c\bar{B}_S^2}{\sqrt{n}} \left[\sqrt{\log(1/\delta)} + B_\tau^2 \int_0^{2(\bar{B}_S+B_\tau)} \sqrt{\log \mathcal{N}(u, \|\cdot\|_{2,\infty}, \mathcal{F})} du \right], \\ \text{approximation error} &:= \inf_{f \in \mathcal{F}} R_{\chi^2}(S_f) - R_{\chi^2}(S_*) \end{aligned}$$

for some absolute constant $c > 0$.

The proof of Theorem 4 is provided in Appendix B.5. Note that we do not assume the boundedness of S_* as in Theorem 1.

4.3.2 Implications of low f-Sufficiency

Similar to the KL case in Section 4.2, the downstream performance of f can be controlled by its f-sufficiency for a broad class of f considered in Definition 1. Recall the CBS form in Definition 1.

Proposition 5 (f-sufficiency bound on downstream performance). *The results in Theorem 2 and 3 hold with $\operatorname{Suff}_{\text{KL}}(f)$ replaced by $c_2^2 \cdot \operatorname{Suff}_f(f)$ for some value $c_2 > 0$ if*

$$\mathbb{E}_{\mathbf{z}^{(1)}} [\mathsf{D}_{\text{TV}}(\mathbb{P}(\cdot | \mathbf{z}^{(1)})) \| \mathbb{P}_{\mathbf{z}^{(2)} | \mathbf{z}^{(1)}}(\cdot | f(\mathbf{z}^{(1)}))] \leq c_2 \cdot \sqrt{\operatorname{Suff}_f(f)}. \quad (13)$$

Proposition 5 follows immediately by noting that, in the proof of Theorem 2 and 3, $\operatorname{Suff}_{\text{KL}}(f)$ is only used as an upper bound of the expected total variation distance (e.g., by Pinsker's inequality). It can be verified that KL-divergence and χ^2 -divergence satisfy Eq. (13) with $c_2 = 1/\sqrt{2}$. Let $r = \mathbb{P}(\mathbf{z}^{(1)}, \mathbf{z}^{(2)}) / [\mathbb{P}(\mathbf{z}^{(1)})\mathbb{P}(\mathbf{z}^{(2)})]$ denote the density ratio. Moreover, for general f, we can choose $c_2 = (2 \inf_{(\mathbf{z}^{(1)}, \mathbf{z}^{(2)})} f''(r))^{-1/2}$, which is bounded when f is strongly convex on the range of the density ratio r. For example, we can choose $c_2 = \sqrt{2}B^{3/4}$ when $f(x) = 1 - \sqrt{x}$ corresponds to squared Hellinger-sufficiency if the density ratio $r \leq B$ for all pairs $(\mathbf{z}^{(1)}, \mathbf{z}^{(2)})$. We refer the readers to Lemma 3 in Appendix A.2 for further details. Combining the results from Sections 4.3.1 and 4.3.2, we provide end-to-end theoretical guarantees for the downstream performance of encoders obtained by minimizing general f-contrastive losses.

5 Examples

In this section, we present concrete examples on linear regression and topic classification to illustrate the applicability of our general results in Section 4.

5.1 Linear regression

Let \mathbf{x} follow a distribution $\mathbb{P}_{\mathcal{X}}$ on $\mathcal{X} \subseteq \mathbb{R}^d$. We consider a downstream linear regression task, where each observed sample takes the form $(\mathbf{x}, \mathbf{y}) \in \mathbb{R}^d \times \mathbb{R}$, with the conditional expectation $\mathbb{E}[\mathbf{y} | \mathbf{x}] = \langle \mathbf{x}, \boldsymbol{\theta}_* \rangle$ for some unknown parameter $\boldsymbol{\theta}_* \in \mathbb{R}^d$. The goal is to predict \mathbf{y} given \mathbf{x} . While fitting a linear model using only the downstream samples yields a risk of order $\mathcal{O}(d/m)$, where m is the number of downstream samples, a smaller risk may be achieved by fitting a linear model on a low-dimensional representation $f(\mathbf{z}) \in \mathbb{R}^p$, where $p \ll d$, that captures sufficient information about \mathbf{x} relevant to the downstream task.

Concretely, suppose we are given a linear encoder $f(\mathbf{z}) = \mathbf{W}\mathbf{z}$ for some $\mathbf{W} \in \mathbb{R}^{p \times d}$ and m i.i.d. downstream samples $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^m$ from the linear model $\mathbf{y} = \langle \mathbf{x}, \boldsymbol{\theta}_* \rangle + \varepsilon$, where $\varepsilon \sim \mathcal{N}(0, \bar{\sigma}^2) \perp\!\!\!\perp \mathbf{x}$.

Suppose $\sup_{\mathbf{x} \in \mathcal{X}} \|\mathbf{x}\|_2 \leq B_{\mathbf{x}}, \|\boldsymbol{\theta}_*\|_2 \leq B_{\boldsymbol{\theta}}$ for some $B_{\mathbf{x}}, B_{\boldsymbol{\theta}} > 0$ and let $B = B_{\mathbf{x}}B_{\boldsymbol{\theta}}$. Also assume that $\mathbb{E}[(\mathbf{I}_d - \mathbf{W}^\dagger \mathbf{W})\mathbf{z}|\mathbf{W}\mathbf{z}] = 0$ almost surely. Theorem 6 below gives a theoretical guarantee for learning the downstream task using a given linear encoder.

Theorem 6 (Linear regression with encoder representation). *Let $p \leq d$. Under the setup and assumptions in Section 5.1, consider fitting a linear model $\mathbf{h}_{\hat{\eta}}(\mathbf{x}) = \langle f(\mathbf{z}), \hat{\eta} \rangle$ by ordinary least squares, i.e.,*

$$\hat{\eta} := \operatorname{argmin}_{\boldsymbol{\eta} \in \mathbb{R}^p} \left\{ \widehat{\mathbf{R}}_{\text{lin}}(\mathbf{h}_{\boldsymbol{\eta}}) := \frac{1}{m} \sum_{i=1}^m (\langle f(\mathbf{z}_i), \boldsymbol{\eta} \rangle - \mathbf{y}_i)^2 \right\},$$

where $\mathbf{z} = g(\mathbf{x}), \mathbf{z}_i = g_i(\mathbf{x}_i)$, and $g, \{g\}_{i=1}^m$ are i.i.d. transformations from $\mathbb{P}_{\mathcal{G}}$. Then the expected risk of the truncated linear model $\mathbf{h}_{\hat{\eta}}(\mathbf{x}) := \operatorname{proj}_{[-B, B]}(\mathbf{h}_{\hat{\eta}}(\mathbf{x}))$ satisfies

$$\mathbb{E}[\mathbf{R}_{\text{lin}}(\tilde{\mathbf{h}}_{\hat{\eta}})] := \mathbb{E}[\mathbb{E}_{\mathbf{x}, \mathbf{y}, g}[(\mathbf{y} - \tilde{\mathbf{h}}_{\hat{\eta}}(\mathbf{x}))^2]] \leq \underbrace{\bar{\sigma}^2}_{\text{irreducible risk}} + c \left((B^2 c_2 \sqrt{\operatorname{Suff}_{\mathbf{f}}(f)} + \epsilon_{\mathcal{G}}) + (\bar{\sigma}^2 + B^2) \frac{p \log m}{m} \right),$$

where $\epsilon_{\mathcal{G}} = \mathbb{E}[\langle \mathbf{x} - \mathbf{z}, \boldsymbol{\theta}_* \rangle^2]$ and the outer expectation is over $\{(\mathbf{x}_i, \mathbf{y}_i, g_i)\}_{i=1}^n$ for some absolute constant $c > 0$. Here, $c_2 > 0$ is any value that satisfies Eq. (13).

The proof of Theorem 6 is contained in Appendix C.1. Compared to fitting a linear model on the raw feature $\mathbf{x} \in \mathbb{R}^d$, which yields an excess risk of $\mathcal{O}(d/m)$, Theorem 6 achieves a smaller excess risk of order $\tilde{\mathcal{O}}(p/m)$ when $p \ll d$ and $f(g(\mathbf{x}))$ is a “good” representation of \mathbf{x} , in the sense that $\operatorname{Suff}_{\mathbf{f}}(f)$ and $\epsilon_{\mathcal{G}}$ are sufficiently small. A similar bound can be established for the risk $\mathbf{R}_{\text{lin}}(\tilde{\mathbf{h}}_{\hat{\eta}})$ with high probability under additional sub-Gaussian assumptions on the representation $f(\mathbf{z}) = \mathbf{W}g(\mathbf{x})$ [HKZ11]. We provide the bound in expectation $\mathbb{E}[\mathbf{R}_{\text{lin}}(\tilde{\mathbf{h}}_{\hat{\eta}})]$ for simplicity of presentation.

The assumption $\mathbb{E}[(\mathbf{I}_d - \mathbf{W}^\dagger \mathbf{W})\mathbf{z}|\mathbf{W}\mathbf{z}] = 0$ essentially states that the information of the augmented view \mathbf{z} discarded by the encoder f does not contain any signal with a non-zero mean. Without this assumption, there may not exist a linear function of $f(\mathbf{z})$ that achieves a small risk $\mathbf{R}_{\text{lin}}(\cdot)$, even though Theorem 2 guarantees the existence of a general function of $f(\mathbf{z})$ with a small risk. Note that the assumption is satisfied when e.g., \mathbf{z} follows the standard normal distribution on \mathbb{R}^d .

5.1.1 A concrete scenario

We now present a scenario in which a linear encoder f with low KL-sufficiency $\operatorname{Suff}_{\mathbf{f}}(f)$ can be obtained through SimCLR loss minimization in Eq. (3). Let $\mathbf{U} = (\mathbf{U}_1, \mathbf{U}_2) \in \mathbb{R}^{d \times d}$, where $\mathbf{U}_1 \in \mathbb{R}^{d \times p}$, be a fixed unitary matrix, and define $\mathbf{A} = \mathbf{U}_1 \mathbf{U}_1^\top$. For $i \in [2]$, define the unit sphere in the column space of \mathbf{U}_i as $\mathbb{S}(\mathbf{U}_i) := \{\mathbf{v} \in \mathbb{R}^d : \|\mathbf{v}\|_2 = 1, (\mathbf{I}_d - \mathbf{U}_i \mathbf{U}_i^\top)\mathbf{v} = \mathbf{0}\}$. Assume $\mathbf{x} \in \mathbb{R}^d \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d/p)$ and consider the random transformation g such that $g(\mathbf{x})|\mathbf{x} \stackrel{d}{=} (\mathbf{A}\mathbf{x} + \eta)|\{\mathbf{A}\mathbf{x} + \eta \in \mathbb{S}(\mathbf{U}_1) \oplus \mathbb{S}(\mathbf{U}_2)\}$, i.e., the conditional distribution $g(\mathbf{x})|\mathbf{x}$ follows the distribution of $\mathbf{A}\mathbf{x} + \eta$ conditioned on $\mathbb{S}(\mathbf{U}_1) \oplus \mathbb{S}(\mathbf{U}_2)$,³ where the noise $\eta \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_d/p)$. A concrete example of this transformation involves zeroing out the second half of the coordinates of the sample $\mathbf{x} \in \mathbb{R}^d$, adding some Gaussian noise to all coordinates, and then normalizing both halves of the noisy sample to have unit norm. In this case, $\mathbf{U}_1, \mathbf{U}_2$ correspond to the first and second halves of the coordinates, respectively.

Under this setup, it is readily verified that the distribution of $(\mathbf{z}^{(1)}, \mathbf{z}^{(2)})$ is supported on $\mathbb{S}(\mathbf{U}_1) \oplus \mathbb{S}(\mathbf{U}_2)$, and conditioned on $\mathbb{S}(\mathbf{U}_1) \oplus \mathbb{S}(\mathbf{U}_2)$, the densities satisfy⁴

$$\begin{aligned} \mathbb{P}(\mathbf{z}^{(1)}, \mathbf{z}^{(2)}) &\propto \exp \left(-\frac{p}{2} \left\langle \begin{pmatrix} \mathbf{z}^{(1)} \\ \mathbf{z}^{(2)} \end{pmatrix}, \begin{bmatrix} \mathbf{U}_1 \mathbf{U}_1^\top + \sigma^2 \mathbf{I}_d & \mathbf{U}_1 \mathbf{U}_1^\top \\ \mathbf{U}_1 \mathbf{U}_1^\top & \mathbf{U}_1 \mathbf{U}_1^\top + \sigma^2 \mathbf{I}_d \end{bmatrix}^{-1} \begin{pmatrix} \mathbf{z}^{(1)} \\ \mathbf{z}^{(2)} \end{pmatrix} \right\rangle \right), \\ \mathbb{P}(\mathbf{z}^{(1)}) &\propto 1 \quad \text{and,} \\ \frac{\mathbb{P}(\mathbf{z}^{(1)}, \mathbf{z}^{(2)})}{\mathbb{P}(\mathbf{z}^{(1)})\mathbb{P}(\mathbf{z}^{(2)})} &\propto \exp(\kappa \langle \mathbf{z}^{(1)}, \mathbf{U}_1 \mathbf{U}_1^\top \mathbf{z}^{(2)} \rangle), \quad \kappa := \frac{p}{\sigma^2(\sigma^2 + 2)} \leq \frac{p}{\sigma^4}. \end{aligned}$$

³ $\mathbb{S}(\mathbf{U}_1) \oplus \mathbb{S}(\mathbf{U}_2) := \{\mathbf{v} \in \mathbb{R}^d : \mathbf{v} = \mathbf{v}_1 + \mathbf{v}_2 \text{ for some } \mathbf{v}_1 \in \mathbb{S}(\mathbf{U}_1), \mathbf{v}_2 \in \mathbb{S}(\mathbf{U}_2)\}.$

⁴All densities are with respect to the Lebesgue measure.

Note that $(\mathbf{z}^{(1)}, \mathbf{z}^{(2)})$ restricting on $\mathbb{S}(\mathbf{U}_1) \times \mathbb{S}(\mathbf{U}_1)$ follows the joint von Mises-Fisher distribution (vMF) [Fis53]. In this case, the optimal score is given by $S_*(\mathbf{z}^{(1)}, \mathbf{z}^{(2)}) = \tau(\langle f_*(\mathbf{z}^{(1)}), f_*(\mathbf{z}^{(2)}) \rangle)$, where $\tau(x) = \kappa x$ and $f_*(\mathbf{z}) = \mathbf{U}_1 \mathbf{z}$. Moreover, we have the following guarantee on the sufficiency of the SimCLR estimator \hat{f} .

Corollary 1 (An upper bound on the sufficiency). *Under the setup in Section 5.1.1, let $\mathcal{F} := \{f : f(\mathbf{z}) = \mathbf{W}\mathbf{z}, \mathbf{W} \in \mathbb{R}^{p \times d} \text{ and } \|\mathbf{W}\|_{op} \leq B_{\mathbf{W}}\}$ for some $B_{\mathbf{W}} \geq 1$, and set $\tau(x) = \kappa x$. Define \hat{f} as the SimCLR empirical risk minimizer obtained from Eq. (3), using batch size K and n samples. Then, with probability at least $1 - \delta$, we have*

$$\text{Suff}_{\text{kl}}(\hat{f}) \leq \left(1 + \frac{C}{K}\right) \cdot \sqrt{\frac{dp \cdot \log B_{\mathbf{W}} + \log(1/\delta)}{n}}$$

for some constant $C > 0$ that depends polynomially on $\exp(\kappa)$.

See Appendix C.2 for the proof. Note that the constant $\exp(\kappa)$ depends on the noise level σ . When $\sigma \gtrsim p^{1/4}$, finding a near-sufficient encoder is relatively easy. Combining Theorem 6 and Corollary 1, we conclude that the learned encoder \hat{f} can achieve a small risk in the downstream linear regression task, provided that there are sufficient pretraining and downstream samples, and that data augmentation does not significantly alter the output of the true linear model (i.e., $\epsilon_{\mathcal{G}}$ is small). See Appendix C.3 for an end-to-end statement and its proof.

5.2 Topic classification

Next, we provide theoretical guarantees for contrastive learning and its downstream performance in a classification setting. Let $\mathcal{Y} = \{1, 2, \dots, M\}$ represent a set of classes. A sample \mathbf{x} is generated by first selecting a class $\mathbf{y} \in \mathcal{Y}$ from some distribution $\mathbb{P}_{\mathcal{Y}}$, and then drawing $\mathbf{x} = (\mathbf{x}^{c_1}, \mathbf{x}^{c_2}) \in [S] \times [S]$ conditioned on \mathbf{y} , with the joint distribution

$$\mathbb{P}(\mathbf{x}|\mathbf{y}) = \mathbb{P}_c(\mathbf{x}^{c_1}|\mathbf{y}) \times \mathbb{P}_c(\mathbf{x}^{c_2}|\mathbf{y}),$$

where $\mathbb{P}_c(\cdot|\mathbf{y})$ is some conditional distribution over $[S]$. For example, in a topic classification task, each sample consists of a two-part sentence (or a two-word phrase), with the class \mathbf{y} representing the topic (e.g., sports, technology, or health). The first and second parts (or words), \mathbf{x}^{c_1} and \mathbf{x}^{c_2} , are independently sampled from a vocabulary of size S , conditioned on the topic \mathbf{y} .

Contrastive learning. We consider learning a near-sufficient encoder f via minimizing the χ^2 -contrastive loss. Namely, we consider the random dropout transformation $g : [S] \times [S] \rightarrow [S]$, which selects one component \mathbf{x}^{c_i} from the pair $(\mathbf{x}^{c_1}, \mathbf{x}^{c_2})$ with equal probability as the augmented view \mathbf{z} and drops the other. With slight abuse of notation, we also denote the augmented view \mathbf{z} using one-hot encoding. We consider encoders f that are linear functions of \mathbf{z} augmented with the one-hot encoding, namely, consider the encoder space

$$\mathcal{F} = \{f_{\text{aug}} : \cup_{i=1}^S \{e_i\} \mapsto \mathbb{R}^{M+S} \mid f_{\text{aug}}(\mathbf{z}) = ((\mathbf{W}\mathbf{z})^\top, w \cdot \mathbf{z}^\top)^\top, \mathbf{W} \in \mathbb{R}^{M \times S}, w \in \mathbb{R}, \|\mathbf{W}\|_{2,\infty} \vee |w/\sqrt{S}| \leq B_{\mathbf{W}}\}$$

with $B_{\mathbf{W}} = M$. To learn an encoder \hat{f}_{aug} , we minimize the χ^2 -contrastive loss computed using n i.i.d. pairs of augmented views via Eq. (10). Importantly, class labels $\{\mathbf{y}_i\}_{i=1}^n$ remain *unobservable* during contrastive learning. We note that a similar data distribution was studied in [TKH21a], where the augmented views correspond deterministically to the first and second components of the sample.

Downstream classification. We consider a downstream task in which we are given i.i.d. samples $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^m$ from the joint distribution of (\mathbf{x}, \mathbf{y}) , and the goal is to learn the conditional topic distribution $\mathbb{P}(\mathbf{y} = y|\mathbf{x})_{y \in [M]} \in \mathbb{R}^M$ using an encoder. Let $\hat{f}_{\text{aug}}(\mathbf{z}) = ((\hat{\mathbf{W}}\mathbf{z})^\top, \hat{w} \cdot \mathbf{z}^\top)^\top$ be the representation learned from contrastive learning, and define the encoder as $\hat{f}(\mathbf{z}) := \hat{\mathbf{W}}\mathbf{z} \in \mathbb{R}^M$. We train a multi-class linear classifier on \hat{f} to predict the topic distribution.

Define the gold representation $\mathbf{E}_* \in \mathbb{R}^{M \times S}$ whose j 'th column gives $\mathbf{E}_{*,j} = \left(\frac{\mathbb{P}_c(\mathbf{y}=1|\mathbf{x}^{c_1}=j)}{\sqrt{\mathbb{P}_{\mathcal{Y}}(\mathbf{y}=1)}}, \dots, \frac{\mathbb{P}_c(\mathbf{y}=M|\mathbf{x}^{c_1}=j)}{\sqrt{\mathbb{P}_{\mathcal{Y}}(\mathbf{y}=M)}} \right)^\top$ for $j \in [S]$. We also make the following regularity assumptions:

- (a) The marginal distributions of \mathbf{y} and \mathbf{x}^{c_1} are uniform over $[M]$ and $[S]$, respectively.
- (b) The minimum singular value of $\mathbf{E}_\star \mathbf{E}_\star^\top$ satisfies $\sigma_{\min}(\mathbf{E}_\star \mathbf{E}_\star^\top) \geq \sigma_{\mathbf{E}_\star}^2$ for some $\sigma_{\mathbf{E}_\star} > 0$.
- (c) $S \geq 4M$ and $\inf_{y \in [M], s \in [S]} \mathbb{P}_c(y|s) \geq \exp(-B)$ for some $B > 0$.

Assumption (a) assumes uniform topic and word (or sentence) distributions, simplifying the analysis of χ^2 -contrastive learning. Assumption (b) is a technical assumption that allows us to transform the learned embedding $\hat{f}(\mathbf{z})$ to the gold representation $\mathbf{E}_\star(\mathbf{z})$. Assumption (c) ensures the vocabulary size S is large compared with the number of topics M and all topics have non-vanishing conditional probability in \mathbb{P}_c . With these assumptions at hand, we have

Theorem 7 (Classification using the χ^2 -trained encoder). *Under the setup and assumptions in Section 5.2 and let \hat{f}_{aug} be the ERM in Eq. (10). Then, with probability at least $1 - \delta$ over $\{(\mathbf{z}_i^{(1)}, \mathbf{z}_i^{(2)})\}_{i=1}^n$,*

$$\text{Suff}_{\chi^2}(\hat{f}_{\text{aug}}) \leq R_f(S_{\hat{f}_{\text{aug}}}) - R_f(S_\star) =: \text{Suff}_{\chi^2}(S_{\hat{f}_{\text{aug}}}) \leq \frac{cS^2M^4}{\sqrt{n}} \left[\sqrt{\log(1/\delta)} + \sqrt{SM}^{1.5} \right] \quad (14)$$

for some absolute constant $c > 0$.

In downstream classification, given m i.i.d. samples $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^m$, consider fitting a multi-class classifier $\mathbf{h}_{\hat{\Gamma}}(\mathbf{x}) = \bar{\mathbf{h}}_{\hat{\Gamma}}(\hat{f}(\mathbf{z})) := \text{softmax}(\log \text{trun}(\hat{\Gamma}_w \hat{f}(\mathbf{z}) + \hat{\Gamma}_b))$ with

$$\hat{\Gamma} := \underset{\Gamma_w \in \mathbb{R}^{M \times M}, \Gamma_b \in \mathbb{R}^M}{\operatorname{argmin}} \left\{ \hat{R}_{\text{cls}}(\mathbf{h}_{\Gamma}) := -\frac{1}{m} \sum_{i=1}^m \log \bar{\mathbf{h}}_{\Gamma}(\hat{f}(\mathbf{z}_i))_{\mathbf{y}_i} \right\}, \quad (15)$$

where $\mathbf{z} = g(\mathbf{x})$, $\mathbf{z}_i = g_i(\mathbf{x}_i)$ and $g, \{g\}_{i=1}^m$ are i.i.d. dropout transformations, $B_\Gamma \geq 4\sqrt{SM}/\sigma_{\mathbf{E}_\star}$, and $\text{trun}(x) := \text{proj}_{[\exp(-B), 1]}(x)$. Then there exists some absolute constants $c, c' > 0$ such that, given the encoder \hat{f} and suppose $\text{Suff}_{\chi^2}(S_{\hat{f}_{\text{aug}}}) \leq c' \frac{\sigma_{\mathbf{E}_\star}^2}{S^2M}$, with probability at least $1 - \delta_1$

$$\begin{aligned} R_{\text{cls}}(\bar{\mathbf{h}}_{\hat{\Gamma}}) &:= \mathbb{E}_{\mathbf{x}, \mathbf{y}, g} [\text{D}_{\text{KL}}(\mathbb{P}(\mathbf{y}|\mathbf{x}) || \mathbf{h}_{\hat{\Gamma}}(\hat{f}(g(\mathbf{x}))))] \\ &\leq c \left(\underbrace{\left[\epsilon_G^{\text{cls}} + \frac{S \exp(B)}{\sigma_{\mathbf{E}_\star}^2} \cdot \text{Suff}_{\chi^2}(S_{\hat{f}_{\text{aug}}}) \right]}_{\text{approximation error}} + \underbrace{\frac{B}{\sqrt{m}} \left[\sqrt{\log(1/\delta_1)} + M(\sqrt{\log B_\Gamma} + \sqrt{B}) \right]}_{\text{generalization error}} \right). \end{aligned}$$

See the proof in Appendix C.4. Note that the bound on downstream classification depends on the sufficiency of the score function $\text{Suff}_{\chi^2}(S_{\hat{f}_{\text{aug}}})$, introduced in Appendix A.3, rather than $\text{Suff}_{\chi^2}(\hat{f})$. This distinction arises because we restrict ourselves to linear classifiers, whereas Theorem 3 considers arbitrary measurable functions, leading to an additional approximation error term.

6 Experiments

We also conduct synthetic experiments to learn data representations via contrastive learning using two-layer neural networks, and evaluate them on downstream linear regression.

In the contrastive learning stage, we generate n i.i.d. samples $\mathbf{x}_i \sim \mathcal{N}(0, \mathbf{I}_d)$. The augmentation g adds i.i.d. $\mathcal{N}(0, \sigma_1^2)$ noise to the first $s < d$ coordinates of \mathbf{x}_i , and replaces the remaining coordinates with i.i.d. $\mathcal{N}(0, 1)$ noise. We apply KL and χ^2 -contrastive learning (Eq. 3 and 10) with link function $\tau(x) = x$, and encoder $f(\cdot)$ being a two-layer ReLU neural network mapping \mathbb{R}^d to \mathbb{R}^s . We set $s = 10, d = 100, n = 500$, hidden dimension 64, and batch size $K = 64$. The encoder is trained using Adam (learning rate 0.001) for 1000 epochs until convergence.

For downstream regression, we generate m i.i.d. samples $(\mathbf{x}_i, \mathbf{y}_i)$, where $\mathbf{x}_i \sim \mathcal{N}(0, \mathbf{I}_d)$ and $\mathbf{y}_i = \langle \mathbf{x}_i, \boldsymbol{\theta}_\star \rangle + \varepsilon_i$, with $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ independent of \mathbf{x}_i . We choose $\boldsymbol{\theta}_\star = (\mathbf{1}_s^\top / \sqrt{s}, \mathbf{0}_{d-s}^\top)^\top$ and $\sigma = 1$. Using the learned representation $\hat{f}(\mathbf{x}_i) \in \mathbb{R}^s$ from KL (or χ^2)-contrastive learning, we fit a downstream linear model to predict \mathbf{y}_i . We define the excess risk of any predictor h as $\mathbb{E}[(\mathbf{y}_i - h(\mathbf{x}_i))^2] - \sigma^2$, and evaluate the excess risk of the linear model trained on $\hat{f}(\mathbf{x}_i)$. For comparison, we also report the excess risk of a linear model trained directly on the original features \mathbf{x}_i (denoted as Direct LR). Results for various downstream sample size m and the standard deviation over 10 runs are shown in Figure 1.

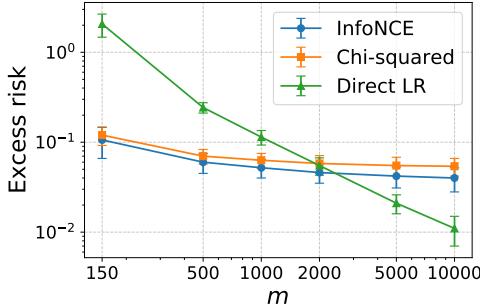


Figure 1: Excess risk for various downstream sample sizes m . The errorbars represent the standard deviation over 10 runs.

From the Figure, we observe that linear regression based on KL (i.e., InfoNCE) or χ^2 -pretrained representations achieve comparable excess risks, both much lower than that of direct linear regression when the sample size m is relatively small (e.g., $m = 150, 500$). This suggests that both KL and χ^2 -contrastive learning can learn a “good” low-dimensional representation for the downstream task. As the sample size increases, the excess risk of direct linear regression converges to zero, while those of KL and χ^2 -pretrained representations converge to non-zero constants. This is consistent with our theoretical results, which attribute the excess risk to the non-zero sufficiency of \hat{f} and the augmentation error ϵ_G . More results comparing KL (i.e., InfoNCE) and χ^2 -contrastive learning in the CLIP setting are provided in Appendix D.

7 Conclusion

In this work, we present a new theoretical framework for data augmentation-based contrastive learning, with SimCLR as a representative example. Based on the extended concept of approximate sufficient statistics, we establish a connection between minimizing the f-contrastive losses and minimizing the conditional Bregman sufficiency (CBS) of the encoder. Moreover, we show that the learned encoders can be effectively applied to downstream tasks with performance depending on their sufficiency and the error on the downstream task induced by data augmentation.

Our work opens up many directions for future research. First, as seen in Definition 1, the concept of approximate sufficient statistics is not limited to contrastive learning; exploring its applicability to other self-supervised and supervised learning paradigms is a promising direction. Second, while approximate sufficiency quantifies the information preserved by the encoder, it does not reflect the redundancy in its representation. Thus, it would be interesting to generalize the concept of minimal sufficient statistics and develop practical algorithms for finding representations that are both approximately sufficient and minimal. Lastly, our work mainly focuses on the empirical risk minimizers in contrastive learning. Understanding what representations are learned and how training algorithms influence the learned representation remains another exciting avenue for future research.

Acknowledgement

This project was supported by NSF grants DMS-2210827, CCF-2315725, CAREER DMS-2339904, ONR grant N00014-24-S-B001, DARPA AIQ grant HR001124S0029-AIQ-FP-003, an Amazon Research Award, a Google Research Scholar Award, an Okawa Foundation Research Grant, and a Sloan Research Fellowship.

References

- [AC10] Jean-Yves Audibert and Olivier Catoni, *Linear regression through pac-bayesian truncation*, arXiv preprint arXiv:1010.0072 (2010).

- [AGKM21] Jordan T Ash, Surbhi Goel, Akshay Krishnamurthy, and Dipendra Misra, *Investigating the role of negatives in contrastive representation learning*, arXiv preprint arXiv:2106.09943 (2021).
- [BS16] Mark Bun and Thomas Steinke, *Concentrated differential privacy: Simplifications, extensions, and lower bounds*, Theory of cryptography conference, Springer, 2016, pp. 635–658.
- [BZMA20] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli, *wav2vec 2.0: A framework for self-supervised learning of speech representations*, Advances in neural information processing systems **33** (2020), 12449–12460.
- [CKNH20] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton, *A simple framework for contrastive learning of visual representations*, International conference on machine learning, PMLR, 2020, pp. 1597–1607.
- [DBP⁺23] Imant Daunhawer, Alice Bizeul, Emanuele Palumbo, Alexander Marx, and Julia E Vogt, *Identifiability results for multimodal contrastive learning*, arXiv preprint arXiv:2303.09166 (2023).
- [DS93] Virginia De Sa, *Learning classification with unlabeled data*, Advances in neural information processing systems **6** (1993).
- [Fis53] Ronald Aylmer Fisher, *Dispersion on a sphere*, Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences **217** (1953), no. 1130, 295–305.
- [GH10] Michael Gutmann and Aapo Hyvärinen, *Noise-contrastive estimation: A new estimation principle for unnormalized statistical models*, Proceedings of the thirteenth international conference on artificial intelligence and statistics, JMLR Workshop and Conference Proceedings, 2010, pp. 297–304.
- [GKKW06] László Györfi, Michael Kohler, Adam Krzyzak, and Harro Walk, *A distribution-free theory of nonparametric regression*, Springer Science & Business Media, 2006.
- [GSA⁺20] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al., *Bootstrap your own latent-a new approach to self-supervised learning*, Advances in neural information processing systems **33** (2020), 21271–21284.
- [GSK18] Spyros Gidaris, Praveer Singh, and Nikos Komodakis, *Unsupervised representation learning by predicting image rotations*, arXiv preprint arXiv:1803.07728 (2018).
- [HFLM⁺18] 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).
- [HFW⁺20] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick, *Momentum contrast for unsupervised visual representation learning*, Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 9729–9738.
- [HKZ11] Daniel Hsu, Sham M Kakade, and Tong Zhang, *An analysis of random design linear regression*, arXiv preprint arXiv:1106.2363 **6** (2011).
- [HWGM21] Jeff Z HaoChen, Colin Wei, Adrien Gaidon, and Tengyu Ma, *Provable guarantees for self-supervised deep learning with spectral contrastive loss*, Advances in Neural Information Processing Systems **34** (2021), 5000–5011.
- [HYZJ21] Weiran Huang, Mingyang Yi, Xuyang Zhao, and Zihao Jiang, *Towards the generalization of contrastive self-supervised learning*, arXiv preprint arXiv:2111.00743 (2021).

- [JYX⁺21] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig, *Scaling up visual and vision-language representation learning with noisy text supervision*, International conference on machine learning, PMLR, 2021, pp. 4904–4916.
- [Kee10] Robert W Keener, *Theoretical statistics: Topics for a core course*, Springer Science & Business Media, 2010.
- [Lin88] Ralph Linsker, *Self-organization in a perceptual network*, Computer **21** (1988), no. 3, 105–117.
- [LLSH23] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi, *Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models*, International conference on machine learning, PMLR, 2023, pp. 19730–19742.
- [LLSZ21] Jason D Lee, Qi Lei, Nikunj Saunshi, and Jiacheng Zhuo, *Predicting what you already know helps: Provable self-supervised learning*, Advances in Neural Information Processing Systems **34** (2021), 309–323.
- [LZS⁺24] Yiwei Lu, Guojun Zhang, Sun Sun, Hongyu Guo, and Yaoliang Yu, *f-micl: Understanding and generalizing infonce-based contrastive learning*, arXiv preprint arXiv:2402.10150 (2024).
- [MJ09] Kanti V Mardia and Peter E Jupp, *Directional statistics*, John Wiley & Sons, 2009.
- [NF16] Mehdi Noroozi and Paolo Favaro, *Unsupervised learning of visual representations by solving jigsaw puzzles*, European conference on computer vision, Springer, 2016, pp. 69–84.
- [NGD⁺23] Ryumei Nakada, Halil Ibrahim Gulluk, Zhun Deng, Wenlong Ji, James Zou, and Linjun Zhang, *Understanding multimodal contrastive learning and incorporating unpaired data*, International Conference on Artificial Intelligence and Statistics, PMLR, 2023, pp. 4348–4380.
- [NS21] Kento Nozawa and Issei Sato, *Understanding negative samples in instance discriminative self-supervised representation learning*, Advances in Neural Information Processing Systems **34** (2021), 5784–5797.
- [OLCM25] Kazusato Oko, Licong Lin, Yuhang Cai, and Song Mei, *A statistical theory of contrastive pre-training and multimodal generative ai*, arXiv preprint arXiv:2501.04641 (2025).
- [OLV18] Aaron van den Oord, Yazhe Li, and Oriol Vinyals, *Representation learning with contrastive predictive coding*, arXiv preprint arXiv:1807.03748 (2018).
- [POVDO⁺19] Ben Poole, Sherjil Ozair, Aaron Van Den Oord, Alex Alemi, and George Tucker, *On variational bounds of mutual information*, International Conference on Machine Learning, PMLR, 2019, pp. 5171–5180.
- [RKH⁺21] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al., *Learning transferable visual models from natural language supervision*, International conference on machine learning, PMLR, 2021, pp. 8748–8763.
- [SCL⁺23] Zhenmei Shi, Jiefeng Chen, Kunyang Li, Jayaram Raghuram, Xi Wu, Yingyu Liang, and Somesh Jha, *The trade-off between universality and label efficiency of representations from contrastive learning*, arXiv preprint arXiv:2303.00106 (2023).
- [SDGS18] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut, *Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning*, Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2018, pp. 2556–2565.
- [Soh16] Kihyuk Sohn, *Improved deep metric learning with multi-class n-pair loss objective*, Advances in neural information processing systems **29** (2016).

- [SPA⁺19] Nikunj Saunshi, Orestis Plevrakis, Sanjeev Arora, Mikhail Khodak, and Hrishikesh Khandeparkar, *A theoretical analysis of contrastive unsupervised representation learning*, International Conference on Machine Learning, PMLR, 2019, pp. 5628–5637.
- [SZL24] Ravid Shwartz Ziv and Yann LeCun, *To compress or not to compress—self-supervised learning and information theory: A review*, Entropy **26** (2024), no. 3, 252.
- [SZZ⁺23] Liangliang Shi, Gu Zhang, Haoyu Zhen, Jintao Fan, and Junchi Yan, *Understanding and generalizing contrastive learning from the inverse optimal transport perspective*, International conference on machine learning, PMLR, 2023, pp. 31408–31421.
- [TKH21a] Christopher Tosh, Akshay Krishnamurthy, and Daniel Hsu, *Contrastive estimation reveals topic posterior information to linear models*, Journal of Machine Learning Research **22** (2021), no. 281, 1–31.
- [TKH21b] ———, *Contrastive learning, multi-view redundancy, and linear models*, Algorithmic Learning Theory, PMLR, 2021, pp. 1179–1206.
- [TKI20] Yonglong Tian, Dilip Krishnan, and Phillip Isola, *Contrastive multiview coding*, Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16, Springer, 2020, pp. 776–794.
- [VEG24] Anna Van Elst and Debarghya Ghoshdastidar, *Tight pac-bayesian risk certificates for contrastive learning*, arXiv preprint arXiv:2412.03486 (2024).
- [VKSG⁺21] Julius Von Kügelgen, Yash Sharma, Luigi Gresele, Wieland Brendel, Bernhard Schölkopf, Michel Besserve, and Francesco Locatello, *Self-supervised learning with data augmentations provably isolates content from style*, Advances in neural information processing systems **34** (2021), 16451–16467.
- [VR88] Dietrich Von Rosen, *Moments for the inverted wishart distribution*, Scandinavian Journal of Statistics (1988), 97–109.
- [Wai19] Martin J. Wainwright, *High-dimensional statistics: A non-asymptotic viewpoint*, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, 2019.
- [WI20] Tongzhou Wang and Phillip Isola, *Understanding contrastive representation learning through alignment and uniformity on the hypersphere*, International conference on machine learning, PMLR, 2020, pp. 9929–9939.
- [WL21] Zixin Wen and Yuanzhi Li, *Toward understanding the feature learning process of self-supervised contrastive learning*, International Conference on Machine Learning, PMLR, 2021, pp. 11112–11122.
- [WZW⁺22] Yifei Wang, Qi Zhang, Yisen Wang, Jiansheng Yang, and Zhouchen Lin, *Chaos is a ladder: A new theoretical understanding of contrastive learning via augmentation overlap*, arXiv preprint arXiv:2203.13457 (2022).
- [XZ24] Xiangxiang Xu and Lizhong Zheng, *Dependence induced representations*, 2024 60th Annual Allerton Conference on Communication, Control, and Computing, IEEE, 2024, pp. 1–8.
- [ZIE16] Richard Zhang, Phillip Isola, and Alexei A Efros, *Colorful image colorization*, Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part III 14, Springer, 2016, pp. 649–666.
- [ZMKB23] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer, *Sigmoid loss for language image pre-training*, Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 11975–11986.

- [ZSS⁺21] Roland S Zimmermann, Yash Sharma, Steffen Schneider, Matthias Bethge, and Wieland Brendel, *Contrastive learning inverts the data generating process*, International Conference on Machine Learning, PMLR, 2021, pp. 12979–12990.

Contents

1	Introduction	1
2	Related work	2
3	Approximate sufficient statistics	3
4	Statistical properties of contrastive learning	4
4.1	Setup and the ERM estimator	4
4.2	Using the encoder for downstream tasks	6
4.3	General f-contrastive learning	7
4.3.1	Finding encoders with low f-sufficiency	7
4.3.2	Implications of low f-Sufficiency	8
5	Examples	8
5.1	Linear regression	8
5.1.1	A concrete scenario	9
5.2	Topic classification	10
6	Experiments	11
7	Conclusion	12
A	Properties of approximate sufficient statistics	18
A.1	Equivalence in Definition 1	18
A.2	Properties and examples	19
A.3	Sufficiency of similarity scores	21
B	Proofs in Section 4	22
B.1	Proof of Eq. (6)	22
B.2	Proof of Theorem 1	22
B.3	Proof of Theorem 2	26
B.4	Proof of Theorem 3	27
B.5	Proof of Theorem 4	28
C	Proofs in Section 5	30
C.1	Proof of Theorem 6	30
C.2	Proof of Corollary 1	30
C.3	An end-to-end result on downstream linear regression	31
C.4	Proof of Theorem 7	34
C.4.1	Proof of Eq. (14)	34
C.4.2	Proof of Eq. (15)	36
C.5	An auxiliary lemma	38
D	Additional experiments	40

A Properties of approximate sufficient statistics

In this section, we discuss some properties of approximate sufficient statistics introduced in Definition 1 and provide some concrete examples.

A.1 Equivalence in Definition 1

Lemma 1 (Equivalence of three forms of sufficiency). *The ILS, VFS, CBS definitions in Definition 1 are equivalent, i.e., for any statistic T*

$$\text{Suff}_{\text{il},f}(T) = \text{Suff}_{\text{vf},f}(T) = \text{Suff}_{\text{cb},f}(T) =: \text{Suff}_f(T).$$

Proof of Lemma 1. **(ILS) \Leftrightarrow (VFS).** Note that by the variational form of f -divergence, we have

$$\begin{aligned} & -I_f(X, Y) \\ &= \inf_{S: \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}} \mathbb{E}_{\mathbb{P}(x,y)}[-S(x,y)] + \mathbb{E}_{\mathbb{P}(x)\mathbb{P}(y)}[f^*(S(x,y))] \\ &= \inf_{S_x: \mathcal{X} \rightarrow \mathbb{R}, S: \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}} \mathbb{E}_{\mathbb{P}(x,y)}[S_x(x) - S(x,y)] + \mathbb{E}_{\mathbb{P}(x)\mathbb{P}(y)}[f^*(S(x,y) - S_x)] \\ &= \inf_{S: \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}} \mathbb{E}_{\mathbb{P}(x,y)}[-S(x,y)] + \inf_{S_x: \mathcal{X} \rightarrow \mathbb{R}} \mathbb{E}_{\mathbb{P}(x)\mathbb{P}(y)}[f^*(S(x,y) - S_x) + S_x(x)] = \inf_{S: \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}} R_f(S). \end{aligned}$$

Similarly,

$$\begin{aligned} & -I_f(T(X), Y) \\ &= \inf_{S: T(\mathcal{X}) \times \mathcal{Y} \rightarrow \mathbb{R}} \mathbb{E}_{\mathbb{P}(T(x),y)}[-S(T(x),y)] + \mathbb{E}_{\mathbb{P}(T(x))\mathbb{P}(y)}[f^*(S(T(x),y))] \\ &= \inf_{S: T(\mathcal{X}) \times \mathcal{Y} \rightarrow \mathbb{R}} \mathbb{E}_{\mathbb{P}(T(x),y)}[-S(T(x),y)] + \inf_{S_x: T(\mathcal{X}) \rightarrow \mathbb{R}} \mathbb{E}_{\mathbb{P}(T(x))\mathbb{P}(y)}[f^*(S(T(x),y) - S_x(T(x))) + S_x(T(x))] \\ &= \inf_{S: T(\mathcal{X}) \times \mathcal{Y} \rightarrow \mathbb{R}} \mathbb{E}_{\mathbb{P}(x,y)}[-S(T(x),y)] + \inf_{S_x: T(\mathcal{X}) \rightarrow \mathbb{R}} \mathbb{E}_{\mathbb{P}(x)\mathbb{P}(y)}[f^*(S(T(x),y) - S_x(T(x))) + S_x(T(x))] \\ &= \inf_{S: T(\mathcal{X}) \times \mathcal{Y} \rightarrow \mathbb{R}} R_f(S \circ T). \end{aligned}$$

Combining the two results yields the equivalence between **(ILS)** and **(VFS)**.

(ILS) \Leftrightarrow (CBS). By definition of the **(ILS)**

$$\begin{aligned} \text{Suff}_{\text{il},f}(T) &= I_f(X, Y) - I_f(T(X), Y) \\ &= \int f\left(\frac{\mathbb{P}(x,y)}{\mathbb{P}(x)\mathbb{P}(y)}\right) \mathbb{P}(x)\mathbb{P}(y)d\mu - \int f\left(\frac{\mathbb{P}(T(x),y)}{\mathbb{P}(T(x))\mathbb{P}(y)}\right) \mathbb{P}(T(x))\mathbb{P}(y)d\mu \\ &= \int f\left(\frac{\mathbb{P}(y|x)}{\mathbb{P}(y)}\right) \mathbb{P}(x)\mathbb{P}(y)d\mu - \int f\left(\frac{\mathbb{P}(y|T(x))}{\mathbb{P}(y)}\right) \mathbb{P}(x)\mathbb{P}(y)d\mu \\ &= \mathbb{E}_{\mathbb{P}(x)\mathbb{P}(y)}\left[f\left(\frac{\mathbb{P}(y|x)}{\mathbb{P}(y)}\right) - f\left(\frac{\mathbb{P}(y|T(x))}{\mathbb{P}(y)}\right)\right] = \text{Suff}_{\text{cb},f}(T), \end{aligned}$$

where the last equality follows since

$$\begin{aligned} & \mathbb{E}_{\mathbb{P}(x)\mathbb{P}(y)}\left[f'\left(\frac{\mathbb{P}(y|T(x))}{\mathbb{P}(y)}\right)\left(\frac{\mathbb{P}(y|x)}{\mathbb{P}(y)} - \frac{\mathbb{P}(y|T(x))}{\mathbb{P}(y)}\right)\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[f'\left(\frac{\mathbb{P}(y|T(x))}{\mathbb{P}(y)}\right)\left(\frac{\mathbb{P}(y|x)}{\mathbb{P}(y)} - \frac{\mathbb{P}(y|T(x))}{\mathbb{P}(y)}\right) \middle| T(x)\right]\right] \\ &= \mathbb{E}\left[\frac{1}{\mathbb{P}(y)}f'\left(\frac{\mathbb{P}(y|T(x))}{\mathbb{P}(y)}\right)\left[\mathbb{E}[\mathbb{P}(y|x)|T(x)] - \mathbb{P}(y|T(x))\right]\right] = 0. \end{aligned} \tag{16}$$

An equivalent expression of (CBS). We now show that

$$\mathbb{E}_{\mathbb{P}(x) \times \mathbb{P}(y)}\left[B_f\left(\frac{\mathbb{P}(y|x)}{\mathbb{P}(y)}, \frac{\mathbb{P}(y|T(x))}{\mathbb{P}(y)}\right)\right] = \inf_{\mathbb{Q}: T(\mathcal{X}) \rightarrow \Delta(\mathcal{Y})} \mathbb{E}_{\mathbb{P}(x) \times \mathbb{P}(y)}\left[B_f\left(\frac{\mathbb{P}(y|x)}{\mathbb{P}(y)}, \frac{\mathbb{Q}(y|T(x))}{\mathbb{P}(y)}\right)\right].$$

This follows immediately as for any $\mathbb{Q} : T(\mathcal{X}) \mapsto \Delta(\mathcal{Y})$

$$\begin{aligned} & \mathbb{E}_{\mathbb{P}(x) \times \mathbb{P}(y)} \left[B_f \left(\frac{\mathbb{P}(y|x)}{\mathbb{P}(y)}, \frac{\mathbb{Q}(y|T(x))}{\mathbb{P}(y)} \right) \right] - \mathbb{E}_{\mathbb{P}(x) \times \mathbb{P}(y)} \left[B_f \left(\frac{\mathbb{P}(y|x)}{\mathbb{P}(y)}, \frac{\mathbb{P}(y|T(x))}{\mathbb{P}(y)} \right) \right] \\ &= \mathbb{E}_{\mathbb{P}(x) \times \mathbb{P}(y)} \left[f \left(\frac{\mathbb{P}(y|T(x))}{\mathbb{P}(y)} \right) - f \left(\frac{\mathbb{Q}(y|T(x))}{\mathbb{P}(y)} \right) - f' \left(\frac{\mathbb{Q}(y|T(x))}{\mathbb{P}(y)} \right) \left(\frac{\mathbb{P}(y|x)}{\mathbb{P}(y)} - \frac{\mathbb{Q}(y|T(x))}{\mathbb{P}(y)} \right) \right] \\ &\geq \mathbb{E}_{\mathbb{P}(x) \times \mathbb{P}(y)} \left[f' \left(\frac{\mathbb{Q}(y|T(x))}{\mathbb{P}(y)} \right) \left(\frac{\mathbb{P}(y|T(x))}{\mathbb{P}(y)} - \frac{\mathbb{P}(y|x)}{\mathbb{P}(y)} \right) \right] = 0, \end{aligned}$$

where the first equality uses Eq. (16). \square

A.2 Properties and examples

Lemma 2 (Global minimizers of $R_f(S)$). *Recall*

$$R_f(S) = \mathbb{E}_{\mathbb{P}(x,y)}[-S(x,y)] + \inf_{S_x : \mathcal{X} \mapsto \mathbb{R}} \mathbb{E}_{\mathbb{P}(x)\mathbb{P}(y)}[f^*(S(x,y) - S_x(x)) + S_x(x)].$$

For f that is strictly convex and differentiable, the following results hold for $R_f(\cdot)$.

- (1). The infimum in the definition of $R_f(\cdot)$ is obtained by $S_x(x)$ such that $\mathbb{E}_{\mathbb{P}(y)}[(f')^{-1}(S(x,y) - S_x(x))] = 1$ for all x .
- (2). Let $S_*(x,y) := f' \left(\frac{\mathbb{P}(x,y)}{\mathbb{P}(x)\mathbb{P}(y)} \right)$. The global minimizers of $R_f(\cdot)$ form the set

$$\mathcal{M}_f := \left\{ S : \mathcal{X} \times \mathcal{Y} \mapsto \mathbb{R}, S(x,y) = S_*(x,y) + S_x(x) \text{ for some } S_x : \mathcal{X} \mapsto \mathbb{R} \right\}.$$

Proof of Lemma 2. For any fixed x , we have

$$\nabla_c \mathbb{E}_{\mathbb{P}(y)}[f^*(S(x,y) - c) + c] = \mathbb{E}_{\mathbb{P}(y)}[-\nabla f^*(S(x,y) - c) + 1]$$

Claim (1) follows immediately from setting the derivative equal to zero and noting that $\nabla f^* = (f')^{-1}$.

To prove claim (2), we first note that adding any function $S_x(x)$ to $S(x,y)$ does not change the value of $R_f(S)$ due to the infimum inside the definition of $R_f(S)$. Therefore, it suffices to show that the unique minimizer of

$$\bar{R}_f(S) := \mathbb{E}_{\mathbb{P}(x,y)}[-S(x,y)] + \mathbb{E}_{\mathbb{P}(x)\mathbb{P}(y)}[f^*(S(x,y))]$$

is $S_* = f' \left(\frac{\mathbb{P}(x,y)}{\mathbb{P}(x)\mathbb{P}(y)} \right)$. Write $S = S_* + ch$. It can be verified that $\bar{R}_f(S_* + ch)$ is strictly convex in c . Thus S_* is the unique minimizer of \bar{R}_f if $\nabla_c \bar{R}_f(S_* + ch)|_{c=0} = 0$ for all h . This is true since

$$\begin{aligned} \nabla_c \bar{R}_f(S_* + ch)|_{c=0} &= \mathbb{E}_{\mathbb{P}(x,y)}[-h(x,y)] + \mathbb{E}_{\mathbb{P}(x)\mathbb{P}(y)}[\nabla f^*(S(x,y))h(x,y)] \\ &= \mathbb{E}_{\mathbb{P}(x,y)}[-h(x,y)] + \mathbb{E}_{\mathbb{P}(x)\mathbb{P}(y)} \left[\frac{\mathbb{P}(x,y)}{\mathbb{P}(x)\mathbb{P}(y)} h(x,y) \right] = 0, \end{aligned}$$

where the second inequality uses the property of convex conjugates that $\nabla f^*(f'(x)) = x$. \square

Lemma 3 (A general bound on $D_{TV}(\mathbb{P}(y|x)||\mathbb{P}(y|T(x)))$ based on sufficiency.). *For f in Definition 1 that is twice continuously differentiable, and for any statistic T , we have*

$$\mathbb{E}_{\mathbb{P}(x)}[D_{TV}(\mathbb{P}(y|x)||\mathbb{P}(y|T(x)))] \leq c_2 \cdot \sqrt{\text{Suff}_{cb,f}(T)}, \quad (17)$$

where $c_2 := \left(2 \inf_{(x,y) \in \text{supp}(x,y)} f'' \left(\frac{\mathbb{P}(y|x)}{\mathbb{P}(y)} \right) \right)^{-1/2}$, and $\text{supp}(x,y)$ denotes the support of $\mathbb{P}(x) \times \mathbb{P}(y)$. Notably, when $f(x) = (x-1)^2/2$ (χ^2 -divergence), we have $c_2 = 1/\sqrt{2}$.

Proof of Lemma 3. Using the CBS form of sufficiency, we find that

$$\begin{aligned}\text{Suff}(T) &= \mathbb{E}_{\mathbb{P}(x) \times \mathbb{P}(y)} \left[B_f \left(\frac{\mathbb{P}(y|x)}{\mathbb{P}(y)}, \frac{\mathbb{P}(y|T(x))}{\mathbb{P}(y)} \right) \right] \\ &\geq \frac{1}{2} \mathbb{E}_{\mathbb{P}(x) \times \mathbb{P}(y)} \left[f'' \left(\frac{\mathbb{P}(y|x)}{\mathbb{P}(y)} \right) \cdot \left[\frac{\mathbb{P}(y|x)}{\mathbb{P}(y)} - \frac{\mathbb{P}(y|T(x))}{\mathbb{P}(y)} \right]^2 \right] \\ &\geq \frac{1}{2} \inf_{(x,y) \in \text{supp}(x,y)} f'' \left(\frac{\mathbb{P}(y|x)}{\mathbb{P}(y)} \right) \cdot \mathbb{E}_{\mathbb{P}(x) \times \mathbb{P}(y)} \left[\left[\frac{\mathbb{P}(y|x)}{\mathbb{P}(y)} - \frac{\mathbb{P}(y|T(x))}{\mathbb{P}(y)} \right]^2 \right],\end{aligned}$$

where the first inequality follows from the definition of Bregman divergence and the fact that the range of $\mathbb{P}(y|T(x))$ belongs to the range of $\mathbb{P}(y|x)$. Moreover, by Jensen's inequality, we have

$$\begin{aligned}\left(\mathbb{E}_{\mathbb{P}(x) \times \mathbb{P}(y)} \left[\left[\frac{\mathbb{P}(y|x)}{\mathbb{P}(y)} - \frac{\mathbb{P}(y|T(x))}{\mathbb{P}(y)} \right]^2 \right] \right)^{1/2} &\geq \mathbb{E}_{\mathbb{P}(x) \times \mathbb{P}(y)} \left[\left| \frac{\mathbb{P}(y|x)}{\mathbb{P}(y)} - \frac{\mathbb{P}(y|T(x))}{\mathbb{P}(y)} \right| \right] \\ &= 2\mathbb{E}_{\mathbb{P}(x)} [\mathsf{D}_{\text{TV}}(\mathbb{P}(y|x) || \mathbb{P}(y|T(x)))].\end{aligned}$$

Putting pieces together yields Lemma 3. \square

Example 1 (KL-sufficiency). Take $f(x) = x \log x$ (KL-divergence), then we have

$$\begin{aligned}\text{Suff}_{\text{cb},f}(T) &= \mathbb{E}_{\mathbb{P}(x)} \left[\mathsf{D}_{\text{KL}}(\mathbb{P}(y|x) || \mathbb{P}(y|T(x))) \right], \quad \text{and} \\ R_f(S) &= \mathbb{E}_{\mathbb{P}(x,y)} [-S(x,y)] + \mathbb{E}_{\mathbb{P}(x)} [\log \mathbb{E}_{\mathbb{P}(y)} [\exp(S(x,y))]].\end{aligned}$$

It can be verified that the InfoNCE loss in Eq. (2) is an asymptotically unbiased estimate of $R_f(S)$ as the batch size $K \rightarrow \infty$ (see Eq. 6). Moreover, by Pinsker's inequality

$$\mathbb{E}_{\mathbb{P}(x)} [\mathsf{D}_{\text{TV}}(\mathbb{P}(y|x) || \mathbb{P}(y|T(x)))] \leq \frac{1}{\sqrt{2}} \cdot \sqrt{\text{Suff}_{\text{cb},\text{kl}}(T)}.$$

Example 2 (Chi-sufficiency). Take $f(x) = (x-1)^2/2$ (χ^2 -divergence), then we have

$$\begin{aligned}\text{Suff}_{\text{cb},f}(T) &= \mathbb{E}_{\mathbb{P}(x) \times \mathbb{P}(y)} \left[\frac{1}{2} \left(\frac{\mathbb{P}(y|x)}{\mathbb{P}(y)} - \frac{\mathbb{P}(y|T(x))}{\mathbb{P}(y)} \right)^2 \right], \\ R_f(S) &= \mathbb{E}_{\mathbb{P}(x,y)} [-S(x,y)] + \mathbb{E}_{\mathbb{P}(x)\mathbb{P}(y)} [(S(x,y) - \mathbb{E}_{\mathbb{P}(y)} [S(x,y)])^2/2 + S(x,y)].\end{aligned}$$

Lemma 3 gives

$$\mathbb{E}_{\mathbb{P}(x)} [\mathsf{D}_{\text{TV}}(\mathbb{P}(y|x) || \mathbb{P}(y|T(x)))] \leq \frac{1}{\sqrt{2}} \sqrt{\text{Suff}_{\text{cb},\chi^2}(T)}.$$

Also, we can bound the χ^2 -divergence by the sufficiency:

$$\mathbb{E}_{\mathbb{P}(x)} \chi^2(\mathbb{P}(y|x) || \mathbb{P}(y|T(x))) \leq \text{Suff}_{\text{cb},f}(T) \cdot \left[2 \sup_{(x,y) \in \text{supp}(x,y)} \frac{\mathbb{P}(T(x))\mathbb{P}(y)}{\mathbb{P}(T(x),y)} \right].$$

Example 3 (Squared Hellinger-sufficiency). Take $f(x) = 1 - \sqrt{x}$, then we have $f^*(x) = -1 - \frac{1}{4x}$ for $x < 0$, and

$$\text{Suff}_{\text{cb},f}(T) = \mathbb{E}_{\mathbb{P}(x)} [H^2(\mathbb{P}(y) || \mathbb{P}(y|x)) - H^2(\mathbb{P}(y) || \mathbb{P}(y|T(x)))],$$

where $H^2(p||q) := \int (\sqrt{p(x)} - \sqrt{q(x)})^2 dx/2$ is the squared Hellinger distance. Similarly, the squared Hellinger distance between $\mathbb{P}(y|x), \mathbb{P}(y|T(x))$ can be bounded by the sufficiency of T :

$$\begin{aligned}\mathbb{E}_{\mathbb{P}(x)} [H^2(\mathbb{P}(y|x) || \mathbb{P}(y|T(x)))] &= \frac{1}{2} \mathbb{E}_{\mathbb{P}(x)} \left[\sum_y (\sqrt{\mathbb{P}(y|x)} - \sqrt{\mathbb{P}(y|T(x))})^2 \right] \\ &\leq \left[\sup_{(x,y) \in \text{supp}(x,y)} \sqrt{\frac{\mathbb{P}(T(x),y)}{\mathbb{P}(T(x))\mathbb{P}(y)}} \right] \cdot \mathbb{E}_{\mathbb{P}(x)} \left[\sum_y \sqrt{\mathbb{P}(y)} \frac{(\sqrt{\mathbb{P}(y|T(x))} - \sqrt{\mathbb{P}(y|x)})^2}{2\sqrt{\mathbb{P}(y|T(x))}} \right] \\ &= \left[\sup_{(x,y) \in \text{supp}(x,y)} \sqrt{\frac{\mathbb{P}(T(x),y)}{\mathbb{P}(T(x))\mathbb{P}(y)}} \right] \cdot \text{Suff}_{\text{cb},f}(T),\end{aligned}$$

where the last equality follows from

$$\begin{aligned} & \mathbb{E}\left[\sqrt{\mathbb{P}(y|T(x))} - \sqrt{\mathbb{P}(y|x)} \mid y, T(x)\right] \\ &= \mathbb{E}\left[\left(\sqrt{\mathbb{P}(y|T(x))} - \sqrt{\mathbb{P}(y|x)}\right) \cdot \frac{\sqrt{\mathbb{P}(y|T(x))} - \sqrt{\mathbb{P}(y|x)}}{2\sqrt{\mathbb{P}(y|T(x))}} \mid y, T(x)\right] + \mathbb{E}\left[\frac{\mathbb{P}(y|T(x)) - \mathbb{P}(y|x)}{2\sqrt{\mathbb{P}(y|T(x))}} \mid y, T(x)\right] \\ &= \mathbb{E}\left[\frac{(\sqrt{\mathbb{P}(y|T(x))} - \sqrt{\mathbb{P}(y|x)})^2}{2\sqrt{\mathbb{P}(y|T(x))}} \mid y, T(x)\right]. \end{aligned}$$

A.3 Sufficiency of similarity scores

The definition of approximate sufficiency can be extended to score functions $S : \mathcal{X} \times \mathcal{Y} \mapsto \mathbb{R}$ that measure the similarity between (X, Y) .

Definition 2 (Approximate sufficient score functions). *Let $S : \mathcal{X} \times \mathcal{Y} \mapsto \mathbb{R}$ be a similarity score function. It induces a conditional density \mathbb{P}_S on $\mathcal{X} \times \mathcal{Y}$ w.r.t. the base measure μ via*

$$\mathbb{P}_S(y|x) = \mathbb{P}(y)(f')^{-1}(\bar{S}(x, y)),$$

where $\bar{S}(x, y) = S(x, y) - S_x(x)$ such that $\mathbb{E}_{\mathbb{P}(y)}[(f')^{-1}\bar{S}(x, y)] = 1$ for all x . We define the sufficiency of S in two equivalent forms:

- **Variational Form Sufficiency (VFS):** The variational form sufficiency of T is given by

$$\text{Suff}_{vf,f}(S) = R_f(S) - \inf_{\tilde{S} : \mathcal{X} \times \mathcal{Y} \mapsto \mathbb{R}} R_f(\tilde{S}),$$

and the f-contrastive loss

$$R_f(S) := \mathbb{E}_{\mathbb{P}(x,y)}[-S(x, y)] + \inf_{S_x : \mathcal{X} \mapsto \mathbb{R}} \mathbb{E}_{\mathbb{P}(x)\mathbb{P}(y)}[f^*(S(x, y) - S_x(x)) + S_x(x)], \quad (18)$$

where f^* is the Fenchel-dual of f .

- **Conditional Bregman Sufficiency (CBS):** The conditional Bregman sufficiency of T is defined as

$$\text{Suff}_{cb,f}(S) = \mathbb{E}_{\mathbb{P}(x) \times \mathbb{P}(y)}\left[B_f\left(\frac{\mathbb{P}(y|x)}{\mathbb{P}(y)}, \frac{\mathbb{P}_S(y|x)}{\mathbb{P}(y)}\right)\right],$$

where $B_f(a, b) := f(a) - f(b) - (a - b)f'(b)$ is the Bregman divergence of f .

Note that the excess risk of the contrastive loss equals the sufficiency of S under our definition. Similar to Definition 1, we have

Lemma 4 (Equivalence of two forms of score sufficiency). *For any similarity score $S : \mathcal{X} \times \mathcal{Y} \mapsto \mathbb{R}$, the three forms of sufficiency in Definition 2 are equivalent, i.e.,*

$$\text{Suff}_{vf,f}(S) = \text{Suff}_{cb,f}(S) =: \text{Suff}_f(S).$$

Proof of Lemma 4. **(VFS) \Leftrightarrow (CBS).** Let $S_*(x, y) = f'(\frac{\mathbb{P}(x,y)}{\mathbb{P}(x)\mathbb{P}(y)})$. We have by Lemma 2 that $S_* \in \arg\min_{\tilde{S}} R_f(\tilde{S})$. By the definition of the (VFS), we have

$$\begin{aligned} \text{Suff}_{vf,f}(S) &= R_f(S) - R_f(S_*) \\ &= \mathbb{E}_{\mathbb{P}(x,y)}[S_* - \bar{S}(x, y)] + \mathbb{E}_{\mathbb{P}(x)\mathbb{P}(y)}[f^*(\bar{S}(x, y)) - f^*(S_*(x, y))] \\ &\stackrel{(i)}{=} \mathbb{E}_{\mathbb{P}(x,y)}[S_* - \bar{S}(x, y)] + \mathbb{E}_{\mathbb{P}(x)\mathbb{P}(y)}\left[f\left(\frac{\mathbb{P}(x,y)}{\mathbb{P}(x)\mathbb{P}(y)}\right) - \frac{\mathbb{P}(x,y)}{\mathbb{P}(x)\mathbb{P}(y)}S_*(x, y)\right] \\ &= -\mathbb{E}_{\mathbb{P}(x,y)}[\bar{S}(x, y)] + \mathbb{E}_{\mathbb{P}(x)\mathbb{P}(y)}[f^*(\bar{S}(x, y))] + \mathbb{E}_{\mathbb{P}(x)\mathbb{P}(y)}\left[f\left(\frac{\mathbb{P}(x,y)}{\mathbb{P}(x)\mathbb{P}(y)}\right)\right] \\ &\stackrel{(ii)}{=} -\mathbb{E}_{\mathbb{P}(x,y)}[\bar{S}(x, y)] + \mathbb{E}_{\mathbb{P}(x)\mathbb{P}(y)}\left[f\left(\frac{\mathbb{P}(x,y)}{\mathbb{P}(x)\mathbb{P}(y)}\right) + \frac{\mathbb{P}_S(y|x)}{\mathbb{P}(y)}\bar{S}(x, y) - f\left(\frac{\mathbb{P}_S(y|x)}{\mathbb{P}(y)}\right)\right] \\ &= \mathbb{E}_{\mathbb{P}(x)\mathbb{P}(y)}\left[f\left(\frac{\mathbb{P}(y|x)}{\mathbb{P}(y)}\right) - f\left(\frac{\mathbb{P}_S(y|x)}{\mathbb{P}(y)}\right)\right] - \mathbb{E}_{\mathbb{P}(x) \times \mathbb{P}(y)}\left[\bar{S}(x, y)\left(\frac{\mathbb{P}(y|x)}{\mathbb{P}(y)} - \frac{\mathbb{P}_S(y|x)}{\mathbb{P}(y)}\right)\right], \end{aligned}$$

where step (i) and (ii) use $f((f')^{-1}(z)) + f^*(z) = z(f')^{-1}(z)$ with $z = S_*(x, y)$ and $\bar{S}(x, y)$, respectively. Since $\bar{S}(x, y) = f'(\frac{\mathbb{P}_S(y|x)}{\mathbb{P}(y)})$, it follows immediately that $\text{Suff}_{\text{vf}, f}(S) = \text{Suff}_{\text{cb}, f}(S)$. \square

Example 4. Take $f(x) = x \log x$ (*KL-divergence*). Then $S_*(x, y) = \log(\mathbb{P}(x, y)/[\mathbb{P}(x)\mathbb{P}(y)])$, $B_f(a, b) = a \log(a/b) - (a - b)$, and $\mathbb{P}_S(y|x) = \mathbb{P}(y) \exp(S(x, y))/\mathbb{E}_{\mathbb{P}(y)}[\exp(S(x, y))]$. Also, we have

$$\begin{aligned} \text{Suff}_{\text{kl}}(S) &= R_f(S) - R_f(S_*) = \int \mathbb{P}(y|x) \log \left(\frac{\mathbb{P}(y|x)}{\mathbb{P}_S(y|x)} \right) - (\mathbb{P}(y|x) - \mathbb{P}_S(y|x)) \mathbb{P}(x) dy dx \\ &= \mathbb{E}_{x \sim \mathbb{P}(x)} [\text{D}_{\text{KL}}(\mathbb{P}(y|x) \parallel \mathbb{P}_S(y|x))]. \end{aligned}$$

Example 5. Take $f(x) = (x - 1)^2/2$ (χ^2 -divergence). Then $S_*(x, y) = \mathbb{P}(x, y)/[\mathbb{P}(x)\mathbb{P}(y)] - 1$, $B_f(a, b) = (a - b)^2/2$, and $\mathbb{P}_S(y|x) = \mathbb{P}(y)(S(x, y) - \mathbb{E}_y[S(x, y)]) + 1$. Moreover,

$$\begin{aligned} \text{Suff}_{\chi^2}(S) &= R_f(S) - R_f(S_*) = \frac{1}{2} \mathbb{E}_{\mathbb{P}(x) \times \mathbb{P}(y)} \left[\frac{(\mathbb{P}(y|x) - \mathbb{P}_S(y|x))^2}{\mathbb{P}(y)^2} \right] \\ &= \frac{1}{2} \mathbb{E}_{\mathbb{P}(x)} \sum_y \left[\frac{(\mathbb{P}(y|x) - \mathbb{P}_S(y|x))^2}{\mathbb{P}(y|x)} \cdot \frac{\mathbb{P}(y|x)}{\mathbb{P}(y)} \right] \\ &\geq \inf_{(x,y) \in \text{supp}(x,y)} \frac{\mathbb{P}(x,y)}{\mathbb{P}(x)\mathbb{P}(y)} \cdot \mathbb{E}_{\mathbb{P}(x)} [\chi^2(\mathbb{P}(y|x) \parallel \mathbb{P}_S(y|x))]. \end{aligned}$$

B Proofs in Section 4

B.1 Proof of Eq. (6)

As given in Example 1 (which can be established using Lemma 2), the KL-contrastive loss has the form

$$R_{\text{kl}}(S) = \mathbb{E}_{(\mathbf{z}^{(1)}, \mathbf{z}^{(2)})} [-S(\mathbf{z}^{(1)}, \mathbf{z}^{(2)})] + \mathbb{E}_{\mathbf{z}^{(1)} \sim \mathbb{P}_z} [\log \mathbb{E}_{\mathbf{z}^{(2)} \sim \mathbb{P}_z} [\exp(S(\mathbf{z}^{(1)}, \mathbf{z}^{(2)}))]].$$

Recall the SimCLR loss $\bar{R}_{\text{simclr}, K}(S)$ in Eq. (2). We then have

$$\begin{aligned} &\lim_{K \rightarrow \infty} \bar{R}_{\text{simclr}, K}(S) - \log K \\ &= \frac{1}{2} \lim_{K \rightarrow \infty} \mathbb{E} \left[-\log \frac{\exp(S(\mathbf{z}_1^{(1)}, \mathbf{z}_1^{(2)}))}{\sum_{j \in [K]} \exp(S(\mathbf{z}_1^{(1)}, \mathbf{z}_j^{(2)}))/K} \right] + \frac{1}{2} \mathbb{E} \left[-\log \frac{\exp(S(\mathbf{z}_1^{(1)}, \mathbf{z}_1^{(2)}))}{\sum_{j \in [K]} \exp(S(\mathbf{z}_j^{(1)}, \mathbf{z}_1^{(2)}))/K} \right] \\ &= \lim_{K \rightarrow \infty} \mathbb{E} \left[\log \sum_{j \in [K]} \exp(S(\mathbf{z}_1^{(1)}, \mathbf{z}_j^{(2)}))/K \right] - \mathbb{E}[\exp(S(\mathbf{z}_1^{(1)}, \mathbf{z}_1^{(2)}))] = R_{\text{kl}}(S), \end{aligned}$$

where the second equality follows from the symmetry of S in its arguments and the last equality uses the law of large numbers (note that $\mathbf{z}_1^{(1)}$ is independent of $\mathbf{z}_j^{(2)}$ for $j \neq 1$) and the bounded convergence theorem.

B.2 Proof of Theorem 1

We begin the proof by stating the following proposition that connects the excess risk with sufficiency.

Proposition 8 (Near-minimizers of SimCLR as near-sufficient statistics; Proposition 1 in [OLCM25]). *Suppose Assumption 1 holds and S_* is a global minimizer of $\bar{R}_{\text{simclr}, K}(S)$ as defined in Section 4.1. Then, there exists a constant $C > 0$, which depends polynomially on B_S , such that for any function $f \in \mathcal{F}$, its sufficiency can be bounded by its SimCLR excess risk. Namely, for any $K \geq 2$, we have*

$$\text{Suff}(f) \leq \lim_{K' \rightarrow \infty} \left[\bar{R}_{\text{simclr}, K'}(S_f) - \bar{R}_{\text{simclr}, K'}(S_*) \right] \leq \underbrace{\left[\bar{R}_{\text{simclr}, K}(S_f) - \bar{R}_{\text{simclr}, K}(S_*) \right]}_{\text{SimCLR excess risk}} \cdot \left(1 + \frac{C}{K} \right).$$

A similar version of this result has been established for contrastive language-image pretraining (CLIP) in Proposition 1 in [OLCM25]. The proof of Proposition 8 follows immediately from the proof of Proposition 1 in [OLCM25] as the SimCLR setup can be viewed as a special case of CLIP in which the text and image follows a symmetric distribution conditioned on their shared information.

Adopt the shorthand notation \bar{R}_K for $\bar{R}_{\text{simclr}, K}$. With Proposition 8 at hand, we obtain the following decomposition for some $C > 0$ polynomially dependent on B_S

$$\begin{aligned} \text{Suff}(\hat{f}) &\leq \left[\bar{R}_K(S_{\hat{f}}) - \bar{R}_K(S_{\star}) \right] \cdot \left(1 + \frac{C}{K} \right) \\ &= \left[[\bar{R}_K(S_{\hat{f}}) - \inf_{f \in \mathcal{F}} \bar{R}_K(S_f)] + [\inf_{f \in \mathcal{F}} \bar{R}_K(S_f) - \bar{R}_K(S_{\star})] \right] \cdot \left(1 + \frac{C}{K} \right) \\ &\leq \underbrace{\left[\bar{R}_K(S_{\hat{f}}) - \inf_{f \in \mathcal{F}} \bar{R}_K(S_f) \right]}_{\text{generalization error}} \cdot \left(1 + \frac{C}{K} \right) + \underbrace{\left[\inf_{f \in \mathcal{F}} \bar{R}_K(S_f) - \bar{R}_K(S_{\star}) \right]}_{\text{approximation error}} \cdot \left(1 + \frac{C}{K} \right). \end{aligned}$$

Therefore, it remains to prove the following bound.

(1). With probability at least $1 - \delta$, the excess risk

$$\bar{R}_K(S_{\hat{f}}) - \inf_{f \in \mathcal{F}} \bar{R}_K(S_f) \leq \frac{C}{\sqrt{n}} \left[\sqrt{\log(1/\delta)} + B_{\tau}^2 \int_0^{2(\log B_S + B_{\tau})} \sqrt{\log \mathcal{N}(u, \|\cdot\|_{2,\infty}, \mathcal{F})} du \right] \quad (19)$$

for some constant $C > 0$ that is polynomially dependent on B_S .

Proof of Eq. (19). Recall the definition of $\hat{R}_{\text{simclr}, K}$ in Eq. (3) and adopt the shorthand \hat{R}_K for $\hat{R}_{\text{simclr}, K}$. Let $B_f := \sqrt{B_{\tau}(\log B_S + B_{\tau})}$, $B := c(B_S^6 + 1)B_f B_{\tau}$ for some absolute constant $c > 0$. It can be verified by Assumption 2 that \mathcal{F} must satisfy $\|f\|_{2,\infty} \leq B_f$ for all $f \in \mathcal{F}$ for Assumption 1 to hold. Define the zero-mean random process $X_f := \hat{R}_K(S_f) - \mathbb{E}[\hat{R}_K(S_f)]$, $f \in \mathcal{F}$. We will show that

$$\mathbb{P}\left(\left| \sup_{f \in \mathcal{F}} |X_f| - \mathbb{E}\left[\sup_{f \in \mathcal{F}} |X_f|\right] \right| \geq t\right) \leq 2 \exp\left(-\frac{2nt^2}{9B_S^4}\right), \quad \text{for all } t \geq 0, \quad \text{and} \quad (20a)$$

$$\begin{aligned} \mathbb{E}\left[\sup_{f \in \mathcal{F}} |X_f|\right] &\leq \mathbb{E}[|X_{f_0}|] + \mathbb{E}\left[\sup_{f, \tilde{f} \in \mathcal{F}} |X_f - X_{\tilde{f}}|\right] \\ &\leq c \frac{B_S^2}{\sqrt{n}} + 32 \frac{B}{\sqrt{n}} \cdot \int_0^{2B_f} \sqrt{\log \mathcal{N}(u, \|\cdot\|_{2,\infty}, \mathcal{F})} du \end{aligned} \quad (20b)$$

for any $f_0 \in \mathcal{F}$ and some absolute constant $c > 0$. Combining the two bounds and noting

$$\bar{R}_K(S_{\hat{f}}) - \inf_{f \in \mathcal{F}} \bar{R}_K(S_f) \leq 2 \sup_{f \in \mathcal{F}} |\hat{R}_K(S_f) - \bar{R}_K(S_f)| = 2 \sup_{f \in \mathcal{F}} |\hat{R}_K(S_f) - \mathbb{E}[\hat{R}_K(S_f)]| = 2 \sup_{f \in \mathcal{F}} |X_f| \quad (21)$$

yields claim (1).

Proof of Eq. (20a). Let $\bar{z}_i = (z_i^{(1)}, z_i^{(2)})$. Then $\{\bar{z}_i\}_{i=1}^n$ are i.i.d. pairs of augmented views. For any $i \in [n_1], j \in [K]$, suppose $\bar{z}_{(i-1)K+j}$ is replaced by some alternative sample $\tilde{z}_{(i-1)K+j} = (\tilde{z}_{(i-1)K+j}^{(1)}, \tilde{z}_{(i-1)K+j}^{(2)})$ in the calculation of $\hat{R}_K(S_f)$. Then we have

$$\begin{aligned} &|X_f(\bar{z}_1, \dots, \bar{z}_{(i-1)K+j}, \dots, \bar{z}_n) - X_f(\bar{z}_1, \dots, \tilde{z}_{(i-1)K+j}, \dots, \bar{z}_n)| \\ &= |\hat{R}_K(S_f)(\bar{z}_1, \dots, \bar{z}_{(i-1)K+j}, \dots, \bar{z}_n) - \hat{R}_K(S_f)(\bar{z}_1, \dots, \tilde{z}_{(i-1)K+j}, \dots, \bar{z}_n)| \leq U_1 + U_2, \end{aligned} \quad (22)$$

where (assuming $\tilde{z}_s = \bar{z}_s$ for $j \in [n] \setminus \{(i-1)K+j\}$)

$$U_1 := \frac{1}{n} \left| S_f(z_{(i-1)K+j}^{(1)}, z_{(i-1)K+j}^{(2)}) - S_f(\tilde{z}_{(i-1)K+j}^{(1)}, \tilde{z}_{(i-1)K+j}^{(2)}) \right| \leq \frac{2 \log B_S}{n},$$

and

$$\begin{aligned}
U_2 &:= \frac{1}{2n} \sum_{k=1}^K \left| \left[\log \left(\frac{1}{K} \sum_{l \in [K]} \exp(S_f(z_{(i-1)K+k}^{(1)}, z_{(i-1)K+l}^{(2)})) \right) + \log \left(\frac{1}{K} \sum_{l \in [K]} \exp(S_f(z_{(i-1)K+l}^{(1)}, z_{(i-1)K+k}^{(2)})) \right) \right] \right. \\
&\quad \left. - \left[\log \left(\frac{1}{K} \sum_{l \in [K]} \exp(S_f(\tilde{z}_{(i-1)K+k}^{(1)}, \tilde{z}_{(i-1)K+l}^{(2)})) \right) + \log \left(\frac{1}{K} \sum_{l \in [K]} \exp(S_f(\tilde{z}_{(i-1)K+l}^{(1)}, \tilde{z}_{(i-1)K+k}^{(2)})) \right) \right] \right| \\
&\stackrel{(i)}{\leq} \frac{B_S}{2n} \sum_{k=1}^K \left| \frac{1}{K} \left| \sum_{l \in [K]} \exp(S_f(z_{(i-1)K+k}^{(1)}, z_{(i-1)K+l}^{(2)})) - \sum_{l \in [K]} \exp(S_f(\tilde{z}_{(i-1)K+k}^{(1)}, \tilde{z}_{(i-1)K+l}^{(2)})) \right| \right. \\
&\quad \left. + \frac{1}{K} \left| \sum_{l \in [K]} \exp(S_f(z_{(i-1)K+l}^{(1)}, z_{(i-1)K+k}^{(2)})) - \sum_{l \in [K]} \exp(S_f(\tilde{z}_{(i-1)K+l}^{(1)}, \tilde{z}_{(i-1)K+k}^{(2)})) \right| \right| \\
&\leq \frac{B_S}{nK} \sum_{k=1}^K \sum_{l=1}^K \left| \exp(S_f(z_{(i-1)K+k}^{(1)}, z_{(i-1)K+l}^{(2)})) - \exp(S_f(\tilde{z}_{(i-1)K+k}^{(1)}, \tilde{z}_{(i-1)K+l}^{(2)})) \right| \\
&\stackrel{(ii)}{\leq} \frac{2(B_S^2 - 1)}{n},
\end{aligned}$$

Here, step (i) follows from the triangle inequality, a Taylor expansion of $\log(x)$, and Assumption 1; step (ii) follows from Assumption 1 and noting that $\left| \exp(S_f(z_{(i-1)K+k}^{(1)}, z_{(i-1)K+l}^{(2)})) - \exp(S_f(\tilde{z}_{(i-1)K+k}^{(1)}, \tilde{z}_{(i-1)K+l}^{(2)})) \right| \neq 0$ for at most $2K$ terms with indices $k, l \in [K]$.

Putting pieces together, we find

$$\begin{aligned}
&|\hat{R}_K(S_f)(\bar{z}_1, \dots, \bar{z}_{(i-1)K+j}, \dots, \bar{z}_n) - \hat{R}_K(S_f)(\tilde{z}_1, \dots, \tilde{z}_{(i-1)K+j}, \dots, \bar{z}_n)| \\
&\leq \frac{2 \log B_S + 2B_S^2 - 2}{n} \leq \frac{3B_S^2}{n}
\end{aligned}$$

for any $\tilde{z}_{(i-1)K+j}$ and any $i \in [n_1], j \in [K]$ and all $f \in \mathcal{F}$. Therefore, Eq. (20a) follows from Corollary 2.21 in [Wai19] for functions with bounded differences.

Proof of Eq. (20b). First, we have $\mathbb{E}[|X_{f_0}|] \leq cB_S^2/\sqrt{n}$ by properties of sub-Gaussian variables and the fact that, for any $f_0 \in \mathcal{F}$, X_{f_0} is zero-mean with bounded differences cB_S^2/n , as implied by the proof of Eq. (20a). By Dudley's entropy integral bound (see Theorem 5.22 in [Wai19]), it suffices to show $\{X_f, f \in \mathcal{F}\}$ is a zero-mean sub-Gaussian process with respect to the metric $\rho_X(f, \tilde{f}) := B\|f - \tilde{f}\|_{2,\infty}/\sqrt{n}$.

Let $\|\mathbf{x}\|_\psi := \inf\{t > 0 : \mathbb{E}[\psi(\mathbf{x}/t)] \leq 1\}$ denote the Orlicz norm for random variables and let $\psi_2(u) = \exp(u^2) - 1$. We have

$$\|X_f - X_{\tilde{f}}\|_{\psi_2} = \|\hat{R}_K(S_f) - \hat{R}_K(S_{\tilde{f}}) - \mathbb{E}[\hat{R}_K(S_f) - \hat{R}_K(S_{\tilde{f}})]\|_{\psi_2} \leq c(\|U_3 - \mathbb{E}[U_3]\|_{\psi_2} + \|U_4 - \mathbb{E}[U_4]\|_{\psi_2}) \quad (23)$$

for some absolute constant $c > 0$ (we allow the value of c to vary from place to place), where

$$\begin{aligned}
U_3 &:= \frac{1}{n} \sum_{i=1}^{n_1} \sum_{j=1}^K \left[S_f(z_{(i-1)K+j}^{(1)}, z_{(i-1)K+j}^{(2)}) - S_{\tilde{f}}(z_{(i-1)K+j}^{(1)}, z_{(i-1)K+j}^{(2)}) \right], \\
U_4 &:= \frac{1}{2n} \sum_{i=1}^{n_1} \sum_{j=1}^K \left[\left[\log \left(\frac{1}{K} \sum_{l \in [K]} \exp(S_f(z_{(i-1)K+j}^{(1)}, z_{(i-1)K+l}^{(2)})) \right) + \log \left(\frac{1}{K} \sum_{l \in [K]} \exp(S_f(z_{(i-1)K+l}^{(1)}, z_{(i-1)K+j}^{(2)})) \right) \right] \right. \\
&\quad \left. - \left[\log \left(\frac{1}{K} \sum_{l \in [K]} \exp(S_{\tilde{f}}(z_{(i-1)K+j}^{(1)}, z_{(i-1)K+l}^{(2)})) \right) + \log \left(\frac{1}{K} \sum_{l \in [K]} \exp(S_{\tilde{f}}(z_{(i-1)K+l}^{(1)}, z_{(i-1)K+j}^{(2)})) \right) \right] \right].
\end{aligned}$$

It remains to show both $U_3 - \mathbb{E}[U_3]$ and $U_4 - \mathbb{E}[U_4]$ are $\rho_X(f, \tilde{f})$ sub-Gaussian.

Notice that for any $\mathbf{z}^{(1)}, \mathbf{z}^{(2)} \in \mathcal{X}, f, \tilde{f} \in \mathcal{F}$, by Assumption 2, we have

$$\begin{aligned} |\mathsf{S}_f(\mathbf{z}^{(1)}, \mathbf{z}^{(2)}) - \mathsf{S}_{\tilde{f}}(\mathbf{z}^{(1)}, \mathbf{z}^{(2)})| &\leq B_\tau \cdot |\langle f(\mathbf{z}^{(1)}), f(\mathbf{z}^{(2)}) \rangle - \langle \tilde{f}(\mathbf{z}^{(1)}), \tilde{f}(\mathbf{z}^{(2)}) \rangle| \\ &\leq B_\tau (\|f(\mathbf{z}^{(2)})\|_2 \cdot \|f - \tilde{f}\|_{2,\infty} + \|\tilde{f}(\mathbf{z}^{(1)})\|_2 \cdot \|f - \tilde{f}\|_{2,\infty}) \\ &\stackrel{(i)}{\leq} 2B_f B_\tau \|f - \tilde{f}\|_{2,\infty}, \end{aligned} \quad (24)$$

where step (i) uses $\mathsf{S}_f(\mathbf{z}, \mathbf{z}) = \|f(\mathbf{z})\|_2^2 \leq B_f^2$ for $\mathbf{z} \in \mathcal{X}$. Since $\bar{\mathbf{z}}_i = (\mathbf{z}_i^{(1)}, \mathbf{z}_i^{(2)}), i \in [n]$ are i.i.d., it follows immediately that $U_3 - \mathbb{E}[U_3]$ is $2B_f B_\tau \|f - \tilde{f}\|_{2,\infty}/\sqrt{n}$ -sub-Gaussian, i.e.,

$$\|U_3 - \mathbb{E}[U_3]\|_{\psi_2} \leq \frac{cB_f B_\tau}{\sqrt{n}} \|f - \tilde{f}\|_{2,\infty}. \quad (25)$$

Recall the definition of $\{\bar{\mathbf{z}}_s, \tilde{\mathbf{z}}_s\}_{s=1}^n$ in the proof of Eq. (20a). To bound $\|U_4\|_{\psi_2}$, we start with introducing the shorthands for any fixed indices $i \in [n_1], j \in [K]$

$$\begin{aligned} \mathcal{U}_k(\bar{\mathbf{z}}) &:= \frac{1}{K} \sum_{l \in [K]} \exp(\mathsf{S}_f(\mathbf{z}_{(i-1)K+k}^{(1)}, \mathbf{z}_{(i-1)K+l}^{(2)})), \quad \mathcal{V}_k(\bar{\mathbf{z}}) := \frac{1}{K} \sum_{l \in [K]} \exp(\mathsf{S}_f(\mathbf{z}_{(i-1)K+l}^{(1)}, \mathbf{z}_{(i-1)K+k}^{(2)})), \\ \tilde{\mathcal{U}}_k(\bar{\mathbf{z}}) &:= \frac{1}{K} \sum_{l \in [K]} \exp(\mathsf{S}_{\tilde{f}}(\mathbf{z}_{(i-1)K+k}^{(1)}, \mathbf{z}_{(i-1)K+l}^{(2)})), \quad \tilde{\mathcal{V}}_k(\bar{\mathbf{z}}) := \frac{1}{K} \sum_{l \in [K]} \exp(\mathsf{S}_{\tilde{f}}(\mathbf{z}_{(i-1)K+l}^{(1)}, \mathbf{z}_{(i-1)K+k}^{(2)})) \end{aligned}$$

for all $k \in [K]$. Similar to the proof of Eq. (20a), for any given index $(i-1)K+j$, we have

$$\begin{aligned} &|U_4(\bar{\mathbf{z}}_1, \dots, \bar{\mathbf{z}}_{(i-1)K+j}, \dots, \bar{\mathbf{z}}_n) - U_4(\bar{\mathbf{z}}_1, \dots, \tilde{\mathbf{z}}_{(i-1)K+j}, \dots, \bar{\mathbf{z}}_n)| \\ &= \left| \frac{1}{2n} \sum_{k=1}^K \left[\log \left(\frac{\mathcal{U}_k(\bar{\mathbf{z}})}{\tilde{\mathcal{U}}_k(\bar{\mathbf{z}})} \right) + \log \left(\frac{\mathcal{V}_k(\bar{\mathbf{z}})}{\tilde{\mathcal{V}}_k(\bar{\mathbf{z}})} \right) - \log \left(\frac{\mathcal{U}_k(\tilde{\mathbf{z}})}{\tilde{\mathcal{U}}_k(\tilde{\mathbf{z}})} \right) - \log \left(\frac{\mathcal{V}_k(\tilde{\mathbf{z}})}{\tilde{\mathcal{V}}_k(\tilde{\mathbf{z}})} \right) \right] \right| \\ &\leq \frac{B_S^2}{2n} \sum_{k=1}^K \left[\left| \frac{\mathcal{U}_k(\bar{\mathbf{z}})}{\tilde{\mathcal{U}}_k(\bar{\mathbf{z}})} - \frac{\mathcal{U}_k(\tilde{\mathbf{z}})}{\tilde{\mathcal{U}}_k(\tilde{\mathbf{z}})} \right| + \left| \frac{\mathcal{V}_k(\bar{\mathbf{z}})}{\tilde{\mathcal{V}}_k(\bar{\mathbf{z}})} - \frac{\mathcal{V}_k(\tilde{\mathbf{z}})}{\tilde{\mathcal{V}}_k(\tilde{\mathbf{z}})} \right| \right], \end{aligned}$$

where the last line follows from Assumption 1 and a Taylor expansion of $\log(x)$. Moreover,

$$\begin{aligned} \sum_{k=1}^K \left| \frac{\mathcal{U}_k(\bar{\mathbf{z}})}{\tilde{\mathcal{U}}_k(\bar{\mathbf{z}})} - \frac{\mathcal{U}_k(\tilde{\mathbf{z}})}{\tilde{\mathcal{U}}_k(\tilde{\mathbf{z}})} \right| &= \sum_{k=1}^K \left| \frac{\mathcal{U}_k(\bar{\mathbf{z}}) - \tilde{\mathcal{U}}_k(\bar{\mathbf{z}})}{\tilde{\mathcal{U}}_k(\bar{\mathbf{z}})} - \frac{\mathcal{U}_k(\tilde{\mathbf{z}}) - \tilde{\mathcal{U}}_k(\tilde{\mathbf{z}})}{\tilde{\mathcal{U}}_k(\tilde{\mathbf{z}})} \right| \\ &\stackrel{(ii)}{\leq} B_S^2 \sum_{k=1}^K |(\mathcal{U}_k(\bar{\mathbf{z}}) - \tilde{\mathcal{U}}_k(\bar{\mathbf{z}}))\tilde{\mathcal{U}}_k(\bar{\mathbf{z}}) - (\mathcal{U}_k(\tilde{\mathbf{z}}) - \tilde{\mathcal{U}}_k(\tilde{\mathbf{z}}))\tilde{\mathcal{U}}_k(\bar{\mathbf{z}})| \\ &\leq B_S^2 \sum_{k=1}^K [|((\mathcal{U}_k - \tilde{\mathcal{U}}_k)(\bar{\mathbf{z}}) - (\mathcal{U}_k - \tilde{\mathcal{U}}_k)(\tilde{\mathbf{z}}))\tilde{\mathcal{U}}_k(\bar{\mathbf{z}})| + |(\mathcal{U}_k - \tilde{\mathcal{U}}_k)(\tilde{\mathbf{z}})(\tilde{\mathcal{U}}_k(\tilde{\mathbf{z}}) - \tilde{\mathcal{U}}_k(\bar{\mathbf{z}}))|] \\ &\stackrel{(iii)}{\leq} B_S^3 \sum_{k=1}^K [|(\mathcal{U}_k - \tilde{\mathcal{U}}_k)(\bar{\mathbf{z}}) - (\mathcal{U}_k - \tilde{\mathcal{U}}_k)(\tilde{\mathbf{z}})| + 2B_f B_\tau \|f - \tilde{f}\|_{2,\infty} |\tilde{\mathcal{U}}_k(\tilde{\mathbf{z}}) - \tilde{\mathcal{U}}_k(\bar{\mathbf{z}})|], \end{aligned}$$

where step (ii) uses Assumption 1, step (iii) uses Assumption 1, Eq. (24) and a Taylor expansion of $\exp(x)$. Similar to the proof of Eq. (20a), by counting the number of terms in the summations that are different and using Assumption 1, we find

$$\begin{aligned} \sum_{k=1}^K |\tilde{\mathcal{U}}_k(\tilde{\mathbf{z}}) - \tilde{\mathcal{U}}_k(\bar{\mathbf{z}})| &\leq 2B_S, \quad \text{and} \\ \sum_{k=1}^K |(\mathcal{U}_k - \tilde{\mathcal{U}}_k)(\bar{\mathbf{z}}) - (\mathcal{U}_k - \tilde{\mathcal{U}}_k)(\tilde{\mathbf{z}})| &\leq 4B_S B_f B_\tau \|f - \tilde{f}\|_{2,\infty}. \end{aligned}$$

Similar results hold for \mathcal{V} by symmetry. Putting pieces together, we obtain

$$|U_4(\bar{\mathbf{z}}_1, \dots, \bar{\mathbf{z}}_{(i-1)K+j}, \dots, \bar{\mathbf{z}}_n) - U_4(\bar{\mathbf{z}}_1, \dots, \tilde{\mathbf{z}}_{(i-1)K+j}, \dots, \bar{\mathbf{z}}_n)| \leq \frac{4B_S^6 B_f B_\tau}{n}.$$

Therefore, it follows from Corollary 2.21 in [Wai19] for functions with bounded differences that

$$\|U_4 - \mathbb{E}[U_4]\|_{\psi_2} \leq \frac{cB_S^6 B_f B_\tau}{\sqrt{n}}. \quad (26)$$

Substituting Eq. (25) and (26) into Eq. (23), we obtain that $\{X_f, f \in \mathcal{F}\}$ is a zero-mean sub-Gaussian process with respect to the metric $\rho_X(f, \tilde{f}) := B\|f - \tilde{f}\|_{2,\infty}/\sqrt{n}$. This concludes the proof of Eq. (20b).

B.3 Proof of Theorem 2

Write $\mathbf{z} = g(\mathbf{x})$ with $g \sim \mathbb{P}_G \perp\!\!\!\perp \mathbf{x} \sim \mathbb{P}_{\mathcal{X}}$. Define $h_{\min} := \operatorname{argmin}_h \mathbb{E}_{\mathbf{x} \sim \mathbb{P}_{\mathcal{X}}, g \sim \mathbb{P}_G}[(h(g(\mathbf{x})) - h_*(\mathbf{x}))^2]$ and $\mathbf{h}(\mathbf{u}) := \mathbb{E}[h_{\min}(\mathbf{z}^{(1)})|f(\mathbf{z}^{(1)}) = \mathbf{u}]$. Note that $|h_{\min}(\mathbf{z}^{(1)})| = |\mathbb{E}[h_*(\mathbf{x})|\mathbf{z}^{(1)}]|$ is bounded by B_{h_*} almost surely by the assumption in Theorem 2. We first show that $R_G(h \circ f)$ satisfies bound (7a) with ϵ_G replaced by $\tilde{\epsilon}_G = \inf_h \mathbb{E}_{\mathbf{x} \sim \mathbb{P}_{\mathcal{X}}, g \sim \mathbb{P}_G}[(h(g(\mathbf{x})) - h_*(\mathbf{x}))^2]$. The original bound (7a) follows immediately since $\tilde{\epsilon}_G \leq \epsilon_G$.

Since $(a + b)^2 \leq 2a^2 + 2b^2$, we have

$$R_G(h \circ f) = \mathbb{E}_{\mathbf{x}, \mathbf{z}^{(1)}, \mathbf{z}^{(2)}}[(h(f(\mathbf{z}^{(1)})) - h_*(\mathbf{x}))^2] \leq 2\mathbb{E}_{\mathbf{x}, \mathbf{z}^{(1)}, \mathbf{z}^{(2)}}[(h(f(\mathbf{z}^{(1)})) - h_{\min}(\mathbf{z}^{(2)}))^2] + 2\tilde{\epsilon}_G. \quad (27a)$$

Introduce a random variable which follows the distribution of $\mathbf{z}^{(1)}$ conditioned on $f(\mathbf{z}^{(1)})$ and is independent of $(\mathbf{z}^{(1)}, \mathbf{z}^{(2)})$ when conditioned on $f(\mathbf{z}^{(1)})$, i.e., $[\tilde{\mathbf{z}}^{(1)} \sim \mathbb{P}_{\mathbf{z}}(\mathbf{z}^{(1)}|f(\mathbf{z}^{(1)})) \perp\!\!\!\perp (\mathbf{z}^{(1)}, \mathbf{z}^{(2)})|f(\mathbf{z}^{(1)})]$. Consider the joint distribution of the tuple $(\tilde{\mathbf{z}}^{(1)}, \mathbf{z}^{(1)}, \mathbf{z}^{(2)})$. By Bayes' formula, we have $\tilde{\mathbf{z}}^{(1)} \stackrel{d}{=} \mathbf{z}^{(1)} \sim \mathbb{P}_{\mathbf{z}}$ and $\mathbf{z}^{(2)}|\tilde{\mathbf{z}}^{(1)} \sim \mathbb{P}(\mathbf{z}^{(2)}|f(\mathbf{z}^{(1)}) = f(\tilde{\mathbf{z}}^{(1)}))$ and therefore

$$\begin{aligned} \mathbb{E}[(h(f(\mathbf{z}^{(1)})) - h_{\min}(\mathbf{z}^{(2)}))^2] &\stackrel{(i)}{\leq} \mathbb{E}[(h_{\min}(\tilde{\mathbf{z}}^{(1)}) - h_{\min}(\mathbf{z}^{(2)}))^2] \\ &= \mathbb{E}_{\tilde{\mathbf{z}}^{(1)} \sim \mathbb{P}_{\mathbf{z}}, \mathbf{z}^{(2)} \sim \mathbb{P}(\mathbf{z}^{(2)}|f(\mathbf{z}^{(1)}) = f(\tilde{\mathbf{z}}^{(1)}))}[(h_{\min}(\tilde{\mathbf{z}}^{(1)}) - h_{\min}(\mathbf{z}^{(2)}))^2], \end{aligned} \quad (27b)$$

where step (i) follows from

$$\mathbb{E}[(h(f(\mathbf{z}^{(1)})) - h_{\min}(\mathbf{z}^{(2)}))^2|f(\mathbf{z}^{(1)})] \leq \mathbb{E}[(h_{\min}(\tilde{\mathbf{z}}^{(1)}) - h_{\min}(\mathbf{z}^{(2)}))^2|f(\mathbf{z}^{(1)})],$$

which uses Jensen's inequality, the independence of $\tilde{\mathbf{z}}^{(1)}$ and $\mathbf{z}^{(2)}$ conditioned on $f(\mathbf{z}^{(1)})$, and the fact that $\mathbb{E}[h_{\min}(\tilde{\mathbf{z}}^{(1)})|f(\mathbf{z}^{(1)})] = h(f(\mathbf{z}^{(1)}))$. Moreover,

$$\begin{aligned} &\mathbb{E}_{\tilde{\mathbf{z}}^{(1)} \sim \mathbb{P}_{\mathbf{z}}, \mathbf{z}^{(2)} \sim \mathbb{P}(\mathbf{z}^{(2)}|f(\mathbf{z}^{(1)}) = f(\tilde{\mathbf{z}}^{(1)}))}[(h_{\min}(\tilde{\mathbf{z}}^{(1)}) - h_{\min}(\mathbf{z}^{(2)}))^2] \\ &\stackrel{(ii)}{\leq} \mathbb{E}_{\tilde{\mathbf{z}}^{(1)} \sim \mathbb{P}_{\mathbf{z}}, \mathbf{z}^{(2)} \sim \mathbb{P}(\mathbf{z}^{(2)}|\mathbf{z}^{(1)} = \tilde{\mathbf{z}}^{(1)})}[(h_{\min}(\tilde{\mathbf{z}}^{(1)}) - h_{\min}(\mathbf{z}^{(2)}))^2] \\ &\quad + \sqrt{2}B_{h_*}^2 \cdot \mathbb{E}_{\tilde{\mathbf{z}}^{(1)} \sim \mathbb{P}_{\mathbf{z}}} \left[\sqrt{\mathsf{D}_{\text{KL}} \left(\mathbb{P}_{\mathbf{z}^{(2)}|\mathbf{z}^{(1)}(\cdot|\tilde{\mathbf{z}}^{(1)})} \middle\| \mathbb{P}_{\mathbf{z}^{(2)}|\mathbf{z}^{(1)}(\cdot|f(\tilde{\mathbf{z}}^{(1)}))} \right)} \right] \\ &\stackrel{(iii)}{\leq} \mathbb{E}_{\tilde{\mathbf{z}}^{(1)} \sim \mathbb{P}_{\mathbf{z}}, \mathbf{z}^{(2)} \sim \mathbb{P}(\mathbf{z}^{(2)}|\mathbf{z}^{(1)} = \tilde{\mathbf{z}}^{(1)})}[(h_{\min}(\tilde{\mathbf{z}}^{(1)}) - h_{\min}(\mathbf{z}^{(2)}))^2] + \sqrt{2}B_{h_*}^2 \cdot \sqrt{\text{Suff}_{\text{cb}, \text{kl}}(f)} \\ &= \mathbb{E}_{\mathbf{z}^{(1)}, \mathbf{z}^{(2)}}[(h_{\min}(\mathbf{z}^{(1)}) - h_{\min}(\mathbf{z}^{(2)}))^2] + \sqrt{2}B_{h_*}^2 \cdot \sqrt{\text{Suff}_{\text{cb}, \text{kl}}(f)}, \end{aligned} \quad (27c)$$

where step (ii) follows from the variational form of total variation distance and Pinsker's inequality, while step (iii) uses the (CBS) definition of $\text{Suff}_{\text{kl}}(f)$ in Definition 1 and Jensen's inequality. Lastly, we have from a triangle inequality that

$$\begin{aligned} &\mathbb{E}_{\mathbf{z}^{(1)}, \mathbf{z}^{(2)}}[(h_{\min}(\mathbf{z}^{(1)}) - h_{\min}(\mathbf{z}^{(2)}))^2] \\ &\leq 2(\mathbb{E}_{\mathbf{x}, \mathbf{z}^{(1)}}[(h_{\min}(\mathbf{z}^{(1)}) - h_*(\mathbf{x}))^2] + \mathbb{E}_{\mathbf{x}, \mathbf{z}^{(2)}}[(h_{\min}(\mathbf{z}^{(2)}) - h_*(\mathbf{x}))^2]) = 4\tilde{\epsilon}_G. \end{aligned} \quad (27d)$$

Combining Eq. (27a)–(27d) yields Eq. (7a) in Theorem 2. Eq. (7b) in Theorem 2 follows immediately by noting

$$\begin{aligned}\mathsf{R}(\mathsf{h} \circ f) &= \mathbb{E}[(\mathsf{h}(f(\mathbf{x})) - h_\star(\mathbf{x}))^2] = \mathbb{E}_{\mathbf{z}^{(1)}}[(\mathsf{h}(f(\mathbf{z}^{(1)})) - h_\star(\mathbf{z}^{(1)}))^2] \\ &\leq 2\mathbb{E}_{\mathbf{z}^{(1)}, \mathbf{z}^{(2)}}[(\mathsf{h}(f(\mathbf{z}^{(1)})) - h_\star(\mathbf{x}))^2] + 2\mathbb{E}_{\mathbf{z}^{(1)}, \mathbf{z}^{(2)}}[(h_\star(\mathbf{z}^{(1)}) - h_\star(\mathbf{x}))^2] \\ &= 2\mathbb{E}_{\mathbf{z}^{(1)}, \mathbf{z}^{(2)}}[(\mathsf{h}(f(\mathbf{z}^{(1)})) - h_\star(\mathbf{x}))^2] + 2\epsilon_G\end{aligned}$$

and using Eq. (7a).

Comments on Theorem 2. Following the same proof strategy, it can be verified that Eq. (7a) and (7b) also hold when choosing $\mathsf{h}(\mathbf{u}) := \mathbb{E}[h_\star(\mathbf{z}^{(1)})|f(\mathbf{z}^{(1)}) = \mathbf{u}]$. The main difference in the proof is to replace h_{\min} by h_\star in Eq. (27a)–(27d).

Moreover, although we consider the expected squared loss (i.e., $\ell(x, y) = (x - y)^2$) for simplicity, it can be seen from the proof that a similar version of Eq. (7a) and (7b) hold for any semimetric $\ell(x, y)$ that is convex in x for all y . This includes the absolute loss, Huber loss, losses induced by norms, etc.

B.4 Proof of Theorem 3

For any densities \mathbb{P}, \mathbb{Q} , define α -Rényi divergence

$$D_\alpha(\mathbb{P}||\mathbb{Q}) := \frac{1}{\alpha - 1} \log \left(\mathbb{E}_{x \sim \mathbb{P}} \left[\left(\frac{\mathbb{P}(x)}{\mathbb{Q}(x)} \right)^{\alpha-1} \right] \right)$$

for any $\alpha > 0$. Note that the 1-Rényi divergence corresponds to the KL divergence. For any densities $\mathbb{P}, \mathbb{Q}, \mathbb{T}$, we have the following triangle-like inequality which we will repeatedly use in the proof.

Lemma 5 (Triangle-like inequality for Rényi divergence (Lemma 26 in [BS16])). *Let \mathbb{P}, \mathbb{Q} , and \mathbb{T} be probability densities w.r.t. the same measure. Then*

$$D_\alpha(\mathbb{P}||\mathbb{Q}) \leq \frac{k\alpha}{k\alpha - 1} D_{\frac{k\alpha-1}{k-1}}(\mathbb{P}||\mathbb{T}) + D_{k\alpha}(\mathbb{T}||\mathbb{Q})$$

for all $k, \alpha \in (1, \infty)$.

Write $\mathbf{z} = g(\mathbf{x})$ with $g \sim \mathbb{P}_G \perp\!\!\!\perp \mathbf{x} \sim \mathbb{P}_X$ and define $\mathsf{h}(f(\mathbf{z})) := \mathbb{P}(\mathbf{y}|f(\mathbf{z})) \in \Delta([\mathcal{K}])$ as the conditional distribution of \mathbf{y} given $f(\mathbf{z})$, where $\mathbf{z} = g(\mathbf{x})$ for some random transformation $g \sim \mathbb{P}_G$. It can be verified that $\mathsf{h} = \operatorname{argmin}_{\mathbb{Q}: \mathbb{R}^p \mapsto \Delta([\mathcal{K}])} D_{\text{KL}}(\mathbb{P}(\mathbf{y}|\mathbf{x}) || \mathbb{Q}(\mathbf{y}|f(\mathbf{z})))$. Therefore, using Lemma 5 with $k = 4/3, \alpha = 1$ (by taking the limit $\alpha \rightarrow 1$), we obtain

$$\begin{aligned}R_G^{\text{cls}}(\mathsf{h} \circ f) &= \mathbb{E}_{\mathbf{x}, \mathbf{y}, \mathbf{z}^{(1)}}[D_{\text{KL}}(\mathbb{P}(\mathbf{y}|\mathbf{x}) || \mathbb{P}(\mathbf{y}|f(\mathbf{z}^{(1)})))] \\ &\leq 4\mathbb{E}_{\mathbf{x}, \mathbf{y}, \mathbf{z}^{(2)}}[D_{\text{KL}}(\mathbb{P}(\mathbf{y}|\mathbf{x}) || \mathbb{P}(\mathbf{y}|\mathbf{z}^{(2)}))] + \mathbb{E}_{\mathbf{x}, \mathbf{y}, \mathbf{z}^{(1)}, \mathbf{z}^{(2)}}[D_{4/3}(\mathbb{P}(\mathbf{y}|\mathbf{z}^{(2)}) || \mathbb{P}(\mathbf{y}|f(\mathbf{z}^{(1)})))] \\ &\leq 4\epsilon_G^{\text{cls}} + \mathbb{E}_{\mathbf{x}, \mathbf{y}, \mathbf{z}^{(1)}, \mathbf{z}^{(2)}}[D_{4/3}(\mathbb{P}(\mathbf{y}|\mathbf{z}^{(2)}) || \mathbb{P}(\mathbf{y}|f(\mathbf{z}^{(1)})))],\end{aligned}\tag{28a}$$

where the last inequality uses the monotonicity of α -Rényi divergence w.r.t. α . Similar to the proof of Theorem 2, introduce a random variable which follows the distribution of $\mathbf{z}^{(1)}$ conditioned on $f(\mathbf{z}^{(1)})$ and is independent of $(\mathbf{z}^{(1)}, \mathbf{z}^{(2)})$ when conditioned on $f(\mathbf{z}^{(1)})$, i.e., $[\tilde{\mathbf{z}}^{(1)} \sim \mathbb{P}_z(\mathbf{z}^{(1)}|f(\mathbf{z}^{(1)})) \perp\!\!\!\perp (\mathbf{z}^{(1)}, \mathbf{z}^{(2)})] | f(\mathbf{z}^{(1)})$. Consider the joint distribution of the tuple $(\tilde{\mathbf{z}}^{(1)}, \mathbf{z}^{(1)}, \mathbf{z}^{(2)})$. By Bayes' formula, we have $\tilde{\mathbf{z}}^{(1)} \stackrel{d}{=} \mathbf{z}^{(1)} \sim \mathbb{P}_z$ and $\mathbf{z}^{(2)} | \tilde{\mathbf{z}}^{(1)} \sim \mathbb{P}(\mathbf{z}^{(2)} | f(\mathbf{z}^{(1)}) = f(\tilde{\mathbf{z}}^{(1)}))$ and thus

$$\begin{aligned}\mathbb{E}_{\mathbf{x}, \mathbf{y}, \mathbf{z}^{(1)}, \mathbf{z}^{(2)}}[D_{4/3}(\mathbb{P}(\mathbf{y}|\mathbf{z}^{(2)}) || \mathbb{P}(\mathbf{y}|f(\mathbf{z}^{(1)})))] &\stackrel{(i)}{\leq} \mathbb{E}_{\mathbf{x}, \mathbf{y}, \mathbf{z}^{(1)}, \mathbf{z}^{(2)}}[D_{4/3}(\mathbb{P}(\mathbf{y}|\mathbf{z}^{(2)}) || \mathbb{P}(\mathbf{y}|\tilde{\mathbf{z}}^{(1)}))] \\ &= \mathbb{E}_{\tilde{\mathbf{z}}^{(1)} \sim \mathbb{P}_z, \mathbf{z}^{(2)} \sim \mathbb{P}(\mathbf{z}^{(2)} | f(\mathbf{z}^{(1)}) = f(\tilde{\mathbf{z}}^{(1)}))}[D_{4/3}(\mathbb{P}(\mathbf{y}|\mathbf{z}^{(2)}) || \mathbb{P}(\mathbf{y}|\tilde{\mathbf{z}}^{(1)}))],\end{aligned}\tag{28b}$$

where step (i) uses Jensen's inequality, the convexity of Rényi divergence w.r.t. its second argument and the fact that $\mathbb{E}[\mathbb{P}(\mathbf{y}|\tilde{\mathbf{z}}^{(1)})|f(\tilde{\mathbf{z}}^{(1)})] = \mathbb{P}(\mathbf{y}|f(\mathbf{z}^{(1)}) = f(\tilde{\mathbf{z}}^{(1)}))$. Moreover,

$$\begin{aligned}
& \mathbb{E}_{\tilde{\mathbf{z}}^{(1)} \sim \mathbb{P}_{\mathbf{z}}, \mathbf{z}^{(2)} \sim \mathbb{P}(\mathbf{z}^{(2)} | f(\mathbf{z}^{(1)}) = f(\tilde{\mathbf{z}}^{(1)}))} [\mathsf{D}_{4/3}(\mathbb{P}(\mathbf{y}|\mathbf{z}^{(2)}) || \mathbb{P}(\mathbf{y}|\tilde{\mathbf{z}}^{(1)}))] \\
& \stackrel{(ii)}{\leq} \mathbb{E}_{\tilde{\mathbf{z}}^{(1)} \sim \mathbb{P}_{\mathbf{z}}, \mathbf{z}^{(2)} \sim \mathbb{P}(\mathbf{z}^{(2)} | \mathbf{z}^{(1)} = \tilde{\mathbf{z}}^{(1)})} [\mathsf{D}_{4/3}(\mathbb{P}(\mathbf{y}|\mathbf{z}^{(2)}) || \mathbb{P}(\mathbf{y}|\tilde{\mathbf{z}}^{(1)}))] \\
& \quad + \sqrt{2B} \cdot \mathbb{E}_{\tilde{\mathbf{z}}^{(1)} \sim \mathbb{P}_{\mathbf{z}}} \left[\sqrt{\mathsf{D}_{\text{KL}}\left(\mathbb{P}_{\mathbf{z}^{(2)}|\mathbf{z}^{(1)}}(\cdot|\tilde{\mathbf{z}}^{(1)}) \middle\| \mathbb{P}_{\mathbf{z}^{(2)}|\mathbf{z}^{(1)}}(\cdot|f(\tilde{\mathbf{z}}^{(1)}))\right)} \right] \\
& \stackrel{(iii)}{\leq} \mathbb{E}_{\tilde{\mathbf{z}}^{(1)} \sim \mathbb{P}_{\mathbf{z}}, \mathbf{z}^{(2)} \sim \mathbb{P}(\mathbf{z}^{(2)} | \mathbf{z}^{(1)} = \tilde{\mathbf{z}}^{(1)})} [\mathsf{D}_{4/3}(\mathbb{P}(\mathbf{y}|\mathbf{z}^{(2)}) || \mathbb{P}(\mathbf{y}|\tilde{\mathbf{z}}^{(1)}))] + \sqrt{2B} \cdot \sqrt{\text{Suff}_{\text{cb},\text{kl}}(f)} \\
& = \mathbb{E}_{\mathbf{z}^{(1)}, \mathbf{z}^{(2)}} [\mathsf{D}_{4/3}(\mathbb{P}(\mathbf{y}|\mathbf{z}^{(2)}) || \mathbb{P}(\mathbf{y}|\mathbf{z}^{(1)}))] + \sqrt{2B} \cdot \sqrt{\text{Suff}_{\text{cb},\text{kl}}(f)}, \tag{28c}
\end{aligned}$$

where step (ii) follows from the variational form of total variation distance, Pinsker's inequality and the fact that

$$\mathsf{D}_{4/3}(\mathbb{P}(\mathbf{y}|\mathbf{z}^{(2)}) || \mathbb{P}(\mathbf{y}|\tilde{\mathbf{z}}^{(1)})) \leq \mathsf{D}_2(\mathbb{P}(\mathbf{y}|\mathbf{z}^{(2)}) || \mathbb{P}(\mathbf{y}|\tilde{\mathbf{z}}^{(1)})) = \log \mathbb{E}_{\mathbf{y} \sim \mathbb{P}(\cdot|\mathbf{z}^{(2)})} \left[\frac{\mathbb{P}(\mathbf{y}|\mathbf{z}^{(2)})}{\mathbb{P}(\mathbf{y}|\tilde{\mathbf{z}}^{(1)})} \right] \leq B,$$

and step (iii) uses the CBS definition of $\text{Suff}_{\text{kl}}(f)$ and Jensen's inequality. Finally, applying Lemma 5 another time using $\alpha = 4/3$ and $k = 1.5$ yields

$$\begin{aligned}
& \mathbb{E}_{\mathbf{z}^{(1)}, \mathbf{z}^{(2)}} [\mathsf{D}_{4/3}(\mathbb{P}(\mathbf{y}|\mathbf{z}^{(2)}) || \mathbb{P}(\mathbf{y}|\mathbf{z}^{(1)}))] \\
& \leq \mathbb{E}_{\mathbf{x}, \mathbf{z}^{(1)}} [\mathsf{D}_2(\mathbb{P}(\mathbf{y}|\mathbf{z}^{(2)}) || \mathbb{P}(\mathbf{y}|\mathbf{x}))] + \mathbb{E}_{\mathbf{x}, \mathbf{z}^{(2)}} [\mathsf{D}_2(\mathbb{P}(\mathbf{y}|\mathbf{x}) || \mathbb{P}(\mathbf{y}|\mathbf{z}^{(1)}))] \leq \epsilon_{\mathcal{G}}^{\text{cls}}. \tag{28d}
\end{aligned}$$

Combining Eq. (28a)–(28d) yields Theorem 3.

B.5 Proof of Theorem 4

Let $f(x) = (x - 1)^2/2$. The proof largely follows the same arguments as the proof of Theorem 1. Thus we only provide a sketch of the proof here. First, it can be readily verified that the set of minimizers of $R_f(S)$ is

$$\mathcal{M}_S := \left\{ S : S = S_\star + \text{const} \text{ for some const } \in \mathbb{R}, \quad S_\star(\mathbf{z}^{(1)}, \mathbf{z}^{(2)}) := \frac{\mathbb{P}(\mathbf{z}^{(1)}, \mathbf{z}^{(2)})}{\mathbb{P}(\mathbf{z}^{(1)}) \cdot \mathbb{P}(\mathbf{z}^{(2)})} \right\}.$$

Moreover, basic algebra shows that $\hat{R}_{\text{chisq},K}(S_f)$ is an unbiased estimate of $R_f(S_f)$. Thus, by the VFS in Definition 1, we have the decomposition

$$S_{\text{uff},\chi^2}(\hat{f}) \leq R_f(S_f) - R_f(S_\star) \leq \underbrace{\left[R_f(S_{\hat{f}}) - \inf_{f \in \mathcal{F}} R_f(S_f) \right]}_{\text{generalization error}} + \underbrace{\left[\inf_{f \in \mathcal{F}} R_f(S_f) - R_f(S_\star) \right]}_{\text{approximation error}}.$$

Therefore, it remains to show

- (1). With probability at least $1 - \delta$, the excess risk

$$R_f(S_{\hat{f}}) - \inf_{f \in \mathcal{F}} R_f \leq \frac{c \bar{B}_S^2}{\sqrt{n}} \left[\sqrt{\log(1/\delta)} + B_\tau^2 \int_0^{2(\bar{B}_S + B_\tau)} \sqrt{\log \mathcal{N}(u, \|\cdot\|_{2,\infty}, \mathcal{F})} du \right] \tag{29}$$

for some absolute constant $c > 0$.

Proof of Eq. (29). Recall the definition of $\widehat{R}_{\text{chisq},K}$ in Eq. (10) and adopt the shorthand \widehat{R}_K for $\widehat{R}_{\text{chisq},K}$. Let $B_f := \sqrt{B_\tau(\bar{B}_S + B_\tau)}$, $B := c(\bar{B}_S + 1)B_f B_\tau$ for some absolute constant $c > 0$. It can be verified using Assumption 2 that \mathcal{F} must satisfy $\|f\|_{2,\infty} \leq B_f$ for all $f \in \mathcal{F}$ to ensure Assumption 1 holds. Define the zero-mean random process $X_f := \widehat{R}_K(S_f) - \mathbb{E}[\widehat{R}_K(S_f)]$, $f \in \mathcal{F}$. We will prove that for some absolute constant $c > 0$

$$\mathbb{P}\left(\left|\sup_{f \in \mathcal{F}} |X_f| - \mathbb{E}\left[\sup_{f \in \mathcal{F}} |X_f|\right]\right| \geq t\right) \leq 2 \exp\left(-\frac{cnt^2}{\bar{B}_S^4}\right), \quad \text{for all } t \geq 0. \quad (30a)$$

$$\mathbb{E}\left[\sup_{f \in \mathcal{F}} |X_f|\right] \leq \mathbb{E}[|X_{f_0}|] + \mathbb{E}\left[\sup_{f, \tilde{f} \in \mathcal{F}} |X_f - X_{\tilde{f}}|\right] \leq c \frac{\bar{B}_S^2}{\sqrt{n}} + 32 \frac{B}{\sqrt{n}} \cdot \int_0^{2B_f} \sqrt{\log \mathcal{N}(u, \|\cdot\|_{2,\infty}, \mathcal{F})} du. \quad (30b)$$

Combining the two bounds and noting

$$\overline{R}_K(S_{\tilde{f}}) - \inf_{f \in \mathcal{F}} \overline{R}_K(S_f) \leq 2 \sup_{f \in \mathcal{F}} |\widehat{R}_K(S_f) - \overline{R}_K(S_f)| = 2 \sup_{f \in \mathcal{F}} |\widehat{R}_K(S_f) - \mathbb{E}[\widehat{R}_K(S_f)]| = 2 \sup_{f \in \mathcal{F}} |X_f|,$$

yields claim (1).

Proof of Eq. (30a). Similar to the proof of Eq. (20a), we establish the bound using concentration properties for functions with bounded differences. Following the notations in the proof of Theorem 1, we let $\bar{z}_i = (\mathbf{z}_i^{(1)}, \mathbf{z}_i^{(2)})$. For any $i \in [n_1], j \in [K]$, suppose $\bar{z}_{(i-1)K+j}$ is replaced by $\tilde{z}_{(i-1)K+j} = (\tilde{\mathbf{z}}_{(i-1)K+j}^{(1)}, \tilde{\mathbf{z}}_{(i-1)K+j}^{(2)})$ in the calculation of $\widehat{R}_K(S_f)$. It can be verified using Assumption 1 that

$$\begin{aligned} & |X_f(\bar{z}_1, \dots, \bar{z}_{(i-1)K+j}, \dots, \bar{z}_n) - X_f(\bar{z}_1, \dots, \tilde{z}_{(i-1)K+j}, \dots, \bar{z}_n)| \\ &= |\widehat{R}_K(S_f)(\bar{z}_1, \dots, \bar{z}_{(i-1)K+j}, \dots, \bar{z}_n) - \widehat{R}_K(S_f)(\bar{z}_1, \dots, \tilde{z}_{(i-1)K+j}, \dots, \bar{z}_n)| \leq \frac{c\bar{B}_S^2}{n} \end{aligned} \quad (31)$$

for some absolute constant $c > 0$. As a result, Eq. (20a) follows immediately from Corollary 2.21 in [Wai19] for functions with bounded differences.

Proof of Eq. (30b). Similar to the proof of Eq. (20b), $\mathbb{E}[|X_{f_0}|] \leq c\bar{B}_S^2/\sqrt{n}$ by the properties of zero-mean sub-Gaussian variable X_{f_0} , and therefore, to establish Eq. (30b), it remains to show $\{X_f, f \in \mathcal{F}\}$ is a zero-mean sub-Gaussian process with respect to the metric $\rho_X(f, \tilde{f}) := B\|f - \tilde{f}\|_{2,\infty}/\sqrt{n}$.

Let $\|\mathbf{x}\|_\psi := \inf\{t > 0 : \mathbb{E}[\psi(\mathbf{x}/t)] \leq 1\}$ denote the Orlicz norm for random variables and let $\psi_2(u) = \exp(u^2) - 1$. Note that for any $\mathbf{z}^{(1)}, \mathbf{z}^{(2)}, \mathbf{z}^{(2)'} \in \mathcal{X}, f, \tilde{f} \in \mathcal{F}$, we have from Eq. (24) that

$$|S_f(\mathbf{z}^{(1)}, \mathbf{z}^{(2)}) - S_{\tilde{f}}(\mathbf{z}^{(1)}, \mathbf{z}^{(2)})| \leq 2B_f B_\tau \|f - \tilde{f}\|_{2,\infty}, \quad (32a)$$

and

$$\begin{aligned} & |(S_f(\mathbf{z}^{(1)}, \mathbf{z}^{(2)}) - S_f(\mathbf{z}^{(1)}, \mathbf{z}^{(2)'})|^2 - (S_{\tilde{f}}(\mathbf{z}^{(1)}, \mathbf{z}^{(2)}) - S_{\tilde{f}}(\mathbf{z}^{(1)}, \mathbf{z}^{(2)'})|^2| \\ & \stackrel{(i)}{\leq} 4\bar{B}_S (|S_f(\mathbf{z}^{(1)}, \mathbf{z}^{(2)}) - S_{\tilde{f}}(\mathbf{z}^{(1)}, \mathbf{z}^{(2)})| + |S_f(\mathbf{z}^{(1)}, \mathbf{z}^{(2)'}) - S_{\tilde{f}}(\mathbf{z}^{(1)}, \mathbf{z}^{(2)'})|) \\ & \leq 16\bar{B}_S B_f B_\tau \|f - \tilde{f}\|_{2,\infty}, \end{aligned} \quad (32b)$$

where step (i) uses Assumption 1. Then, following the proof of Eq. (20b), it can be verified that

$$\|X_f - X_{\tilde{f}}\|_{\psi_2} = \|\widehat{R}_K(S_f) - \widehat{R}_K(S_{\tilde{f}}) - \mathbb{E}[\widehat{R}_K(S_f) - \widehat{R}_K(S_{\tilde{f}})]\|_{\psi_2} \leq \frac{c(\bar{B}_S + 1)B_f B_\tau}{\sqrt{n}} \|f - \tilde{f}\|_{2,\infty}.$$

C Proofs in Section 5

C.1 Proof of Theorem 6

Recall that $B = B_{\mathbf{x}}B_{\boldsymbol{\theta}}$. For linear regression with misspecified model, by Theorem 11.3 in [GKKW06] (see also e.g., Theorem 1.1 in [AC10]), we have

$$\mathbb{E}[\mathsf{R}_{\text{lin}}(\tilde{\mathbf{h}}_{\tilde{\boldsymbol{\eta}}})] - \bar{\sigma}^2 \leq 8(\inf_{\boldsymbol{\eta} \in \mathbb{R}^p} \mathsf{R}_{\text{lin}}(\mathbf{h}_{\boldsymbol{\eta}}) - \bar{\sigma}^2) + c(B^2 + \bar{\sigma}^2) \frac{p \log m}{m}$$

for some absolute constant $c > 0$.

Thus it suffices to show

$$\inf_{\boldsymbol{\eta} \in \mathbb{R}^p} \mathsf{R}_{\text{lin}}(\mathbf{h}_{\boldsymbol{\eta}}) - \bar{\sigma}^2 \leq c(B^2 c_2 \sqrt{\text{Suff}_f(f)} + \epsilon_G) \quad (33)$$

for some absolute constant $c > 0$. Equivalently, we only need to find some $\boldsymbol{\eta} \in \mathbb{R}^p$ such that $\mathsf{R}_{\text{lin}}(\mathbf{h}_{\boldsymbol{\eta}})$ satisfies the bound in Eq. (33). On the other hand, from the proof of Theorem 2, we see that if we choose $h_{\star}(\mathbf{x}) = \langle \mathbf{x}, \boldsymbol{\theta}_{\star} \rangle$ and $h(\mathbf{u}) := \mathbb{E}[h_{\star}(\mathbf{z})|f(\mathbf{z}) = \mathbf{u}] = \langle \boldsymbol{\theta}_{\star}, \mathbb{E}[\mathbf{z}|f(\mathbf{z}) = \mathbf{u}] \rangle$, then the excess risk

$$\mathsf{R}_{\text{lin}}(h) - \bar{\sigma}^2 \leq c(B^2 c_2 \sqrt{\text{Suff}_f(f)} + \epsilon_G)$$

for some absolute constant $c > 0$ by Theorem 2 and Proposition 5. Therefore, it remains to show h is linear in $f(\mathbf{z})$. Note that $f(\mathbf{z}) = \mathbf{W}\mathbf{z}$. Let $\mathbf{W}^{\dagger} = \mathbf{W}^{\top}(\mathbf{W}\mathbf{W}^{\top})^{-1} \in \mathbb{R}^{d \times p}$ be the generalized inverse of \mathbf{W} and $\tilde{\boldsymbol{\eta}} = \mathbf{W}^{\dagger\top} \boldsymbol{\theta}_{\star} \in \mathbb{R}^p$. In fact, choosing $\tilde{\boldsymbol{\eta}} = \mathbf{W}^{\dagger\top} \boldsymbol{\theta}_{\star} \in \mathbb{R}^p$, we have

$$h(\mathbf{u}) = \langle \boldsymbol{\theta}_{\star}, \mathbb{E}[\mathbf{z}|f(\mathbf{z}) = \mathbf{u}] \rangle = \langle \boldsymbol{\theta}_{\star}, \mathbb{E}[\mathbf{W}^{\dagger}\mathbf{u} + (\mathbf{I}_d - \mathbf{W}^{\dagger}\mathbf{W})\mathbf{z}|f(\mathbf{z}) = \mathbf{u}] \rangle = \langle \boldsymbol{\theta}_{\star}, \mathbf{W}^{\dagger}\mathbf{u} \rangle = \langle \tilde{\boldsymbol{\eta}}, \mathbf{u} \rangle,$$

where the third equality uses the assumption that $\mathbb{E}[(\mathbf{I}_d - \mathbf{W}^{\dagger}\mathbf{W})\mathbf{z}|\mathbf{W}\mathbf{z}] = 0$ almost surely.

C.2 Proof of Corollary 1

It suffices to apply Theorem 1 to the setup in Corollary 1.

By the boundedness of $\mathbf{z}^{(1)}, \mathbf{z}^{(2)}$ and the property that $\mathbb{E}_{\mathbf{z}^{(1)}, \mathbf{z}^{(2)} \sim \mathbb{P}_{\mathbf{z}} \times \mathbb{P}_{\mathbf{z}}} \left[\frac{\mathbb{P}(\mathbf{z}^{(1)}, \mathbf{z}^{(2)})}{\mathbb{P}(\mathbf{z}^{(1)})\mathbb{P}(\mathbf{z}^{(2)})} \right] = 1$, we have

$$\sup_{\mathbf{z}^{(1)}, \mathbf{z}^{(2)}} \frac{\mathbb{P}(\mathbf{z}^{(1)}, \mathbf{z}^{(2)})}{\mathbb{P}(\mathbf{z}^{(1)})\mathbb{P}(\mathbf{z}^{(2)})} \leq \frac{\sup_{\mathbf{z}^{(1)}, \mathbf{z}^{(2)}} \frac{\mathbb{P}(\mathbf{z}^{(1)}, \mathbf{z}^{(2)})}{\mathbb{P}(\mathbf{z}^{(1)})\mathbb{P}(\mathbf{z}^{(2)})}}{\inf_{\mathbf{z}^{(1)}, \mathbf{z}^{(2)}} \frac{\mathbb{P}(\mathbf{z}^{(1)}, \mathbf{z}^{(2)})}{\mathbb{P}(\mathbf{z}^{(1)})\mathbb{P}(\mathbf{z}^{(2)})}} \leq \exp(2\kappa).$$

Similarly we have $\inf_{\mathbf{z}^{(1)}, \mathbf{z}^{(2)}} \frac{\mathbb{P}(\mathbf{z}^{(1)}, \mathbf{z}^{(2)})}{\mathbb{P}(\mathbf{z}^{(1)})\mathbb{P}(\mathbf{z}^{(2)})} \geq \exp(-2\kappa)$.

By properties of the von Mises-Fisher distribution (see e.g., [MJ09]), it can be verified that

$$\frac{\mathbb{P}(\mathbf{z}^{(1)}, \mathbf{z}^{(2)})}{\mathbb{P}(\mathbf{z}^{(1)})\mathbb{P}(\mathbf{z}^{(2)})} = \mathcal{E}_p(\kappa) \cdot \exp(\kappa \langle \mathbf{z}^{(1)}, \mathbf{U}_1 \mathbf{U}_1^{\top} \mathbf{z}^{(2)} \rangle) \cdot \mathbb{1}_{\{\mathbf{z}^{(1)}, \mathbf{z}^{(2)} \in \mathbb{S}(\mathbf{U}_1) \oplus \mathbb{S}(\mathbf{U}_2)\}}, \quad \kappa := \frac{p}{(1 + \sigma^2)^2 - 1},$$

where

$$\begin{aligned} \mathcal{E}_p(\kappa) &:= \frac{\Gamma(p/2) I_{p/2-1}(\kappa)}{(\frac{\kappa}{2})^{p/2-1}} = \Gamma(p/2) \cdot \sum_{m=0}^{\infty} \frac{1}{m! \Gamma(m + p/2)} \left(\frac{\kappa}{2}\right)^{2m} = \sum_{m=0}^{\infty} \frac{(p-2)!!}{(2m)!!(2m+p-2)!!} \kappa^{2m} \\ &< \sum_{m=0}^{\infty} \frac{1}{(2m)!} \kappa^{2m} < e^{\kappa}, \quad \text{and } \mathcal{E}_p(\kappa) > \frac{\Gamma(p/2)}{0! \Gamma(p/2)} \cdot \left(\frac{\kappa}{2}\right)^0 = 1. \end{aligned} \quad (34)$$

Thus, when $\tau(x) = \kappa x$, Assumption 1 and 2 are satisfied with $B_S = \exp(2\kappa), B_T = 2\kappa$ (note that the condition $\kappa^{-1} \leq B_T$ is unnecessary, as from the proof of Theorem 1, we only need $|\tau(\langle f(\mathbf{z}^{(1)}), \mathbf{z}^{(2)} \rangle)| \leq \log B_S$, which follows from the boundedness of \mathcal{F}).

Approximation error. The approximation error $\inf_{f \in \mathcal{F}} \bar{R}_{\text{simclr}, K}(S_f) - \bar{R}_{\text{simclr}, K}(S_*) = 0$ since $S_* + c_1$ is realized by f_* and the link function $\tau(x) = \kappa x$ for some normalizing constant c_1 and $\bar{R}_{\text{simclr}, K}(S_*) = \bar{R}_{\text{simclr}, K}(S_* + c_1)$.

Generalization error. Let $\mathcal{W} := \{\mathbf{W} \in \mathbb{R}^{p \times d}, \|\mathbf{W}\|_{\text{op}} \leq B_{\mathbf{W}}\}$. First, for $f_i(\mathbf{z}) = \mathbf{W}_i \mathbf{z}$ ($i = 1, 2$), since $\|f_1 - f_2\|_{2,\infty} \leq \|\mathbf{W}_1 - \mathbf{W}_2\|_{\text{op}} \cdot \|\mathbf{z}\|_2 \leq 2\|\mathbf{W}_1 - \mathbf{W}_2\|_{\text{op}}$, it follows that

$$\log \mathcal{N}(u, \|\cdot\|_{2,\infty}, \mathcal{F}) \leq \log \mathcal{N}\left(\frac{u}{2}, \|\cdot\|_{\text{op}}, \mathcal{W}\right) \leq cdp \cdot \log\left(1 + \frac{4B_{\mathbf{W}}}{u}\right),$$

where the last inequality follows from the upper bound of the covering number of a unit ball (see e.g., exercise 5.8 in [Wai19]) and the assumption that $p \leq d$. Therefore,

$$B_{\tau} \int_0^{2(\log B_S + B_{\tau})} \sqrt{\log \mathcal{N}(u, \|\cdot\|_{2,\infty}, \mathcal{F})} du \leq c\kappa \int_0^{c\kappa} \sqrt{\log \mathcal{N}(u, \|\cdot\|_{2,\infty}, \mathcal{F})} du \leq c\sqrt{dp\kappa^2} \sqrt{\log B_{\mathbf{W}}}.$$

Combining the result on the approximation error and the generalization error and applying Theorem 1 yields the desired result.

C.3 An end-to-end result on downstream linear regression

Combining Theorem 6 and Corollary 1, we arrive at the following result on the downstream performance of encoder learned by SimCLR.

Theorem 9 (Linear regression using the SimCLR-trained encoder). *Under the setup described in Section 5.1, let \hat{f} be the empirical risk minimizer obtained from Eq. (3) in Corollary 1 on a restricted function space $\mathcal{F}^o := \{f(\mathbf{z}) = \mathbf{W}\mathbf{z} \in \mathcal{F}, \text{span}(\mathbf{W}^\top) = (\text{span}(\mathbf{W}^\top) \cap \text{span}(U_1)) \oplus (\text{span}(\mathbf{W}^\top) \cap \text{span}(U_2))\} \subseteq \mathcal{F}$. In the downstream task, given m i.i.d. samples $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^m$ from $\mathbf{y} = \text{proj}_{[-B, B]}(\langle \mathbf{x}, \theta_* \rangle) + \varepsilon$, where $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d/p)$ follows the same distribution as in contrastive learning, and $\varepsilon \sim \mathcal{N}(0, \bar{\sigma}^2) \perp\!\!\!\perp \mathbf{x}$.*

(a). Consider fitting a (random) linear model $\mathbf{h}_{\boldsymbol{\eta}}(\mathbf{x}) = \langle \hat{f}(\mathbf{z}), \boldsymbol{\eta} \rangle$ by ordinary least squares

$$\hat{\boldsymbol{\eta}} := \underset{\boldsymbol{\eta} \in \mathbb{R}^p}{\text{argmin}} \left\{ \hat{R}_{\text{lin}}(\mathbf{h}_{\boldsymbol{\eta}}) := \frac{1}{m} \sum_{i=1}^m (\langle \hat{f}(\mathbf{z}_i), \boldsymbol{\eta} \rangle - \mathbf{y}_i)^2 \right\},$$

where $\mathbf{z} = g(\mathbf{x})$, $\mathbf{z}_i = g_i(\mathbf{x}_i)$, and $g, \{g\}_{i=1}^m$ are i.i.d. transformations from $\mathbb{P}_{\mathcal{G}}$ as specified in Section 5.1. Then with probability at least $1 - \delta$ over the SimCLR training, the expected risk of the truncated linear model $\tilde{\mathbf{h}}_{\hat{\boldsymbol{\eta}}}(\mathbf{x}) := \text{proj}_{[-B, B]}(\mathbf{h}_{\hat{\boldsymbol{\eta}}}(\mathbf{x}))$ satisfies

$$\begin{aligned} \mathbb{E}[R_{\text{lin}}(\tilde{\mathbf{h}}_{\hat{\boldsymbol{\eta}}})] &:= \mathbb{E}\left[\mathbb{E}_{\mathbf{x}, \mathbf{y}, g}[(\mathbf{y} - \tilde{\mathbf{h}}_{\hat{\boldsymbol{\eta}}}(\mathbf{x}))^2]\right] \\ &\leq \underbrace{\bar{\sigma}^2}_{\text{irreducible risk}} + \underbrace{c\left(B^2\left(1 + \frac{C}{K}\right) \cdot \frac{d^{1/4}p^{1/4}\log^{1/4}B_{\mathbf{W}} + \log^{1/4}(1/\delta)}{n^{1/4}} + \epsilon_{\mathcal{G}}\right)}_{\text{Error from SimCLR training}} + \underbrace{\frac{c(\bar{\sigma}^2 + B^2)\frac{p\log m}{m}}{m}}_{\text{Error from downstream task}}, \end{aligned}$$

where the outer expectation is over $\{(\mathbf{x}_i, \mathbf{y}_i, g_i)\}_{i=1}^n$, $c > 0$ is some absolute constant, $C > 0$ is some constant depending polynomially on $\exp(\kappa)$, and $\epsilon_{\mathcal{G}} \leq \mathbb{E}[\langle \mathbf{x} - \mathbf{z}, \theta_* \rangle^2]$.

(b). In contrast, suppose in addition $\bar{\sigma}^2 \geq 1$, $\|\theta_*\|_2 \leq B_{\theta}$ and $m \geq cd, B \geq c(\bar{\sigma}^2 + B_{\theta}^2)\log m/p$ for some absolute constant $c > 0$, then the truncated ordinary least squares estimator $\tilde{\mathbf{h}}_{\text{ols}}(\mathbf{x}) = \text{proj}_{[-B, B]}(\langle \mathbf{x}, \hat{\theta}_{\text{ols}} \rangle)$ obtained from $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^m$ satisfies

$$\mathbb{E}[R_{\text{lin}}(\tilde{\mathbf{h}}_{\text{ols}})] - \bar{\sigma}^2 := \mathbb{E}\left[\mathbb{E}_{\mathbf{x}, \mathbf{y}}[(\mathbf{y} - \tilde{\mathbf{h}}_{\text{ols}}(\mathbf{x}))^2]\right] - \bar{\sigma}^2 \asymp \bar{\sigma}^2 \frac{d}{m},$$

where \asymp denotes matching upper and lower bounds up to absolute constant factors, and the outer expectation is over $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n$.

We remark that the truncation in the data generation (i.e., $\mathbf{y} = \text{proj}_{[-B, B]}(\langle \mathbf{x}, \boldsymbol{\theta}_\star \rangle) + \varepsilon$) is due to technical difficulties, however, we can choose the threshold B sufficiently large, for example, $B = \mathcal{O}(\log m)$, so that the truncation rarely happens in the generated data. The restriction of the empirical risk minimization to \mathcal{F}° ensures that the condition $\mathbb{E}[(\mathbf{I}_d - \mathbf{W}^\dagger \mathbf{W})\mathbf{z}|\mathbf{W}\mathbf{z}] = 0$ in Theorem 6 holds for any $f(\mathbf{z}) = \mathbf{W}\mathbf{z} \in \mathcal{F}^\circ$. Without this restriction, when $\text{Suff}(\hat{f})$ is sufficiently small, the ERM $\hat{f}(\mathbf{z}) = \widehat{\mathbf{W}}\mathbf{z}$ only satisfies $\mathbb{E}[(\mathbf{I}_d - \widehat{\mathbf{W}}^\dagger \widehat{\mathbf{W}})\mathbf{z}|\widehat{\mathbf{W}}\mathbf{z}] \approx 0$, and the downstream error bound would contain an additional term depending on $\text{Suff}(\hat{f})$.

For the two-step estimator in (a), the first term in the SimCLR training error converges to zero as the pretraining sample size n increases, and the second term ϵ_G is negligible when either the ground truth $\mathbb{E}[\mathbf{y}|\mathbf{x}]$ does not vary significantly (i.e., $\|\boldsymbol{\theta}_\star\|_2$ is small) or the data augmentation introduces negligible error (i.e., $\|\mathbf{x} - \mathbf{z}\|_2$ is small). Thus, compared with the OLS estimator which has a risk of order $\mathcal{O}(d/m)$, the two-step estimator achieves a small risk of order $\mathcal{O}(p/m)$ when the error from SimCLR training is of higher order.

Proof of Theorem 9. First, we have from Corollary 1 that, with probability at least $1 - \delta$, the learned encoder satisfies

$$\text{Suff}(\hat{f}) \leq \left(1 + \frac{C}{K}\right) \cdot \frac{\sqrt{dp \log B_W} + \sqrt{\log(1/\delta)}}{\sqrt{n}},$$

for some constant $C > 0$ depending polynomially on $\exp(\kappa)$. Note that the bound can be directly applied even though we consider the ERM on $\mathcal{F}^\circ \subseteq \mathcal{F}$ since $f_\star \in \mathcal{F}^\circ$ and the proof of Corollary 1 follows from an upper bound on the supremum of an empirical process, which remains valid when restricting to a smaller function space $\mathcal{F}^\circ \subseteq \mathcal{F}$.

Consider the problem of fitting a linear regression using data $\{(\hat{f}(\mathbf{z}_i), \mathbf{y}_i)\}_{i=1}^m$. We have

$$|\mathbb{E}[\mathbf{y}|\hat{f}(\mathbf{z})]| \leq \mathbb{E}[|\mathbb{E}[\mathbf{y}|\mathbf{z}]||f(\mathbf{z})|] = \mathbb{E}[|\mathbb{E}[\text{proj}_{[-B, B]}(\langle \mathbf{x}, \boldsymbol{\theta}_\star \rangle)|\mathbf{z}]||f(\mathbf{z})|] \leq B.$$

Thus the conditions required by Theorem 1.1 in [AC10] are satisfied and we have

$$\mathbb{E}[\mathsf{R}_{\text{lin}}(\tilde{\mathbf{h}}_{\hat{\boldsymbol{\eta}}})] - \bar{\sigma}^2 \leq 8(\inf_{\boldsymbol{\eta} \in \mathbb{R}^p} \mathsf{R}_{\text{lin}}(\mathbf{h}_{\boldsymbol{\eta}}) - \bar{\sigma}^2) + c(B^2 + \bar{\sigma}^2) \frac{p \log m}{m}.$$

Following the proof of Theorem 6, it remains to verify the condition $\mathbb{E}[(\mathbf{I}_d - \widehat{\mathbf{W}}^\dagger \widehat{\mathbf{W}})\mathbf{z}|\widehat{\mathbf{W}}\mathbf{z}] = 0$, where $\widehat{\mathbf{W}}$ is the linear map in \hat{f} (i.e., $\hat{f}(\mathbf{z}) = \widehat{\mathbf{W}}\mathbf{z}$). This follows immediately as \mathbf{z} follows the uniform distribution on $\mathbb{S}(\mathbf{U}_1) \oplus \mathbb{S}(\mathbf{U}_2)$ and the assumption that $\hat{f} \in \mathcal{F}^\circ$.

Ordinary least squares estimator. Adopt the shorthand \mathbf{p} for $\text{proj}_{[-B, B]}$. When applying to a vector, we apply it coordinate-wise. Let $\Sigma = \mathbb{E}[\mathbf{x}\mathbf{x}^\top] = \mathbf{I}_d/p$ be the covariance matrix. For the ordinary least squares (OLS) estimator, let $\mathbf{X} = (\mathbf{x}_1 \ \dots \ \mathbf{x}_m)^\top \in \mathbb{R}^{m \times d}$ denote the sample matrix, $\mathbf{Y} = (\mathbf{y}_1 \ \dots \ \mathbf{y}_m)^\top \in \mathbb{R}^m$ denote the response vector, and $\mathcal{E} = (\varepsilon_1 \ \dots \ \varepsilon_m)^\top \in \mathbb{R}^m$ denote the noise vector. By the definition of OLS, we have $\hat{\boldsymbol{\theta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}$ and

$$\mathbb{E}[\mathsf{R}_{\text{lin}}(\tilde{\mathbf{h}}_{\text{ols}})] - \bar{\sigma}^2 = \mathbb{E}[(\mathbf{p}(\langle \mathbf{x}, (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y} \rangle) - \mathbf{p}(\langle \mathbf{x}, \boldsymbol{\theta}_\star \rangle))^2].$$

We claim two results used later. The proof of them can be found at the end of this section.

$$\mathbb{E}[\text{trace}((\mathbf{X}^\top \mathbf{X})^{-1} \Sigma)] = \frac{d}{m-d-1}, \quad \mathbb{E}[\text{trace}((\mathbf{X}^\top \mathbf{X})^{-1} \Sigma)^2] = \frac{(m-1)d}{(m-d)(m-d-1)(m-d-3)}, \quad (35)$$

$$\mathbb{E}[\|\mathbf{p}(\mathbf{X}\boldsymbol{\theta}_\star) - \mathbf{X}\boldsymbol{\theta}_\star\|_2^4] \leq c \frac{m^2 B_{\boldsymbol{\theta}}^4}{p^2} \cdot \exp(-\frac{B^2}{c B_{\boldsymbol{\theta}}^2/p}) \quad (36)$$

for some absolute constant $c > 0$.

Choose $B \geq c(\bar{\sigma}^2 + B_{\boldsymbol{\theta}}^2) \log m/p$ for some sufficiently large absolute constant $c > 0$. We then have $\mathbb{E}[\|\mathbf{p}(\mathbf{X}\boldsymbol{\theta}_\star) - \mathbf{X}\boldsymbol{\theta}_\star\|_2^4] \leq m^{-4}$. On one hand, to establish the upper bound, we have

$$\begin{aligned} \mathbb{E}[\mathsf{R}_{\text{lin}}(\tilde{\mathbf{h}}_{\text{ols}})] - \bar{\sigma}^2 &\leq \mathbb{E}[(\langle \mathbf{x}, (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y} \rangle - \langle \mathbf{x}, \boldsymbol{\theta}_\star \rangle)^2] \\ &=: T_1 + T_2, \end{aligned}$$

where

$$\begin{aligned}
T_1 &:= \mathbb{E}[\langle \mathbf{x}, (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{p}(\mathbf{X} \boldsymbol{\theta}_*) \rangle - \langle \mathbf{x}, \boldsymbol{\theta}_* \rangle]^2 \\
&= \mathbb{E}[\langle \mathbf{x}, (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top [\mathbf{p}(\mathbf{X} \boldsymbol{\theta}_*) - \mathbf{X} \boldsymbol{\theta}_*] \rangle]^2 \\
&\leq \mathbb{E}[\|\mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \Sigma(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top\|_{\text{op}} \cdot \|[\mathbf{p}(\mathbf{X} \boldsymbol{\theta}_*) - \mathbf{X} \boldsymbol{\theta}_*]\|_2^2] \\
&\stackrel{(i)}{\leq} \sqrt{\mathbb{E}[\text{trace}((\mathbf{X}^\top \mathbf{X})^{-1} \Sigma(\mathbf{X}^\top \mathbf{X})^{-1} \Sigma)]} \cdot \sqrt{\mathbb{E}[\|\mathbf{p}(\mathbf{X} \boldsymbol{\theta}_*) - \mathbf{X} \boldsymbol{\theta}_*\|_2^4]} \stackrel{(ii)}{\leq} \frac{1}{m^2} \leq \frac{\bar{\sigma}^2}{m^2}
\end{aligned}$$

and

$$\begin{aligned}
T_2 &:= \mathbb{E}[\langle \mathbf{x}, (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathcal{E} \rangle]^2 \\
&= \bar{\sigma}^2 \mathbb{E}[\text{trace}((\mathbf{X}^\top \mathbf{X})^{-1} \Sigma)] \stackrel{(iii)}{=} \bar{\sigma}^2 \frac{d}{m-d-1}.
\end{aligned}$$

Here, step (i) uses Cauchy-Schwarz inequality, step (ii) and (iii) follow from claim (35) and (36) and the choice of B . Combining the bounds on T_1, T_2 yields the upper bound $\mathbb{E}[\mathbf{R}_{\text{lin}}(\tilde{\mathbf{h}}_{\text{ols}})] - \bar{\sigma}^2 \leq c\bar{\sigma}^2 \frac{d}{m-d-1}$.

To establish the lower bound, since $\mathbb{E}[a^2] \geq \mathbb{E}[b^2] + \mathbb{E}[(a-b)^2] - 2\sqrt{\mathbb{E}[(a-b)^2]} \cdot \sqrt{\mathbb{E}[b^2]}$, it follows that

$$\begin{aligned}
\mathbb{E}[\mathbf{R}_{\text{lin}}(\tilde{\mathbf{h}}_{\text{ols}})] - \bar{\sigma}^2 &= \mathbb{E}[(\mathbf{p}(\langle \mathbf{x}, (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y} \rangle) - \mathbf{p}(\langle \mathbf{x}, \boldsymbol{\theta}_* \rangle))^2] \\
&= \mathbb{E}[(\mathbf{p}(\langle \mathbf{x}, \boldsymbol{\theta}_* + (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathcal{E} \rangle) - \mathbf{p}(\langle \mathbf{x}, \boldsymbol{\theta}_* \rangle))^2] \\
&\geq T_3 - (T_4 + T_5),
\end{aligned}$$

where

$$\begin{aligned}
T_3 &= \mathbb{E}[\langle \mathbf{x}, \boldsymbol{\theta}_* + (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathcal{E} \rangle - \langle \mathbf{x}, \boldsymbol{\theta}_* \rangle]^2 = \mathbb{E}[\text{trace}((\mathbf{X}^\top \mathbf{X})^{-1} \Sigma)] = \bar{\sigma}^2 \frac{d}{m-d-1}. \\
T_4 &= 2\sqrt{T_3}\sqrt{T_5}, \\
T_5 &:= \mathbb{E}[[\mathbf{p}(\langle \mathbf{x}, \boldsymbol{\theta}_* + (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathcal{E} \rangle) - \langle \mathbf{x}, \boldsymbol{\theta}_* + (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathcal{E} \rangle) - (\mathbf{p}(\langle \mathbf{x}, \boldsymbol{\theta}_* \rangle) - \langle \mathbf{x}, \boldsymbol{\theta}_* \rangle)]^2] \\
&\leq \mathbb{E}[\|\mathbf{p}(\langle \mathbf{x}, \boldsymbol{\theta}_* + (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathcal{E} \rangle) - \langle \mathbf{x}, \boldsymbol{\theta}_* + (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathcal{E} \rangle\|^2] + \bar{\sigma}^2 \frac{1}{m^2},
\end{aligned}$$

where the inequality uses claim (36). To find a further upper bound of T_4, T_5 , we first note that $(\boldsymbol{\theta}_* + (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathcal{E})$ is independent of \mathbf{x} , and

$$\|\boldsymbol{\theta}_* + (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathcal{E}\|_2^2 \leq 2B_\theta^2 + 2\|(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathcal{E}\|_2^2 \leq c\bar{\sigma}^2 \frac{d}{m} + v,$$

where v is some zero-mean $c\bar{\sigma}^2$ -sub-Exponential variable by Theorem 1 in [HKZ11]. Under our choice of B , following the proof of claim (36) and integrating over the sub-Exponential variable v , it can be verified that (when choosing the absolute constant in B sufficiently large) $T_5 \leq 2\bar{\sigma}^2/m^2$. Putting the bounds on T_3, T_5 (and hence T_4) together, we conclude that $\mathbb{E}[\mathbf{R}_{\text{lin}}(\tilde{\mathbf{h}}_{\text{ols}})] - \bar{\sigma}^2 \geq c\bar{\sigma}^2 \frac{d}{m-d-1}$ for some absolute constant $c > 0$.

Proof of claim (35) and (36). Claim (35) follows directly from properties of the inverse Wishart distribution [VR88]. For Claim (36), since each coordinate of $\mathbf{X} \boldsymbol{\theta}_*$ are i.i.d. $\mathcal{N}(0, \|\boldsymbol{\theta}_*\|_2^2)$, w.l.o.g., it suffice to show

$$\mathbb{E}[|\mathbf{p}(z) - z|^4] \leq c \exp(-B^2/c).$$

for $z \sim \mathcal{N}(0, 1)$. Note that this follows immediately since

$$\mathbb{E}[|\mathbf{p}(z) - z|^4] \leq c \int_B^\infty s^4 \exp(-s^2/2) ds \leq cs^3 \exp(-s^2/2) \leq c \exp(-s^2/c).$$

□

C.4 Proof of Theorem 7

We prove Eq. (14) and (15) in Appendix C.4.1 and C.4.2, respectively.

C.4.1 Proof of Eq. (14)

It suffices to apply Theorem 4 to the setup in Theorem 7. With a slight abuse of notation, we use both one-hot vectors in $\cup_{i=1}^S \{e_i\}$ and integers in $[S]$ to represent the augmented views \mathbf{z} and do not distinguish them in the proof. We also occasionally omit the subscripts in $\mathbb{P}_y, \mathbb{P}_c$ when the meaning is clear from the context.

We claim that

$$\frac{\mathbb{P}(\mathbf{z}^{(1)}, \mathbf{z}^{(2)})}{\mathbb{P}(\mathbf{z}^{(1)})\mathbb{P}(\mathbf{z}^{(2)})} = \frac{1}{2} \cdot \sum_{y=1}^M \frac{\mathbb{P}_c(y|\mathbf{z}^{(1)}) \cdot \mathbb{P}_c(y|\mathbf{z}^{(2)})}{\mathbb{P}_y(y)} + \frac{S}{2} \cdot \mathbb{1}_{\{\mathbf{z}^{(1)}=\mathbf{z}^{(2)}\}}. \quad (37)$$

We will prove this claim momentarily. With this claim at hand, we have

Approximation error. Let

$$f_\star(\mathbf{z}) := \frac{1}{\sqrt{2}} \left(\frac{\mathbb{P}_c(\mathbf{y}=1|\mathbf{x}^{c_1}=\mathbf{z})}{\sqrt{\mathbb{P}_y(\mathbf{y}=1)}}, \dots, \frac{\mathbb{P}_c(\mathbf{y}=M|\mathbf{x}^{c_1}=\mathbf{z})}{\sqrt{\mathbb{P}_y(\mathbf{y}=M)}}, \sqrt{S}\mathbf{z}^\top \right)^\top.$$

It can be verified that the parameter $(\mathbf{W}_\star, w_\star)$ corresponding to f_\star lies in Γ . Therefore, the approximation error $\inf_{f \in \mathcal{F}} R_{\chi^2}(\mathbf{S}_f) - R_{\chi^2}(\mathbf{S}_\star) = 0$ since \mathbf{S}_\star is realized by f_\star and the link function $\tau(x) = x$.

Generalization error. Let $\mathcal{W} := \{\mathbf{W} \in \mathbb{R}^{M \times S}, w \in \mathbb{R}, \|\mathbf{W}\|_{2,\infty} \vee |w/\sqrt{S}| \leq B_{\mathbf{W}}\}$ and define the metric $\|(\mathbf{W}_1, w_1) - (\mathbf{W}_2, w_2)\| := \|\mathbf{W}_1 - \mathbf{W}_2\|_{2,\infty} \vee |(w_1 - w_2)/\sqrt{S}|$ on \mathcal{W} .

First, for $f_i(\mathbf{z}) = ((\mathbf{W}_i \mathbf{z})^\top, w_i \cdot \mathbf{z}^\top)^\top$ ($i = 1, 2$), simple calculation shows $\|f_1 - f_2\|_{2,\infty} \leq 2(|w_1 - w_2| \vee \|\mathbf{W}_1 - \mathbf{W}_2\|_{\text{op}})$, and therefore

$$\begin{aligned} \log \mathcal{N}(u, \|\cdot\|_{2,\infty}, \mathcal{F}) &\leq \log \mathcal{N}\left(\frac{u}{2}, \|\cdot\|_{2,\infty}, \mathcal{W}\right) \leq \log \mathcal{N}\left(\frac{u}{2}, \|\cdot\|_{2,\infty}, \mathcal{W}_1\right) + \log \mathcal{N}\left(\frac{u\sqrt{S}}{2}, |\cdot|, \mathcal{W}_2\right) \\ &\leq SM \cdot \log\left(1 + \frac{4B_{\mathbf{W}}}{u}\right) + \log\left(1 + \frac{4B_{\mathbf{W}}}{u}\right), \end{aligned}$$

where $\mathcal{W}_1 := \{\mathbf{W} \in \mathbb{R}^{M \times S}, \|\mathbf{W}\|_{2,\infty} \leq B_{\mathbf{W}}\}$, $\mathcal{W}_2 := \{w \in \mathbb{R}, |w| \leq \sqrt{S}B_{\mathbf{W}}\}$ and the last inequality follows from the upper bound of the covering number of the unit ball (see e.g., Example 5.8 in [Wai19]) and the assumption that $M \leq S$. In addition, it is readily verified that $\mathbf{S}_f(\mathbf{z}^{(1)}, \mathbf{z}^{(2)}) \in [-\bar{B}_S, \bar{B}_S]$ with $\bar{B}_S = 4B_{\mathbf{W}}^2 S = 4M^2 S$ for all $\mathbf{z}^{(1)}, \mathbf{z}^{(2)}$. Consequently,

$$\begin{aligned} &B_\tau \int_0^{B_{\mathbf{W}}} \sqrt{\log \mathcal{N}(u, \|\cdot\|_{2,\infty}, \mathcal{F})} du \\ &\leq c \left(\int_0^{B_{\mathbf{W}}} \sqrt{SM \cdot \log\left(1 + \frac{4B_{\mathbf{W}}}{u}\right)} du + \int_0^{B_{\mathbf{W}}} \sqrt{\log\left(1 + \frac{4B_{\mathbf{W}}}{u}\right)} du \right) \\ &\leq c\sqrt{SM}B_{\mathbf{W}} = c\sqrt{SM^3}. \end{aligned}$$

Combining the result on the approximation error and the generalization error and applying Theorem 4 yields the desired result.

Proof of claim (37). For $\mathbf{z}^{(1)} \neq \mathbf{z}^{(2)}$, by Bayes' formula, we have

$$\begin{aligned} \frac{\mathbb{P}(\mathbf{z}^{(1)}, \mathbf{z}^{(2)})}{\mathbb{P}(\mathbf{z}^{(1)})\mathbb{P}(\mathbf{z}^{(2)})} &= \sum_{\mathbf{x}} \frac{\mathbb{P}(\mathbf{z}^{(2)}|\mathbf{x}) \cdot \mathbb{P}(\mathbf{x}|\mathbf{z}^{(1)})}{\mathbb{P}(\mathbf{z}^{(2)})} = \sum_{\mathbf{x}} \frac{\mathbb{P}(\mathbf{x}|\mathbf{z}^{(2)}) \cdot \mathbb{P}(\mathbf{x}|\mathbf{z}^{(1)})}{\mathbb{P}(\mathbf{x})} \\ &\stackrel{(i)}{=} 2 \frac{\mathbb{P}(\mathbf{x}=(\mathbf{z}^{(1)}, \mathbf{z}^{(2)})|\mathbf{z}^{(2)}) \cdot \mathbb{P}(\mathbf{x}=(\mathbf{z}^{(1)}, \mathbf{z}^{(2)})|\mathbf{z}^{(1)})}{\mathbb{P}(\mathbf{x}=(\mathbf{z}^{(1)}, \mathbf{z}^{(2)}))}, \end{aligned} \quad (38)$$

where step (i) follows from symmetry between $\mathbf{z}^{(1)}, \mathbf{z}^{(2)}$. Moreover,

$$\begin{aligned}\mathbb{P}(\mathbf{x} = (\mathbf{z}^{(1)}, \mathbf{z}^{(2)}) | \mathbf{z}^{(1)}) &= \frac{1}{2} \mathbb{P}(\mathbf{x}^{c_2} = \mathbf{z}^{(2)} | \mathbf{x}^{c_1} = \mathbf{z}^{(1)}) = \frac{1}{2} \sum_{y=1}^M \mathbb{P}_c(\mathbf{x}^{c_2} = \mathbf{z}^{(2)} | y) \cdot \mathbb{P}_c(y | \mathbf{x}^{c_1} = \mathbf{z}^{(1)}) \\ &= \frac{1}{2} \sum_{y=1}^M \frac{\mathbb{P}_c(y | \mathbf{x}^{c_2} = \mathbf{z}^{(2)}) \cdot \mathbb{P}_c(y | \mathbf{x}^{c_1} = \mathbf{z}^{(1)})}{\mathbb{P}_Y(y)} \cdot \mathbb{P}_c(\mathbf{x}^{c_2} = \mathbf{z}^{(2)}) \\ &= \frac{1}{2} \sum_{y=1}^M \frac{\mathbb{P}_c(y | \mathbf{z}^{(2)}) \cdot \mathbb{P}_c(y | \mathbf{z}^{(1)})}{\mathbb{P}_Y(y)} \cdot \mathbb{P}_c(\mathbf{z}^{(2)}),\end{aligned}\tag{39a}$$

$$\mathbb{P}_c(\mathbf{z}) \stackrel{(ii)}{=} \mathbb{P}(\mathbf{z}), \text{ and}\tag{39b}$$

$$\mathbb{P}(\mathbf{z}^{(1)}, \mathbf{z}^{(2)}) \stackrel{(iii)}{=} 2\mathbb{P}((\mathbf{z}^{(1)}, \mathbf{z}^{(2)}), \mathbf{x} = (\mathbf{z}^{(1)}, \mathbf{z}^{(2)})) = \frac{1}{2} \mathbb{P}(\mathbf{x} = (\mathbf{z}^{(1)}, \mathbf{z}^{(2)})),\tag{39c}$$

where step (ii) follows from the generation process of the augmented views $(\mathbf{z}^{(1)}, \mathbf{z}^{(2)})$, and step (iii) follows from symmetry between $\mathbf{z}^{(1)}, \mathbf{z}^{(2)}$. Substituting Eq. (39a) into Eq. (38), we find

$$\begin{aligned}\frac{\mathbb{P}(\mathbf{z}^{(1)}, \mathbf{z}^{(2)})}{\mathbb{P}(\mathbf{z}^{(1)})\mathbb{P}(\mathbf{z}^{(2)})} &= \frac{1}{2} \left(\sum_{y=1}^M \frac{\mathbb{P}_c(y | \mathbf{z}^{(2)}) \cdot \mathbb{P}_c(y | \mathbf{z}^{(1)})}{\mathbb{P}_Y(y)} \right)^2 \cdot \frac{\mathbb{P}_c(\mathbf{z}^{(1)})\mathbb{P}_c(\mathbf{z}^{(2)})}{\mathbb{P}(\mathbf{x} = (\mathbf{z}^{(1)}, \mathbf{z}^{(2)}))} \\ &= \frac{1}{4} \left(\sum_{y=1}^M \frac{\mathbb{P}_c(y | \mathbf{z}^{(2)}) \cdot \mathbb{P}_c(y | \mathbf{z}^{(1)})}{\mathbb{P}_Y(y)} \right)^2 \cdot \frac{\mathbb{P}(\mathbf{z}^{(1)})\mathbb{P}(\mathbf{z}^{(2)})}{\mathbb{P}(\mathbf{z}^{(1)}, \mathbf{z}^{(2)})},\end{aligned}\tag{40}$$

where the second equality uses Eq. (39b) and (39c). Reorganizing Eq. (40), we obtain

$$\frac{\mathbb{P}(\mathbf{z}^{(1)}, \mathbf{z}^{(2)})}{\mathbb{P}(\mathbf{z}^{(1)})\mathbb{P}(\mathbf{z}^{(2)})} = \frac{1}{2} \left(\sum_{y=1}^M \frac{\mathbb{P}_c(y | \mathbf{z}^{(2)}) \cdot \mathbb{P}_c(y | \mathbf{z}^{(1)})}{\mathbb{P}_Y(y)} \right) = \frac{1}{2} \frac{\mathbb{P}_c(\mathbf{z}^{(1)}, \mathbf{z}^{(2)})}{\mathbb{P}_c(\mathbf{z}^{(1)})\mathbb{P}_c(\mathbf{z}^{(2)})},\tag{41}$$

where we recall $\mathbb{P}_c(\cdot)$ is the marginal distribution of \mathbf{x}^{c_1} (or \mathbf{x}^{c_2}) and the second equality follows from Bayes' formula and the fact that $\mathbf{x}^{c_1} \perp\!\!\!\perp \mathbf{x}^{c_2} | \mathbf{y}$.

For $\mathbf{z}^{(1)} = \mathbf{z}^{(2)} = z$, using Eq. (39b) and properties of conditional distribution, we have

$$\sum_{z' \in [S]} \frac{\mathbb{P}_c(\mathbf{z}^{(1)} = z, \mathbf{z}^{(2)} = z')}{\mathbb{P}_c(\mathbf{z}^{(1)} = z)\mathbb{P}_c(\mathbf{z}^{(2)} = z')} = \frac{1}{\mathbb{P}_c(\mathbf{z}^{(2)} = z)} = \frac{1}{\mathbb{P}(\mathbf{z}^{(2)} = z)} = \sum_{z' \in [S]} \frac{\mathbb{P}(\mathbf{z}^{(1)} = z, \mathbf{z}^{(2)} = z')}{\mathbb{P}(\mathbf{z}^{(1)} = z)\mathbb{P}(\mathbf{z}^{(2)} = z')}.$$

Combining this with Eq. (41) for all $\mathbf{z}^{(2)} \neq \mathbf{z}^{(1)}$ and noting that the marginal $\mathbb{P}_c(\cdot)$ is the uniform distribution on $[S]$, we obtain

$$\begin{aligned}\frac{\mathbb{P}(\mathbf{z}^{(1)} = z, \mathbf{z}^{(2)} = z)}{\mathbb{P}(\mathbf{z}^{(1)} = z)\mathbb{P}(\mathbf{z}^{(2)} = z)} &= \frac{1}{2} \cdot \frac{\mathbb{P}_c(\mathbf{x}^{c_1} = z, \mathbf{x}^{c_2} = z)}{\mathbb{P}_c(\mathbf{x}^{c_1} = z)\mathbb{P}_c(\mathbf{x}^{c_2} = z)} + \frac{1}{2} \sum_{z' \in [S]} \frac{\mathbb{P}_c(\mathbf{x}^{c_1} = z, \mathbf{x}^{c_2} = z')}{\mathbb{P}_c(\mathbf{x}^{c_1} = z)\mathbb{P}_c(\mathbf{x}^{c_2} = z')} \\ &= \frac{1}{2} \cdot \frac{\mathbb{P}_c(\mathbf{x}^{c_1} = z, \mathbf{x}^{c_2} = z)}{\mathbb{P}_c(\mathbf{x}^{c_1} = z)\mathbb{P}_c(\mathbf{x}^{c_2} = z)} + \frac{S}{2} \\ &= \frac{1}{2} \cdot \sum_{y=1}^M \frac{\mathbb{P}_c(y | z) \cdot \mathbb{P}_c(y | z)}{\mathbb{P}_Y(y)} + \frac{S}{2}.\end{aligned}$$

C.4.2 Proof of Eq. (15)

Write $\mathbf{z} = g(\mathbf{x})$. By a standard risk decomposition, we have

$$\begin{aligned}\mathsf{R}_{\text{cls}}(\mathbf{h}_{\hat{\Gamma}}) &= \mathbb{E}[\hat{\mathsf{R}}_{\text{cls}}(\mathbf{h}_{\hat{\Gamma}})] - \mathbb{E}[\hat{\mathsf{R}}_{\text{cls}}(\mathbb{P}_{\mathbf{y}|\mathbf{x}}(\cdot|\mathbf{x}))] \\ &= \mathbb{E}[\hat{\mathsf{R}}_{\text{cls}}(\mathbf{h}_{\hat{\Gamma}})] - \inf_{\mathbf{h}} \mathbb{E}[\hat{\mathsf{R}}_{\text{cls}}(\mathbf{h})] \\ &= \underbrace{\inf_{\Gamma: \|\Gamma_w\|_{\text{op}} \vee \|\Gamma_b\|_2 \leq B_\Gamma} \mathbb{E}_{\mathbf{x},g}[\mathsf{D}_{\text{KL}}(\mathbb{P}_{\mathbf{y}|\mathbf{x}}(\cdot|\mathbf{x}) \parallel \bar{\mathbf{h}}_{\Gamma}(f(\mathbf{z})))]}_{\text{approximation error}} \\ &\quad + \underbrace{\mathbb{E}[\hat{\mathsf{R}}_{\text{cls}}(\mathbf{h}_{\hat{\Gamma}})] - \inf_{\Gamma: \|\Gamma_w\|_{\text{op}} \vee \|\Gamma_b\|_2 \leq B_\Gamma} \mathbb{E}[\hat{\mathsf{R}}_{\text{cls}}(\mathbf{h}_{\Gamma})]}_{\text{generalization error}}.\end{aligned}$$

We will prove that for some absolute constant $c > 0$,

1.

$$\text{approximation error} \leq c \left(\epsilon_{\mathcal{G}}^{\text{cls}} + \frac{S \exp(B)}{\sigma_{\mathbf{E}_*}^2} \cdot (R_f(S_{\hat{f}_{\text{aug}}}) - R_f(S_*)) \right), \quad (42a)$$

and

2. with probability at least $1 - \delta$,

$$\text{generalization error} \leq \frac{cB}{\sqrt{m}} \left[\sqrt{\log(1/\delta)} + M(\sqrt{\log B_\Gamma} + \sqrt{B}) \right]. \quad (42b)$$

Approximation error. Let $\mathbf{E}_* \in \mathbb{R}^{M \times S}$ be the representation where

$$\mathbf{E}_{*,j} = \frac{1}{\sqrt{2}} \left(\frac{\mathbb{P}_c(\mathbf{y} = 1 | \mathbf{x}^{c_1} = j)}{\sqrt{\mathbb{P}_{\mathcal{Y}}(\mathbf{y} = 1)}}, \dots, \frac{\mathbb{P}_c(\mathbf{y} = M | \mathbf{x}^{c_1} = j)}{\sqrt{\mathbb{P}_{\mathcal{Y}}(\mathbf{y} = M)}} \right)^\top$$

for $j \in [S]$ and let $\mathbf{E}_*(\mathbf{z})$ denote the \mathbf{z} -th column of \mathbf{E}_* . Let $\hat{\mathbf{E}} := (\hat{f}(1) \ \dots \ \hat{f}(S)) \in \mathbb{R}^{M \times S}$.

Given a representation $\hat{f}(\mathbf{z})$, consider the classifier

$$\bar{\mathbf{h}}_{\Gamma}(\hat{f}(\mathbf{z})) = \text{softmax}(\log \mathbf{trun}(\mathbf{\Gamma}_w \hat{f}(\mathbf{z}) + \mathbf{\Gamma}_b)), \quad \text{where}$$

$$\mathbf{\Gamma}_w := \sqrt{2} \mathbf{P}_{\mathcal{Y}}^{1/2} (\mathbf{E}_* \mathbf{E}_*^\top)^{-1} \mathbf{E}_* (\mathbf{I}_S - \mathbf{P}_{\mathbf{1}_S}) \hat{\mathbf{E}}^\top, \quad \mathbf{\Gamma}_b := \frac{1}{\sqrt{2}} \mathbf{P}_{\mathcal{Y}}^{1/2} (\mathbf{E}_* \mathbf{E}_*^\top)^{-1} \mathbf{E}_* \mathbf{1}_S, \quad (43)$$

and $\mathbf{P}_{\mathcal{Y}} := \text{diag}\{\mathbb{P}_{\mathcal{Y}}(\mathbf{y} = 1), \dots, \mathbb{P}_{\mathcal{Y}}(\mathbf{y} = M)\}$. It can be verified that $\|\mathbf{\Gamma}_w\|_{\text{op}} \leq 2\sqrt{SM}/\sigma_{\mathbf{E}_*} \leq B_\Gamma$ and $\|\mathbf{\Gamma}_b\|_2 \leq \sqrt{S}/\sigma_{\mathbf{E}_*} \leq B_\Gamma$. Moreover, we have by Lemma 5 that

$$\begin{aligned}\mathbb{E}_{\mathbf{x},g}[\mathsf{D}_{\text{KL}}(\mathbb{P}_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\mathbf{x}) \parallel \bar{\mathbf{h}}_{\Gamma}(\hat{f}(\mathbf{z})))] &\leq 2\mathbb{E}_{\mathbf{x},g}[\mathsf{D}_{\text{KL}}(\mathbb{P}_{\mathbf{y}|\mathbf{x}}(\cdot|\mathbf{x}) \parallel \mathbb{P}_{\mathbf{y}|\mathbf{z}}(\cdot|\mathbf{z})) + \mathbb{E}_{\mathbf{x},g}[\mathsf{D}_2(\mathbb{P}_{\mathbf{y}|\mathbf{z}}(\cdot|\mathbf{z}) \parallel \bar{\mathbf{h}}_{\Gamma}(\hat{f}(\mathbf{z})))]] \\ &\leq 2\epsilon_{\mathcal{G}}^{\text{cls}} + \mathbb{E}_{\mathbf{x},g}[\mathsf{D}_2(\mathbb{P}_{\mathbf{y}|\mathbf{z}}(\cdot|\mathbf{z}) \parallel \bar{\mathbf{h}}_{\Gamma}(\hat{f}(\mathbf{z})))].\end{aligned}$$

Therefore, it remains to prove

$$\mathbb{E}_{\mathbf{x},g}[\mathsf{D}_2(\mathbb{P}_{\mathbf{y}|\mathbf{z}}(\cdot|\mathbf{z}) \parallel \bar{\mathbf{h}}_{\Gamma}(\hat{f}(\mathbf{z})))] \leq \frac{cS \exp(B)}{\sigma_{\mathbf{E}_*}^2} \cdot (R_f(S_{\hat{f}_{\text{aug}}}) - R_f(S_*)). \quad (44)$$

Since

$$\begin{aligned}
\mathbb{E}_{\mathbf{x},g}[\mathsf{D}_2(\mathbb{P}_{\mathbf{y}|\mathbf{z}}(\cdot|\mathbf{z})||\bar{\mathsf{h}}_{\mathbf{\Gamma}}(\hat{f}(\mathbf{z})))] &\leq \mathbb{E}_{\mathbf{x},g}\left[\mathbb{E}_{y \sim \mathbb{P}_{\mathbf{y}|\mathbf{z}}(\cdot|\mathbf{z})} \frac{\mathbb{P}_{\mathbf{y}|\mathbf{z}}(y|\mathbf{z}) - \bar{\mathsf{h}}_{\mathbf{\Gamma}}(\hat{f}(\mathbf{z}))_y}{\bar{\mathsf{h}}_{\mathbf{\Gamma}}(\hat{f}(\mathbf{z}))_y}\right] \\
&= \mathbb{E}_{\mathbf{x},g}\left[\sum_{y \in [M]} \frac{(\mathbb{P}_{\mathbf{y}|\mathbf{z}}(y|\mathbf{z}) - \bar{\mathsf{h}}_{\mathbf{\Gamma}}(\hat{f}(\mathbf{z}))_y)^2}{\bar{\mathsf{h}}_{\mathbf{\Gamma}}(\hat{f}(\mathbf{z}))_y}\right] \\
&\leq \exp(B) \cdot \mathbb{E}_{\mathbf{x},g}\left[\sum_{y \in [M]} (\mathbb{P}_{\mathbf{y}|\mathbf{z}}(y|\mathbf{z}) - \bar{\mathsf{h}}_{\mathbf{\Gamma}}(\hat{f}(\mathbf{z}))_y)^2\right] \\
&= \exp(B) \cdot \mathbb{E}_{\mathbf{x},g}\left[\sum_{y \in [M]} (\mathbb{P}_c(y|\mathbf{x}^{c_1} = \mathbf{z}) - \bar{\mathsf{h}}_{\mathbf{\Gamma}}(\hat{f}(\mathbf{z}))_y)^2\right], \tag{45}
\end{aligned}$$

where the third line uses the definition of `trun` and claim (46) in the proof of Lemma 6, and the last line uses the fact that $\mathbb{P}_c(\mathbf{y} = y|\mathbf{x}^{c_1} = j) = \mathbb{P}_{\mathbf{y}|\mathbf{z}}(\mathbf{y} = y|\mathbf{z} = j)$ for all $y \in [M], j \in [S]$. Eq. (44) follows immediately from Lemma 6 which gives an upper bound on the term in Eq. (45).

Generalization error. The proof follows from a standard analysis of empirical process similar to the proof of Eq. (29) in the proof of Theorem 4. Thus, we only provide a sketch of the proof here.

Let $\Gamma := \{\Gamma : \|\Gamma_w\|_{\text{op}} \vee \|\Gamma_b\|_2 \leq B_\Gamma\}$ and define the norm $\|\Gamma - \tilde{\Gamma}\| := \|\Gamma_w - \tilde{\Gamma}_w\|_{\text{op}} \vee \|\Gamma_b - \tilde{\Gamma}_b\|_2$. First, by a triangle inequality, the fact that $\|\log \mathsf{h}_{\mathbf{\Gamma}}\|_{\infty} \leq 2B$ (which follows from the definition of `trun`), and Corollary 2.21 in [Wai19] for functions with bounded differences, we have

$$\text{generalization error} \leq 2\mathbb{E}\left[\sup_{\Gamma \in \Gamma} |\hat{R}_{\text{cls}}(\mathsf{h}_{\mathbf{\Gamma}}) - \mathbb{E}[\hat{R}_{\text{cls}}(\mathsf{h}_{\mathbf{\Gamma}})]|\right] + 2B \frac{\sqrt{\log(1/\delta)}}{\sqrt{m}}$$

with probability at least $1 - \delta$. Let $X_{\mathbf{\Gamma}} := \hat{R}_{\text{cls}}(\mathsf{h}_{\mathbf{\Gamma}}) - \mathbb{E}[\hat{R}_{\text{cls}}(\mathsf{h}_{\mathbf{\Gamma}})]$. Then we have

$$\mathbb{E}\left[\sup_{\Gamma \in \Gamma} |\hat{R}_{\text{cls}}(\mathsf{h}_{\mathbf{\Gamma}}) - \mathbb{E}[\hat{R}_{\text{cls}}(\mathsf{h}_{\mathbf{\Gamma}})]|\right] \leq \mathbb{E}[|X_{\mathbf{\Gamma}_0}|] + \mathbb{E}\left[\sup_{\Gamma, \tilde{\Gamma} \in \Gamma} |X_{\mathbf{\Gamma}} - X_{\tilde{\Gamma}}|\right] \leq \frac{2B}{\sqrt{m}} + \mathbb{E}\left[\sup_{\Gamma, \tilde{\Gamma} \in \Gamma} |X_{\mathbf{\Gamma}} - X_{\tilde{\Gamma}}|\right].$$

Moreover, the process $\{X_{\mathbf{\Gamma}}\}_{\mathbf{\Gamma} \in \Gamma}$ is a zero-mean sub-Gaussian process with respect to the metric $\rho_X(\mathbf{\Gamma}, \tilde{\mathbf{\Gamma}}) := 2\|\log \mathsf{h}_{\mathbf{\Gamma}} - \log \mathsf{h}_{\tilde{\mathbf{\Gamma}}}\|_{\infty}/\sqrt{m}$ since $X_{\mathbf{\Gamma}}$ is the average of i.i.d. random variables bounded by

$$\begin{aligned}
&2 \sup_{i \in [m]} |\langle e_{\mathbf{y}_i}, \log \bar{\mathsf{h}}_{\mathbf{\Gamma}}(\hat{f}(\mathbf{z}_i)) \rangle - \langle e_{\mathbf{y}_i}, \log \bar{\mathsf{h}}_{\tilde{\mathbf{\Gamma}}}(\hat{f}(\mathbf{z}_i)) \rangle| \\
&\leq 2\|\log \bar{\mathsf{h}}_{\mathbf{\Gamma}}(\hat{f}(\mathbf{z}_i)) - \log \bar{\mathsf{h}}_{\tilde{\mathbf{\Gamma}}}(\hat{f}(\mathbf{z}_i))\|_{\infty} \leq \rho_X(\mathbf{\Gamma}, \tilde{\mathbf{\Gamma}}) \cdot \sqrt{m}, \quad \text{and moreover} \\
&\rho_X(\mathbf{\Gamma}, \tilde{\mathbf{\Gamma}}) \stackrel{(i)}{\leq} c\|\log \mathsf{trun}(\mathbf{\Gamma}_w \hat{f}(\mathbf{z}) + \mathbf{\Gamma}_b) - \log \mathsf{trun}(\tilde{\mathbf{\Gamma}}_w \hat{f}(\mathbf{z}) + \tilde{\mathbf{\Gamma}}_b)\|_{\infty}/\sqrt{m}, \\
&\stackrel{(ii)}{\leq} c \exp(B) \cdot \|\mathbf{\Gamma} - \tilde{\mathbf{\Gamma}}\|/\sqrt{m} =: \bar{B}\|\mathbf{\Gamma} - \tilde{\mathbf{\Gamma}}\|/\sqrt{m},
\end{aligned}$$

where step (i) uses $\|\log \text{softmax}(\mathbf{u}) - \log \text{softmax}(\mathbf{v})\|_{\infty} \leq 2\|\mathbf{u} - \mathbf{v}\|_{\infty}$ and step (ii) follows from Taylor expansion of $s(x) = \log x$, the assumption that $\|\hat{f}(\mathbf{z})\|_2 \leq B_{\mathbf{W}} = M$. Therefore, we have by Dudley's integral bound (see e.g., Theorem 5.22 in [Wai19]) that

$$\begin{aligned}
\mathbb{E}\left[\sup_{\mathbf{\Gamma}, \tilde{\mathbf{\Gamma}} \in \Gamma} |X_{\mathbf{\Gamma}} - X_{\tilde{\mathbf{\Gamma}}}| \right] &\leq c \int_0^{cB/\sqrt{m}} \sqrt{\log \mathcal{N}(u, \rho_X, \{X_{\mathbf{\Gamma}}, \mathbf{\Gamma} \in \Gamma\})} du \leq c \int_0^{cB/\sqrt{m}} \sqrt{\log \mathcal{N}\left(u, \frac{\bar{B}}{\sqrt{m}} \|\cdot\|, \Gamma\right)} du \\
&\leq c \int_0^{cB/\sqrt{m}} \sqrt{\log \mathcal{N}\left(\frac{\sqrt{m} \cdot u}{\bar{B}}, \|\cdot\|_{\text{op}}, \Gamma_w\right)} du \\
&\leq c \int_0^{cB/\sqrt{m}} \left(\sqrt{\log \mathcal{N}\left(\frac{\sqrt{m} \cdot u}{\bar{B}}, \|\cdot\|_{\text{op}}, \Gamma_w\right)} + \sqrt{\log \mathcal{N}\left(\frac{\sqrt{m} \cdot u}{\bar{B}}, \|\cdot\|_2, \Gamma_b\right)} \right) du \\
&\leq c \int_0^{cB/\sqrt{m}} \sqrt{M^2 \cdot \log\left(1 + 4\frac{B_{\mathbf{\Gamma}} \bar{B}}{\sqrt{mu}}\right)} du \leq c \frac{BM \log^{1/2}(B_{\mathbf{\Gamma}} \bar{B})}{\sqrt{m}} \leq c \frac{BM(\log^{1/2} B_{\mathbf{\Gamma}} + \sqrt{\bar{B}})}{\sqrt{m}},
\end{aligned}$$

where $\Gamma_w := \{\boldsymbol{\Gamma}_w \in \mathbb{R}^{M \times M} : \|\boldsymbol{\Gamma}_w\|_{\text{op}} \leq B_\Gamma\}$ and $\Gamma_b := \{\boldsymbol{\Gamma}_b \in \mathbb{R}^M : \|\boldsymbol{\Gamma}_b\|_2 \leq B_\Gamma\}$, and the last line uses the covering number bound of unit balls. Putting pieces together yields the desired bound.

C.5 An auxiliary lemma

Lemma 6 (Upper bound on the term in Eq. (45)). *Let the assumptions in Theorem 3 and the notations in its proof in Appendix C.4.2 hold. Assume $R_f(S_{\hat{f}_{\text{aug}}}) - R_f(S_\star) \leq c\sigma_{\mathbf{E}_\star}^2/(S^2M)$ for some absolute constant $c > 0$, then*

$$\mathbb{E}_{\mathbf{x},g} \left[\sum_{y \in [M]} (\mathbb{P}_c(y|\mathbf{x}^{c_1} = \mathbf{z}) - \bar{h}_\Gamma(\hat{f}(\mathbf{z}))_y)^2 \right] \leq \frac{c'S}{\sigma_{\mathbf{E}_\star}^2} \cdot (R_f(S_{\hat{f}_{\text{aug}}}) - R_f(S_\star))$$

for some absolute constant $c' > 0$.

Proof of Lemma 6. The proof consists of two steps. First, we plug the definition of \bar{h}_Γ into Eq. (6) and simplify the expression. Then, we demonstrate that the simplified expression can be further bounded using the excess risk $R_f(S_{\hat{f}_{\text{aug}}}) - R_f(S_\star)$ of the learned encoder \hat{f}_{aug} .

Step 1: simplify the notation. Since

$$\|\nabla_{\mathbf{u}} \text{softmax}(\log \mathbf{u})\|_{\text{op}} = \left\| \frac{1}{\|\mathbf{u}\|_1} \mathbf{I}_M - \frac{\mathbf{u}}{\|\mathbf{u}\|_1} \mathbf{1}_M^\top \right\|_{\text{op}} \leq \frac{1}{\|\mathbf{u}\|_1} + 1$$

for any $\mathbf{u} \in \mathbb{R}_{>0}^M$, we have

$$\begin{aligned} \mathbb{E}_{\mathbf{x},g} \left[\sum_{y \in [M]} (\mathbb{P}_c(y|\mathbf{x}^{c_1} = \mathbf{z}) - \bar{h}_\Gamma(\hat{f}(\mathbf{z}))_y)^2 \right] &\leq c \mathbb{E}_{\mathbf{x},g} \left[\sum_{y \in [M]} (\mathbb{P}_c(y|\mathbf{x}^{c_1} = \mathbf{z}) - \text{trun}(\boldsymbol{\Gamma}_w \hat{f}(\mathbf{z}) + \boldsymbol{\Gamma}_b)_y)^2 \right] \\ &\leq c \mathbb{E}_{\mathbf{x},g} \left[\sum_{y \in [M]} (\mathbb{P}_c(y|\mathbf{x}^{c_1} = \mathbf{z}) - (\boldsymbol{\Gamma}_w \hat{f}(\mathbf{z}) + \boldsymbol{\Gamma}_b)_y)^2 \right], \end{aligned}$$

where in the first inequality we use the claim that

$$|1 - \|\text{trun}(\boldsymbol{\Gamma}_w \hat{f}(\mathbf{z}) + \boldsymbol{\Gamma}_b)\|_1| \leq 1/2. \quad (46)$$

The proof of this claim is deferred to the end of the proof of the lemma. The second inequality follows from a Taylor expansion of $s(x) = \log x$, the boundedness assumption that $\mathbb{P}_c(y|\mathbf{x}^{c_1} = \mathbf{z}) \in [\exp(-B), 1]$, and noting the truncation $\text{trun}(\cdot)$ reduces the ℓ_2 error. Moreover, for any $\mathbf{z} \in [S]$, by the definition of $(\boldsymbol{\Gamma}_w, \boldsymbol{\Gamma}_b)$ in Eq. (43)

$$\begin{aligned} &\boldsymbol{\Gamma}_w \hat{f}(\mathbf{z}) + \boldsymbol{\Gamma}_b \\ &= \sqrt{2} \mathbf{P}_y^{1/2} (\mathbf{E}_\star \mathbf{E}_\star^\top)^{-1} \mathbf{E}_\star [(\mathbf{I}_S - \mathcal{P}_{\mathbf{1}_S}) \hat{\mathbf{E}}^\top \hat{f}(\mathbf{z}) + \mathbf{1}_S/2] \\ &= \sqrt{2} \mathbf{P}_y^{1/2} (\mathbf{E}_\star \mathbf{E}_\star^\top)^{-1} \mathbf{E}_\star \mathbf{E}_\star^\top \mathbf{E}_\star(\mathbf{z}) \\ &\quad + \sqrt{2} \mathbf{P}_y^{1/2} (\mathbf{E}_\star \mathbf{E}_\star^\top)^{-1} \mathbf{E}_\star [(\mathbf{I}_S - \mathcal{P}_{\mathbf{1}_S}) \hat{\mathbf{E}}^\top \hat{f}(\mathbf{z}) + \mathbf{1}_S/2 - \mathbf{E}_\star^\top \mathbf{E}_\star(\mathbf{z})] \\ &= \sqrt{2} \mathbf{P}_y^{1/2} \mathbf{E}_\star(\mathbf{z}) + \sqrt{2} \mathbf{P}_y^{1/2} (\mathbf{E}_\star \mathbf{E}_\star^\top)^{-1} \mathbf{E}_\star [(\mathbf{I}_S - \mathcal{P}_{\mathbf{1}_S}) \hat{\mathbf{E}}^\top \hat{f}(\mathbf{z}) + \mathbf{1}_S/2 - \mathbf{E}_\star^\top \mathbf{E}_\star(\mathbf{z})]. \end{aligned}$$

Since $\sqrt{2} \mathbf{P}_y^{1/2} \mathbf{E}_\star(\mathbf{z}) = (\mathbb{P}_c(y|\mathbf{x}^{c_1} = \mathbf{z}))_{y \in [M]}$ and $\mathbf{z} \stackrel{d}{=} \mathbf{x}^{c_1}$ follows the uniform distribution on $[S]$ by

assumption, it follows that

$$\begin{aligned}
& \mathbb{E}_{\mathbf{x}, g} \left[\sum_{y \in [M]} (\mathbb{P}_c(y | \mathbf{x}^{c_1} = \mathbf{z}) - (\mathbf{\Gamma}_w \hat{f}(\mathbf{z}) + \mathbf{\Gamma}_b)_y)^2 \right] \\
& \leq 2\mathbb{E}_{\mathbf{z}} [\|(\mathbf{E}_* \mathbf{E}_*^\top)^{-1} \mathbf{E}_*[(\mathbf{I}_S - \mathcal{P}_{\mathbf{1}_S}) \hat{\mathbf{E}}^\top \hat{f}(\mathbf{z}) + \mathbf{1}_S/2 - \mathbf{E}_*^\top \mathbf{E}_*(\mathbf{z})]\|_2^2] \\
& \leq \frac{2}{\sigma_{\mathbf{E}_*}^2} \mathbb{E}_{\mathbf{z}} [\|(\mathbf{I}_S - \mathcal{P}_{\mathbf{1}_S}) \hat{\mathbf{E}}^\top \hat{f}(\mathbf{z}) + \mathbf{1}_S/2 - \mathbf{E}_*^\top \mathbf{E}_*(\mathbf{z})\|_2^2] \\
& \leq \frac{2}{S\sigma_{\mathbf{E}_*}^2} \|(\mathbf{I}_S - \mathcal{P}_{\mathbf{1}_S}) \hat{\mathbf{E}}^\top \hat{\mathbf{E}} + \mathbf{1}_S \mathbf{1}_S^\top / 2 - \mathbf{E}_*^\top \mathbf{E}_*\|_{\text{fro}}^2 \\
& = \frac{2}{S\sigma_{\mathbf{E}_*}^2} \|(\mathbf{I}_S - \mathcal{P}_{\mathbf{1}_S}) \hat{\mathbf{E}}^\top \hat{\mathbf{E}} - (\mathbf{I}_S - \mathcal{P}_{\mathbf{1}_S}) \mathbf{E}_*^\top \mathbf{E}_*\|_{\text{fro}}^2,
\end{aligned} \tag{47}$$

where the last equality follows since $\mathbf{E}_*^\top(\mathbf{z}^{(1)}) \mathbf{E}_*(\mathbf{z}^{(2)}) = \frac{\mathbb{P}_c(\mathbf{z}^{(1)}, \mathbf{z}^{(2)})}{2\mathbb{P}_c(\mathbf{z}^{(1)})\mathbb{P}_c(\mathbf{z}^{(2)})}$ for any $(\mathbf{z}^{(1)}, \mathbf{z}^{(2)}) \in [S]$, and $\frac{1}{S} \sum_{\mathbf{z}^{(2)} \in [S]} \frac{\mathbb{P}_c(\mathbf{z}^{(1)}, \mathbf{z}^{(2)})}{\mathbb{P}_c(\mathbf{z}^{(1)})\mathbb{P}_c(\mathbf{z}^{(2)})} = 1$ for all $\mathbf{z}^{(1)} \in [S]$.

Step 2: bound the expression by excess risk. We claim that for some absolute constant $c > 0$,

$$\begin{aligned}
& \|(\mathbf{I}_S - \mathcal{P}_{\mathbf{1}_S}) \hat{\mathbf{E}}^\top \hat{\mathbf{E}} - (\mathbf{I}_S - \mathcal{P}_{\mathbf{1}_S}) \mathbf{E}_*^\top \mathbf{E}_*\|_{\text{fro}}^2 \\
& \leq c \|(\mathbf{I}_S - \mathcal{P}_{\mathbf{1}_S})(\hat{\mathbf{E}}^\top \hat{\mathbf{E}} + \hat{w}\mathbf{I}_S) - (\mathbf{I}_S - \mathcal{P}_{\mathbf{1}_S})(\mathbf{E}_*^\top \mathbf{E}_* + S \cdot \mathbf{I}_S/2)\|_{\text{fro}}^2, \quad \text{and}
\end{aligned} \tag{48a}$$

$$\begin{aligned}
& \|(\mathbf{I}_S - \mathcal{P}_{\mathbf{1}_S})(\hat{\mathbf{E}}^\top \hat{\mathbf{E}} + \hat{w}\mathbf{I}_S) - (\mathbf{I}_S - \mathcal{P}_{\mathbf{1}_S})(\mathbf{E}_*^\top \mathbf{E}_* + S \cdot \mathbf{I}_S/2)\|_{\text{fro}}^2 \\
& \leq S^2 \cdot (R_f(S_{\hat{f}_{\text{aug}}}) - R_f(S_*)). \tag{48b}
\end{aligned}$$

Combining claim (48a) and (48b) and bound (47) yields the desired bound. Now, it remains to prove these two claims.

Proof of claim (48a). Adopt the shorthand notation $\Delta = (\hat{\mathbf{E}}^\top \hat{\mathbf{E}} + \hat{w}\mathbf{I}_S) - (\mathbf{E}_*^\top \mathbf{E}_* + S \cdot \mathbf{I}_S/2)$. First, by the triangle inequality, it suffices to show

$$\|(\mathbf{I}_S - \mathcal{P}_{\mathbf{1}_S})(\hat{w} - S/2)\|_{\text{fro}}^2 \leq c \|(\mathbf{I}_S - \mathcal{P}_{\mathbf{1}_S})\Delta\|_{\text{fro}}^2$$

for some absolute constant $c > 0$. Note that $\text{rank}(\hat{\mathbf{E}}^\top \hat{\mathbf{E}} - \mathbf{E}_*^\top \mathbf{E}_*) \leq 2M$, therefore, there are at least $S/2$ singular values of Δ which equal $|\hat{w} - S/2|/2$. As a result, we have

$$\begin{aligned}
\|(\mathbf{I}_S - \mathcal{P}_{\mathbf{1}_S})\Delta\|_{\text{fro}}^2 &= \text{trace}(\Delta(\mathbf{I}_S - \mathcal{P}_{\mathbf{1}_S})\Delta) = \|\Delta\|_{\text{fro}}^2 - \frac{1}{S} \mathbf{1}_S^\top \Delta^2 \mathbf{1}_S \\
&\geq \|\Delta\|_{\text{fro}}^2 - \|\Delta\|_{\text{op}}^2 \geq \frac{1}{4} \|(\hat{w} - S/2)\mathbf{I}_S\|_{\text{fro}}^2 \geq \frac{1}{4} \|(\mathbf{I}_S - \mathcal{P}_{\mathbf{1}_S})(\hat{w} - S/2)\|_{\text{fro}}^2.
\end{aligned}$$

Proof of claim (48b). Adpot the shorthands $S_{\hat{f}_{\text{aug}}}^m := \left(S_{\hat{f}_{\text{aug}}}(\mathbf{z}^{(1)}, \mathbf{z}^{(2)}) \right)_{\mathbf{z}^{(1)}, \mathbf{z}^{(2)} \in [S]} \in \mathbb{R}^{S \times S}$ and $S_*^m := \left(S_*(\mathbf{z}^{(1)}, \mathbf{z}^{(2)}) \right)_{\mathbf{z}^{(1)}, \mathbf{z}^{(2)} \in [S]} \in \mathbb{R}^{S \times S}$, where $S_*(\mathbf{z}^{(1)}, \mathbf{z}^{(2)}) = \frac{\mathbb{P}_z(\mathbf{z}^{(1)}, \mathbf{z}^{(2)})}{\mathbb{P}_z(\mathbf{z}^{(1)})\mathbb{P}_z(\mathbf{z}^{(2)})}$. Since we assume $\mathbf{z} \stackrel{d}{=} \mathbf{x}^{c_1}$ follows the uniform distribution on $[S]$, by the definition of \hat{f}_{aug} and claim (37) in the proof of Eq. (14)

$$\begin{aligned}
& \|(\mathbf{I}_S - \mathcal{P}_{\mathbf{1}_S})(\hat{\mathbf{E}}^\top \hat{\mathbf{E}} + \hat{w}\mathbf{I}_S) - (\mathbf{I}_S - \mathcal{P}_{\mathbf{1}_S})(\mathbf{E}_*^\top \mathbf{E}_* + S \cdot \mathbf{I}_S/2)\|_{\text{fro}}^2 \\
& = \|(\mathbf{I}_S - \mathcal{P}_{\mathbf{1}_S})(S_{\hat{f}_{\text{aug}}}^m - S_*^m)\|_{\text{fro}}^2 \\
& = S^2 \cdot T_1,
\end{aligned}$$

where

$$T_1 := \mathbb{E}_{\mathbf{z}^{(1)}, \mathbf{z}^{(2)} \sim \mathbb{P}_z \times \mathbb{P}_z} [((S_{\hat{f}_{\text{aug}}} - S_*)(\mathbf{z}^{(1)}, \mathbf{z}^{(2)}) - \mathbb{E}_{\mathbf{z}^{(2)} \sim \mathbb{P}_z} [(S_{\hat{f}_{\text{aug}}} - S_*)(\mathbf{z}^{(1)}, \mathbf{z}^{(2)})])^2].$$

Finally, by a second-order Taylor expansion of $R_f(S)$ at S_* , we have

$$R_f(S_{\hat{f}_{\text{aug}}}) - R_f(S_*) = T_1.$$

Combining the two equalities yields the claim.

Proof of claim (46). Note that for any $\mathbf{z} \in [S]$,

$$\begin{aligned} |1 - \|\text{trun}(\mathbf{\Gamma}_w \hat{f}(\mathbf{z}) + \mathbf{\Gamma}_b)\|_1| &\leq \sum_{y \in [M]} |\mathbb{P}_c(y | \mathbf{x}^{c_1} = \mathbf{z}) - \text{trun}(\mathbf{\Gamma}_w \hat{f}(\mathbf{z}) + \mathbf{\Gamma}_b)_y| \\ &\leq \sum_{y \in [M]} |\mathbb{P}_c(y | \mathbf{x}^{c_1} = \mathbf{z}) - (\mathbf{\Gamma}_w \hat{f}(\mathbf{z}) + \mathbf{\Gamma}_b)_y| \\ &\leq \sqrt{MS} \cdot \sqrt{\mathbb{E}_{\mathbf{x}, g} \left[\sum_{y \in [M]} (\mathbb{P}_c(y | \mathbf{x}^{c_1} = \mathbf{z}) - (\mathbf{\Gamma}_w \hat{f}(\mathbf{z}) + \mathbf{\Gamma}_b)_y)^2 \right]}, \end{aligned}$$

where the last line follows from the assumption that \mathbf{x}^{c_1} (and hence \mathbf{z}) follows the uniform distribution on $[S]$. Thus, combining Eq. (47), claim (48a) and (48b) yields

$$|1 - \|\text{trun}(\mathbf{\Gamma}_w \hat{f}(\mathbf{z}) + \mathbf{\Gamma}_b)\|_1| \leq c \frac{S\sqrt{M}}{\sigma_{\mathbf{E}_*}} \cdot \sqrt{R_f(\mathbf{S}_{\hat{f}^{\text{aug}}}) - R_f(\mathbf{S}_*)} \leq \frac{1}{2}.$$

□

D Additional experiments

We also conducted a small-scale experiment in the CLIP setting (language-image pretraining, [RKH⁺21]) to compare the contrastive learning losses. Namely, we use the CLIP model (RN50-quickgelu, which consists of a ResNet-50 image encoder and 12-layer Transformer text encoder) on a 100K subsample of the `cc3m-wds` dataset [SDGS18] using both KL (i.e., InfoNCE) and χ^2 -contrastive losses (Eq. 3 and 10). The original dataset contains about 3.3M image-text pairs, but due to limited compute, we trained on the subsample for 32 epochs.

We evaluated the models based on their zero-shot classification performance on the ImageNet-1k validation set (1000 classes, 500 images per class). For KL and χ^2 -contrastive losses, we set the link functions $\tau(x)$ to x/t and $e^{x/t}$, respectively, with trainable temperature t initialized to 1. We used a batch size of 128 and the AdamW optimizer with weight decay 0.02, and selected the best learning rate via grid search from $\{3e-5, 1e-4, 3e-4, 1e-3\}$. The optimal learning rate for both losses is 3×10^{-4} .

We repeated the experiments three times and report the top-5 accuracy on the ImageNet-1k validation set. From Table 1, we observe that in this small-scale experiment, the model trained with χ^2 -contrastive loss achieves zero-shot performance comparable to that of InfoNCE. We do not claim that the χ^2 -contrastive loss is superior, as both methods could benefit from further hyperparameter tuning (e.g., initial temperature) or larger datasets. However, we note that χ^2 -contrastive loss is able to learn representations that are useful for downstream tasks, which is consistent with our theoretical findings. We leave more extensive experiments in the CLIP setting to future work.

Table 1: Top-5 zero-shot classification accuracy on ImageNet-1k.

Method	Accuracy (%)
InfoNCE	7.5 ± 0.3
Chi-squared	9.4 ± 0.1